Transportation Engineering and Planning

Third Edition

C. S. Papacostas
P. D. Prevedouros
*Jniversity of Hawaii at Manoa
Honolulu, Hawaii*

PHI Learning Private Limited
New Delhi-110001
2009

This Indian Reprint—Rs. 395.00
(Original U.S. Edition—Rs. 3260.00)

TRANSPORTATION ENGINEERING AND PLANNING, 3rd ed.
by C.S. Papacostas and P.D. Prevedouros

Original edition, entitled *Transportation Engineering and Planning*, 3rd ed. by C.S. Papacostas and P.D. Prevedouros, published by Pearson Education, Inc., publishing as Pearson Prentice Hall.

ISBN-978-81-203-2154-0

Indian edition published by PHI Learning Private Limited.

This edition is manufactured in India, and is authorized for sale in India, Pakistan, Sri Lanka, Bhutan, Bangladesh, Nepal and the Maldives only.

Published by Asoke K. Ghosh, PHI Learning Private Limited, M-97, Connaught Circus,
New Delhi-110001 and Printed by Jay Print Pack Private Limited, New Delhi-110015.

To the memory of my father Symeon
C.S.P.

To my parents Dimitrios and Toula Prevedouros
P.D.P.

About the Authors

Dr. Constantinos S. Papacostas is Professor of Civil Engineering and Director of the Hawaii Local Technical Assistance Program at the University of Hawaii at Manoa, and is Technical Director for Model Development for the Oahu MPO. He received his B.E. degree (magna cum laude) from Youngstown State University and his M.S. and Ph.D. from Carnegie-Mellon University. Dr. Papacostas teaches undergraduate courses in traffic engineering and urban and regional transportation planning, and graduate courses in applications of advanced computer-based techniques to transportation planning and geographic information systems. His research work includes transportation planning, air quality, bus operations, traffic safety, driver behavior, and population, energy, and infrastructure systems.

He was awarded the ASCE 1990 and 1998 Outstanding Faculty Award. He served as President of the Hawaii Chapters of ASCE and ITE, and Civil Engineering Graduate Program Chair. In addition to being an author of numerous articles published in professional journals, Dr. Papacostas is a reviewer for the *Transportation Research Record, Transportation Research,* and *Transportation Quarterly.* He is a member of AAAS, ASCE, APA, ITE, and TRB, and a member of Sigma Tau, Sigma Chi, and Phi Kappa Phi honor societies. [www.eng.hawaii.edu/~csp/]

Dr. Panos D. Prevedouros is Associate Professor of Civil Engineering and Graduate Program Chair at the University of Hawaii at Manoa. He received his B.S. degree from Aristotle University, Greece, and his M.S. and Ph.D. from Northwestern University. He teaches undergraduate courses in transportation and traffic engineering and graduate courses in advanced demand modeling, transportation economic and operational efficiency, and intelligent transport systems.

Dr. Prevedouros was awarded the ASCE 1996 Outstanding Faculty Award and the 1996 A1F04 Best Paper Award (with C. S. Papacostas). He is a principal investigator of research projects in the areas of freeway management, traffic modeling, traffic software evaluation, transportation noise, airport operations, ITS deployment, and traffic detector testing. He publishes extensively in professional journals, and is a member of several TRB committees and ASCE's committee on traffic operations. Also, Dr. Prevedouros is a reviewer for ASCE Journals, the *Transportation Research Record,* and *Transportation Research,* as well as a member of ASCE, ITE, and TRB. [www.eng.hawaii.edu/~panos/]

Contents

Part 4: Supporting Elements

Preface

Great effort was devoted to enrich and update the third edition of this most successful transportation engineering textbook. The book has been restructured to provide a better fit into undergraduate curricula and a better progression between engineering and planning topics. The entire book has been organized into four sections:

1. Design and operations
2. Transportation systems
3. Transportation impacts
4. Supporting elements

Several of the topics in Parts 2 and 3 are appropriate for senior-level transportation courses and introductory graduate courses.

Some notable additions include congestion quantification and management strategies, extensive coverage on intelligent transportation systems, random utility discrete choice modeling, land-use modeling, an exclusive chapter on traffic software, and coverage of HCM 2000, traffic calming, roundabouts, Superpave, TEA-21, and other contemporary topics.

We are most appreciative of the many instructors and departments who have chosen to use this textbook for their transportation classes making it one of the most popular textbooks worldwide. We believe that this edition will be even more helpful in providing comprehensive, unbiased, state-of-the-art knowledge of transportation engineering and planning. We are looking forward to your continued support and, as always, we welcome your comments and suggestions for improvements.

C. S. Papacostas wishes to thank the Oahu Metropolitan Planning Organization (OMPO) and its Executive Director, Gordon G. W. Lum, for the opportunity to serve as technical director for OMPO's land-use and transportation model development during the second half of the 1990s. The wide range of perspectives brought to the table by the project's peer review group (PRG), project consultants, and local and state agency staff helped to strike a reasonable balance in the preparation of the material dealing with land use, transportation planning, and travel-demand modeling. His association with the Hawaii Local Technical Assistance Program was helpful in the areas of transportation materials, ITS architecture and deployment, and other engineering topics covered in the book.

Panos D. Prevedouros is indebted to the Development Programmes Department of INTRACOM, S.A. for providing him with plentiful resources during his sabbatical there, which permitted him to redo Chapter 6 essentially from scratch, draft the chapter on traffic software, and work on other improvements and additions. The cooperation with INTRACOM's Dr. N. Skayannis and Mr. Gabriel Hatoglou on ITS, electronic road pricing, and automated vehicle classification was most beneficial. Dr. Kimon Proussaloglou (Cambridge Systematics), Dr. Haitham Al-Deek (University of Central Florida), and Dr. Asad Khattak (University of North Carolina at Chapel Hill) made valuable suggestions for improvements. The late Dr. Eric Pas (Duke University) gave useful advise on improving the structure of the book and enriching the urban systems section.

We must thank the Transportation Research Board for providing a spacious, multifaceted forum as well as wide accessibility to the ever-expanding and changing body of transportation information and knowledge.

C. S. Papacostas
P. D. Prevedouros
Honolulu, Hawaii

1

Introduction and Background

1.1 THE TRANSPORTATION SYSTEM

1.1.1 Definition and Scope

A transportation system may be defined as consisting of the *fixed facilities*, the *flow entities*, and the *control system* that permit *people and goods* to overcome the friction of geographical space *efficiently* in order to *participate* in a *timely* manner in some *desired activity*.

At first glance this definition may appear to be either trivial or pretentious. After all, "overcoming the friction of geographic space"* is a very awkward way of saying "to move from point *A* to point *B*"! However, this definition reveals the breadth of transportation engineering and delineates the purpose and scope of this introductory text. It identifies the functional components of a transportation system (i.e., the fixed facilities, the flow entities, and the control system) and encapsulates the fact that transportation provides the connectivity that facilitates other societal interactions.

1.1.2 Fixed Facilities

Fixed facilities are the physical components of the system that are fixed in space and constitute the *network* of *links* (e.g., roadway segments, railway track, and pipes) and *nodes* (e.g., intersections, interchanges, transit terminals, harbors, and airports) of the transportation system. Their design, traditionally within the realm of civil engineering, includes soil and foundation engineering, structural design, the design of drainage systems, and *geometric design,* which is concerned with the physical proportioning of the elements of

*Australians refer to the "tyranny of distance" due to the large size and remoteness of their country and continent.

fixed facilities. Although related, geometric design is different from other aspects of design (e.g., structural design, which is concerned with the strength of structures to withstand efficiently the expected forces or loads), which are covered elsewhere in the typical civil engineering curriculum.

1.1.3 Flow Entities and Technology

Flow entities are the units that traverse the fixed facilities. They include vehicles, container units, railroad cars, and so on. In the case of the highway system the fixed facilities are expected to accommodate a wide variety of vehicle types, ranging from bicycles to large tractor-trailer combinations. For the purposes of geometric design the American Association of State Highway and Transportation Officials (AASHTO) has specified a set of *design vehicles*, each describing a typical class of highway vehicles [1.1].

In this book flow entities are considered only in terms of their generic characteristics, such as size, weight, and acceleration and deceleration capabilities, rather than in terms of their specific technological design, which is normally undertaken by mechanical and electrical engineers. Thus vehicular motion and vehicle flow equations are expressed as general relationships between the generic variables and can be applied to many vehicle technologies once their specific attributes are determined.

1.1.4 Control System

The *control system* consists of *vehicular control* and *flow control*. Vehicular control refers to the technological way in which individual vehicles are guided on the fixed facilities. Such control can be manual or automated. The proper geometric design of the fixed facilities must incorporate, in addition to the characteristics of the vehicle, the characteristics of the vehicular control system. In the case of highway facilities, where the vehicles are manually controlled, these include driver characteristics, such as the time a driver takes to perceive and react to various stimuli; examples of such *human factors* are contained in this book. In the case of automated systems similar but more precisely definable response times exist as well.

The flow control system consists of the means that permit the efficient and smooth operation of streams of vehicles and the reduction of conflicts between vehicles. This system includes various types of signing, marking, and signal systems and the underlying rules of operation. Traffic signal control is discussed in Chapter 4 and advanced systems, known as Intelligent Transportation Systems (ITS), are covered in Chapter 6.

1.1.5 Transportation Demand

The definition of a transportation system given earlier addresses another consideration that is of concern to transportation specialists: Transportation systems are constructed as neither pure expressions of engineering ingenuity nor monuments of purely aesthetic quality. They are built to serve people in undertaking their economic, social, and cultural activities. In the jargon of the economist, the demand for transportation is *derived*, or *indirect;* that is, people do not normally travel or move their possessions for the sake of movement but to fulfill certain needs, such as going to school, to work, to shop, or to visit with friends. By the same token, workers do not place themselves in the middle of the morning and evening rush hours because they enjoy traffic congestion but because their work schedules require it. Transportation engineers are among the professionals concerned with accommodating these societal activities by providing efficient ways to satisfy the population's needs for mobility. As used in the foregoing definition of a transportation system, the word *efficient* stands for the balancing of a variety of often conflicting requirements that society in general considers to be important. These requirements include, but are not limited to, cost considerations, convenience, protection of environmental quality, and protection of individual rights, which may have a variable priority, depending on the issue. To be responsive to these needs, transportation engineers often cooperate with other professionals, including economists, planners, and social scientists.

1.1.6 Quantification versus Valuation

Suppose that the following question was posed to a classical physicist and to an Aristotelian philosopher:

> An object is let go from a height of 20 feet directly above the head of a person. What will the value of the object's velocity be at the instant when it comes in contact with the person's head?

It would not be surprising if after mentally applying the appropriate equation, the classical physicist were to reply: "Well . . . the object's velocity will be about 36 feet per second." However, an engineering student may be somewhat surprised at the philosopher's response along this line: "I believe that, to the person, the object's velocity at that instant will be of no value whatsoever."

The difference between the two answers lies in the meaning that each of the respondents attached to the term "value." The philosopher's use of the word is related to the quality of a thing being useful or desirable to someone, or perhaps how much the thing is desirable or undesirable. Clearly, the assignment of such value is subjective: It depends on the *value system* of the person making the assessment. On the other hand, the physicist's response involved an attempt to *quantify* objectively the state of the object's velocity, which is independent of the person who attempts to assess it. Of course, the physicist could have given the wrong answer by either using the wrong equation (i.e., not understanding how gravity works) or making a calculation error when using the right equation.

Engineers often encounter both meanings of value in their work. For example, suppose that an engineer is asked to estimate the reduction in carbon monoxide emissions that would result from a public policy that aims to encourage people to form car pools. Using the best available mathematical formulation of the problem, the engineer would produce an estimate in essentially the same way as the physicist.

Now consider that the implementation of the public policy requires the expenditure of a certain level of funding and that an estimate of this level has been obtained as objectively as the current understanding of the subject allows. Having quantified these estimates does not in itself reveal whether or not implementation of the policy is desirable. Before such a decision can be made, it is necessary to place relative values on the costs associated with the implementation of the policy and on the benefits that will be derived from it. Simply stated, to make this "apples-and-oranges" decision, someone or some group must assess whether reducing pollutant emissions (by x parts per million) is worth the expenditure of y dollars.

In the private sector of the economy people frequently make such judgments based on their own value systems. By contrast, decisions made in the public sector generally involve compromises between the often conflicting values of the groups that constitute the community (e.g., those in the construction industry versus environmentalists vis-à-vis the construction of a freeway).

This book emphasizes the basic methods and techniques that are presently available to the practitioner for the purpose of *quantifying* the impacts or consequences of transportation-related proposals. The chapter on *evaluation* includes some techniques that are often used to aid in the selection of the most suitable course of action from a set of alternatives. The real-world application of these evaluative techniques, however, *presupposes* the existence of a value system. Analysts tend to valuate the consequences of transportation proposals based on their analyses of the economic choices of consumers, or other philosophical perspectives. One of the fundamental purposes of government is to provide the mechanism for the resolution of such differences.

1.2 TRANSPORTATION SYSTEM CLASSIFICATION

1.2.1 Classification Schemes

Transportation systems can be categorized in several ways. For example, they may be classified according to the types of technology they employ, the function or type of service they provide, who owns or is responsible for their implementation and operation, and so forth. Each of these diverse typologies views transportation systems from a different perspective and is useful in making distinctions that are relevant to different types of transportation-related decisions.

The definition of the transportation system given earlier makes a distinction between *passenger* and *freight* transportation. Both are necessary to satisfy human needs and both constitute a significant portion of the U.S. gross national product (GNP). During the past few decades the total U.S. expenditures for passenger and freight transportation have fluctuated, respectively, around 8 and 12% of the GNP [1.2].

The transportation system is further categorized into four major subsystems according to the medium on which the flow elements are supported. These subsystems are commonly referred to as *modes*. Chapter 5 provides an overview of the principal characteristics of modes. It should be understood that this term is also used to make finer distinctions among the various means of travel. For example, driving alone and forming car pools are sometimes considered to be different modes. The four major subsystems are:

1. Land transportation
   a. Highway
   b. Rail
2. Air transportation
   a. Domestic
   b. International
3. Water transportation
   a. Inland
   b. Coastal
   c. Ocean

4. Pipelines
   a. Oil
   b. Gas
   c. Other

1.2.2 Private and Public Transportation

Transportation services are also classified as either *for-hire* or *not-for-hire* services. These categories are also known, respectively, as *public* and *private* transportation, but these terms refer to their availability to the general public and to private parties, respectively, not to their ownership. For example, a city bus system may be owned either privately or publicly. In either case the service provided is public transportation because the system is available for use by the general public. For-hire systems are further classified into *contract carriers* and *common carriers*. The former stand ready to provide service to the public under individual contractual arrangements. Common carriers, on the other hand, generally offer scheduled service and are open to all members of the public willing to pay the posted fare. The terms *mass transportation* or *mass transit* usually refer to the common carriage of passengers. Taxis, car rentals, and certain other individually arranged services belong to the category of contract public transportation.

1.3 THE ROLE OF GOVERNMENT

1.3.1 Governmental Participation

A characteristic of human social organizations is the establishment of a "government," which in an objective sense may be defined as consisting of the rules of conduct, the collective decision-making processes, and the means of enforcing the rules that attempt to impart social and economic order and to maintain the cohesiveness of a society.

A transportation system provides the necessary connectivity that enhances the interaction between people. It is a historical fact that by facilitating the movement of peoples and the spreading of ideas advances in transportation technology have been closely related to the evolution of civilization as we know it. Since ancient times cities have developed in locations that took advantage of the availability of transportation connections such as rivers and protected harbors. The Roman Empire was held together by a very elaborate system of roadways, some of which (e.g., the Appian Way) remain to the present day. Catanese and Snyder [1.3] state that in eighteenth-century England:

> Transportation was the key to industrialization. Unless raw materials could be brought to the factories and finished products distributed to market areas, the industrial revolution could not happen [1.3].

Similarly, the westward expansion in nineteenth-century America would not have been possible without the construction of the transcontinental railroads; many modern American cities have had their origins at the junctures of railroad lines. Because of the profound role that transportation plays in society, governments have always become involved in the provision, operation, and regulation of transportation systems through both the enactment of laws and the establishment of public planning processes.

The specific actions that a government takes at any given time as well as the method by which it chooses to implement those actions reflect the contemporary value system of the society it represents. Conceptually, there exists a continuum of governmental forms ranging from anarchy (i.e., complete lack of governmental intervention in the affairs of people) to totalitarianism (i.e., complete control by government). Actual governmental structures lie somewhere between the two extremes.

The U.S. governmental structure places a high value on individual freedom and civil rights. Individuals and groups are permitted to pursue what they consider to be in their best interests. They are also afforded relatively greater opportunities to vie with others in persuading the government to take actions favoring what they value. Citizen participation is, in fact, a requirement of public planning law. Dissent and difference of opinion are tolerated and permitted to find expression in the political arena.

1.3.2 Instruments of Governmental Involvement

In rough outline, the typical ways by which the government intervenes in the marketplace to accomplish objectives that, in its representational role, the government finds to be in the public interest include *promotion, regulation,* and *investment.* Incidentally, at any given time the meaning of the term *public interest* is largely implicit in the specific actions that the government takes and thus is itself in a state of flux. Also, differences of opinion as to what is in the public interest frequently arise.

Promotion refers to attempts by the government to encourage or discourage certain situations without legally requiring them. An advertising campaign favoring carpooling aimed at reducing rush-hour congestion and obviating the need for costly highway construction or used as a strategy to reduce energy consumption is an example of promotion.

Regulation refers to those government actions that place legal requirements on individuals and firms to satisfy the public interest. Transportation-related examples of regulation include the establishment of automobile bumper standards to reduce fatalities, automobile air-pollution-emission standards to improve environmental quality, and reduced freeway speed limits to conserve energy. Other examples are the regulation of airline route structures to ensure the availability of service to all and the regulation of the rates that trucking companies can charge their customers.

Investment involves the financial support, public financing, or even public ownership of various systems or services. Subsidies to privately owned bus companies to ensure service to mobility-disadvantaged groups, public ownership of highways to maintain a comprehensive level of accessibility, and participation in the construction of airports and harbors are but a few examples of investment actions.

1.3.3 Arguments for and Against Government Intervention

Involvement of the government in the marketplace is predicated on the proposition that the proposed actions are in the public's interest. In other words, there is a justifiable public purpose for action. Often there are those in the community who favor the proposed actions and those who oppose them. In the public debate that ensues these groups present their arguments for or against the actions.

In pure terms, those that are in favor of government intervention typically advance one or both of two types of argument. These are known as the *welfare* argument and the *social-cost* argument. The welfare argument typically supports specific government actions to protect the rights and privileges of *individuals or specific groups of individuals*. The social-cost argument, on the other hand, usually claims that the government should be involved in order either to avoid impacts that would be detrimental to *society at large* or to bring about conditions that benefit everybody.

In his book on government involvement in the area of housing, Friedman [1.4] points out that the two approaches

> . . . are not unique to housing. Social legislation in general is proposed and defended on one or both of these approaches, either that it helps the poor or some worthy class or that by helping the poor [or some worthy class] it helps all of us. Most frequently, perhaps, both justifications are used [1.4].

Consider, for example, a proposal to provide public subsidies for special transportation services to disadvantaged groups such as the poor. The welfare argument favoring such government action would, at its core, make the case that these particular groups of individuals have the same rights to mobility as everyone else and that by taking such action the government ensures that these rights are protected. One social-cost argument, on the other hand, would likely state that by enhancing the mobility of these groups the government would provide them access to employment opportunities that would result in strengthening the economy and thus benefiting society at large.

The basic argument usually advanced against government interference is *the free-market* argument and its variants. In its strict form it states that a free market is the most efficient and fair way to allocate resources. According to this argument, competitive forces in the private sector can result in lower costs than what the public (i.e., government) sector can deliver in the absence of competition. A related assertion in this vein is that government actions force people and other economic entities to act in ways that they do not otherwise judge to be in their best interest. This is considered coercive as it violates the *freedom of choice* of individuals.

Continuing with the example of transportation subsidies to the poor, arguments against the proposal would likely state that the services should be left to the private sector, because if the market demands transportation services to these groups, the private sector can deliver them more efficiently than the government sector. Moreover, by expending "taxpayers' money" for the benefit of a particular group, the government, in effect, forces people to spend their income in ways that they would not otherwise choose.

Even when legislation is enacted favoring some government action, questions relating to the appropriate and equitable degree of intervention also require resolution. Suppose, for example, that those who argued in favor of transportation subsidies to the poor prevailed in the related public debate. Even then, the question of how much subsidization is appropriate needs to be resolved. In other words, a decision must be made as to whether the proposed services should be totally free to the subject groups or whether they should cover only a certain percentage (and specifically what percentage) of the costs associated with providing the service.

1.4 TOOLS AND APPLICATIONS

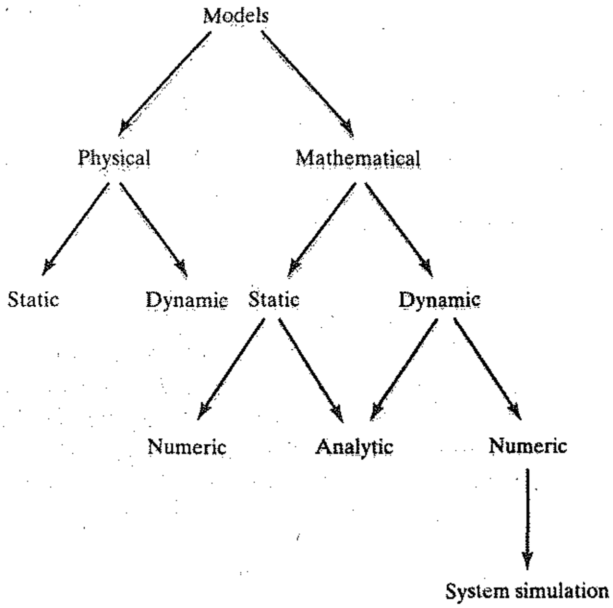
1.4.1 Background

The typical program of study leading to the first course in transportation engineering includes the basic sciences, mathematics, and a range of computer skills. The subject matter of those courses of study stresses the basic tools necessary for work in the field of engineering. The latter differs from the pure sciences in that it is more concerned with the *application* of scientific knowledge. When seeking solutions to real-world problems, questions of economy and other considerations prescribe a need to employ appropriate simplifying assumptions. In order to be useful, such simplifications must render a problem amenable to efficient solution while retaining the essential aspects of the real-world conditions. The importance of making judicious assumptions that are based on a clear understanding of the problem at hand cannot be overemphasized. These assumptions are based on the current state of the art and are themselves subject to change as our understanding of systems is enhanced through additional experience and research. The reader should always be attentive to the fundamental assumptions that are involved in a particular situation and the extent to which these assumptions can affect the results.

1.4.2 Mathematical Models

Transportation engineers and planners employ models to study and analyze the systems of concern. A *model* may be defined as the representation of a part of reality. Figure 1.4.1 shows that models may be classified as *physical* or *mathematical* on one hand and as *static* or *dynamic* on the other [1.5]. Static models represent the structure of a system, whereas

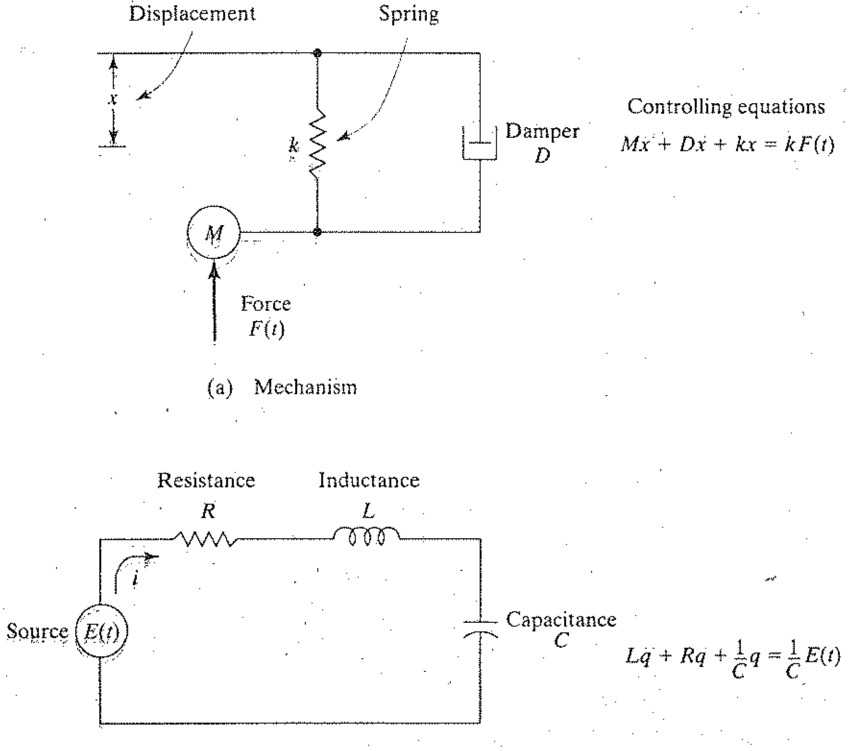
Figure 1.4.1 Types of models. (From Gordon [1.5].)



dynamic models also incorporate a representation of the system's process, that is, the way in which it changes over time. The familiar models of molecular structures consisting of small spheres and pegs are examples of physical static models. Physical dynamic models include wind tunnels, where facsimiles of systems based on the laws of similitude are tested before implementation. These models also include models relying on analogy, such as those representing the vibration of a mechanism via an equivalent electrical circuit (see Fig. 1.4.2). *Analog computers* are well suited for this type of modeling. It is interesting to recognize that the two diagrams shown in Fig. 1.4.2 are, in fact, static representations of a mechanical and an electrical system. When constructed, the electrical circuit can be used as a physical dynamic model of the mechanical system.

A mathematical model employs one or more equations to represent a system and its behavior. Thus Newton's second law and Bernoulli's equation are examples of mathematical models. In addition to being either static or dynamic, mathematical models may be classified according to the method of solution employed, for example, analytical or numerical. Numerical models have proliferated in recent years because they are amenable to solution by *digital computers*.

All models are abstractions of the systems they represent. In other words, a system and its model are not identical in all respects: The model is a simplified representation of



(a) Mechanism

(b) Circuit

Figure 1.4.2 Model analogy.
(From Gordon [1.5].)

the system. Consequently, a number of different models can be used to describe the same system. The appropriate model to a particular endeavor should be selected so as to strike a balance between the ease of application on one hand and the realistic representation of the subject system on the other. The principle of *parsimony,* also known as *Occam's Razor,* states that if two theories (or models) explain the same observations equally well, the simpler (i.e., more parsimonious) of the two is better. For example, Newton's theory is sufficiently accurate for the cases examined in engineering mechanics to be given preference over the more complicated theory of relativity. The same is not true, however, for the study of interplanetary motion.

1.4.3 Components of Mathematical Models

A mathematical model can have one of an infinite number of mathematical forms. It can be linear, nonlinear, exponential, differential, and so forth. Most of the mathematical models that are familiar to readers have been simply presented to them, and therefore the fact that someone (usually the person whose name is associated with the model) had to *postulate* its mathematical form possibly remains hidden. Researchers are constantly faced with the problem of model postulation, and in many situations, including certain areas of transportation, analysts are also required to select the mathematical forms of their models.

Selecting a mathematical form, however, is not the same as having a useful model. Consider, for example, the following mathematical form relating four variables, X, Y, Z, and W:

$$Y = a X^b Z^c W^d \tag{1.4.1}$$

In this equation variable Y is a function of the other three variables. In other words, Y can be computed if the numerical values of X, Z, and W are known. For this reason Y is called the *dependent,* or *explained,* variable and the other three are called the *independent,* or *explanatory,* variables. The model also contains four constants, a, b, c, and d, which must also be known. These constants are referred to as the *parameters* of the model.

Model estimation is the process by which the numerical values of the parameters of a postulated model are determined. It is accomplished through the use of statistical methods and based on *experimental* knowledge, that is, observations, of the dependent and independent variables. This means that once the nature of the variables is identified, a series of experiments are conducted to obtain a set of N simultaneous observations on the dependent and independent variables, in our case Y_i, X_i, Z_i, W_i, with i varying from 1 to N. These observations are then employed to estimate the numerical values of the model parameters that render the postulated model capable of reproducing the experimental data. To avoid misspecification of the mathematical form that relates the variables, alternative functions may be postulated and estimated. After several statistical goodness-of-fit tests, the one that best describes the experimental data can then be selected. In this manner it is ensured that the selected model is realistic. In a strict sense, the term *calibration* refers to procedures that are used to *adjust* the values of a model's parameters to make them consistent with observations. Many authors, however, use the terms *estimation* and *calibration* interchangeably and the practice is followed in this book.

Model validation refers to the testing of a calibrated model using empirical data other than those used to calibrate the model in the first place. This is how scientific theories (models) are tested, modified, or replaced. An improved model must be able to explain everything that the old model can plus something that the old model cannot. As mentioned earlier, the old model may still be a useful tool as long as the scope of its applicability is clearly understood.

To comprehend the importance of the previous modeling steps, consider that X and Z in Eq. 1.4.1 stand for the masses M_1 and M_2 of two bodies, W stands for the distance R separating them, and the dependent variable Y represents their mutual attraction force F. Moreover, assume that the calibration of this postulate vis-à-vis experimental observations has resulted in the following numerical values of the model's parameters:

$$a = K = 3.442(10^{-8})\,\frac{\text{ft}^4}{\text{lb-s}^4}$$

$$b = c = 1 \text{ and } d = -2$$

Equation 1.4.1 becomes

$$F = \frac{KM_1M_2}{R^2} \tag{1.4.2}$$

which is Newton's law of gravitation.

1.4.4 Transportation Models

The study of many physical phenomena, most appropriately, tends to concentrate on systems of physical objects under the influence of forces. Stripped to its essentials, the transportation system involves a physical phenomenon, the motion of flow entities on the fixed facilities. Hence mathematical *models of the physical system* are required, including the equations of motion.

Additionally, transportation engineering must explicitly incorporate the human dimension, which consists of *human factors* and *human behavior*. As used in this book, the term "human factors" refers to those measurable characteristics of human beings that are relatively difficult to modify, such as sensory, perceptual, and kinesthetic characteristics. A driver's perception-reaction time that elapses between the instant when a stimulus is first displayed (e.g., the onset of the yellow phase of a traffic signal) and the instant when a driver reacts to the stimulus (e.g., applying the brakes) is a specific example of a human factor that can be measured in either the field or the laboratory. *Human-factor models* are needed in addition to models of the physical transportation system.

The line of demarcation between human factors and human behavior is not very clear. Nevertheless, the use of the term *human behavior* in this book generally refers to the way in which people act, the types of choices they make, and so forth. Specifically, *travel behavior* includes the way people decide whether and when to travel, where to go, how to get there, and the like. The transportation analyst's arsenal of mathematical models includes *models of human behavior*. Human behavior is in a continuous state of flux and is affected by technological changes: The behavioral patterns of the ancient Greeks were very different from those of modern Americans, but the range of visual perception of the two populations has remained about the same. In fact, no universally applicable calibrated model of travel behavior exists. Even when using the same mathematical form, transportation studies for different cities must estimate their models to conform to local conditions.

1.5 SUMMARY

In this introduction we have defined a transportation system in terms of its fixed facilities, flow entities, and control system and identified its general function as that of providing the necessary connectivity that facilitates other societal activities. A brief outline of several transportation classification schemes were followed. The means by which government participates in the transportation sector (i.e., promotion, regulation, and investment) were presented to place the study of transportation engineering and planning within its larger societal context. The fundamental concepts of mathematical modeling (i.e., model postulation, calibration, and validation) were introduced and the three categories of quantitative models used by transportation engineers and planners (i.e., models of the physical system, human factor models, and travel behavior models) were explained. A distinction between quantification and valuation was also drawn.

REFERENCES

1.1 American Association of State Highway and Transportation Officials, *A Policy on Geometric Design of Highways and Streets,* AASHTO, Washington, DC, 1990.

1.2 Transportation Policy Associates, *Transportation in America, A Statistical Analysis of Transportation in the United States,* 2nd ed., TPA, Washington, DC, 1984.

1.3 Catanese, A. J., and J. C. Snyder, eds., *Introduction to Urban Planning,* McGraw-Hill, New York, 1979.

1.4 Friedman, L. M., *Government and Slum Housing: A Century of Frustration,* Rand McNally & Company, Chicago, 1971.

1.5 Gordon, G., *System Simulation,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1978.

PART 1

Design and Operation

2

Roadway Design

2.1 INTRODUCTION

A fundamental characteristic of any transportation system is the motion of vehicles. It is, therefore, important to review the basic *kinematic* and *kinetic* equations of motion. These two branches of dynamics are related: *Kinematics* is the study of motion irrespective of the forces that cause it, whereas *kinetics* accounts for these forces. The motion of a body can be *rectilinear* or *curvilinear* and can be investigated in relation to a fixed coordinate system (i.e., *absolute* motion) or in relation to a moving coordinate system (i.e., *relative* motion). In this chapter the basic equations of motion of a single vehicle are cast in the form in which they are used for the purposes of design. Three examples of human factor models are then introduced to illustrate how human characteristics can be incorporated in design. The important differences between the *maximum* technological capabilities of vehicles on one hand and the *practical* design levels necessitated by considerations of passenger *safety* and *comfort* on the other are explained. Finally, the methods of *geometric design* are presented for the horizontal and vertical alignment of highways and for the channelization of intersections and interchanges. Geometric design is concerned only with the size, shape, and geometry of transportation facilities. Other aspects of engineering design are also necessary for the successful implementation of these facilities, including the proper selection of pavement and structural materials, geotechnical design, and the structural design of various components such as bridges. In addition, complex decisions related to whether a new facility is needed, or whether an existing facility should be improved, must precede the design phase.

2.2 EQUATIONS OF MOTION

2.2.1 Rectilinear Motion

The rectilinear *position* x of a particle is measured from a reference point and has units of length. The displacement of the particle is the difference in its position between two instants.

Velocity v is the displacement of the particle divided by the time over which the displacement occurs. In the limit the instantaneous value of velocity is given by the first derivative of displacement with respect to time:

$$v = \frac{dx}{dt} \tag{2.2.1}$$

Speed is a scalar quantity and is equal to the magnitude of velocity, which is a vector. However, the two terms are used interchangeably in this book when the meaning is clear from the context.

Acceleration a is the rate of change of velocity with respect to time:

$$a = \frac{dv}{dt} \tag{2.2.2}$$

It can be positive, zero, or negative. Negative acceleration (i.e., *deceleration*) is often denoted by d, and its magnitude is given in the positive. Thus a deceleration of 16 ft/s^2 is the same as an acceleration of -16 ft/s^2.

Applying the chain rule yields

$$a = \frac{dv}{dx}\left(\frac{dx}{dt}\right),$$

or

$$a = \left(\frac{dv}{dx}\right)v$$

which leads to

$$v\,dv = a\,dx \tag{2.2.3}$$

Often the given variables are expressed as functions of time or of each other [e.g., $a = f(v)$], and the specific relationships between pairs of variables are derived through the application of the calculus. These relationships are frequently plotted to aid the visualization of the particle's motion.

The simplest case of rectilinear motion is the case of *constant acceleration*, where

$$\frac{dv}{dt} = a = \text{constant}$$

Separating variables and integrating over the limits $t = 0$ to t gives

$$\int_{v_0}^{v} dv = \int_0^t a\,dt$$

$$v = at + v_0 \tag{2.2.4}$$

The velocity of the particle can also be expressed as a function of distance by integrating Eq. 2.2.3 over the appropriate limits of integration to yield

$$\tfrac{1}{2}(v^2 - v_0^2) = a(x - x_0) \tag{2.2.5}$$

which, upon the rearrangement of terms, becomes

$$x - x_0 = \frac{v^2 - v_0^2}{2a} \tag{2.2.6}$$

This expression is useful for computing the distance traveled by a vehicle at constant acceleration (or deceleration) from an initial velocity v_0 to a final velocity v.

In view of Eq. 2.2.1, Eq. 2.2.4 may be integrated to express x as a function of time:

$$x = \frac{1}{2}at^2 + v_0t + x_0 \tag{2.2.7}$$

Example 2.1: Constant Acceleration

A vehicle approaches an intersection at 30 mi/h. At time $t = 0$ it begins to decelerate at $d = 16$ ft/s^2. Calculate the time it would take the vehicle to stop. Given that at the beginning of deceleration the vehicle was located 55 ft away from the stopping line, determine whether it was able to stop legally (i.e., behind the stopping line). Plot the relationships of acceleration, velocity, and position as functions of time and the relationship between velocity and position.

Solution Set the positive x-axis in the direction of motion, with the origin at the initial position of the vehicle. Thus at $t = 0, x = 0, v = 44$ ft/s, and $a = -16$ ft/s^2. This is the case of constant acceleration. The time it took the vehicle to stop from an initial velocity of 44 ft/s is given by Eq. 2.2.4:

$$0 = -16t + 44 \qquad \text{or} \qquad t = 2.75 \text{ s}$$

The distance covered during deceleration may be computed from either Eq. 2.2.6 or Eq. 2.2.7:

$$x = 60.5 \text{ ft}$$

Because this is greater than the available distance of 55 ft, the vehicle was not able to stop before reaching the stopping line. The required plots are shown in Fig. 2.2.1.

Discussion Since acceleration is the derivative of velocity with respect to time, the slope of the $v-t$ diagram is equal to the acceleration at time t. In this example the slope of the $v-t$ diagram is negative, constant, and equal to -16 ft/s^2. Similarly, the particle's velocity is equal to the slope of the $x-t$ diagram. Thus when velocity is equal to zero, the $x-t$ diagram attains a critical point, in this case a maximum. The plots of the mathematical functions derived earlier extend beyond the instant when the vehicle comes to a complete stop (see dashed lines). This range is not applicable in this situation because the subject vehicle will not continue to decelerate at 16 ft/s^2 beyond that instant, that is, reverse its direction by backing up! The assumption of an "average" constant deceleration is an idealized but often acceptable approximation.

Example 2.2: Acceleration as a Function of Velocity

The acceleration of a vehicle from an initial speed v_0 is given by the relationship

$$a = \frac{dv}{dt} = A - Bv \tag{2.2.8}$$

where A and B are constant [2.1]. Derive and plot the $x-t$, $a-t$, and $v-t$ relationships assuming that at $t = 0, x = 0$.

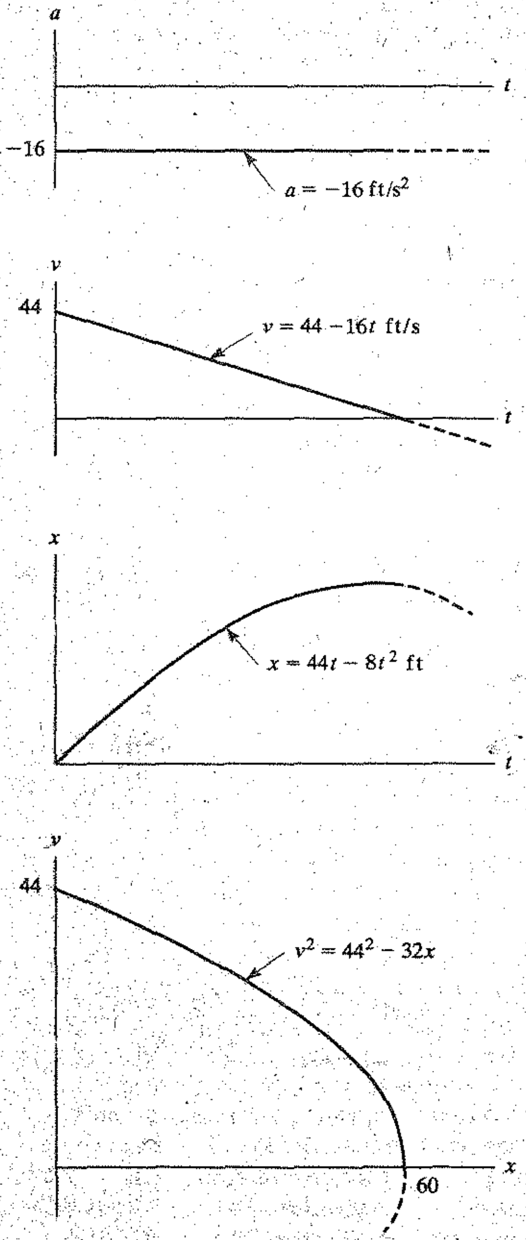


Figure 2.2.1 Examples of rectilinear motion relationships.

Solution This is a case of variable acceleration. Consequently, the equations employed in Example 2.1 do not apply. Separating the variables of Eq. 2.2.8 and integrating over the appropriate limits, we obtain

$$\int_{v_0}^{v} \frac{dv}{A - Bv} = \int_0^t dt$$

or

$$\frac{-1}{B} \ln(A - Bv)\Big|_{v_0}^{v} = t$$

and

$$\frac{A - Bv}{A - Bv_0} = e^{-Bt}$$

Solving for v as a function of t yields

$$v = \frac{A}{B}(1 - e^{-Bt}) + v_0 e^{-Bt}$$

Substitution of this expression into Eq. 2.2.8 results in the needed relationship of acceleration as a function of time:

$$a = (A - Bv_0)e^{-Bt}$$

Finally, substituting $v = dx/dt$ into the $v-t$ equation and integrating leads to the following $x-t$ relationship:

$$x = \left(\frac{A}{B}\right)t - \frac{A}{B^2}(1 - e^{-Bt}) + \frac{v_0}{B}(1 - e^{-Bt})$$

The required plots are shown in Fig. 2.2.2.

Discussion Equation 2.2.8 closely approximates the situation where a vehicle traveling at an initial speed v_0 attempts to accelerate as quickly as possible to its maximum speed ("pressing the accelerator to the floor"). Examination of the value of acceleration at $t = 0$ shows that the initial value of acceleration depends on the initial speed. Moreover, A has the dimensions of acceleration and is the maximum acceleration that the vehicle can attain starting from rest (i.e., $v_0 = 0$). Theoretically, the maximum speed attainable by the vehicle is A/B. This can be verified by examining the $v-t$ curve in the limit. The values ~~of~~ the constants A and B depend on the technological design of the subject vehicle and can be measured experimentally. Dimensional consistency requires B to have ~~the units~~ of 1/time. The equations developed so far are based on kinematics. They are used ~~in the~~ study of kinetics where the relationship among force, mass, and acceleration is of prime concern. Newton's second law provides the fundamental equation relating the three variables:

$$F = ma \tag{2.2.9}$$

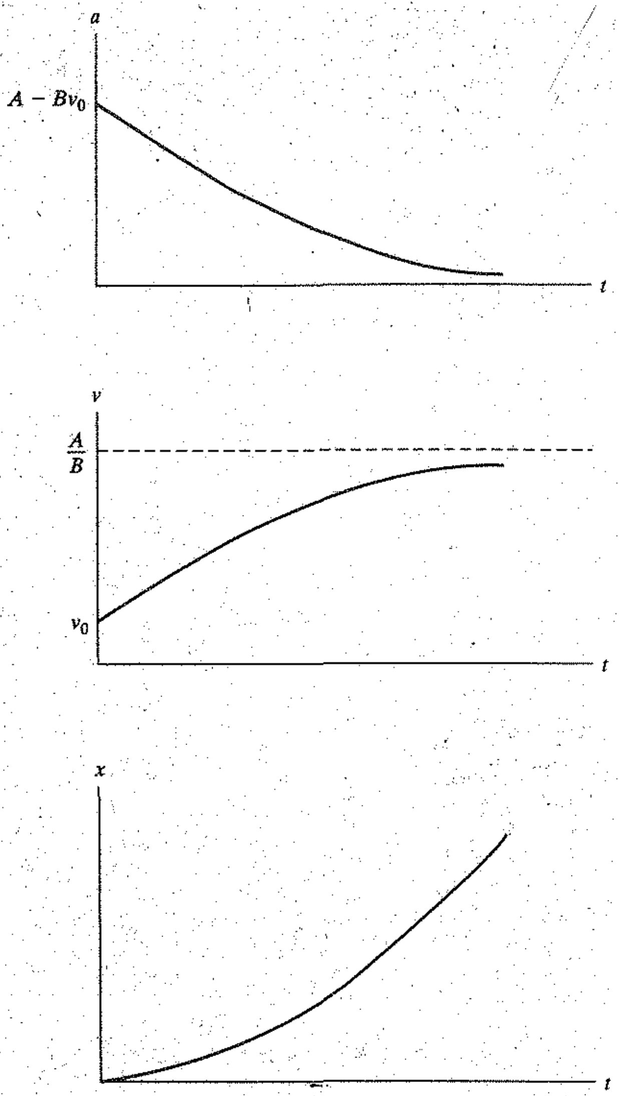


Figure 2.2.2 Acceleration, speed, and distance as a function of time.

2.2.2 Braking Distance

A very common case of rectilinear motion is the case of a vehicle braking on a grade, that is, while moving either uphill or downhill. Figure 2.2.3 shows the major forces acting on a vehicle as it climbs uphill. Ignoring all resistances except friction and grade resistance, the free-body diagram of the vehicle becomes that as shown in Fig. 2.2.4.

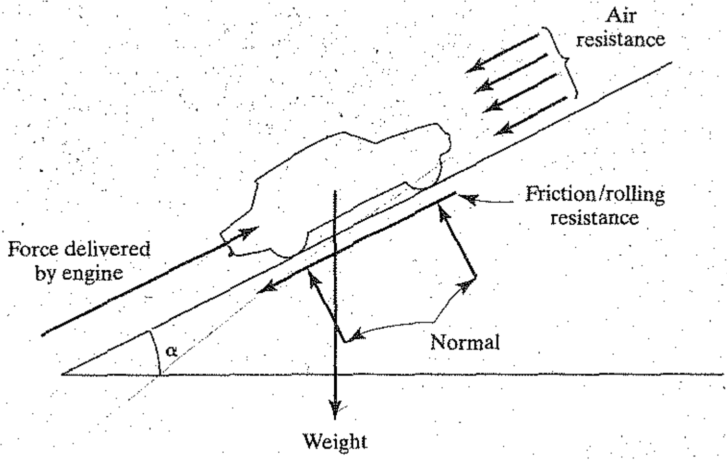


Figure 2.2.3 Forces acting on a moving vehicle.

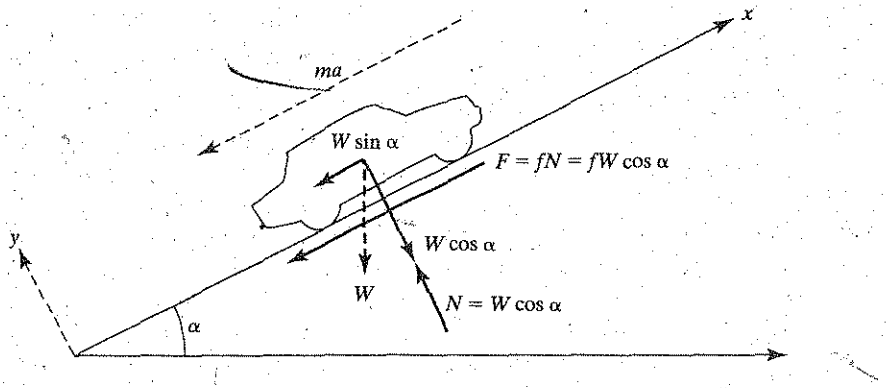


Figure 2.2.4 Vehicle braking.

It is customary to designate the *braking distance* D_b in the horizontal direction rather than along the incline, which is taken as the x-axis. For small incline angles α the difference between the two distances is very small, as their relationship verifies:

$$D_b = x \cos \alpha \tag{2.2.10}$$

where both x and D_b are measured from the point at which braking commences.

The condition of static equilibrium is satisfied in the y-direction with the y-component of the vehicle's weight counteracted by the normal force $N = W \cos \alpha$. Equation 2.2.9 in the x-direction yields

$$\left(\frac{W}{g}\right)a + Wf \cos \alpha + W \sin \alpha = 0 \tag{2.2.11}$$

Substituting Eq. 2.2.10 into Eq. 2.2.6 and solving for acceleration, we have

$$a = (v^2 - v_0^2)\frac{\cos\alpha}{2D_b} \tag{2.2.12}$$

Substituting Eq. 2.2.12 in Eq. 2.2.11, dividing by $W\cos\alpha$, and solving for D_b gives

$$D_b = \frac{v_0^2 - v^2}{2g(f + G)} \tag{2.2.13}$$

where $G = \tan\alpha$, or the *percent grade* divided by 100.

Repeating the earlier solution for the case when a vehicle is braking while traveling downhill yields an equation identical to Eq. 2.2.13 except for a reversal of the sign of G in the denominator. Equation 2.2.13 is often expressed to cover both situations as

$$D_b = \frac{v_0^2 - v^2}{2g(f \pm G)} \tag{2.2.14}$$

where the plus and minus signs correspond to *uphill* and *downhill* braking, respectively. It is easy to remember the proper sign by remembering that the uphill braking distance is shorter than the downhill braking distance because of the effect of gravity.

To compute the total braking distance from an initial speed v_0 to a complete stop, simply substitute $v = 0$. For level paths the gradient G is equal to zero. The fact that Eq. 2.2.6 was used in the development of the preceding model implies that the acceleration rate is assumed to be constant. This assumption is reflected in the value of the coefficient of friction f, which is considered to represent the average effect of friction during the entire braking maneuver. The coefficient of friction, of course, depends on the characteristics of the contacting surfaces, that is, the vehicle's tires and the pavement. Because of a wide range of possible pavement–tire combinations and conditions, the coefficient of friction is often calculated experimentally as follows: At the location of interest where G is known, the braking distance needed to stop a vehicle from a known speed is measured. These values are then substituted in Eq. 2.2.13 to obtain the value of f. For the purposes of statistical reliability, the test is repeated a number of times. As a rule of thumb, f is approximately equal to 0.6 when the pavement is dry and about 0.3 when the pavement is wet. On ice, of course, f is much lower. The coefficient of friction has also been found to decrease somewhat with increasing initial speed [2.2]. Engineering design is normally based on wet rather than dry conditions.

Equation 2.2.14 can also aid in the estimation of initial speed v_0 at which a vehicle was traveling prior to a collision based on the length of the skid marks left on the pavement. However, the speed at impact v must also be estimated. This is accomplished by considering the kinetic energy dissipated for the damage or deformation sustained by the vehicle(s) involved in the collision.

Example 2.3: Braking Distance

A driver of a car applied the brakes and barely avoided hitting an obstacle on the roadway. The vehicle left skid marks of 88 ft. Assuming that $f = 0.6$, determine whether the driver was in violation of the 45-mi/h speed limit at that location if she was traveling (a) uphill on a 3° incline, (b) downhill on a 2.3° incline, or (c) on a level roadway. Also, compute the average deceleration developed in each case.

Solution The stopping distance D_b is computed from the length of the skid marks using Eq. 2.2.10, and the initial velocity is calculated by Eq. 2.2.14 because the final velocity is zero in all three cases. The kinematic relationship of Eq. 2.2.6 can then be solved to compute the corresponding deceleration:

Case	G	D_b (ft)	v_0 (ft/s)	d (ft/s^2)
(a)	0.05	87.88	60.65	20.90
(b)	0.04	87.93	56.30	18.03
(c)	0.00	88.00	58.31	19.32

Because the speed limit was 45 mi/h, or 66 ft/s, the driver was not speeding in any of the three cases.

Discussion The kinematic equation 2.2.6 and the kinetic equation 2.2.14 describe the same phenomenon. Comparison of these equations shows that the deceleration of the braking vehicle can be expressed in terms of two components: The first is due to the friction developed between the tires and the pavement and the second is due to the effect of grade. The difference between D_b and the distance traveled along the incline x is very small for typical highway grades. Finally, the coefficient of friction given in this problem implies a dry pavement.

2.2.3 Curvilinear Motion

Vehicles do not traverse straight paths exclusively but must also negotiate curved paths, as illustrated in Fig. 2.2.5(a). The figure shows that the direction of velocity is always tangent to the path. The vehicle's acceleration may be resolved into two components in the tangential and normal directions, respectively. The magnitude of the tangential component is

$$a_t = \frac{dv}{dt} \tag{2.2.15}$$

The normal component of the acceleration acts toward the center of curvature and has a magnitude of

$$a_n = \frac{v^2}{\rho} \tag{2.2.16}$$

where ρ is the radius of curvature of the path. For a constant velocity v the tangential component of the acceleration vanishes but the normal component remains. For circular paths the radius of curvature is constant and equal to the radius of the circular path R.

Figure 2.2.5(b) shows the tangential and normal components of the forces acting on a vehicle as it traverses a curved path. Applying Newton's second law to the two directions, we have

$$\Sigma F_t = m\left(\frac{dv}{dt}\right) \tag{2.2.17}$$

and

$$\Sigma F_n = \frac{mv^2}{\rho} \tag{2.2.18}$$

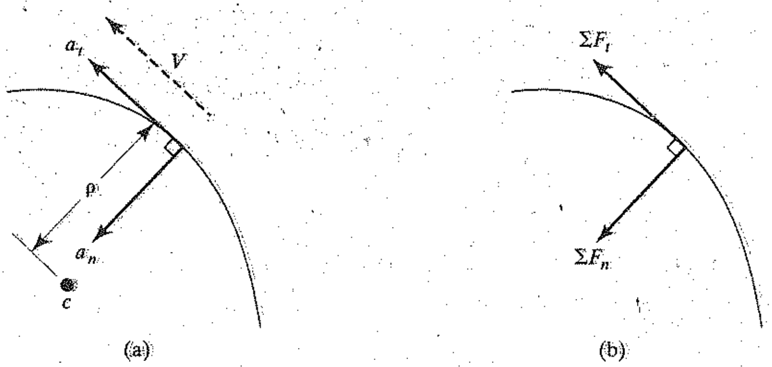


Figure 2.2.5 Curvilinear motion.

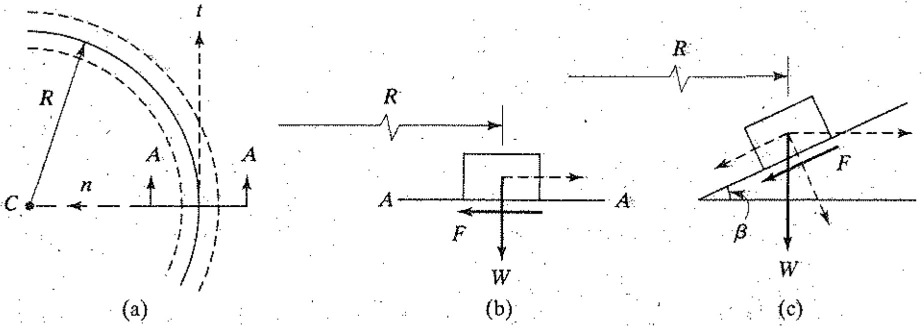


Figure 2.2.6 Lateral effect.

For a horizontal roadway cross section *AA*, as shown in Fig. 2.2.6(b), the only force in the normal direction is due to the *side friction* between the vehicle's tires and the pavement, which resists the tendency of the vehicle to slide. To minimize this tendency, highway design provides for the banking, or *superelevation*, of the cross section of the roadway, as shown in Fig. 2.2.6(c): The cross section is tilted by an angle β so that the component of the vehicle's weight along the tilted pavement surface also resists the sliding tendency of the vehicle. This effect is extremely pronounced in the design of car-racing tracks because of the large normal accelerations developed at racing speeds. The slope to which a highway cross section is tilted is known as the *rate of superelevation*. It is denoted by the letter *e* and equals the tangent of the angle β. The banking angle β should not be confused with the grade angle α discussed in Section 2.2.2.

Figure 2.2.7 shows the free-body diagram of the vehicle as it travels along a circular path at the verge of sliding. In the *y*-direction perpendicular to the surface of the pavement, static equilibrium yields $N = W\cos\beta + ma_n \sin\beta$. Writing Newton's second law for the *x*-direction gives

$$W\sin\beta + f_s W\cos\beta + \frac{Wv^2}{gR} = ma_n \cos\beta$$

$$= \frac{W}{g}\left(\frac{v^2}{R}\right)\cos\beta$$

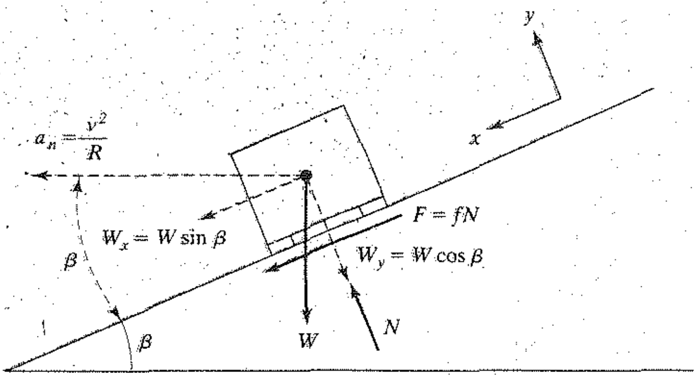


Figure 2.2.7 Free-body diagram of turning vehicle.

Dividing both sides by $W \cos \beta$ yields

$$e + f_s = \frac{v^2}{gR}(1 - f_s e) \qquad (2.2.19)$$

where $e = \tan \beta$ and f_s is the *coefficient of side friction*. For typical highway conditions $f_s\, e$ is close to zero and may be dropped.

Example 2.4: Curvilinear Motion

A 2000-lb vehicle is traveling along a horizontal circular path of radius $R = 500$ ft. At the instant of interest the vehicle is traveling at 88 ft/s while decelerating at 8 ft/s^2. Determine the total horizontal force acting on the vehicle.

Solution As Fig. 2.2.5(b) illustrates, the total force can be resolved into a tangential and a normal component as follows:

$$F_t = ma_t = \frac{2000}{32.2}(-8) = -497 \text{ lb}$$

and

$$F_n = ma_n = m\left(\frac{v^2}{R}\right) = \frac{2000}{32.2}\left(\frac{88^2}{500}\right) = 962 \text{ lb}$$

The total horizontal force F is 1083 lb, as shown in Fig. 2.2.8.

Discussion The direction of the tangential force in this case is in the negative direction because the vehicle is decelerating. The normal force, however, is still in the direction toward the center of curvature. The total force, of course, is given by the vector addition of the two components.

Example 2.5: Superelevation

A vehicle is traveling along a horizontal circular curve of radius $R = 1000$ ft at the legal speed limit of 60 mi/h. Given that the coefficient of side friction is 0.2, determine the angle β at which the pavement should be banked to avoid outward sliding.

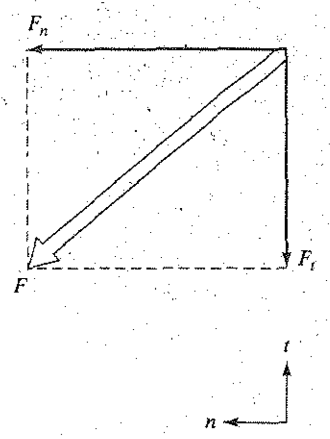


Figure 2.2.8 Resultant force.

Solution From Eq. 2.2.19

$$e = \frac{v^2}{gR} - f_s = 0.04 \text{ ft/ft}$$

Hence $\tan \beta = e = 0.04$ and $\beta = 2.3°$.

Discussion The friction developed between the tires and the pavement is aided by gravity so that the curve can be safely negotiated at 60 mi/h, or 88 ft/s. Solving Eq. 2.2.19 for v with $e = 0$ shows that without superelevating the pavement cross section, the maximum safe speed would be 80 ft/s, or about 55 mi/h.

When the center of gravity of the vehicle is high above the pavement, there exists the danger of overturning; Fig. 2.2.9 is used to examine this situation. Note that at the instant when overturning or tipping is imminent, the normal force is acting on the outside wheel idealized by point A. The location of the center of gravity of the vehicle is given by $\bar{x} = X$ and $\bar{y} = Y$ in relation to point A. Taking moments about point A gives

$$XW \cos \beta + YW \sin \beta = \left(\frac{Y}{e} - X\right)(\sin \beta) m a_n$$

Dividing both sides by $W \cos \beta$ and rearranging terms, we have

$$\frac{v^2}{gR} = \frac{X + Ye}{Y - Xe} \tag{2.2.20}$$

Example 2.6: Slipping and Overturning

A truck with a center of gravity at $X = 4$ ft and $Y = 5$ ft is traveling on a circular path of radius $R = 600$ ft and superelevation $e = 0.05$. Determine the maximum safe speed to avoid both slipping and overturning, assuming that the coefficient of side friction is 0.2.

Solution Equation 2.2.19 applies to slipping. The maximum speed to avoid slipping is

$$v^2 = gR(e + f) \quad \text{or} \quad v = 69.5 \text{ ft/s}$$

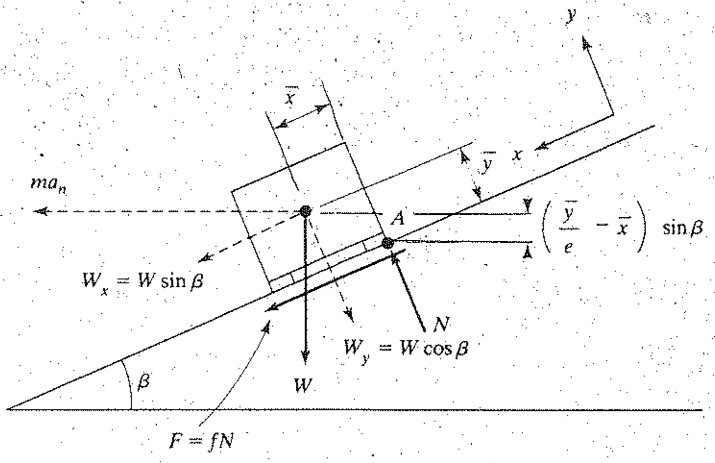


Figure 2.2.9 Case of overturning.

The maximum speed to avoid overturning is given by Eq. 2.2.20:

$$v = 130.8 \text{ ft/s}$$

The maximum safe speed is 69.5 ft/s, the smaller of the two.

Discussion The proper design of highways involves the selection of a design speed, the radius of curvature, and the superelevation rate. Determining the chance of overturning requires knowledge of the dimensions of the vehicles using the roadway. If the dimensions of vehicles change subsequent to the construction of the roadway, the highway engineer is left with a number of choices, including roadway reconstruction, changing the speed limit, or prohibiting certain vehicles from using the roadway.

2.2.4 Relative Motion

It is often practical to examine the motion of one particle in relation to another. For example, the motion of vehicles on a highway may be studied from the point of view of the driver of a moving vehicle. The simplest case of *relative motion* involves the motion of one particle B relative to a coordinate system (x, y, z) that is *translating* but not rotating with respect to a fixed coordinate system (X, Y, Z), as shown by Fig. 2.2.10.

The relationship between the position vectors of the two particles in relation to the fixed system, $\mathbf{r}_A$ and $\mathbf{r}_B$, and the position vector $\mathbf{r}_{B/A}$ of B with respect to the moving particle A is

$$\mathbf{r}_B = \mathbf{r}_A + \mathbf{r}_{B/A} \qquad (2.2.21a)$$

Differentiating with respect to time gives

$$\mathbf{v}_B = \mathbf{v}_A + \mathbf{v}_{B/A} \qquad (2.2.21b)$$

and

$$\mathbf{a}_B = \mathbf{a}_A + \mathbf{a}_{B/A} \qquad (2.2.21c)$$

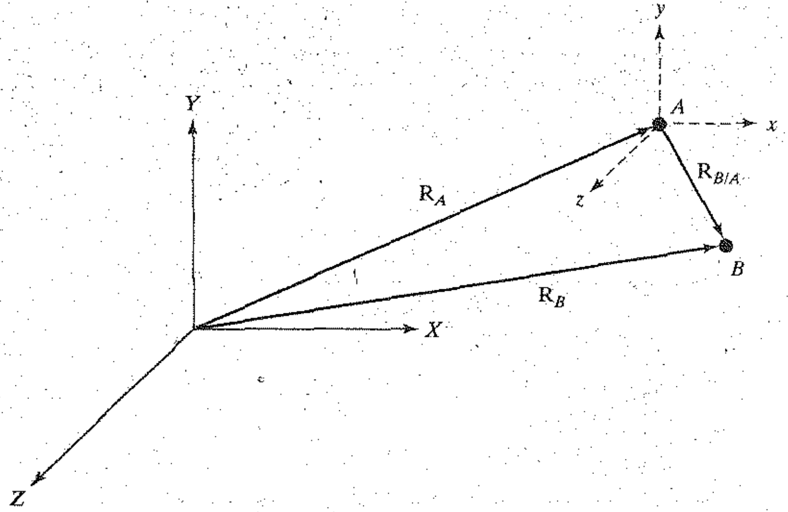


Figure 2.2.10 Relative position of two particles.

Example 2.7: Relative Motion

A police car, *A*, equipped with a radar capable of measuring the relative speed and the relative acceleration between it and another vehicle, *B*, is following a suspected speeding vehicle in a 40-mi/h straight roadway (Fig. 2.2.11). At the instant of interest the police car is accelerating at 8 ft/s^2 from a speed of 50 mi/h. The radar reads $v_{B/A} = -5$ mi/h and $a_{B/A} = -16$ ft/s^2. Determine the absolute speed and acceleration of the vehicle *B*.

Solution

$$v_{B/A} = v_B - v_A$$

or

$$-5 = v_B - 50 \quad \text{and} \quad v_B = 45 \text{ mi/h}$$

Similarly,

$$a_{B/A} = a_B - a_A$$

or

$$-16 = a_B - 8 \quad \text{and} \quad a_B = -8 \text{ ft/s}^2$$

Discussion The driver, *B*, was going 5 mi/h above the speed limit and was decelerating at 8 ft/s^2, perhaps to minimize the consequences of the transgression.

Example 2.8: Polar Coordinates

Car *A* is traveling at $v = 88$ ft/s. At the instant shown in Fig. 2.2.12, $d\mathbf{r}/dt = -25.4\mathbf{n}_r$ and $d\boldsymbol{\theta}/dt = 1.47\mathbf{n}_\theta$ rad/s, where $\mathbf{n}_r$ and $\mathbf{n}_\theta$ are unit vectors in the *r*- and θ-directions, respectively. Determine the absolute velocity of car *B*.

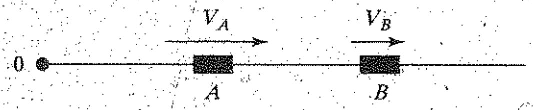


Figure 2.2.11 Relative speeds of vehicles.

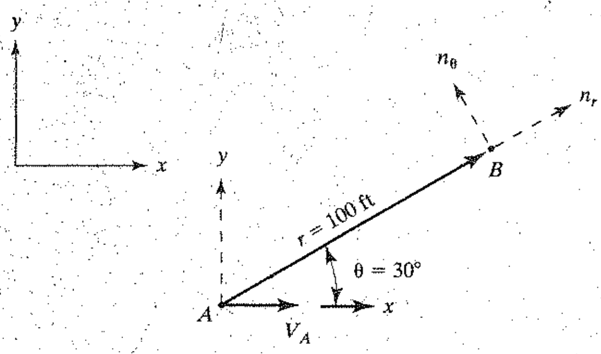


Figure 2.2.12 Polar coordinates *r* and θ.

Solution 1 The relative velocity expressed in the *r*- and θ-directions is

$$\mathbf{v}_{B/A} = \left(\frac{dr}{dt}\right)\mathbf{n}_r + \left(\frac{r\,d\theta}{dt}\right)\mathbf{n}_\theta$$

or

$$\mathbf{v}_{B/A} = -25.4\mathbf{n}_r + (100)(0.147)\mathbf{n}_\theta$$

Because

$$\mathbf{n}_r = \cos 30°\mathbf{i} + \sin 30°\mathbf{j} = 0.866\mathbf{i} + 0.500\mathbf{j}$$

and

$$\mathbf{n}_\theta = -\sin 30°\mathbf{i} + \cos 30°\mathbf{j} = -0.500\mathbf{i} + 0.866\mathbf{j}$$

where **i** and **j** are unit vectors in the *x*- and *y*-directions,

$$\begin{aligned}\mathbf{v}_{B/A} &= -25.4(0.866\mathbf{i} + 0.500\mathbf{j}) + 14.7(0.500\mathbf{i} + 0.866\mathbf{j})\\ &= -29.3\mathbf{i}\end{aligned}$$

Hence

$$\mathbf{v}_B = \mathbf{v}_A - \mathbf{v}_{B/A} = 88\mathbf{i} - 29.3\mathbf{i} = 58.7\mathbf{i}$$

Solution 2

$$\mathbf{r}_{B/A} = r(\cos\theta)\mathbf{i} + r(\sin\theta)\mathbf{j}$$

Recalling that for a purely translating frame, $d\mathbf{i}/dt = d\mathbf{j}/dt = 0$, we have

$$\begin{aligned} \mathbf{v}_{B/A} &= \frac{d\mathbf{r}_{B/A}}{dt} \\ &= (-r \sin \theta\theta' + r' \cos \theta)\mathbf{i} + (-r \cos \theta\theta' + r' \sin \theta)\mathbf{j} \\ &= 29.3\mathbf{i} \end{aligned}$$

as before.

2.3 HUMAN FACTORS

2.3.1 Perception-Reaction

The equations developed so far are based purely on the equations of motion without taking into account the effect of driver performance on the motion described. For example, Eq. 2.2.14 gives the braking distance for a vehicle from the moment when the brakes take effect to the moment when the vehicle reaches its final speed. Normally a driver undertakes such a maneuver in response to a stimulus, for example, avoiding an object on the roadway. When a stimulus appears, a driver requires a certain amount of time to perceive and comprehend it, to decide on the appropriate response, and to react accordingly. The vehicle *braking* distance or time constitutes only a portion of the overall *stopping* distance or time. In many applications the overall maneuver may be divided into two parts: *perception-reaction,* which includes the occurrences up to the beginning of the vehicular response, and braking, which is described by the equations of motion developed in the previous section. If a driver takes 1.5 s to perceive and react to a hazard in the vehicle's path at a speed of 60 mi/h (88 ft/s), the vehicle would cover 132 ft before the braking phase begins.

Figure 2.3.1 presents the findings of a study conducted by Johannson and Rumar [2.3] regarding driver response times to anticipated braking. The continuous curve at the low end of the histogram represents the reaction time of the person who took the measurements and which was accounted for in computing the driver data shown. Johannson and Rumar also found that the response times were longer than those shown when the drivers were surprised. The figure illustrates the presence of considerable variability between individuals. In order to enhance safety, engineering designs that incorporate driver characteristics are typically based on values in the 85th to 95th percentile range.

Driver response is related to driver characteristics and conditions, such as age, medical condition, alcohol and drug use, fatigue, sleep deprivation, and emotional condition. It also depends on the complexity of the stimulus and the complexity of the required response. Good traffic engineering designs attempt to minimize the stimuli and driving tasks to which the driver must attend at the same time. This is the "one task at a time" rule, which, due to the complexity of the driving environment, is not always possible.

Figure 2.3.2 offers two useful insights. The first is that the time to react to unexpected information is clearly longer than the time to react to expected information (e.g., since the traffic signal ahead turned red, the vehicles ahead are expected to slow down and stop). The other insight is that the complexity of the given information has a positive relationship with reaction time (and conceivably a positive relationship with accident risk). Thus the larger the quantity is and the more complex the information is, the longer it will take drivers to

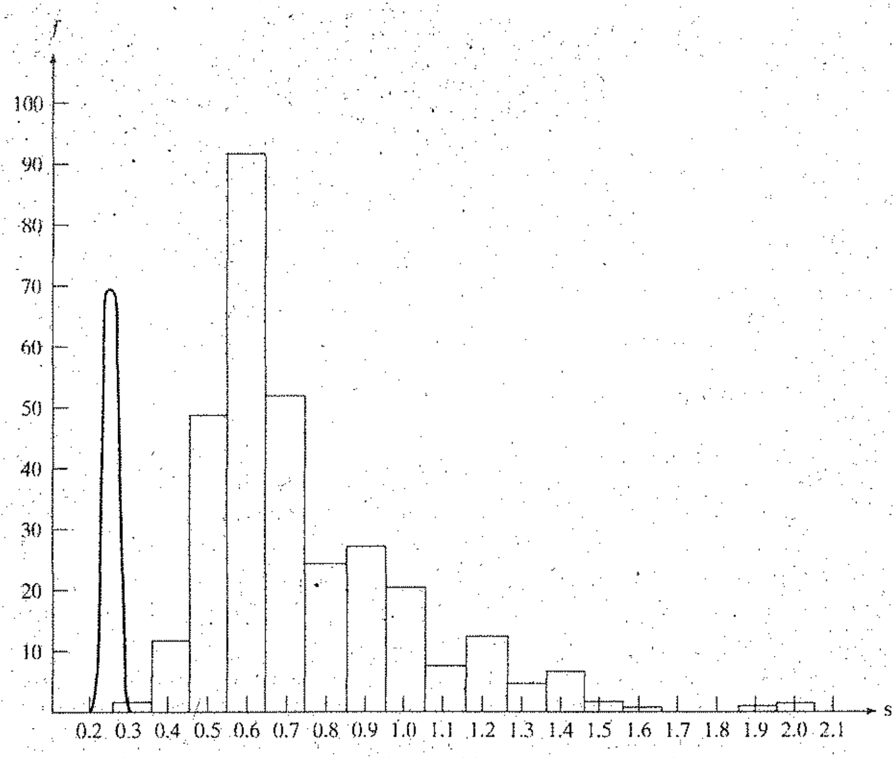


Figure 2.3.1 Distribution of brake reaction times. (From Johannson and Rumar [2.3].)

comprehend the information and react accordingly. This principle is useful to remember when placing regulatory or information traffic signs: They should be clear and properly spaced to avoid giving drivers too much information at a time (this is euphemistically called the *information pollution* phenomenon).

Example 2.9

Using the data of Example 2.3, determine the stopping distances horizontally (D_s) and along the pavement (X_s) in each of the three cases, given that the driver's perception-reaction time was $\delta = 1.5$ s.

Solution The distance traveled during the perception-reaction time must be added to the braking distance to compute the total stopping distance. *Assuming* that the vehicle was not accelerating during the time interval δ, the distances traveled during δ were $X_r = \delta\, v_0$ along the pavement and $D_r = X_r \cos \alpha$ horizontally.

Case	D_s (ft)	X_s
(a)	178.73	178.98
(b)	172.31	172.45
(c)	175.47	175.47

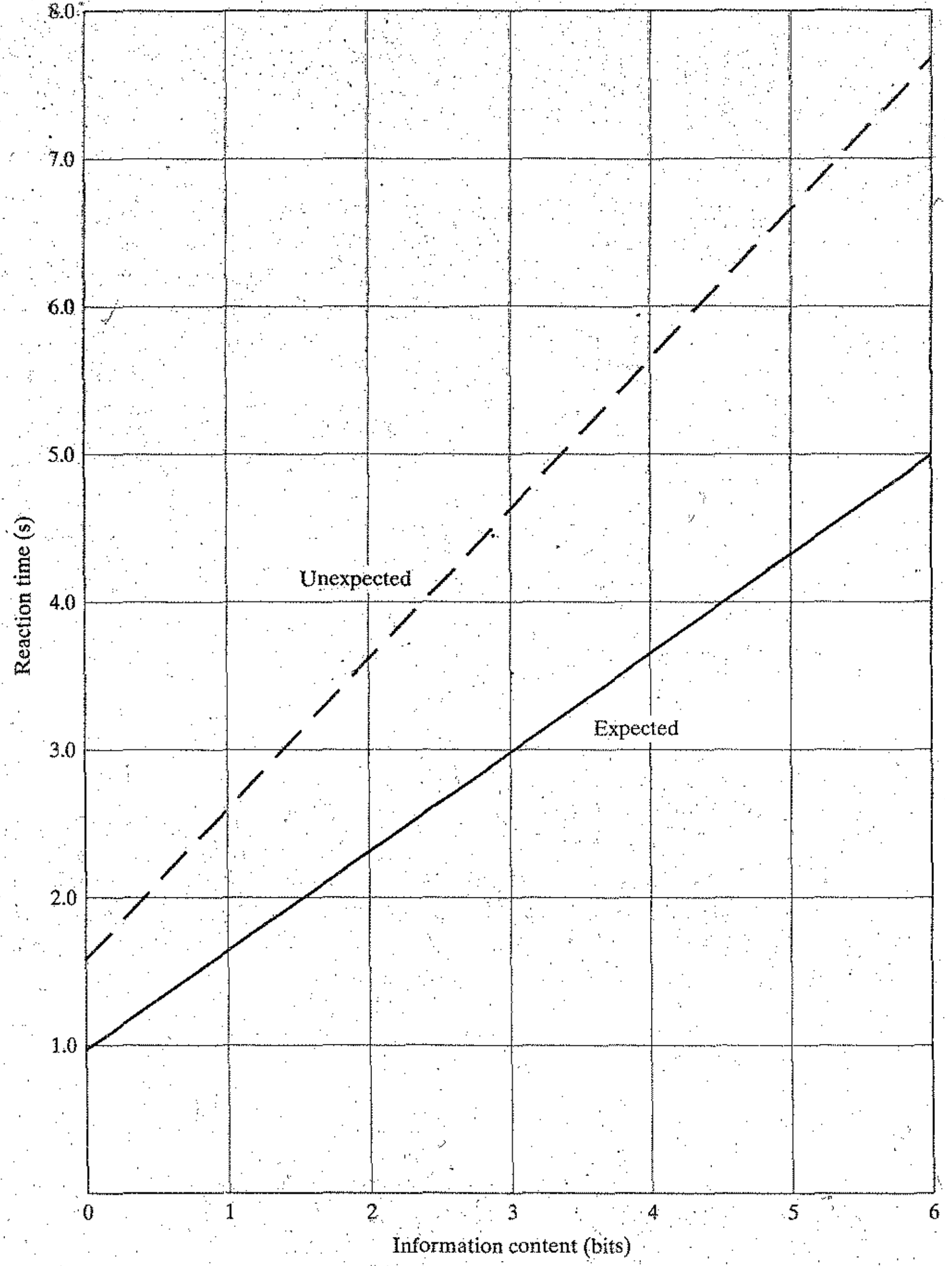


Figure 2.3.2 Eighty-fifth percentile driver reaction time to expected and unexpected information.
(From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2-2] (Fig. 11-19, p. 48). Used by permission.)

Discussion The difference between the two distances D_s and X_s is insignificant considering typical highway grades. The reason computed stopping distances are longer for steeper grades is that of differences in the initial speeds required to stop within the 88 ft specified *in this case.* The assumption of constant speed prior to the stopping maneuver has an effect on the results.

2.3.2 Dilemma Zones

Most probably, the reader has encountered the situation of approaching a signalized intersection just when the traffic signal turned yellow and has faced the decision of whether to apply the brakes in order to stop for the red signal or to attempt to clear the intersection on yellow. On occasion the reader may have felt that it was impossible to execute safely either maneuver. The duration of the yellow phase of the traffic signal, τ, is related to this situation. A properly selected yellow duration that incorporates the motion of the vehicle during the driver's perception-reaction time can eliminate this problem, and a design formula has been developed by Gazis et al. [2.4] as follows.

Figure 2.3.3 shows a vehicle approaching a signalized intersection at a speed v_0. When the signal turns yellow, the vehicle is located at a distance x from the stop line. The driver must then decide whether to stop or go. The stopping maneuver requires that the vehicle can travel no more than the distance x to the stop line. Clearing the intersection, on the other hand, requires that the vehicle must travel a distance of at least $(x + w + L)$, where w is the width of the intersection and L is the length of the vehicle. Moreover, this distance must be covered prior to the onset of red (i.e., during yellow). Employing the subscripts 1 and 2 to represent the clearing and the stopping maneuvers, respectively, we obtain

$$x - v_0\delta_2 \geqslant \frac{v_0^2}{2a_2} \tag{2.3.1}$$

which is necessary for a successful stopping maneuver.

The left-hand side of this inequality is the difference between the total distance to the stopping line minus the distance traveled during perception-reaction time at the approach speed v_0, and therefore it specifies the maximum braking distance available to the vehicle. For a successful stopping maneuver this distance should be equal to or greater than the braking distance required by the vehicle at a deceleration a_2. The smallest deceleration rate to accomplish the task is given by the solution of Eq. 2.3.1 as

$$a_2 = \frac{v_0^2}{2(x - v_0\delta_2)} \tag{2.3.2}$$

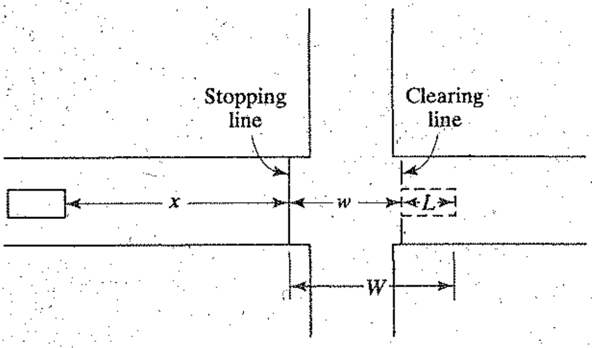


Figure 2.3.3 Vehicle approaching a signalized intersection.
(From Gazis et al. [2.4].)

Considering v_0 and δ_2 to be known, the relationship between x and a_2 is plotted on Fig. 2.3.4. It is a rectangular parabola with an asymptote at $x = v_0\delta_2$, or the perception-reaction distance. This is reasonable because if the vehicle was closer to the stop line at the onset of yellow, it would enter the intersection before the commencement of braking. The mathematical relationship shows the deceleration a_2 to be unbounded. However, there exists a practical upper limit a_2 (max) to the deceleration that a real vehicle can develop. Furthermore, this limit is often higher than the deceleration rate that the drivers and passengers of the vehicle would consider comfortable. The comfortable deceleration rate a_2^* is normally in the vicinity of 8 to 10 ft/s^2 when passengers are seated and around 4 or 5 ft/s^2 when passengers are standing, as in a transit vehicle. The distinction between the *maximum attainable level* and a *desired lower level* must always be kept in mind. The corresponding distance x_c represents the minimum distance for which the vehicle can be stopped comfortably. For shorter distances it would be uncomfortable, unsafe, or impossible to stop. This critical distance is

$$x_c = v_0\delta_2 + \frac{v_0^2}{2a_2^*} \tag{2.3.3}$$

A successful clearing maneuver is represented by

$$x + w + L - v_0\delta_1 \leq v_0(\tau - \delta_1) + \tfrac{1}{2}a_1(\tau - \delta_1)^2 \tag{2.3.4}$$

The right-hand side of Eq. 2.3.4 represents the distance traveled from an initial speed v_0 at constant acceleration a_1 during the time interval $(\tau - \delta_1)$, that is, subsequent to the perception-reaction time and before the onset of the red. The left-hand side of Eq. 2.3.4

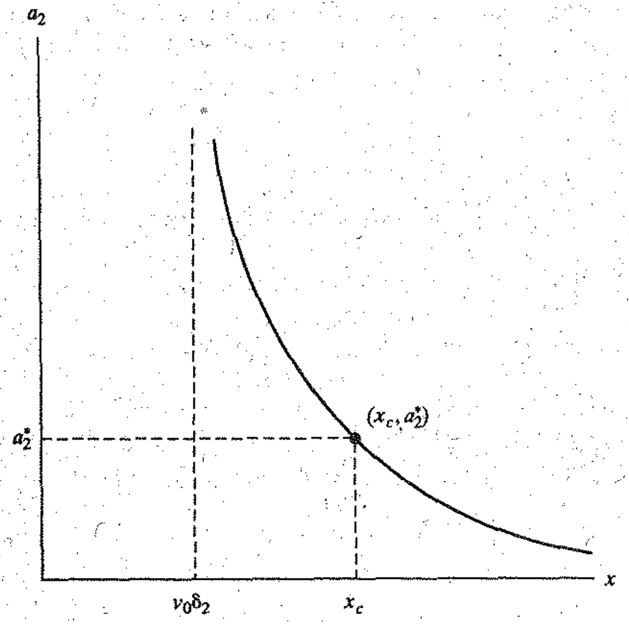


Figure 2.3.4 Acceleration requirements for stopping.
(From Gazis et al. [2.4].)

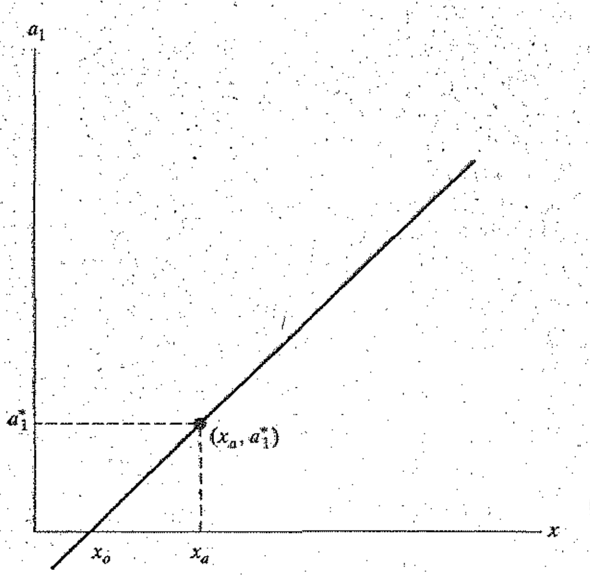


Figure 2.3.5 Acceleration requirements for clearing. (From Gazis et al. [2.4].)

represents the distance available for the clearing maneuver. The acceleration needed just to clear the intersection is

$$a_1 = \frac{2x}{(\tau - \delta_1)^2} + \frac{2(w + L - v_0 \tau)}{(\tau - \delta_1)^2} \tag{2.3.5}$$

which for known values of w, L, v_0, and δ_1 represents a straight line, as shown in Fig. 2.3.5. The distance x_a corresponds to the maximum comfortable acceleration rate a_2^*, and the x-intercept specifies the maximum distance between the vehicle and the stop line from which the vehicle can clear the intersection without accelerating.

The distance

$$x_o = v_0 \tau - (w + L) \tag{2.3.6}$$

is relevant to this analysis because a vehicle that approaches the intersection at the speed limit should not be required to accelerate in order to clear the intersection and thus to break the law. The distance x_o defines a point beyond which a vehicle traveling at the speed limit would not be able to clear safely or legally the intersection on yellow.

For a particular site the relative magnitudes of the two critical distances x_o and x_c determine whether a vehicle can or cannot safely execute either or both maneuvers, as illustrated by Fig. 2.3.6. In part (a), $x_c < x_o$ and the driver can execute either maneuver no matter where the vehicle is located at the onset of yellow. The zone between x_c and x_o is known as the *option zone*. The limiting case is represented by part (b). A problem becomes apparent when $x_c > x_o$, when a *dilemma zone* of length $x_c - x_o$ exists: A vehicle approaching the intersection at the legal speed limit can execute neither of the two maneuvers safely, legally, and comfortably if it happens to be located within the dilemma zone at the onset of yellow.

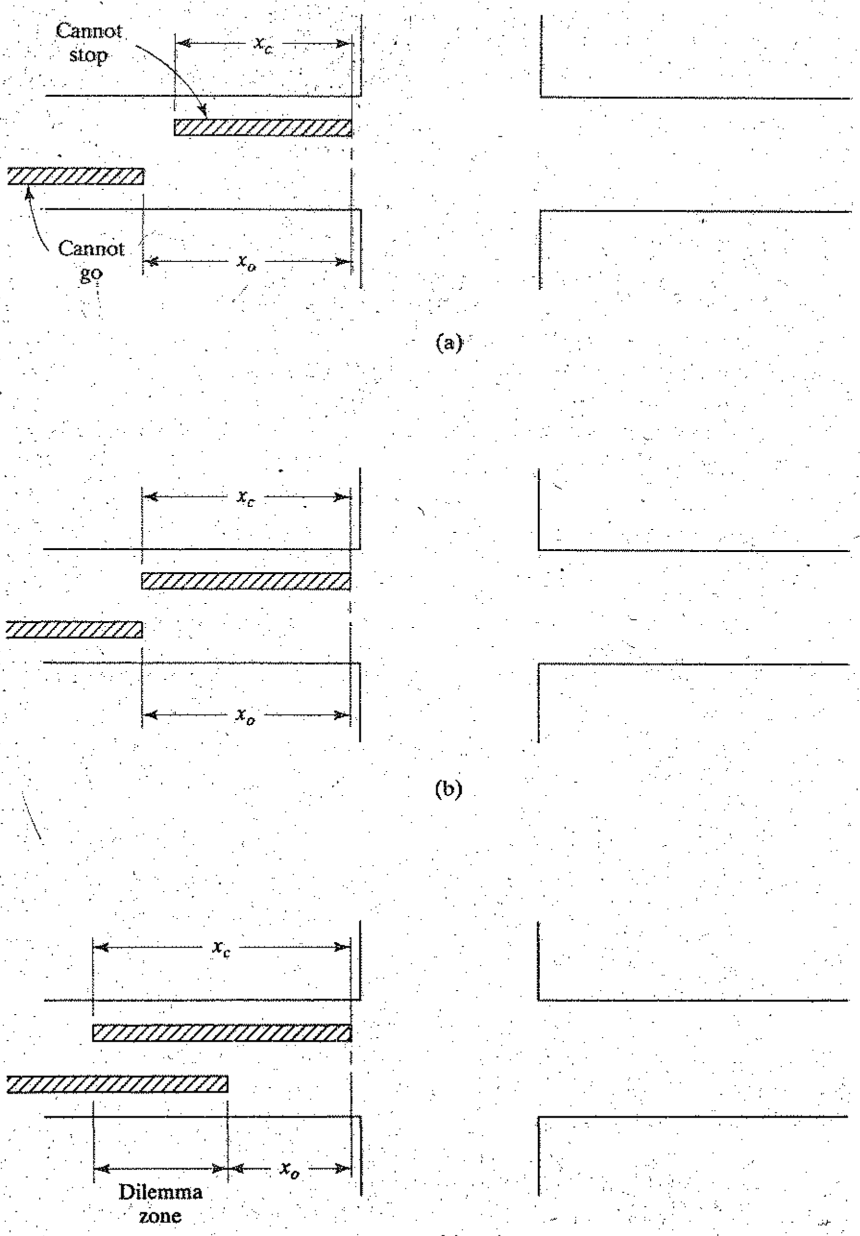


Figure 2.3.6 Dilemma zone.
(From Gazis et al. [2.4].)

The dilemma zone may be eliminated by either changing the speed limit, which in certain locations may be undesirable, or selecting an appropriate minimum duration for the yellow signal phase that results in $x_c = x_o$. In this case Eqs. 2.3.3 and 2.3.6 yield

$$\tau_{\min} = \delta_2 + \frac{v_0}{2a_2^*} + \frac{w + L}{v_0} \tag{2.3.7}$$

Thus properly selected values of a vehicle length, human factors (i.e., comfortable deceleration and sufficient perception-reaction time), and speed limit v_0 specify the *minimum* yellow duration, which, barring *driver error,* ensures that if the vehicle cannot stop, it can clear the intersection. Vehicles traveling at other speeds, however, may still experience a dilemma zone problem, depending on their position at the onset of yellow.

The selected value of τ (which should not be less than the $\tau_{\min}$ calculated by Eq. 2.3.7) is often referred to as the *change interval* [2.5]. It is the time period that elapses between the green displays for two conflicting traffic movements. Papacostas and Kasamoto [2.6] called this time period the *intergreen interval.* In a study of the change interval and its possible subdivision (see below) they interpreted the results obtained by superimposing the plots of Eqs. 2.3.3 and 2.3.6 while allowing the value of the approach speed to vary (i.e., x_c and x_o as functions of speed v). They identified three distinct cases depending on whether the two curves have 0, 1, or 2 points of intersection.

Figure 2.3.7 shows the case involving two common points at speeds v_1 and v_2. In this figure the variable X_s is the same as x_c. Assuming zero acceleration, vehicles approaching the intersection at speeds less than v_1 or greater than v_2 run the risk of facing the dilemma zone problem, depending on the distance from the stop line at the onset of yellow (i.e., regions E and D, respectively). Incidentally, the part of region E that lies below the horizontal axis represents slow-moving vehicles already within the intersection at the onset of the change interval that cannot clear the intersection without accelerating, perhaps because of intersection blockage by vehicles ahead.

When a vehicle's speed and location combination at the onset of yellow places it in region C, an option zone situation would apply. In region A the vehicle cannot clear the intersection prior to the onset of red but can come to a safe and comfortable stop, whereas from within region B a vehicle cannot stop safely or comfortably but can clear the intersection without having to accelerate.

According to the straight line Eq. 2.2.6, the slope of the v versus x_0 function is equal to the change interval: The longer the interval is, the steeper the slope is. The change interval duration illustrated in the figure could have resulted from Eq. 2.3.7, with either v_1 or v_2 substituted for v_0. In either case a range of speeds exists between v_1 and v_2 for which there is no dilemma zone irrespectively of vehicle location.

Papacostas and Kasamoto also showed that when the two curves are tangent to each other (i.e., when they have only one point in common), the option zone shown on Fig. 2.3.7 disappears. This means that the possibility of a vehicle being in a dilemma situation exists for all but one value of v. When the two curves fail to intersect, the value of τ (corresponding to a shallow slope) is smaller than the $\tau_{\min}$ given by Eq. 2.3.7. In this case the dilemma situation can occur at any approach speed. Of course, this condition would not occur if Eq. 2.3.7 were applied properly. Nevertheless, it can be encountered when an insufficient change interval is picked perhaps by adjusting the duration of yellow without referring to Eq. 2.3.7. Papacostas and Kasamoto [2.6] recommend as good practice the habit of

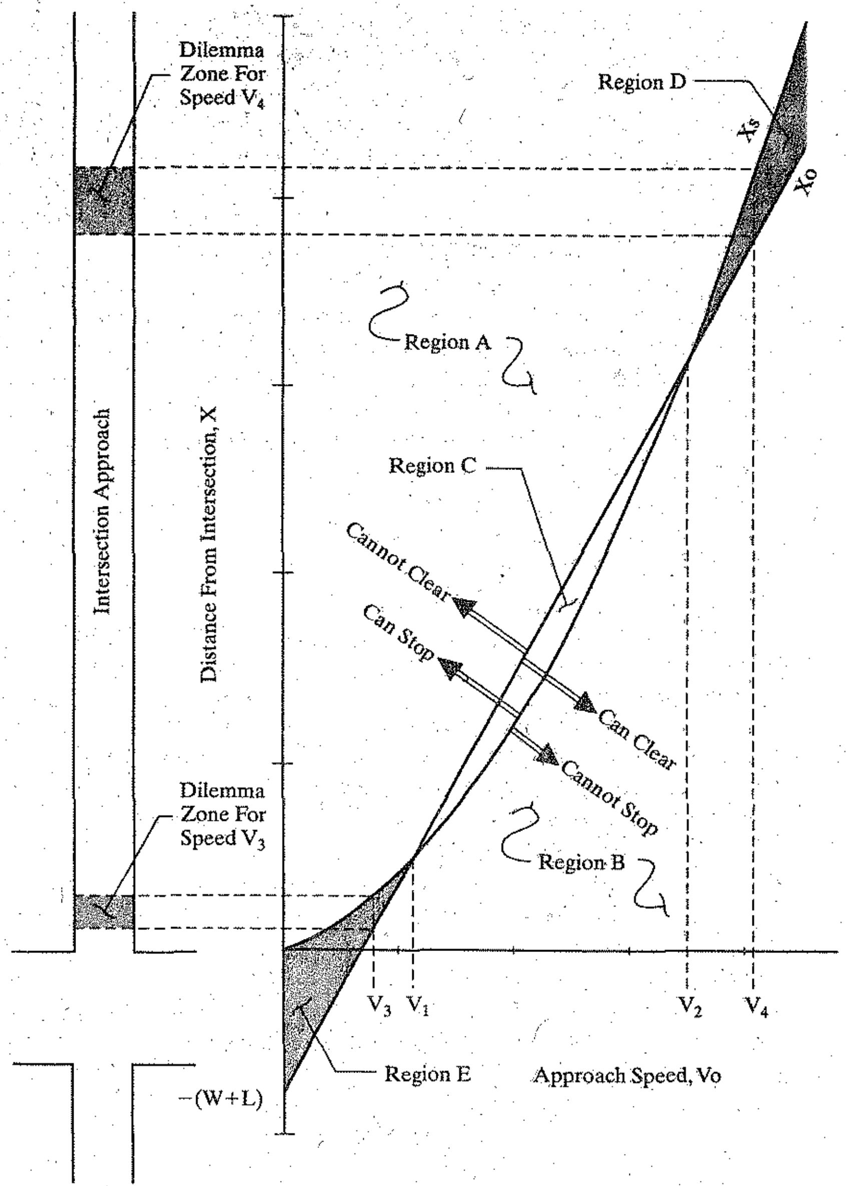


Figure 2.3.7 The case of intersecting X_o and X_s plots. (From Papacostas and Kasamoto [2.5].)

preparing a plot similar to Fig. 2.3.7 after selecting τ if only to assess visually the implied design conditions.

The Institute of Transportation Engineers (ITE) recommends adopting the larger of the two change interval values obtained by applying Eq. 2.3.7 with v_0 set at the measured 85th (i.e., high) and 25th (i.e., low) percentile speeds. This is apparently motivated by the possibility of encountering dilemma zones at high as well as low approach speeds. The ITE-recommended practice further calls for subdividing the selected change interval into two parts as explained next.

Note that the third term of the right-hand side of Eq. 2.3.7 represents the time needed by a vehicle traveling at v_0 to cover a length equal to the width of the intersection plus a vehicle length. For this reason the term has been dubbed the *clearance interval*. This, however, does not assume that the vehicle will always be located exactly at the stop line at the onset of the clearance interval and actually uses this entire interval for clearance.

ITE [2.5] suggests that the change interval be subdivided into an *all-red clearance interval* equal to the last term of Eq. 2.3.7 and the balance be reserved for the yellow phase. The clearance interval is called "all-red" because during this time interval *all* approaching traffic movements face a red display simultaneously. This, however, is not the only possible (or defensible) way to subdivide the intergreen interval. Some local agencies opt to keep the same length of yellow at all intersections within their jurisdiction (with a few exceptions where local conditions, such as extremely wide intersections, warrant otherwise) and to reserve the remaining intergreen duration for the all-red interval. This practice ensures that drivers will always know what yellow duration to expect. Another approach is to keep a constant all-red interval of about 1 to 2 s and to devote the rest of the time to yellow. This is done to avoid the possibility of having signals with excessively long all-red intervals that would encourage "red-light running." In some cases a conservative approach is taken by allowing the short all-red interval in addition to yellow as computed by Eq. 2.3.7.

Local policies relating to the split between yellow and all-red must be based on prevailing conditions, including the applicable traffic code. For example, any policy allowing for the use of all-red as a portion of the change interval given by Eq. 2.3.7 will be inappropriate in cases where the local law requires vehicles to *clear* the intersection totally before the onset of red. Another important consideration is the fact that drivers who are unfamiliar with the signal timings of a particular intersection are unaware of the duration of either the change interval or its splits. The guiding rule would be that, barring driver error, if the driver cannot stop, he or she should be able to clear the intersection.

Example 2.10

A driver traveling at the speed limit of 30 mi/h was cited for crossing an intersection on red. He claimed that he was innocent because the duration of the yellow display was improper, and consequently a dilemma zone existed at that location. Using the following data, determine whether the driver's claim was correct.

$$\text{Yellow duration} = 4.5 \text{ s}$$

$$\text{Perception-reaction time} = 1.5 \text{ s}$$

$$\text{Comfortable deceleration} = 10 \text{ ft/s}^2$$

$$\text{Car length} = 15 \text{ ft}$$

$$\text{Intersection width} = 50 \text{ ft}$$

Solution The required minimum duration of the yellow phase is

$$\tau_{\min} = 1.5 + \frac{44}{20} + \frac{65}{44} = 5.18 \text{ s}$$

Because the actual duration was 4.5 s, the driver's claim cannot be dismissed. There was a dilemma zone, the length of which was

$$x_c - x_o = v_0\delta_2 - v_0\tau + \frac{v_0^2}{2a_2^*} + w + L$$
$$= 29.8 \text{ ft}$$

Whether the vehicle was within the dilemma zone at the onset of yellow and whether the driver was not speeding cannot be proven.

Example 2.11

A car stalled 50 ft from the stopping line at an approach to a signalized intersection of $w = 40$ ft. The driver managed to start it again at the moment the traffic signal turned yellow and decided to clear the intersection. Given that the car accelerated according to

$$a = 4.8 - 0.06v \text{ ft/s}^2$$

and that $\tau = 4.5$ s and $\delta_1 = 1.0$ s, determine whether the driver was able to clear the intersection on yellow.

Solution Of the available 4.5 s of yellow, 1 s elapsed during perception-reaction. According to Example 2.2, during the remaining 3.5 s the vehicle covered a distance of

$$x = \frac{4.8}{0.06}(3.5) - \frac{4.8}{0.06^2}[1 - e^{-(0.06)(3.5)}] + 0 = 27.45 \text{ ft}$$

Because $27.45 < (50 + w + L)$, the driver was unable to clear the intersection on yellow.

Discussion In this case the given acceleration was a function of speed and implicitly of time. Therefore the distance traveled had to be computed accordingly. The design in Eq. 2.3.7 is based on a vehicle approaching at the speed limit and either decelerating at an average rate or clearing the intersection without having to accelerate.

2.3.3 Visual Acuity

A driver visually perceives the actions of other vehicles, the location of objects, traffic control devices, and the general traffic environment. *Visual acuity* refers to the sharpness with which a person can see an object [2.1, 2.7]. One measurement of visual acuity is the *recognition acuity* obtained by the use of the standard Snellen chart, which is familiar to anyone who has visited an ophthalmologist for an eye examination: The person is asked to read letters of different heights from a specified distance. The result of the test is specified in relation to a subject of normal vision. Normal vision is taken to mean that in a well-lit environment a person can recognize a letter of about $\frac{1}{3}$ in. in height at a distance of 20 ft; the visual acuity of this person is given as 20/20. A person with worse vision must be closer to the display in order to recognize the same letter. This relative visual acuity is designated by a ratio such as 20/40, meaning that the person can clearly see an object at a distance of 20 ft when a distance of 40 ft is sufficient for a person with normal vision. Alternatively, the person with 20/40 vision requires an object twice as large as the one that a person with normal vision can clearly discern from the same distance.

Visual acuity is affected by factors, such as the contrast and brightness of the object, the level of illumination, and the relative motion between the observer and the object. Visual acuity is termed *static* in the absence of relative motion and *dynamic* when relative motion

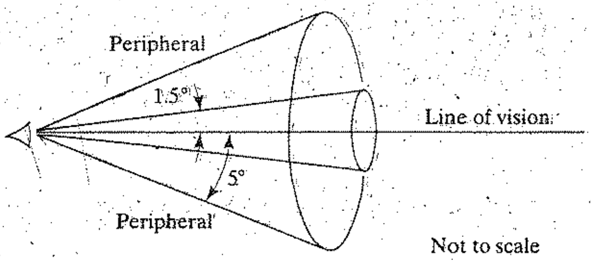


Figure 2.3.8 Cones of vision.

exists. Night driving requires artificial illumination of signs by either permanent fixtures or reliance on the vehicle's headlights. In addition, acuity decreases with increasing visual angles, as illustrated in Fig. 2.3.8. The most clear vision occurs within a cone of vision in the vicinity of 3°. The clarity of vision is fairly good up to approximately 10°, beyond which lies the region of peripheral vision, which may extend up to 160°. For practical design, traffic signs should be placed within the 10° cone and at locations permitting ample distance for perception-reaction and maneuver execution.

Visual acuity and perception-reaction tend to deteriorate with age. In the United States the needs of older drivers are increasingly influencing all elements of highway design. This is because the median age of the population has been on the increase for several decades [2.7, 2.8].

Example 2.12

A driver with 20/20 vision can read a sign from a distance of 90 ft. If the letter size is 2 in., how close would a person with 20/50 vision have to be in order to read the same sign? For the given definition of normal vision, calculate the height of the lettering that a driver with 20/60 vision can read from a distance of (a) 90 ft and (b) 36 ft.

Solution The distance x from the location of the sign can be computed by simple proportions as follows:

$$x = (90\text{ ft})(20/50) = 36\text{ ft}$$

Similarly, the required letter heights can be obtained by proportioning as

(a) $h = (2\text{ in.})\frac{60}{20} = 6\text{ in.}$

(b) $h = (2\text{ in.})\frac{60}{50} = 2.4\text{ in.}$

Example 2.13

Assume that a driver with normal vision can read a sign from a distance of 50 ft for each inch of letter height and that the "design driver" has 20/40 vision. Determine how far away from an exit ramp a directional sign should be located to allow a safe reduction of speed from 60 to 30 mi/h, given a perception-reaction time of 1.5 s, a coefficient of friction of 0.30, a letter size of 8 in., and a level freeway.

Solution As specified, a driver with normal vision can recognize the sign from a distance of 50 × 8 = 400 ft. A driver with 20/40 vision must be no more than 200 ft away. Including perception-reaction, the distance traveled to decelerate from 60 to 30 mi/h is

$$x = v_0\delta + \frac{v_0^2 - v^2}{2g(f + G)} = (88)(1.5) + \frac{88^2 - 44^2}{2(32.2)(0.30)} = 433\text{ ft}$$

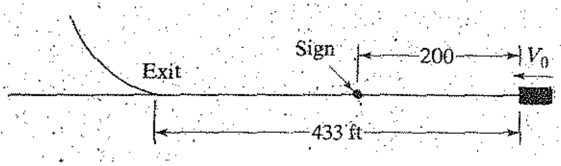


Figure 2.3.9 Determination of sign location for good visibility.

Hence the sign must be located at least 433 − 200 = 233 ft, or about 250 ft, in advance of the exit ramp (Fig. 2.3.9).

Discussion The solution to this problem brings to bear the perception-reaction phenomenon discussed in Section 2.3.1 and the braking distance covered in Section 2.2.2. Moreover, the "design driver" does not represent the best performer, and the design conditions assume a wet pavement (i.e., $f = 0.30$).

2.3.4 Lateral Displacement

When approaching an object located near their paths, as shown in Fig. 2.3.10, drivers show a tendency to displace laterally away from the object even though it may not be on their direct path. Taragin [2.9] reported a set of experiments which measured this tendency: Various objects were placed at different lateral distances on two-lane and multilane highways of various pavement widths, and the effects of these objects were compared to cases where no object was present. The measured effects consisted of speed adjustments, the longitudinal distance l at which vehicles were seen to displace laterally, and the magnitude of the observed lateral displacement.

The major results of the experiments included the following: The narrower the pavement and the closer the object to the pavement edge were, the greater was the magnitude of lateral displacement. When the object was placed at the edge of pavement, the lateral displacement was found to be 3.3 ft in the case of two-lane highways with 8-ft lanes and 1.8 ft for 12-ft lanes. In certain cases speed reductions became apparent.

Subsequent research by Michaels and Gozan [2.10] compared two mathematical models of this phenomenon and concluded that "a model of lateral displacement based on the rate of change of visual angle accounts best for the obtained results." Mathematically this model can be derived as follows.

The relationship between the longitudinal distance l, the lateral placement of the object a, and the visual angle θ is

$$l = a \cot \theta \tag{2.3.8}$$

$$\frac{dl}{dt} = -a \csc^2\theta \frac{d\theta}{dt} \tag{2.3.9}$$

Because $dl/dt = -v$, the vehicle's velocity and $\csc^2\theta = (a^2 + l^2)/a^2$,

$$\frac{d\theta}{dt} = \frac{va}{a^2 + l^2} \tag{2.3.10}$$

Thus, given the vehicle's speed, the longitudinal distance l, and the rate of change of the visual angle, the driver can estimate the lateral placement of the object to judge whether or not it lies in the vehicle's path. If the object lies directly in the vehicle's path (i.e., $a = 0$), the driver cannot detect any angular change. According to this human factor model, each driver

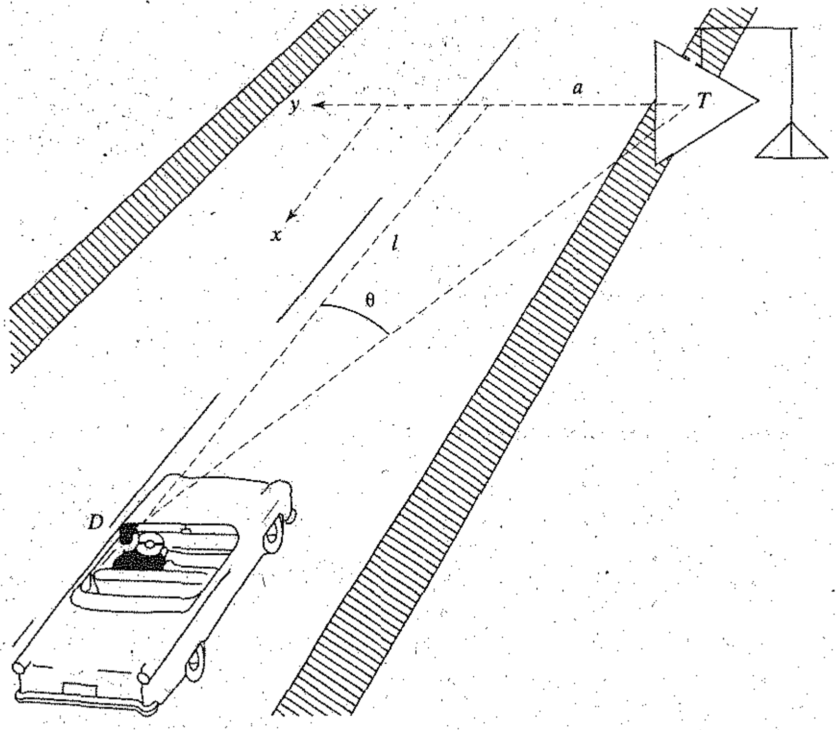


Figure 2.3.10 Geometry of lateral clearance.
(From Michaels and Gozan [2.10], with slight modification.)

has a subjective *critical rate of change in visual angle,* below which the driver presumes that the vehicle is in a collision path and displaces away from the object in the lateral direction.

Michaels and Gozan pointed out that this model explains Taragin's findings with regard to speed adjustments, and they provide additional information regarding other factors affecting the magnitude of lateral displacement, including the size, shape, and brightness of the object. This model can be extended to situations where the object is another moving vehicle on the roadway, in which case the equations of relative motion must be employed.

The understanding of the driver characteristic described in this section can aid in controlling vehicular speeds in highway construction zones through the proper placement of cones or barricades and in design decisions relating to the placement of objects (e.g., signs, bridge abutments, and raised medians) along highways.

Example 2.14

A vehicle traveling at 40 mi/h was observed to displace laterally when it was located 300 ft away from a bridge abutment placed 6 ft to the right of its path. At what longitudinal distance from the same abutment would you expect the same driver to displace laterally when traveling at 60 mi/h?

Solution By Eq. 2.3.10 the critical rate of change in visual angle for this driver is

$$\left(\frac{d\theta}{dt}\right)_{cr} = 0.0039 \text{ rad/s}$$

For the case of $v = 60$ mi/h $= 88$ ft/s

$$0.0039 = \frac{(88)(6)}{6^2 + l^2}$$

and

$$l = 368 \text{ ft}$$

2.4 GEOMETRIC DESIGN OF HIGHWAYS

2.4.1 Background

Geometric design refers to the physical proportioning of facilities, as distinguished from other aspects of design, such as structural design. In this section we address the basic components of geometric design, with the emphasis placed on highway facilities. The five elements examined are the cross section, horizontal alignment, superelevation, vertical alignment, and channelization. A discussion on pavement design is included.

2.4.2 Functional Classification of Highways

In the United States highways are classified according to the function they serve (*functional classification*) and with respect to the entity (private, municipal, state, or federal) responsible for their construction, maintenance, and operation (i.e., *jurisdictional classification*). Of the two, the functional classification is more relevant to geometric design. Table 2.4.1 lists the major functional categories of highways in the United States [2.2]. Figure 2.4.1 is a conceptual description of the relative emphasis that each highway category places on the functions of providing "mobility" (i.e., continuous travel) on one hand and "accessibility" (i.e., direct access to abutting property) on the other: Local streets are predominantly designed for accessibility rather than mobility, whereas high-level facilities such as expressways and freeways are predominantly designed for high-speed continuous movement. The technical difference between freeways and expressways is that the former are characterized by *full control of access;* that is, access to and egress from these facilities are permitted only at controlled locations such as entrance and exit ramps, whereas the latter have *partial control of access;* that is, access or egress may also be permitted directly from or to abutting

TABLE 2.4.1 Highway Functional Classification

Rural		Urban	
Principal arterials	Freeways	Principal arterials	Interstate freeways
			Other freeways/expressways
	Other		Other
Minor arterials		Minor arterials	
Collectors	Major	Collector streets	
	Minor		
Local roads		Local streets	

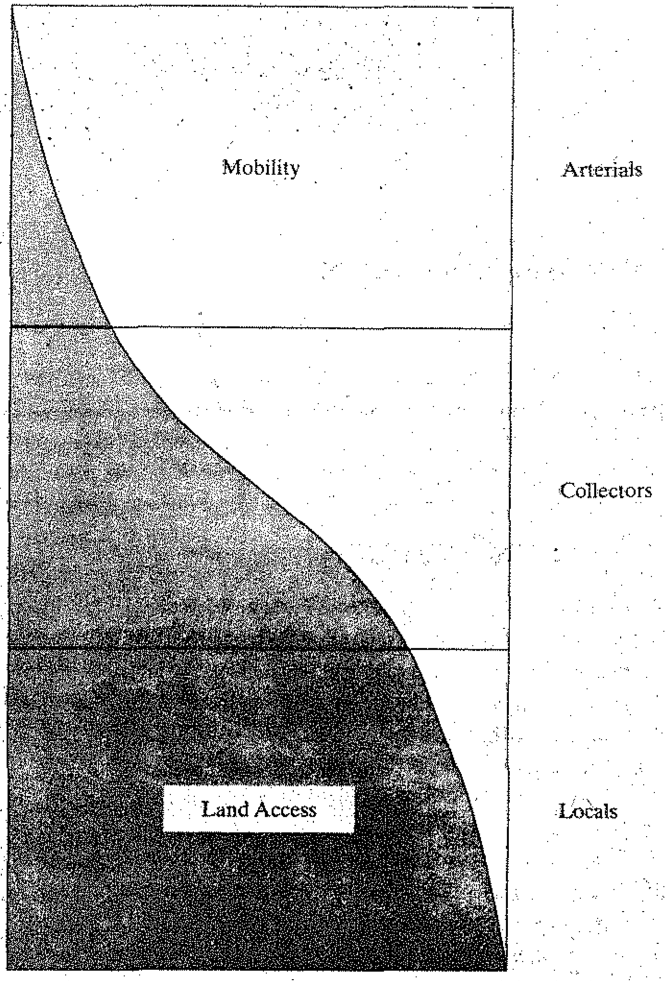


Figure 2.4.1 Relationship of functionally classified highways to mobility and land access.
(From *A Policy on Geometric Design of Highways and Streets*, Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. 1–5, p. 9.) Used by permission.)

property or via a limited number of at-grade intersections. The functional hierarchies of rural and urban highways are schematically illustrated in Fig. 2.4.2.

Generally the design requirements for the various highway types follow the functions served. At one extreme, local roads and streets are designed primarily for light, low-speed traffic to gain access to residences and other land uses; they are closely spaced and often designed to discourage through traffic. At the other extreme, freeways are designed for high traffic levels at high speeds; they are sparsely spaced and designed to facilitate continuous travel between major activity centers. Urban and rural principal arterials are interconnected to serve continuous intercity and interstate movements.

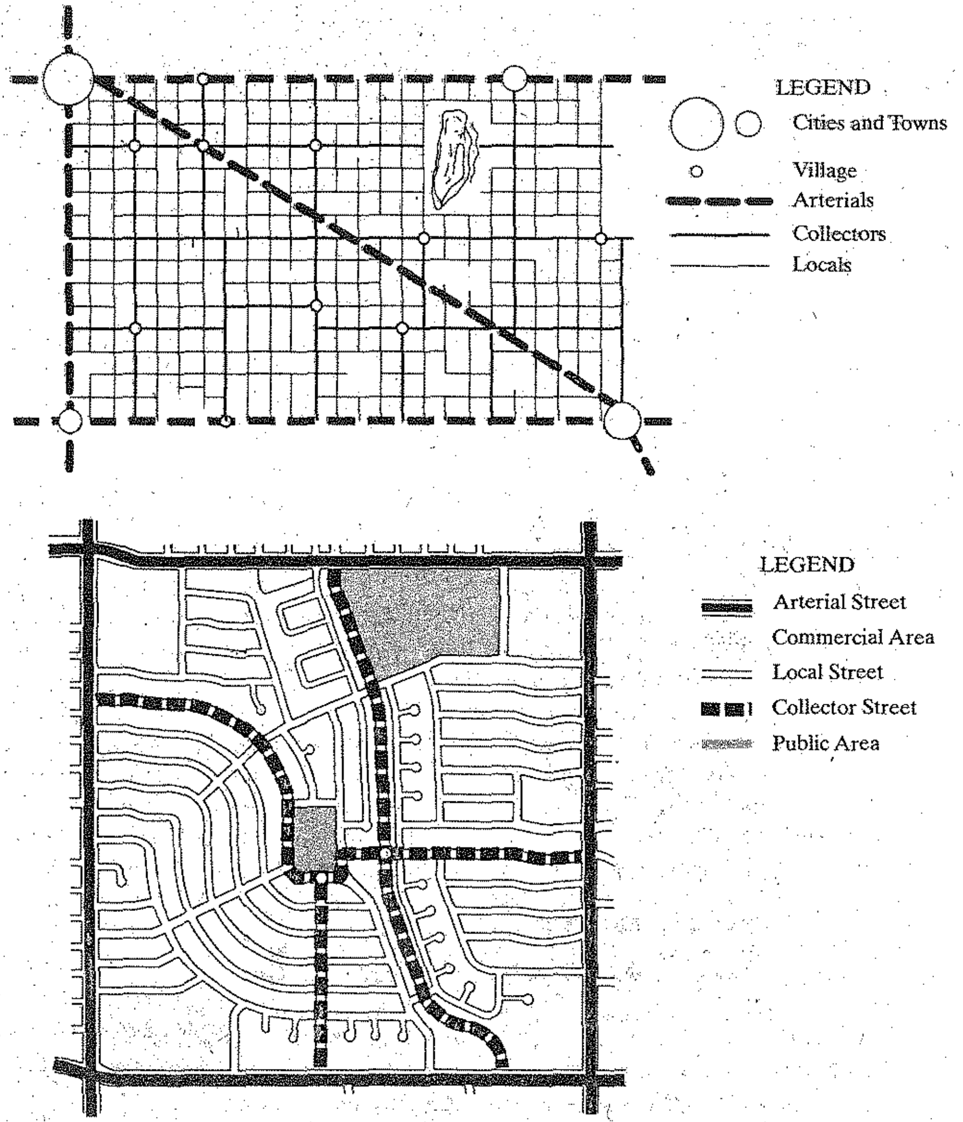


Figure 2.4.2 Illustration of functionally classified highways. (From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Figs. 1–3 and 1–4, pp. 7–8.) Copyright 1990. Used with permission.)

2.4.3 Cross-Section Design

Cross-section design refers to the profile of the facility that is perpendicular to the centerline and extends to the limits of the right-of-way within which the facility is constructed. Figure 2.4.3 illustrates the cross section of a typical *undivided* two-lane rural highway with a lane in each direction of travel. Lane separation is designated by longitudinal *pavement*

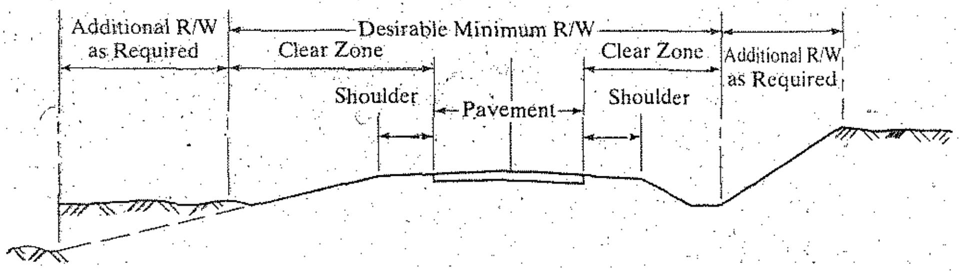


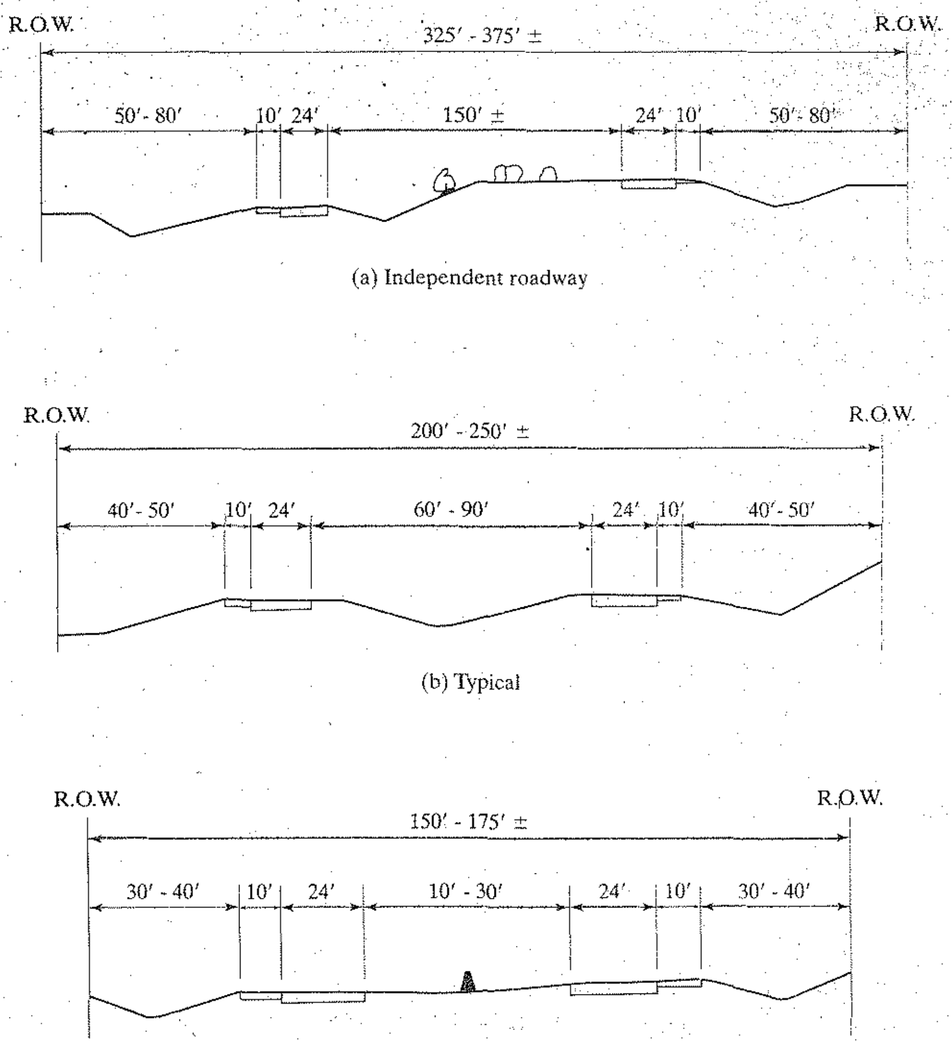
Figure 2.4.3 Cross section of a two-lane rural highway. (From *A Policy on Geometric Design of Highways and Streets*, Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. VII-1, p. 501.) Copyright 1990. Used by permission.)

markings. A *normal crown*, that is, a mild slope in the pavement on either side of the centerline, is provided to facilitate the removal of water. Depending on drainage requirements, crowns in the range of $\frac{1}{8}$ to about $\frac{1}{4}$ in./ft of width are typical. Paved or unpaved *shoulders* are provided at either end of the *travel-way* pavement for emergency situations. Beyond the shoulders, drainage ditches are provided with cut or filled side slopes at appropriate angles to ensure slope stability. Figure 2.4.4 shows typical types of *divided* multilane rural highways. The separation of the two directions of travel may be accomplished by constructing independent roadways and by utilizing raised or depressed *medians*. Various types of *barriers* (including guardrails and concrete barriers) may be used along the median and at the end of the *clear zone* beyond the shoulders. Depending on their function, urban facilities may also be either undivided or divided. Urban roadways often incorporate drainage ditches or gutters and raised curbs. Urban arterials can be at ground level (i.e., *at grade*), *elevated*, or *depressed*; they may also contain special bus lanes and rail-transit ways within their rights-of-way.

2.4.4 Horizontal Alignment

The *horizontal alignment* of a highway, railway, or transit guideway represents the projection of the facility on a horizontal plane. It generally consists of straight-line segments (*tangents*) connected by *circular curves* either directly (*simple curves*) or via intermediate *transition curves*. Figure 2.4.5 illustrates these two common geometric arrangements. The length of the facility is measured along the horizontal alignment of a control line, such as the centerline of a highway, and is usually expressed in terms of 100-ft *stations* from a reference point. Thus a point on the alignment designated as *sta. 14* is located at a distance of 14 stations (i.e., 14 × 100 = 1400 ft) from the reference point. Similarly, a point identified as *sta. 14 + 56.70* is located at a distance of 1456.70 ft from the reference point.

Figure 2.4.6 shows the horizontal alignment of the centerline of a simple curve. A simple circular curve connects two tangents, which when projected meet at a *point of intersection*, or PI. Proceeding in the direction of increasing station values, point *A* is designated as the *point of curvature* (PC), that is, the point where the curve begins. Point *B*, or the end of the curve, is denoted as the *point of tangency*, or PT. At these two points, of course, the radii of the circular curve are perpendicular to the tangents. The length of the curve *AB* equals



(a) Independent roadway

(b) Typical

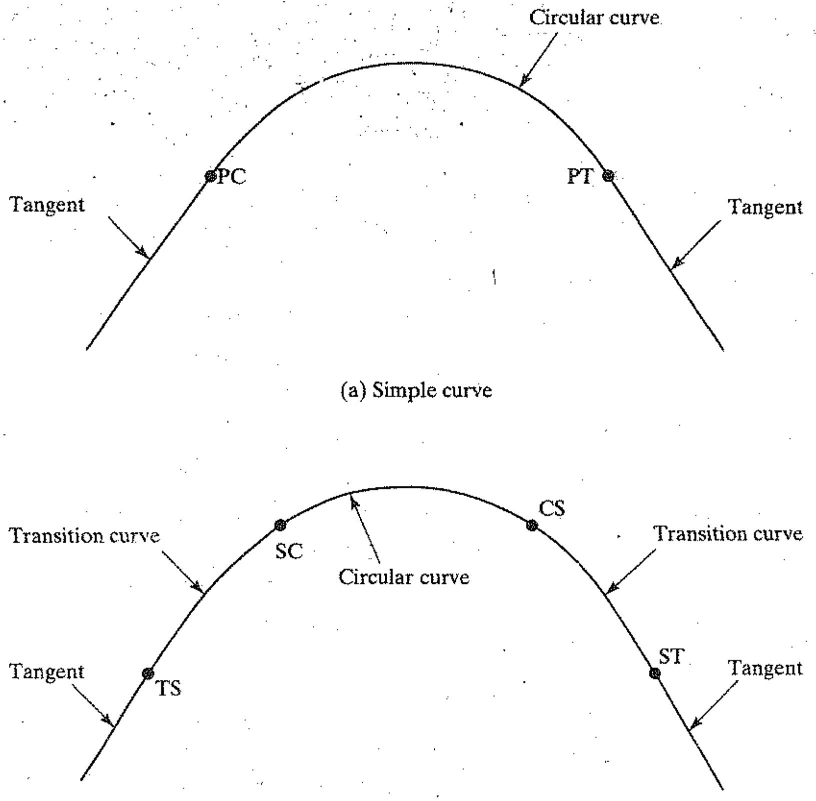
(c) Restricted

Figure 2.4.4 Cross sections of four-lane rural highways. (From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. VIII-48, p. 662.) Copyright 1990. Used by permission.)

$$L = 2\pi R\left(\frac{\Delta}{360}\right) \tag{2.4.1}$$

where R is the curve radius. Other important distances and equations are shown in Fig. 2.4.6.

Two straight lines intersecting at the PI can be connected by an infinite number of circular curves. Each of these curves may be defined by its radius R or by its *degree of curve* D, for which two alternative definitions are encountered in practice. The *arc definition* is



(a) Simple curve

(b) Curve with transition spiral

Figure 2.4.5 Typical horizontal curves.

preferred by highway engineers and is equal to the central angle (in degrees) subtended by an arc of 100 ft. In this case the radius of the curve and the degree of curve are related by the following proportion:

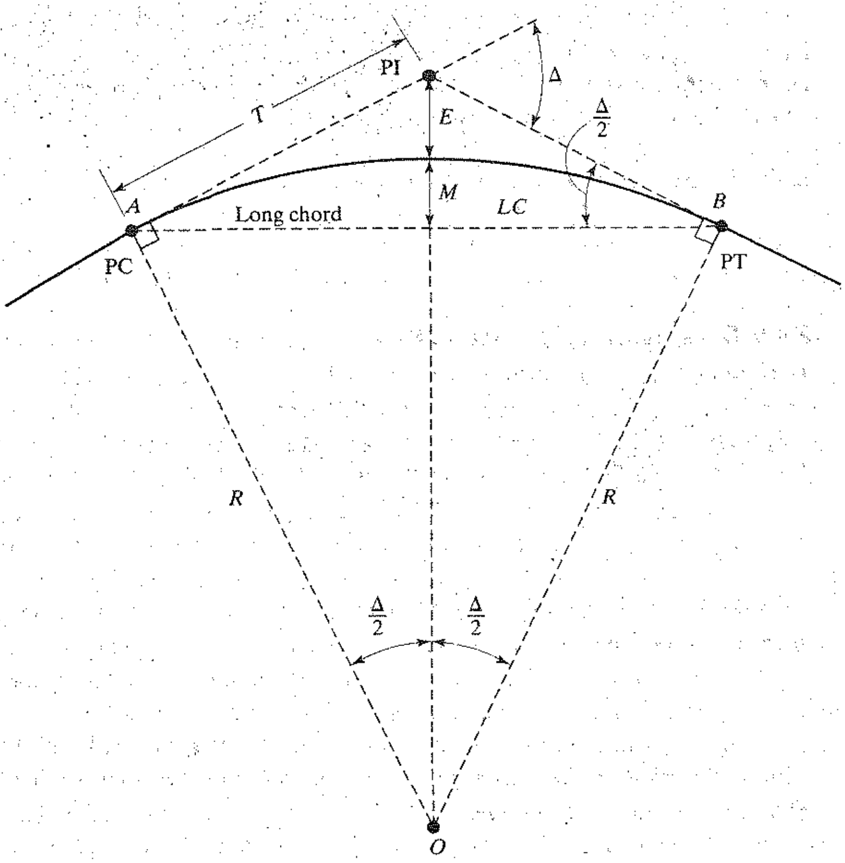
$$\frac{100}{2\pi R} = \frac{D}{360} \tag{2.4.2}$$

or

$$D = \left(\frac{5729.58}{R}\right)^{\circ}$$

Railway design, on the other hand, uses the *chord definition* for the degree of curve, which is equal to the angle subtended by a chord of 100 ft, in which case the relationship between the radius and the degree of curve is

$$\sin\left(\frac{D}{2}\right) = \frac{50}{R} \tag{2.4.3}$$



D: Degree of curve (see text)

E: External distance $= R\left(\sec\frac{\Delta}{2} - 1\right)$

M: Middle ordinate distance $= R\left(1 - \cos\frac{\Delta}{2}\right)$

T: Length of tangent $= R\tan\frac{\Delta}{2}$

L: Length of curve $= 100\frac{\Delta}{D}$

LC: Long chord $= 2R\sin\frac{\Delta}{2}$

Figure 2.4.6 Simple circular curve.

Figure 2.4.7 illustrates the two definitions of the degree of curve. In either case specifying the degree of curve is equivalent to specifying the radius. The degree of curve D must not be confused with the *external angle* of deflection (Δ) between the tangents, which is equal to the total *central angle* subtended by the entire length of the curve AB. Using the arc definition for the degree of curve, the following relationship among the length of curve L, the degree of curve D, and the external angle Δ becomes apparent:

$$L = \frac{100\Delta}{D} \tag{2.4.4}$$

2.4.5 Determination of Design Radius

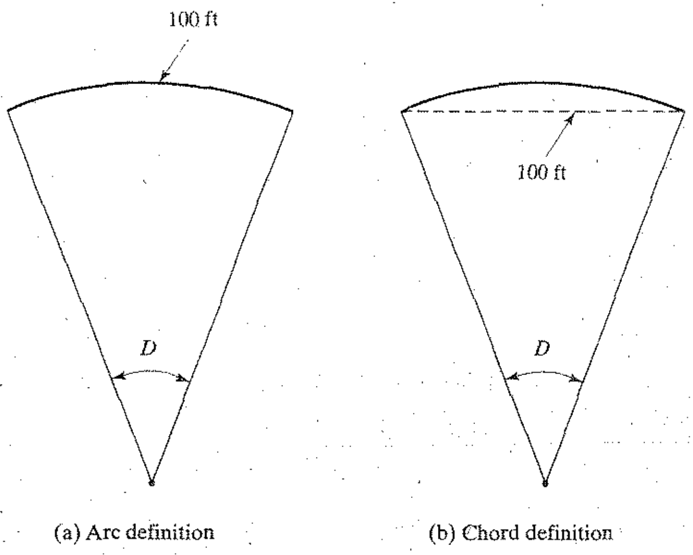
The requirement for lateral banking or *superelevating* the cross section of curved paths (discussed in Section 2.2.3) imposes a constraint on the *minimum radius* that the curve may have. Equation 2.2.19 (reproduced as Eq. 2.4.5) expresses the relationship among the superelevation rate e, the design speed v, the coefficient of side friction f_s, and the curve radius R:

$$e + f_s = \frac{v^2}{gR} \tag{2.4.5}$$

An alternative specification of this formula found in U.S. design manuals is the following mixed-unit equation:

$$e + f_s = \frac{v^2}{15R} \tag{2.4.6}$$

where the design speed has units of miles per hour (mi/h) and the radius is specified in feet (ft).



(a) Arc definition (b) Chord definition

Figure 2.4.7 Degree of curve.

The maximum allowable design value for e [2.2] is 0.12 ft/ft and the suggested maximum is set at 0.10, but special conditions may override these values. For example, a maximum superelevation rate of 0.08 ft/ft may be more appropriate at localities where snow and ice conditions occur. The *design* values for the coefficient of side friction depend on design speed and range from about 0.17 at 20 mi/h to 0.10 at 70 mi/h. According to Eq. 2.4.5, the *minimum radius* for the selected design speed is

$$R_{\text{min}} = \frac{v^2}{g(e_{\text{max}} + f_s)} \tag{2.4.7}$$

and the corresponding *maximum degree of curve* is given by Eq. 2.4.2.

From the perspective of the driver, the longer the radius is, the better the design curve will be. Thus the minimum radius does not represent the desired design radius. Where conditions permit the selection of a design radius that is longer than the minimum, the design value of e can be computed by

$$e_{\text{des}} = \frac{v^2}{gR} - f_s \qquad \text{for } R > R_{\text{min}} \tag{2.4.8}$$

Example 2.15

Calculate the maximum degree of curve and the minimum radius of a simple circular curve with an external angle of 100°. The design speed is 50 mi/h, the corresponding value of f_{max} is 0.14, and the maximum design value for e is 0.10. Also, calculate the design value for e for a curve that has a radius of 800 ft.

Solution By either Eq. 2.4.5 or Eq. 2.4.6

$$R = 695 \text{ ft}$$

and by Eq. 2.4.2

$$D = \frac{5729.58}{695} = 8.24°$$

The external angle does not enter these calculations.

For a radius of 800 ft, Eq. 2.4.8 yields

$$e_{\text{des}} = 0.21 - 0.14 = 0.07 \text{ ft/ft}$$

2.4.6 Superelevation Design

Banking the cross section is needed on the curved portion of the facility but is not necessary along the tangent segments of the horizontal alignment. Consequently a transition of the cross section from the normal crown on the tangent to a fully superelevated pavement on the curve must be developed.

As an illustration of superelevation design, consider a simple circular curve for the two-lane highway of Fig. 2.4.8. The cross section is at the normal crown at point A and fully superelevated at point E. Point B represents the intermediate condition, where the outside edge of the travel way has been rotated to the level of the centerline; point C represents the condition where the outside edge, the centerline, and the inside edge are aligned at a slope equal to the normal crown. Since the normal crown is milder than the design superelevation

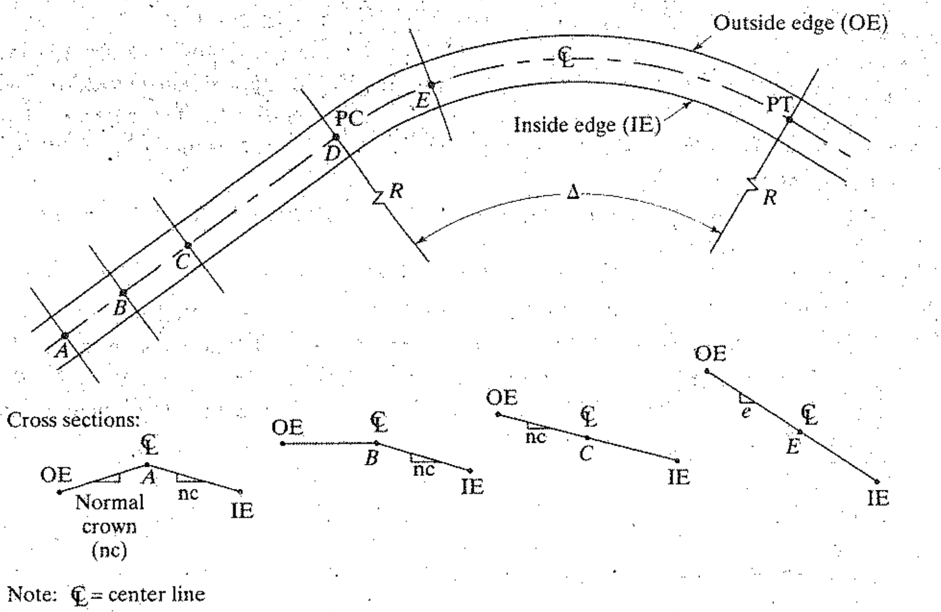


Figure 2.4.8 Development of superelevation.

rate, the cross section must be further rotated until it reaches full superelevation at point E with an intermediate slope at the PC (i.e., point D). The distance AB and BE along the horizontal alignment are called the *tangent runout* and the *superelevation runoff*, respectively. The length of the superelevation runoff depends on the rate at which the cross section is rotated.

The selection of the length of superelevation runoff is not an exact science. Table 2.4.2 presents minimum lengths for *two-lane* rural highways having either 10- or 12-ft lanes. The superelevation runoff lengths for three-, four-, and six-lane highways should be 1.2, 1.5, and 2.0 times, respectively, those calculated for two-lane highways.

On simple curves about two-thirds of the superelevation runoff is typically placed on the tangent and the rest of the curve. When transition curves are used, the superelevation runoff is developed on them. Transition curves (see Fig. 2.4.5) are usually introduced on high-speed curves. Most often, they are appropriate lengths of spirals with end radii of curvature that are consistent with those of the tangent (i.e., infinite at the TS and the ST) and the circular curve (i.e., R at the SC and the CS); TS is tangent to spiral, CS is curve to spiral, and so on (see Figs. 2.4.5 and 2.4.17).

Figure 2.4.9 illustrates four common methods of developing the transition to full superelevation. For ease of presentation the curved alignment is shown to be stretched out into a straight line. The first method rotates the pavement about the centerline, the second rotates about the inside edge, and the third rotates about the outside edge. The fourth method applies to pavements that begin with a straight cross-section slope and are revolved about the outside edge. This type of cross slope may be found on the separate roadways that make up a divided multilane facility (see Fig. 2.4.4). The longitudinal profile of the pave-

TABLE 2.4.2 Required Length of Superelevation Runoff for Two-Lane Roads

Superelevation	Length of runoff (ft) for design speed (mi/h) of:							
rate, *e*	20	30	40	50	55	60	65	70
				12-ft lanes				
0.02	50	100	125	150	160	175	190	200
0.04	60	100	125	150	160	175	190	200
0.06	95	110	125	150	160	175	190	200
0.08	125	145	170	190	205	215	230	240
0.10	160	180	210	240	255	270	290	300
0.12	195	215	250	290	305	320	350	360
				10-ft lanes				
0.02	50	100	125	150	160	175	190	200
0.04	50	100	125	150	160	175	190	200
0.06	80	100	125	150	160	175	190	200
0.08	105	120	140	160	170	180	190	200
0.10	130	150	175	200	215	225	240	250
0.12	160	180	210	240	255	270	290	300

Source: (From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2]. (Table III-15, p. 178.)

ment along the length of the highway corresponding to each of the four methods of obtaining full superelevation is shown in Fig. 2.4.9. In part (a) the location of the inside edge, the centerline, and the outside edge are shown relative to the elevation of the centerline. In parts (b) and (c) the edge and centerline profiles are shown relative to the unrotated centerline, that is, the "theoretical centerline profile." At point *A* of the first three diagrams the outside edge is as far below the centerline as the inside edge, the difference in elevation being equal to the normal crown times the pavement width in each travel direction. At point *B* the outside edge has reached the level of the centerline, and at point *C* the outside edge is located as far above as the inside edge is below the centerline. Finally, at point *E* the cross section is fully superelevated. The reverse of these profiles is found at the other end of the circular curve. As the note at the bottom of Fig. 2.4.9 suggests, the angular breaks should be rounded by smooth curves.

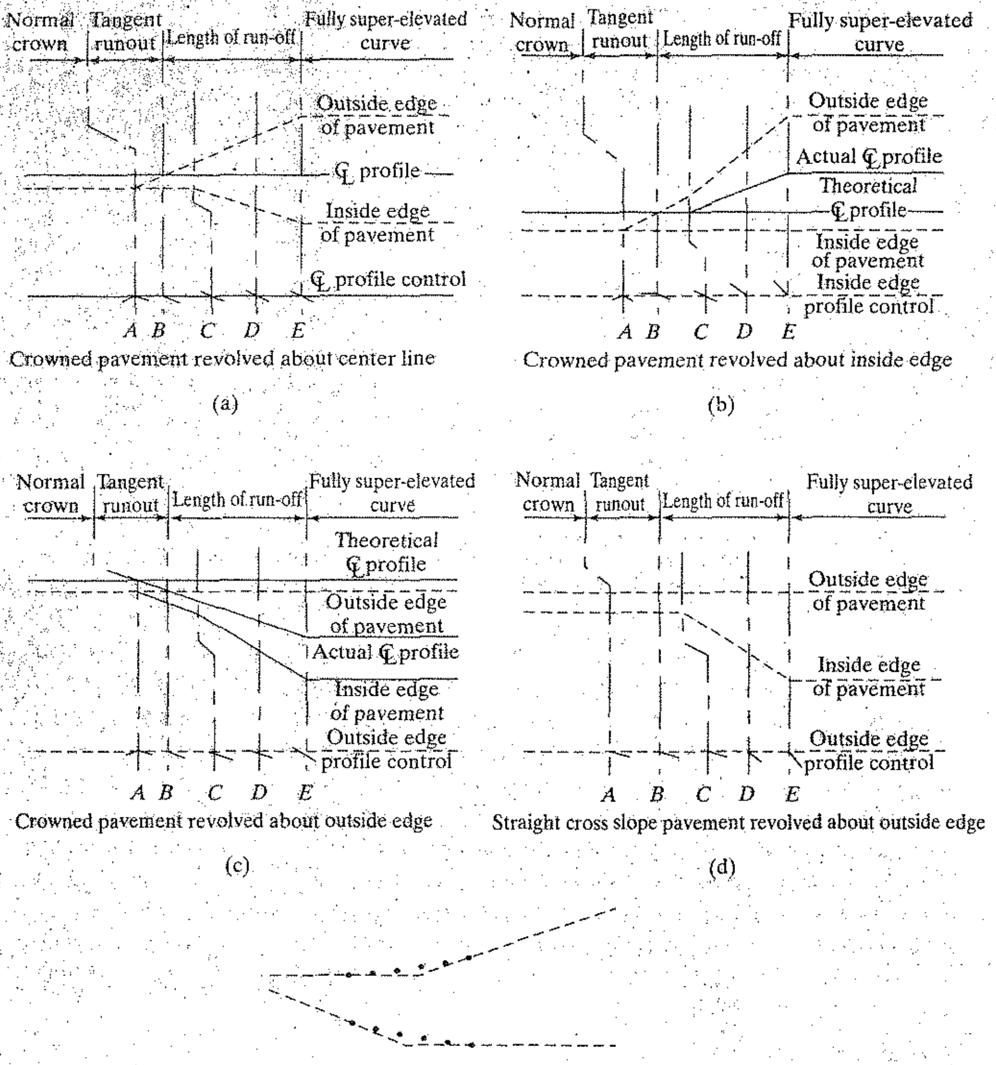
Example 2.16

Draw the longitudinal profile of the curve of Example 2.15 using the minimum radius for a two-lane rural highway given a normal crown of $\frac{1}{4}$ in./ft and a lane width $W = 12$ ft.

Solution With the calculated radius of 695 ft and an external angle of 100°, Eq. 2.4.1 gives a curve length of

$$L = 1213 \text{ ft}$$

For $e = 0.10$, a design speed $v = 50$ mi/h, and 12-ft lanes, the suggested minimum length of superelevation runoff (Table 2.4.2) is 240 ft. Place two-thirds (or 160 ft) of the runoff on each tangent at either end and the rest (80 ft) on the curve. This leaves $1213 - 2 \times 80 = 1053$ ft of the curve's length at full superelevation. Rotation about the centerlines means that at full superelevation the inside and outside edges are offset by $W \times e = 12 \times 0.10 = 1.2$ ft from the



Crowned pavement revolved about center line

(a)

Crowned pavement revolved about inside edge

(b)

Crowned pavement revolved about outside edge

(c)

Straight cross slope pavement revolved about outside edge

(d)

Note: Angular breaks to be appropriately rounded as shown by dotted line.

Figure 2.4.9 Methods of attaining full superelevation. (From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. III-16, p. 183.) Copyright 1990. Used by permission.)

centerline. At the normal cross section the two edges are 12 × 0.25 = 3 in. or 0.25 ft, below the centerline. Simple calculations lead to a longitudinal profile, as shown in Fig. 2.4.10.

2.4.7 Vertical Alignment

The vertical alignment of highways and railways consists of *grade tangents* connected with *parabolic vertical curves.* The desirable maximum design grades and gradient change depend on both the facility type and vehicular characteristics. For highways the desirable

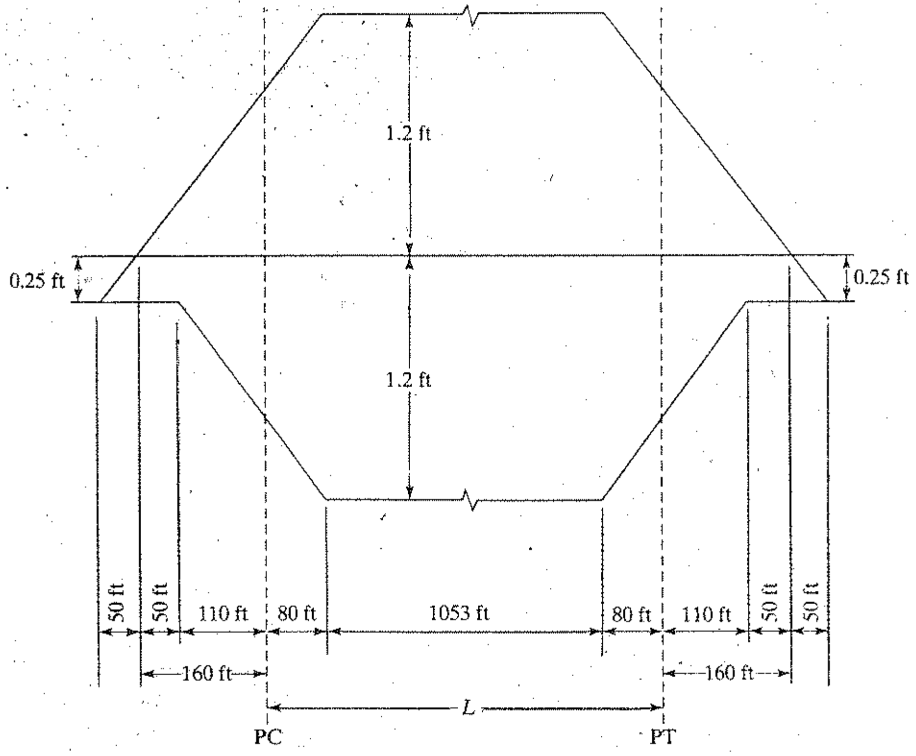


Figure 2.4.10 Cross-section drawing of superelevation along a curve. (*Caution: x* and *y* scales are different.)

maximum grades range from about 2% for freeways to about 6% for local streets. Higher grades may be unavoidable at locations of difficult topography, and the combined effect of gradient and the length over which it is sustained (i.e., the *length of grade*) must also be considered, especially at locations frequently used by heavy vehicles with limited climbing capability. Railroad design tolerates much smaller maximum grades, with about 4% representing the limit corresponding to the worst topography. The maximum grades for fixed-guideway transit systems are a function of the tract and wheel combinations employed. Steel-wheel on steel-rail systems are similar to railroads, whereas rubber-tire systems approach the highway case. Some systems are specifically designed for very steep inclines.

The *length of a vertical curve is measured along the horizontal alignment,* and a point on the curve is specified by its station location on the horizontal alignment and its elevation from a datum. The beginning and end of a vertical curve are denoted, respectively, as the *vertical point of curvature* (VPC) and the *vertical point of tangency* (VPT), and the point where the grade lines intersect is known as the *vertical point of intersection* (VPI). Figure 2.4.11 describes a *symmetrical* vertical curve for which the grade tangents are equal; asymmetrical vertical curves are used only in places of unusual constraints. The figure applies to both *crest* curves, as shown, and *sag* curves, the latter being merely a reflection of the former with the curve lying above the VPI. In the case of symmetrical curves a vertical line passing through the VPI bisects the length of the curve, but the high

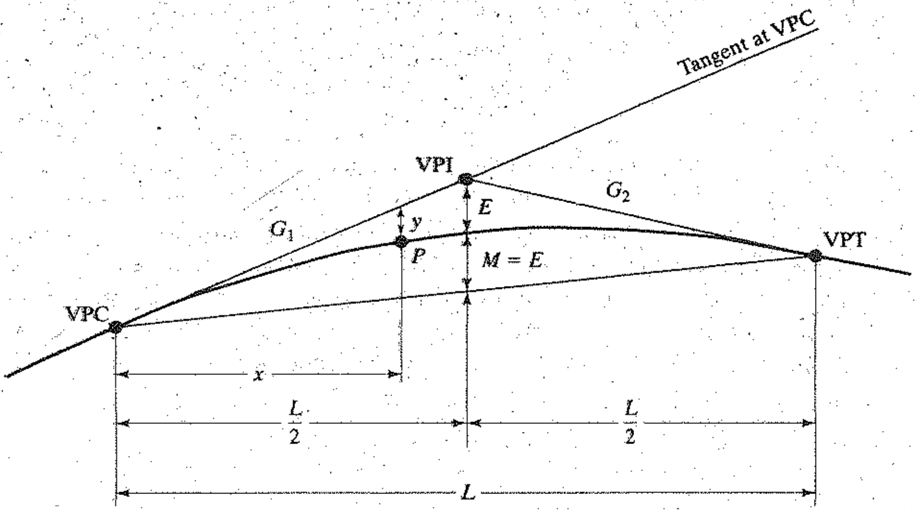


Figure 2.4.11 Symmetric vertical curve.

(or low) point of the curve does not necessarily lie directly below (or above) the VPI. For convenience the horizontal alignment is shown as a straight line, but in reality it may be prescribing a curved path. Moreover, there is no necessary coincidence among the VPC, the VPI, and the VPT on one hand and the PC, the PI, and the PT on the other.

Denoting the *percent grade* at the VPC as G_1 and the percent grade at the VPT as G_2, the total change in grade

$$A = G_2 - G_1 \text{ percent} \tag{2.4.9}$$

is negative in the case of crest curves and positive in the case of sag curves. The ratio of the curve's length to the absolute value of the change in grade

$$K = \frac{L}{|A|} \tag{2.4.10}$$

specifies the *vertical curvature* of the curve. Special attention to drainage design is warranted when K is greater than 167. The *external distance* E from the VPI to the middle of the curve is

$$E = \frac{AL}{800} \text{ ft} \tag{2.4.11}$$

where L is the length of the curve in feet. Note that E is positive in the case of sag curves and negative in the case of crest curves, indicating that the midpoint of the curve lies above and below the VPI, respectively.

Other vertical *offsets* y between the grade tangent passing through the VPC and the vertical curve are calculated by

$$y = 4E\left(\frac{x}{L}\right)^2 \tag{2.4.12}$$

where x is the distance along the horizontal alignment from the VPC to the point of interest. The high (or low) point is located at a distance

$$X = \frac{LG_1}{G_1 - G_2} \qquad X \geqslant 0 \tag{2.4.13}$$

from the VPC. The *curve elevation* of any point P is computed by

$$\text{Elevation of } P = \left[\text{elevation of VPC} + \left(\frac{G_1}{100}\right)x\right] + y \tag{2.4.14}$$

where the term in brackets represents the *tangent elevation* on the vertical tangent passing through the VPC. In practice, curve elevations are computed as 25- or 50-ft intervals. In addition, the elevations are calculated for critical points, such as the high or low points and points where necessary clearances below the pavement (e.g., drainage facilities such as culverts) or above the pavement (e.g., overpasses) are present.

Example 2.17

A 600-ft vertical curve connects a +4% grade to a −2% grade at station 25 + 60.55 and elevation 648.64 ft. Calculate the location and elevation of the VPC, the middle of the curve, the VPT, and the curve elevation at stations 24 + 00 and 27 + 00.

Solution The curve is a crest with $A = -2 - (+4) = -6\%$ and $K = 600/6 = 100$. The middle distance is $E = -4.5$ ft. The middle point of the curve is 4.5 ft below the VPI at sta. 25 + 60.55. The VPC and the VPT are located at either side of the VPI at distances of $L/2 = 300$ ft, or three stations. Hence the VPC is $3 \times 4 = 12$ ft below the VPI at sta. 22 + 60.55, and the VPT is $3 \times 2 = 6$ ft below the VPI at sta. 28 + 60.55. The high point is located at a distance $X = 400$ ft, or four stations from the VPC (i.e., at sta. 26 + 60.55). The following table illustrates the use of Eqs. 2.4.12 and 2.4.14 to calculate the required curve elevations.

Point P (sta.)	x (ft)	Tangent elevation	Offset y (ft)	Curve elevation
22 + 60.55 (VPC)	000.00	636.64	0.00	636.64
24 + 00.00	139.45	642.22	−0.97	641.25
25 + 60.55	300.00	648.64	−4.50	644.14
26 + 60.55 (high)	400.00	652.64	−8.00	644.64
27 + 00.00	439.45	654.22	−9.66	644.56
28 + 60.55 (VPT)	600.00	660.64	−18.00	642.64

Discussion Figure 2.4.12 shows the vertical curve. Offsets are measured in relation to the tangent at the VPC irrespective of whether the subject point P is to the right or to the left of the VPI. The negative sign of the offsets indicates that the curve elevations are below the tangent elevations. The application of Eq. 2.4.12 in the given table resulted in the same curve elevations for the middle point and the VPT as those obtained by the calculations preceding the table. All the tabulated results could have been obtained by viewing the curve from the VPT. In that case G_1 and G_2 would be +2% and −4%, respectively, the distance x would be measured to the left of the VPT, and the offsets would be measured from the tangent at the VPT.

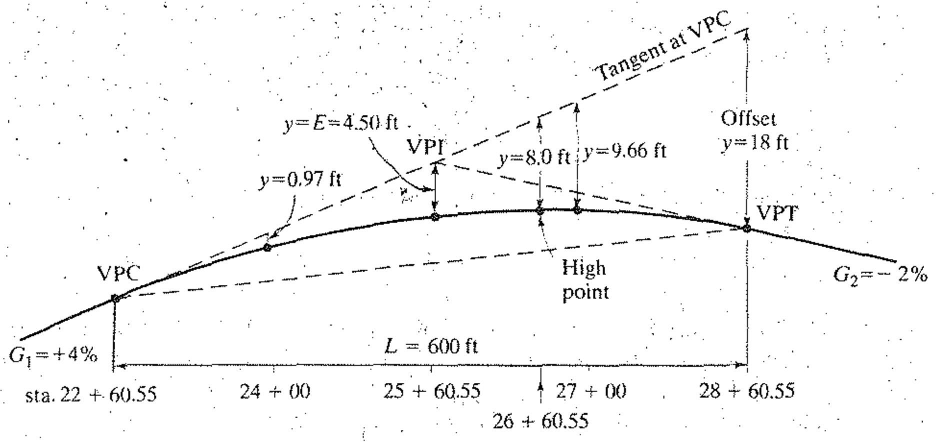


Figure 2.4.12 Example of vertical curve connection.

TABLE 2.4.3 Stopping Sight Distance

Design speed (mi/h)	Assumed speed for condition (mi/h)	Brake reaction		Coefficient of friction *f*	Braking distance on level[a] (ft)	Stopping sight distance	
		Time (s)	Distance (ft)			Computed[a] (ft)	Rounded for design (ft)
20	20–20	2.5	73.3–73.3	0.40	33.3–33.3	106.7–106.7	125–125
25	24–25	2.5	88.0–91.7	0.38	50.5–54.8	138.5–146.5	150–150
30	28–30	2.5	102.7–110.0	0.35	74.7–85.7	177.3–195.7	200–200
35	32–35	2.5	117.3–128.3	0.34	100.4–120.1	217.7–248.4	225–250
40	36–40	2.5	132.0–146.7	0.32	135.0–166.7	267.0–313.3	275–325
45	40–45	2.5	146.7–165.0	0.31	172.0–217.7	318.7–382.7	325–400
50	44–50	2.5	161.3–183.3	0.30	215.1–277.8	376.4–461.1	400–475
55	48–55	2.5	176.0–201.7	0.30	256.0–336.1	432.0–537.8	450–550
60	52–60	2.5	190.7–220.0	0.29	310.8–413.8	501.5–633.8	525–650
65	55–65	2.5	201.7–238.3	0.29	347.7–485.6	549.4–724.0	550–725
70	58–70	2.5	212.7–256.7	0.28	400.5–583.3	613.1–840.0	625–850

[a]Different values for the same speed result from using unequal coefficients of friction.

Source: American Association of State Highway and Transportation Officials, "A Policy on Geometric Design of Highways and Streets" [2.2] (Table III-1, p. 120.) Copyright 1990. Used with permission.

2.4.8 Stopping and Passing Sight Distance

The design of a facility must ensure that drivers are provided with adequate *sight distances* to perceive dangerous situations ahead and to take preventive action. If an object is present in the vehicle's path, warranting that the vehicle be stopped to avoid a mishap, the object must be visible to the driver from a distance at least equal to a minimum *stopping sight distance*, S_s. Table 2.4.3 presents the design values suggested by AASHTO [2.2], which incorporate several practical approximations.

To overtake another vehicle safely on two-lane highways (i.e., one lane in each direction), a driver must consider the relative speeds and positions of the driver's own vehicle, the vehicle to be overtaken, and an oncoming vehicle in the opposite direction. A successful passing maneuver involves the elements shown in Fig. 2.4.13. The minimum *passing sight*

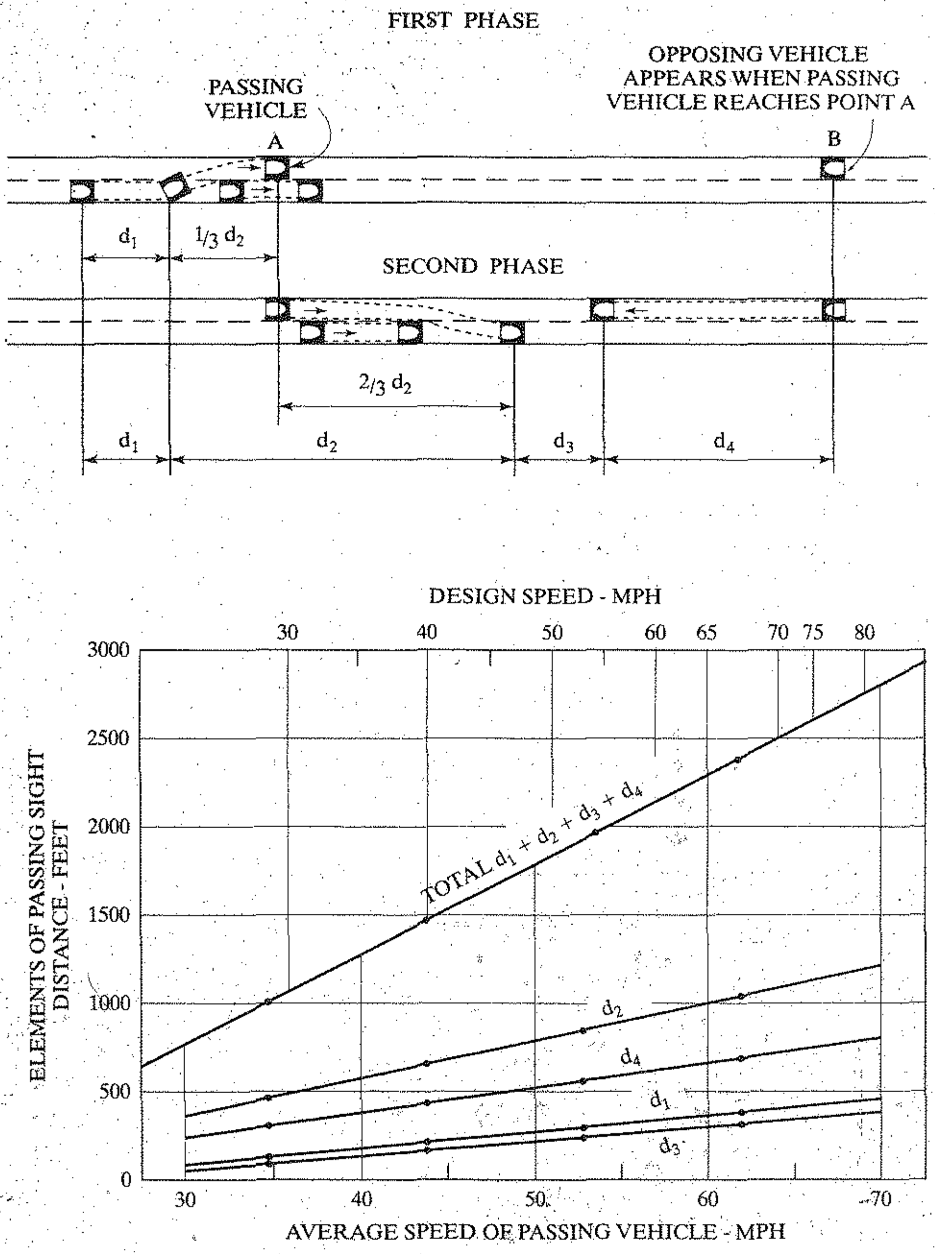


Figure 2.4.13 Passing sight distance.
(From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. III-2, p. 130.) Copyright 1990. Used by permission.)

distances suggested by AASHTO for various design speeds are also shown. Safe stopping sight distances must be adhered to at all points along the alignment, and where adequate passing distances are not possible, no-passing zones must be established. Reducing the speed limit to assure adequate sight distances is also a possibility, but frequent speed limit changes should be avoided.

2.4.9 Geometrics of Sight Distance

On the horizontal plane the available sight distance is affected by the presence of objects, embankments, and other restrictions, as shown in Fig. 2.4.14. The effect of curvature (i.e., either R or D) is captured by the inserted equations. This figure corresponds to the low range of stopping sight distance shown in Table 2.4.3.

On crests the vertical curvature of the facility itself causes sight restrictions, as illustrated in Fig. 2.4.15(a), which shows the line of sight between the driver's eyes (located at a distance h_1 above the pavement) and the top of an object of height h_2. For many years the design specifications promulgated by AASHTO provided for an eye height of 3.75 ft, an object height of 6 in. for the computation of stopping sight distance, and an object height of 4.5 ft for the computation of passing sight distance. The rationale for these values was that an object of less than 6 in. in height is lower than the undercarriage height of the vast majority of vehicles on the roadway, and a height of 4.5 ft represents the height of oncoming vehicles that are relevant to the passing maneuver. Because of changing vehicular dimensions, subsequent research has recommended a modification of these values. In 1984 AASHTO lowered the values for the driver's eye height and the height of oncoming vehicles to 3.50 and 4.25 ft, respectively. The 6-in. object height involved in the calculation of stopping sight distances has been retained. On sag curves the worst situation occurs at night when the line of sight is limited within the area of headlight illumination [Fig. 2.4.15(b)].

In either case the sight distance may be shorter, equal to, or longer than the length of the curve. The following equations can be used to calculate the minimum length of a curve that satisfies a given sight-distance requirement:

Crest vertical curves:

$$L = \frac{|A|S^2}{200(\sqrt{h_1} + \sqrt{h_2})^2} \quad \text{for } S \leqslant L \tag{2.4.15a}$$

$$L = 2S - \frac{200(\sqrt{h_1} + \sqrt{h_2})^2}{|A|} \quad \text{for } S \geqslant L \tag{2.4.15b}$$

Sag vertical curves:

$$L = \frac{|A|S^2}{200(h + S \tan \beta)} \quad \text{for } S \leqslant L \tag{2.4.16a}$$

$$L = 2S - \frac{200(h + S \tan \beta)}{|A|} \quad \text{for } S \geqslant L \tag{2.4.16b}$$

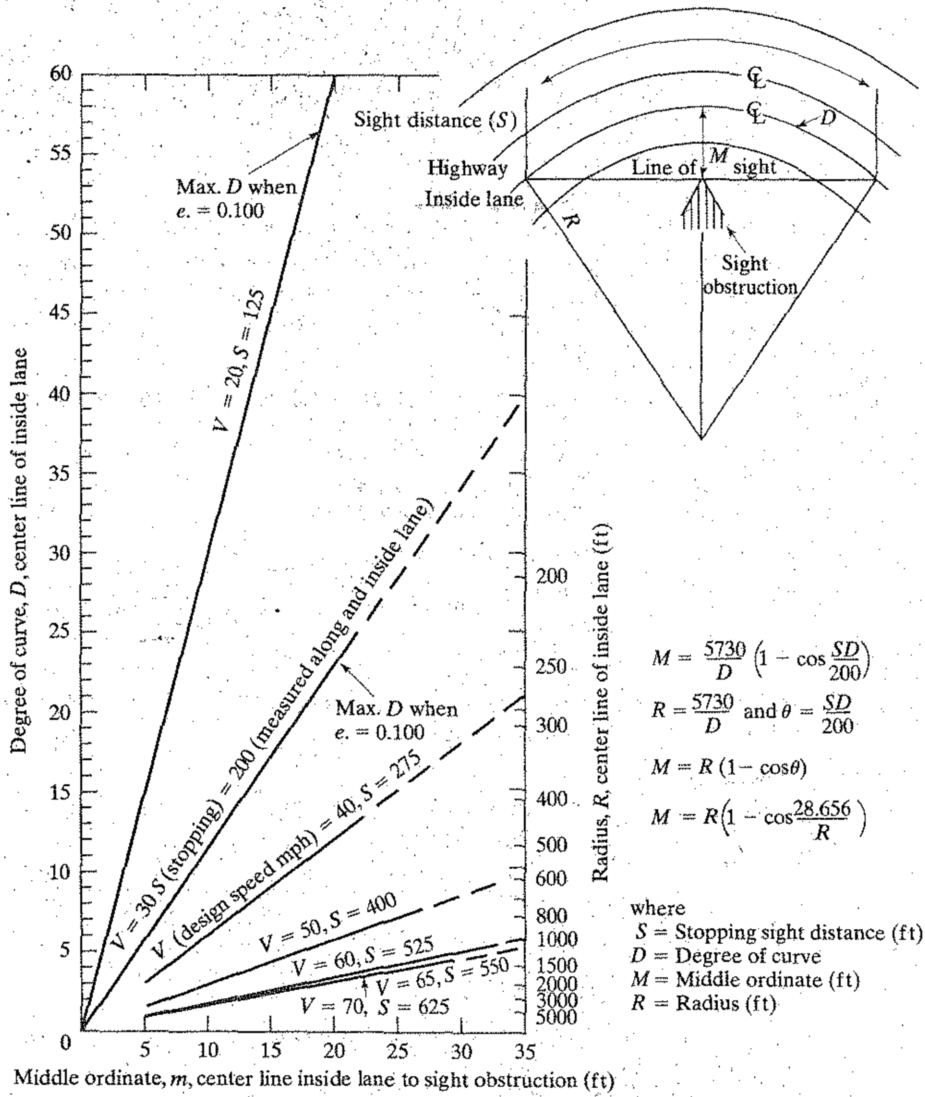


Figure 2.4.14 Geometry of horizontal sight distance. (From *A Policy on Geometric Design of Highways and Streets*, Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. III-26A, p. 222.) Copyright 1990. Used by permission.)

where

L = length of curve, in ft

S = sight distance, in ft

$|A| = |G_2 - G_1|$, in %

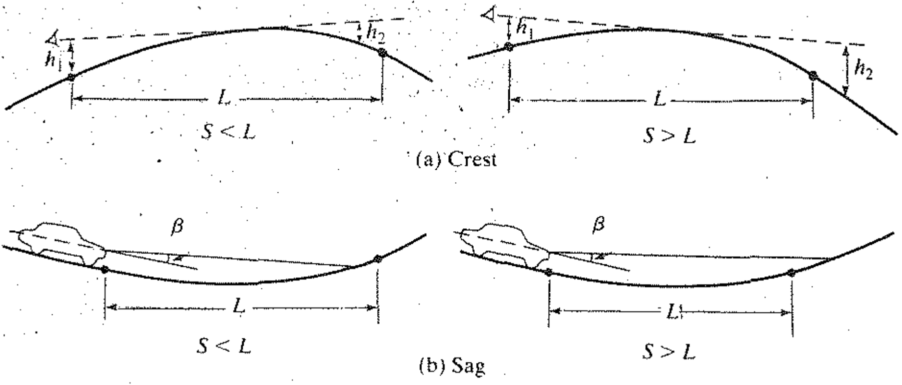


Figure 2.4.15 Geometry of vertical sight distance.

h_1 = height of driver's eyes, in ft

h_2 = height of object, in ft

h = headlight height: approximately 2 ft

β = beam angle: approximately 1°

Figure 2.4.16 illustrates that stopping and passing distances may be measured directly on scaled horizontal and vertical profiles.

Example 2.18

For a design speed of 50 mi/h, determine the minimum length of a crest vertical curve with $A = -4\%$ that meets the post-1984 AASHTO criteria for (a) stopping and (b) passing.

Solution (a) Since A is negative, the curve is a crest. Conservatively, a minimum stopping sight distance of 475 ft is obtained from Table 2.4.3 and substituted in Eqs. 2.4.15 along with the AASHTO recommended heights of $h_1 = 3.50$ and $h_2 = 0.5$ ft:

$$\text{For } S \leqslant L: \quad L = 679 \text{ ft}$$

$$\text{For } S \geqslant L: \quad L = 618 \text{ ft}$$

The first answer is selected because it satisfies the constraint, that is, 475 < 679.

(b) The minimum recommended passing sight distance of about 1700 ft is obtained from Fig. 2.4.13 for a design speed of 50 mi/h. With $h_1 = 3.50$ and $h_2 = 4.25$ ft, the crest equations yield

$$\text{For } S \leqslant L: \quad L = 3738 \text{ ft}$$

$$\text{For } S \geqslant L: \quad L = 2627 \text{ ft}$$

Since the first satisfies the condition $S \leqslant L$, it gives the proper answer.

2.4.10 Discussion of Alignment Design

So far the basic elements of the horizontal and vertical design have been discussed separately. These two aspects of the three-dimensional control line must be integrated and drawn to a suitable scale to aid in the eventual layout and construction of the facility. Figure 2.4.17 illustrates the standard plan and profile drawings of the centerline of a two-lane highway.

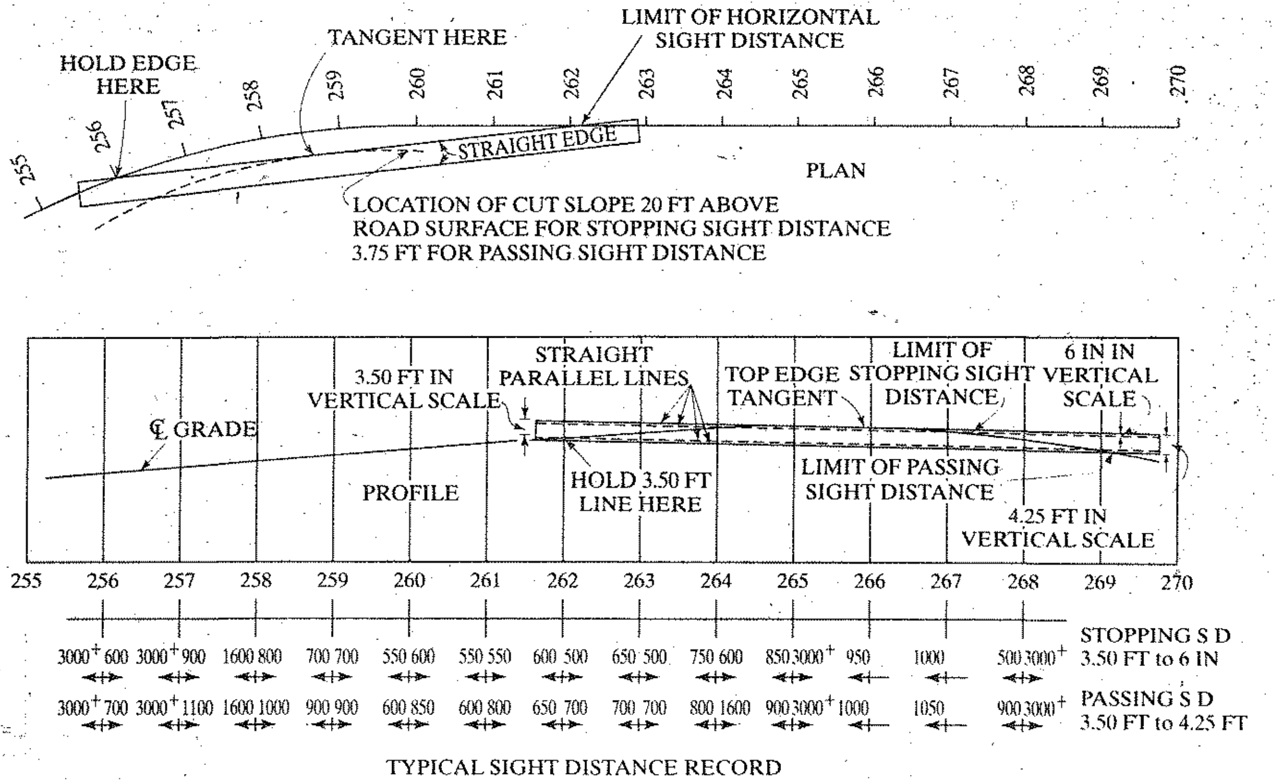


Figure 2.4.16 Scaling and recording sight distances.
(From *A Policy on Geometric Design of Highways and Streets.* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. III-3, p. 139.) Copyright 1990. Used by permission.)

On the horizontal plane the highway segment prescribes a circular curve. The station locations of important points are clearly marked, and the characteristics of the curve are specified.

The "stretched-out" vertical profile of the baseline is shown on the lower portion of the figure: It consists of a crest curve connecting a +1.76% grade and a −4.17% grade, followed by a sag curve between the −4.17% grade and a +6.77% grade. The irregular dotted line represents the existing ground elevation to which the vertical alignment or grade line conforms as much as possible. Fitting the grade line to the existing ground, however, must meet maximum grade and adequate sight-distance criteria as explained earlier. Another major consideration in grade control is related to the amounts of earthwork required. Economic considerations warrant that *cut* and *fill* should be balanced within the limits of the construction area as much as possible to avoid the cost of both bringing extra material (*borrow*) to the site and removing excess excavated quantities to locations that lie outside the site (*overhaul*).

The process of selecting, designing, and locating the final alignment of a facility connecting two points is a highly complex undertaking. It begins with a determination that such a facility is, in fact, needed. Given that the need for a facility has been established, a sequence of interrelated steps follows. These entail the collection and study of the necessary information including topographic maps and photogrammetric reconnaissance surveys, the identification of alternative alignments, the preliminary selection of the preferred alignment, the surveying and mapping of the corridor through which the preferred alignment passes, and the design of the final alignment. These activities take place within a variety of economic,

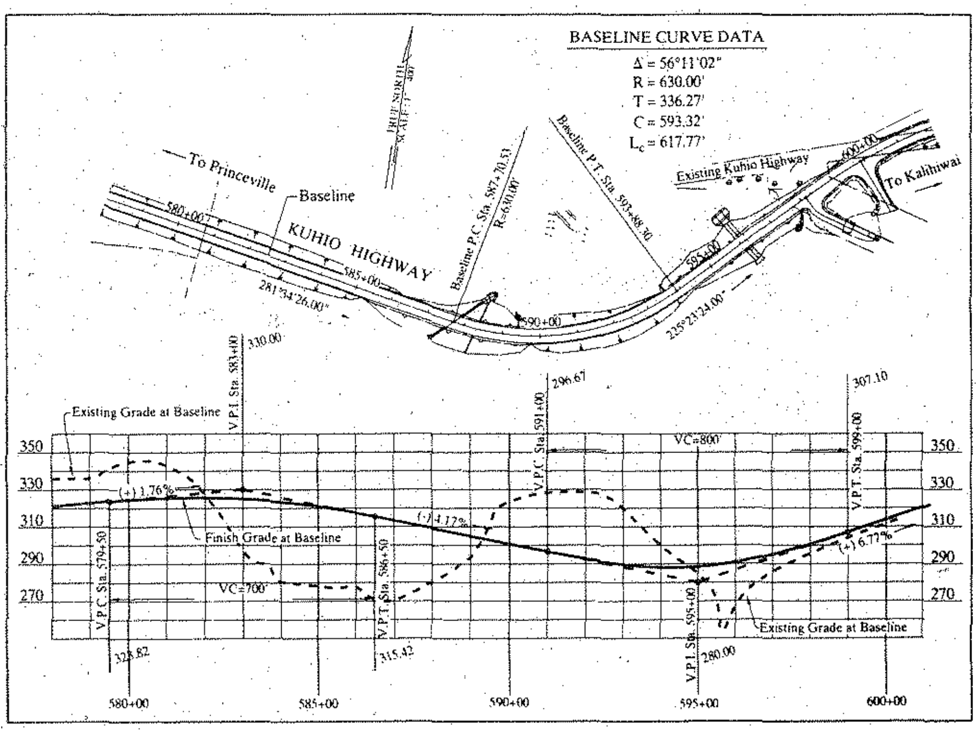


Figure 2.4.17 Typical plan and profile of a highway.
(From Hawaii State Department of Transportation, Highway Design Branch.)

legal, and environmental constraints. For example, federal legislation requires that proposed facilities should have a minimum impact on environmentally sensitive areas, natural habitats of endangered species, and sites of historical and archaeological significance. Only after these and other socioeconomic requirements have been met in a satisfactory manner would the phases of detailed surveying and route layout, and finally, construction, proceed.

2.4.11 Delineation of Vehicular Paths

In addition to the design of the horizontal and vertical alignment of a highway, geometric design includes the delineation of vehicular paths within the travel way to conform to both the physical (space) requirements of vehicles and the steering tendencies of drivers. Figure 2.4.18 illustrates the practice of *curve widening* that is recommended for sharp curves to allow for the fact that the vehicle's front and rear wheels do not track exactly the same trajectories and in response to a tendency on the part of drivers to steer away from the pavement's edge.

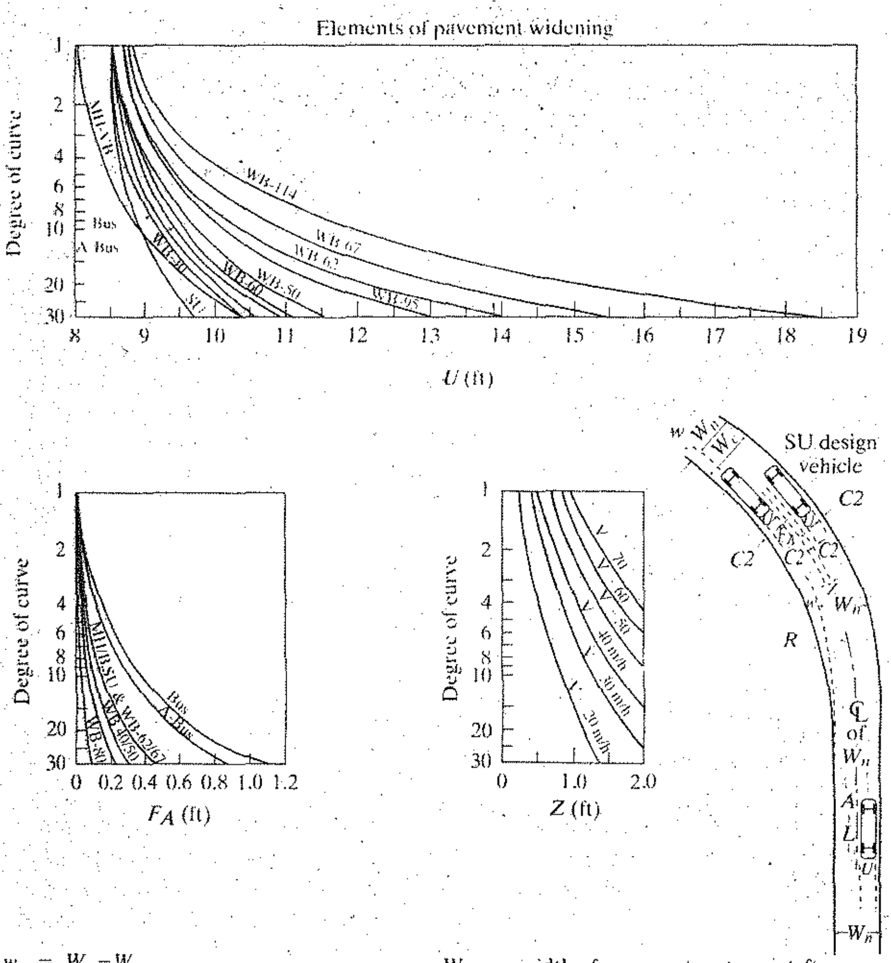
The proper delineation of vehicular paths is accomplished by a number of devices, including longitudinal and transverse pavement markings, raised medians and islands, curbing, guardrails, and the like. These devices are accompanied by appropriate directional signs and other pavement markings, such as painted turning arrows [2.11, 2.12].

The most common pavement markings in the direction of travel are *yellow* and *white lines*, the former separating paths in opposing directions and the latter delineating paths in the same direction. In either case *solid lines* designate segments where path changes for the purpose of either passing or lane changing are prohibited. *Broken lines* permit the execution of such maneuvers.

At intersection and interchange areas where conflicting paths are found special design efforts are necessary either to minimize the number of conflicting movements or to reduce their severity. Figure 2.4.19 shows the major types of freeway interchanges, where by introducing special directional roadways, the more severe *crossing conflicts* are eliminated in favor of the less severe *merging* and *diverging conflicts*. This practice often necessitates the construction of *grade-separated* facilities (i.e., underpasses and overpasses). Additionally, *auxiliary lanes*, such as acceleration and deceleration lanes, permit safe speed changes and reduce the speed differences between conflicting vehicles. The conflicts in the area of *at-grade* intersections are typically reduced by both the spatial separation of paths (*channelization*) and the temporal separation of conflicting movements (e.g., by *signal control*).

2.4.12 Design Vehicles

Whether designing special roadways at an interchange or channelizing an at-grade intersection, it is important to consider the turning characteristics of the vehicles that are expected to use the facility. For purposes of design consistency the weights, dimensions, and operating characteristics of various types of vehicles have been selected to represent the wide variety of vehicular types that normally use the highway system. Table 2.4.4 summarizes the dimensions of these design vehicles, and Fig. 2.4.20 shows the space requirements for a W-60 design vehicle when executing turns between 30 and 180° at the minimum turning radius for that vehicle that correspond to low speeds. The selection of the proper design vehicle for a particular application must allow for the great majority of the vehicles that are expected to use the facility. For example, while the BUS design vehicle may be adequate for downtown streets, the largest WB-60 vehicle is appropriate for the design of truck routes. Table 2.4.5 shows the minimum turning radii of design vehicles.



(1) $w = W_c - W_n$

(2) $W_c = N(U+C) + (N-1)F_A + Z$

N = number of lanes

w = widening for pavement on curve, ft

W_c = width of pavement on curve, ft

(3) $U = u+R - \sqrt{R^2-L^2}$

(4) $\text{FA} = \sqrt{R^2+A(2L+A)} - R$

(5) $\text{Z} = V/\sqrt{R}$

W_n = width of pavement on tangent, ft

U = track width of vehicle (out-to-out tires), ft

C = lateral clearance per vehicle; assumed 2, 2.5 and 3 ft for W_n of 20, 22 and 24 ft, respectively

F_A = width of front overhang, ft

Z = extra width allowance for difficulty of driving on curves, ft

u = track width on tangent (out-to-out) 8.5 ft

R = radius on centerline of two-lane pavement, ft

L = wheelbase

A = front overhang

V = design speed of highway, mi/h

Figure 2.4.18 Curve-widening.
(From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. III-25, p. 215.) Copyright 1990. Used by permission.)

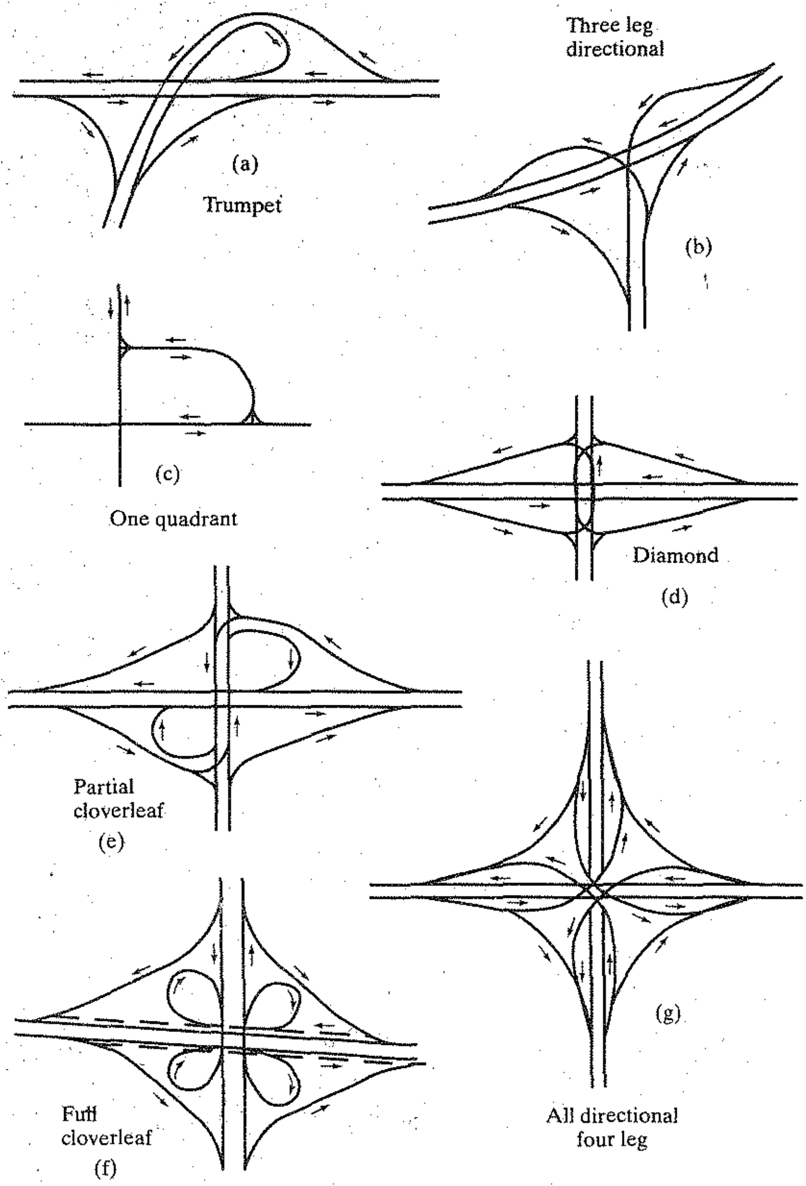


Figure 2.4.19 Interchange types.
(From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. X-1, p. 854.) Copyright 1990. Used by permission.)

2.4.13 Channelization of At-Grade Intersections

Channelization has been defined as "the separation or regulation of conflicting traffic movements into definite paths by means of traffic islands or pavement markings to facilitate the safe and orderly movements of both vehicles and pedestrians" [2.2].

TABLE 2.4.4 Specifications of Design Vehicles

		Dimension[a] (ft)										
			Overall		Overhang							
Design vehicle type	Symbol	Height	Width	Length	Front	Rear	WB$_1$	WB$_2$	S	T	WB$_3$3	WB$_4$
Passenger car	P	4.25	7	19	3	5	11					
Single-unit truck	SU	13.5	8.5	30	4	6	20					
Single-unit bus	BUS	13.5	8.5	40	7	8	25					
Articulated bus	A-BUS	10.5	8.5	60	8.5	9.5	18		4[b]	20[b]		
Combination trucks												
Intermediate semitrailer	WB-40	13.5	8.5	50	4	6	13	27				
Large semitrailer	WB-50	13.5	8.5	55	3	2	20	30				
"Doublebottom" semitrailer-full trailer	WB-60	13.5	8.5	65	2	3	9.7	20	4[c]	5.4[c]	20.9	
Interstate semitrailer	WB-62[d]	13.5	8.5	69	3	3	20	40–42				
	WB-67[e]	13.5	8.5	74	3	3	20	45–47				
Triple semitrailer	WB-96	13.5	8.5	102	2.5	3.3	13.5	20.7	3.3[f]	6[f]	21.7	21.7
Turnpike double semitrailer	WB-114	13.5	8.5	118	2	2	22	40	2[g]	6[g]	44	
Recreation vehicles												
Motor home	MH		8	30	4	6	20					
Car and camper trailer	P/T		8	49	3	10	11	18	5			
Car and boat trailer	P/B		8	42	3	8	11	15	5			
Motor home and boat trailer	MH/B		8	53	4	8	20	21	6			

[a] WB$_1$, WB$_2$, WB$_3$, WB$_4$ are effective vehicle wheel bases. *S* is the distance from the rear effective axle to the hitch point. *T* is the distance from the hitch point to the lead effective axle of the following unit.

[b] Combined dimension 24, split is estimated.

[c] Combined dimension 9.4, split is estimated.

[d] Design vehicle with a 48-ft trailer as adopted in 1982 STAA (Surface Transportation Assistance Act).

[e] Design vehicle with a 53-ft trailer as grandfathered in 1982 STAA (Surface Transportation Assistance Act).

[f] Combined dimension 9.3, split is estimated.

[g] Combined dimension 8, split is estimated.

Source: American Association of State Highway and Transportation Officials, "A Policy on Geometric Design of Highways and Streets" [2.2], (Table II-1, p. 21). Copyright 1990. Used with permission.

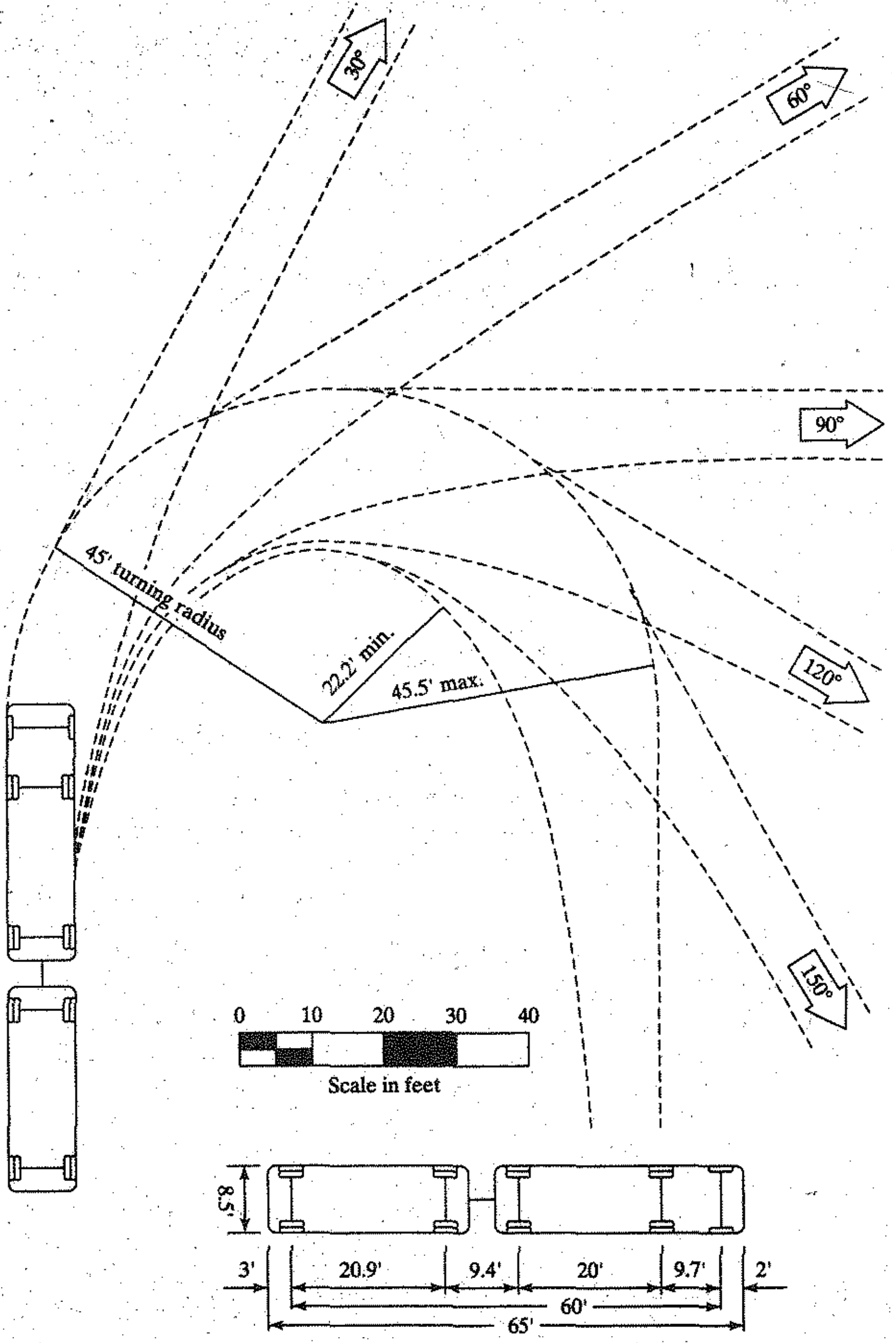


Figure 2.4.20 Minimum turning paths for a W-60 design vehicle.
(From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. II-7, p. 31.) Copyright 1990. Used by permission.)

TABLE 2.4.5 Minimum Turning Radii of Design Vehicle

Design vehicle type	Symbol	Minimum design turning radius (ft)	Minimum inside radius (ft)
Passenger car	P	24	13.8
Single-unit truck	SU	42	27.8
Single-unit bus	BUS	42	24.4
Articulated bus	A-BUS	38	14.0
Semitrailer, intermediate	WB-40	40	18.9
Semitrailer, combination, large	WB-50	45	19.2
Semitrailer–full trailer combination	WB-60	45	22.2
Interstate semitrailer	WB-62[a]	45	9.1
Interstate semitrailer	WB-67[b]	45	00
Triple semitrailer	WB-96	50	20.7
Turnpike double semitrailer	WB-114	60	17
Motor home	MH	40	26.0
Passenger car with travel trailer	P/T	24	2.0
Passenger car with boat and trailer	P/B	24	6.5
Motor home and boat trailer	MH/B	50	35

[a]Design vehicle with 48-ft trailer as adopted in 1982 STAA (Surface Transportation Assistance Act).

[b]Design vehicle with 53-ft trailer as grandfathered in 1982 STAA (Surface Transportation Assistance Act).

Source: (From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Table II-2, p. 22). Copyright 1990. Used by permission.

At-grade intersections are classified into Y-, T-, four-leg, multileg, and rotary intersections. Within each category a very large number of variations is possible and, in fact, no two intersections are exactly the same. Consequently each intersection must be treated individually as a separate design problem. Each design consists of the placement of combinations of triangular and elongated islands, the latter including medians and median treatments, edge-of-pavement treatments, pavement markings, and associated signing and traffic controls.

The channelization devices must be of sufficient number and size to command the attention of motorists, but cluttered designs containing too many small islands and signs must be avoided. The design should provide natural and well-defined paths to minimize vehicle wander. It must enhance the confidence and convenience of drivers by affording them adequate sight distances, clearly guiding them into the proper channels of movements, and preventing the choice of prohibited paths. The possibility of multiple paths between the same two points must be eliminated, and drivers should not be required to make many decisions at the same time.

Whenever possible, especially at high-speed locations, auxiliary storage lanes should be provided for turning vehicles that are required either to slow down or to stop. Moreover, the angles and areas of conflict must be controlled. The length over which merging and diverging movements are accommodated must be of adequate length, and these movements must be confined to low angles. Crossing paths should minimize the area of conflict, and hence must be as close to right angles as possible.

Among the channelization concepts illustrated in Fig. 2.4.21 are the rounded or bullet-shaped median openings and the corner curb radii that are designed to fit the paths

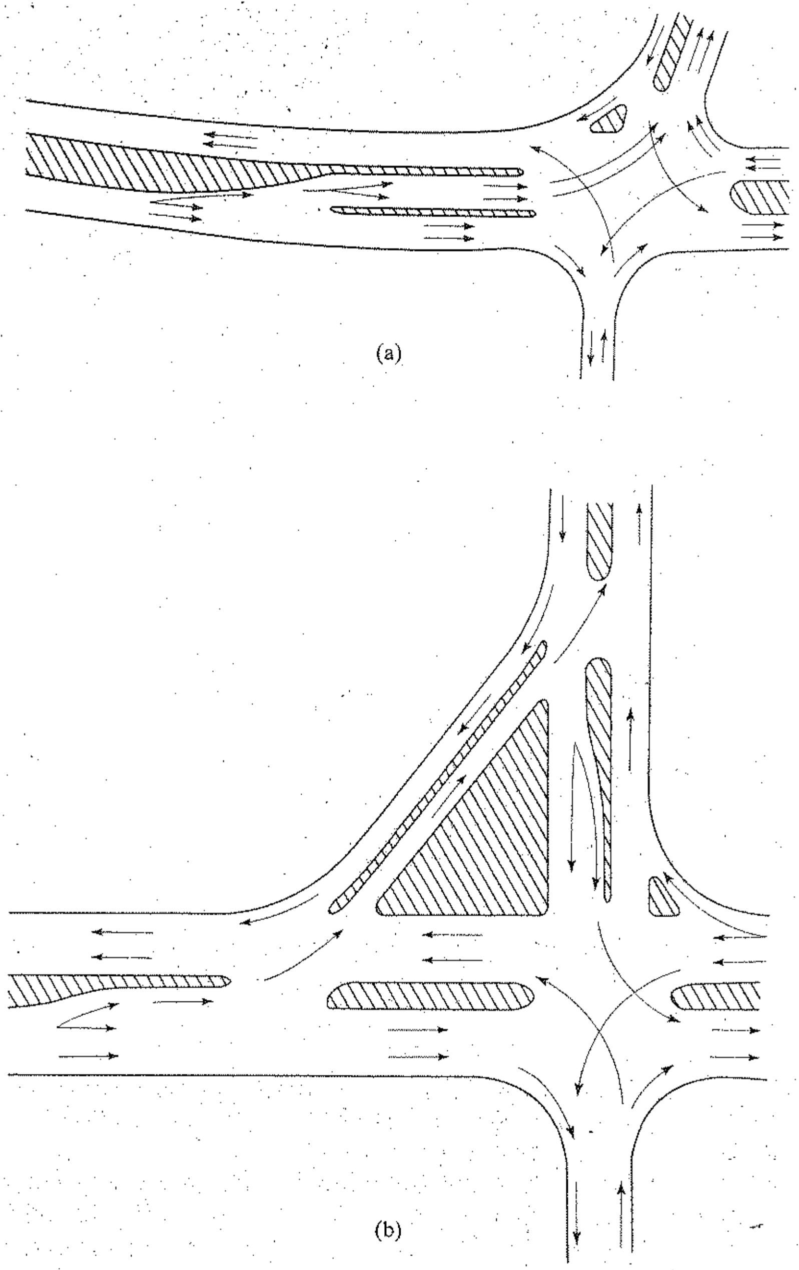


Figure 2.4.21 Example of channelization.
(From *A Policy on Geometric Design of Highways and Streets,* Copyright 1990, by the American Association of State Highway and Transportation Officials, Washington, DC [2.2] (Fig. IX-7, p. 682.) Copyright 1990. Used by permission.)

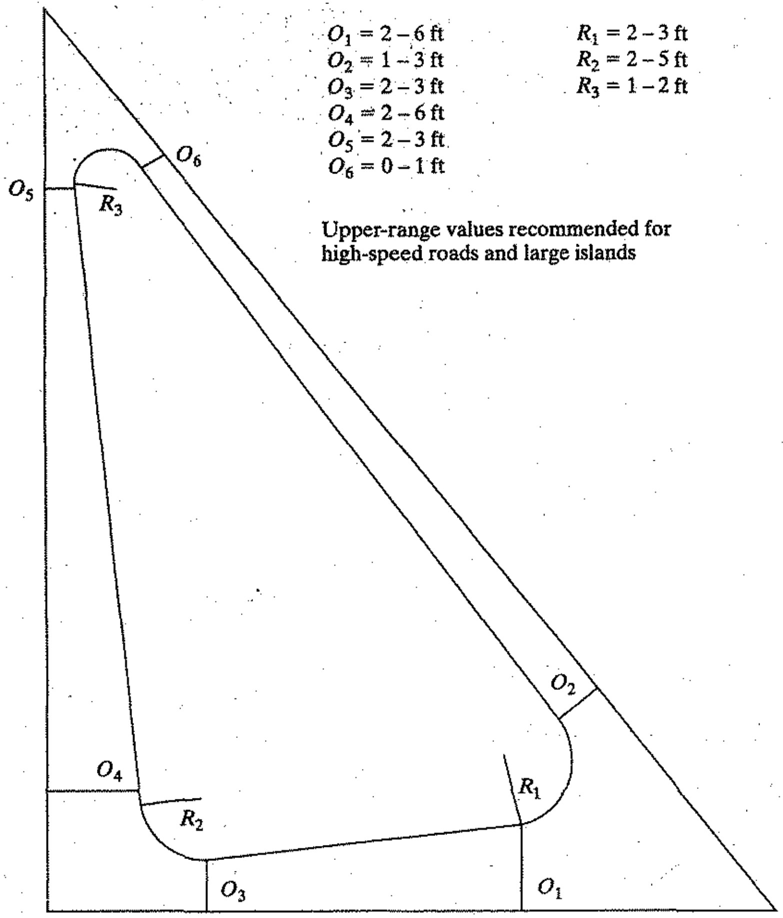


Figure 2.4.22 Radii and offsets of traffic islands. Offsets measured from the geometric island fitting the channelization specifications.
(From Transportation Research Board [2.11].)

of vehicles as closely as possible. The corners of islands are also rounded, and the island edges on the approach side are slightly offset to guide, or "catch," the vehicle into the proper channel. Figure 2.4.22 illustrates this requirement.

AASHTO [2.2] provides specific guidelines for adherence to channelization principles and for the design of appropriate treatments. In addition, the Transportation Research Board issued a report [2.11] in which nine principles of channelization and proven techniques for the cost-effective design of channelized intersections are described. Table 2.4.6 summarizes how various design elements can be implemented to address the nine principles of channelization. Figure 2.4.23 illustrates how approach alignment and physical channelization devices can be used to define clearly proper vehicle paths. The dashed lines in the figure delineate the space requirements of design vehicles. Figure 2.4.24 shows how the proper placement of channelization devices can be used to discourage undesirable or

TABLE 2.4.6 Objectives of Channelization and Design Elements to Achieve Them

	Objectives of channelization								
Design elements	Prohibitions of movements	Definition of vehicle paths	Promotion of safe speeds	Separation of conflicts	Angles of cross and merge	Facilitation of high priority movements	Facilitation of traffic control	Accommodation of slow or decelerating vehicles	Safe refuge for pedestrians
Traffic lanes						●	●	●	
Traffic islands	●	●		●	●		●		●
Median dividers	●	●		●	●		●		●
Corner radii	●	●	●			●			●
Approach geometry	●	●	●	●	●	●			
Pavement tapers or transitions		●	●		●			●	
Traffic control devices	●			●		●	●		●

Source: Transportation Research Board [2.11].

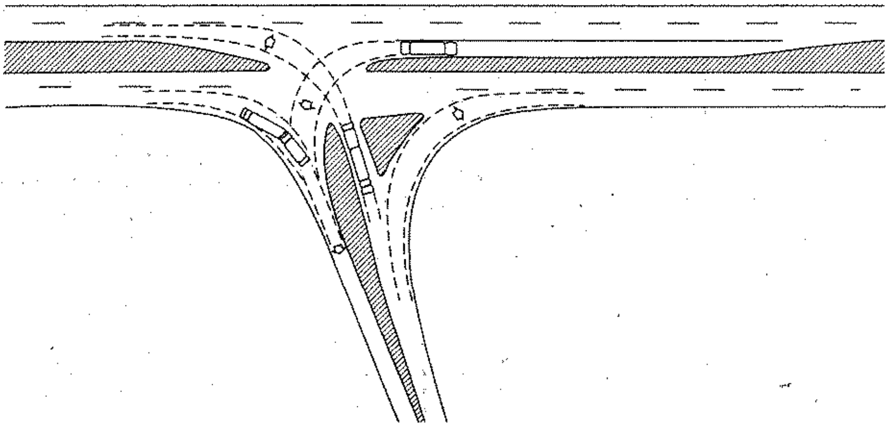
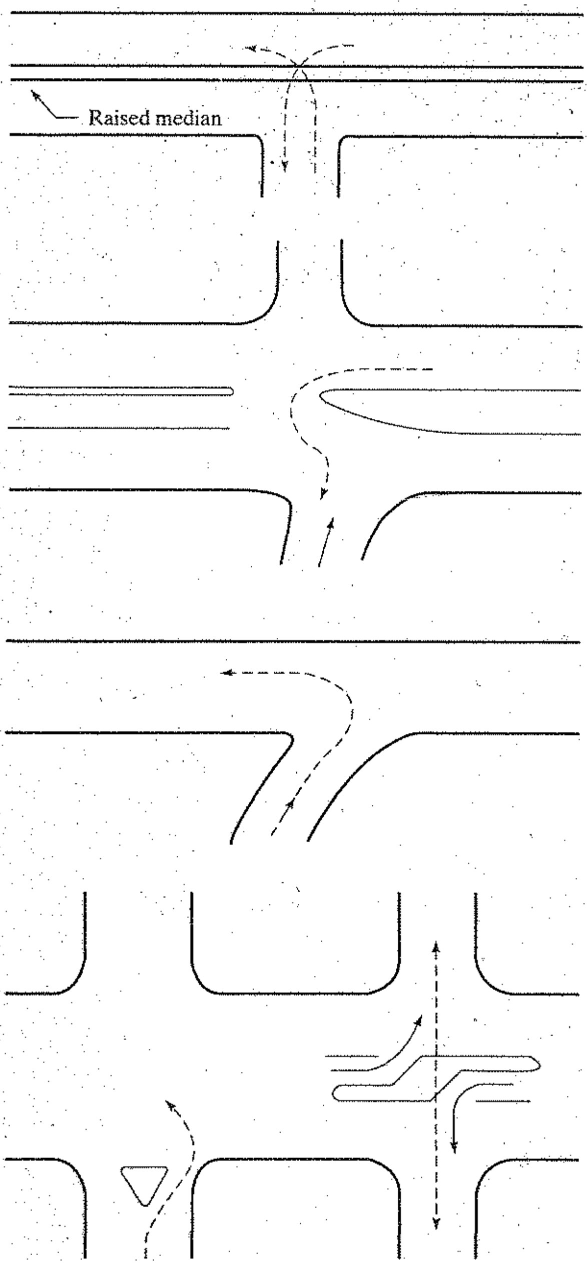


Figure 2.4.23 Design of all median elements must consider the natural paths of design vehicles.
(From Transportation Research Board [2.11].)



Raised medians block left turns to and from minor streets or driveways. Such treatment may be appropriate at locations where left turns are dangerous or cause congestion.

Placement of median channelization and design of corner radii can effectively discourage dangerous wrong-way movements onto freeway ramps without hindering other intended movements.

Alignment of the approach and design of corner radii can encourage right-turn-only movements and discourage undesirable left turns.

Raised traffic islands can block through movements or undesirable turning movements without hindering other intersection movements.

Figure 2.4.24 Sample designs to restrict or prevent undesirable or wrong movements. (From Transportation Research Board [2.11].)

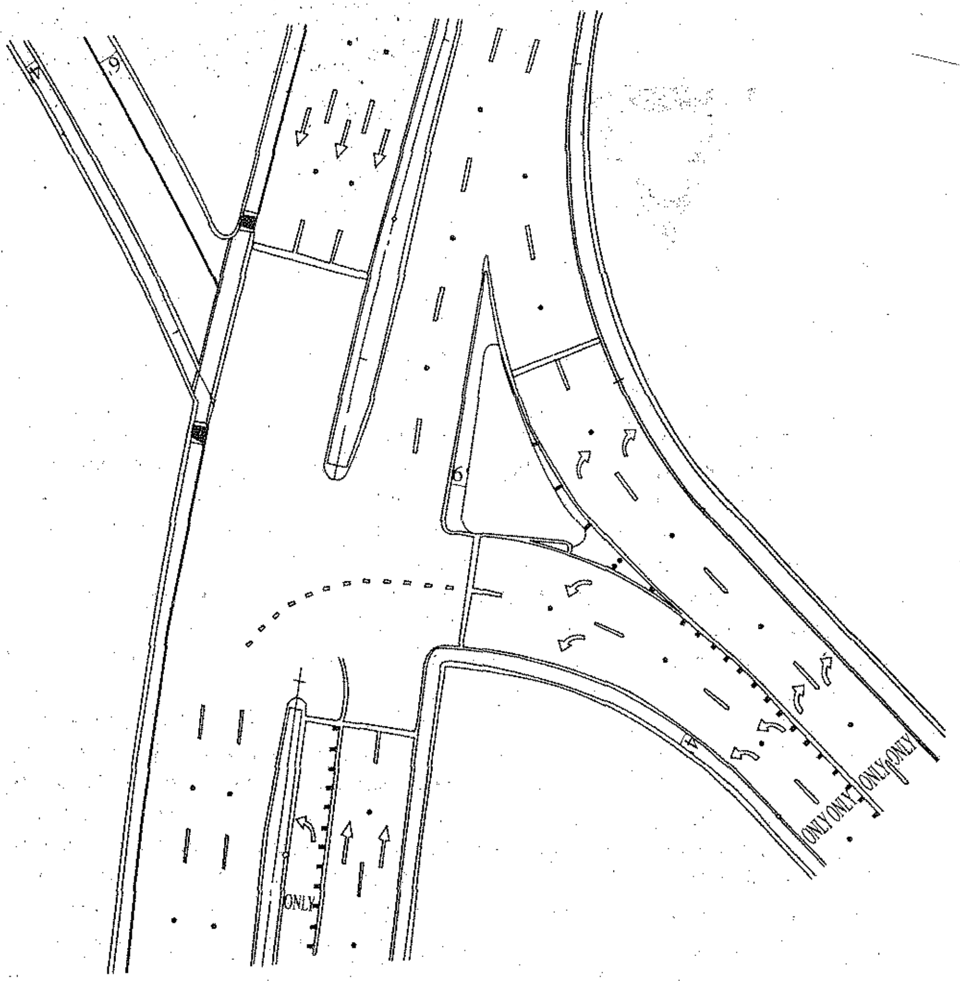


Figure 2.4.25 Sample design and application of channelization and pavement markings at a complex intersection.
(From Transportation Research Board [2.11].)

wrong-way movements (shown by dashed lines). Figure 2.4.25 shows an application of physical channelization and pavement markings at an intersection involving a double left-turn movement. The pavement markings shown are consistent with the provisions of the *Manual on Uniform Traffic Control Devices* [2.12].

2.4.14 Modern Roundabouts

A relatively recent development in the United States was the introduction of modern roundabouts as a means of enhancing the capacity and safety of at-grade intersecting streets. Roundabouts have been used extensively for decades in European countries and more recently in Australia. This section describes the basic function and geometric characteristics of modern roundabouts and distinguishes them from older-type traffic circles.

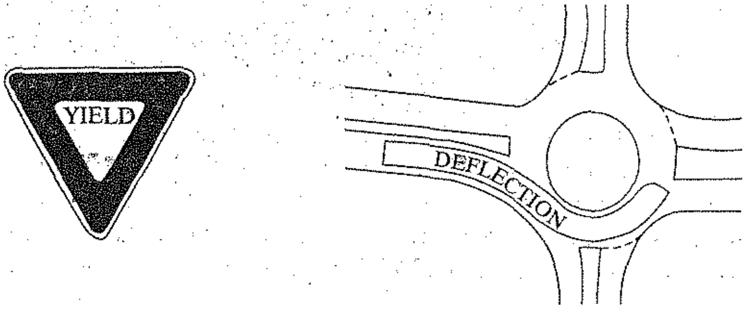


Figure 2.4.26 Yield-at-entry and deflection of entering traffic [2.13].

Many ancient cities had a radial structure focused in a central square where typically a monument of cultural significance was erected. Approaching people and carriages would circulate around the central monument when transferring between roadways. The same concept was retained over the years even after the introduction of mechanized transportation systems. One distinguishing characteristic of the latter was the requirement that centrally circulating traffic would move in one direction (counterclockwise in the United States). Known variably as *traffic circles, rotaries,* or plainly *roundabouts,* such designs are found in older U.S. cities. It is reported that the first one-way rotary system was established in 1904 at New York City's Columbus Circle [2.13].

These old-style (or "nonconforming") circles contain certain combinations of geometric and traffic control characteristics that are avoided with the use of modern roundabouts. Examples of such characteristics are included in the following discussion.

Many nonconforming circles are designed to give priority to entering vehicles approaching the circle on major streets. This is often accomplished by interrupting the circulating traffic through the use of traffic signals or stop signs. Moreover, the preferred through movement is allowed to enter and cross at high speeds. In some cases the design allows relatively straight paths for these preferred movements (i.e., paths tangential to the central circle or even cutting through the central island). Larger circles are designed so as to provide for long weaving sections between entrances and exits and this imposes a capacity limitation at those locations. Other features found at some nonconforming circles include permitting vehicle parking within the circle and allowing pedestrians to cross onto the central island.

By contrast, the basic philosophy of modern roundabouts is to give priority to vehicles that are already within the circulatory roadway. In 1966 Great Britain officially adopted this "priority-to-the-circle" (also known as "off-side" priority or "yield-at-entry") rule. The major implication of this rule is that all entering traffic can enter the circulatory roadway only when a safe (or "acceptable") gap is found. Entering vehicles are consequently controlled by a YIELD sign and are deflected by the central island to the right as illustrated in Fig. 2.4.26. Taken together, these two treatments induce entering vehicles to reduce their speeds and thus increase the safety level.

Figure 2.4.27 illustrates the major features of modern roundabouts. The marked *yield line* is where entering vehicles are required to wait, if necessary, for a gap. Raised *splitter islands* separate entering from exiting traffic, induce entering vehicles to take a deflecting (rather than a straight) path, and serve as a refuge to pedestrians who are not allowed to cross the circulatory roadway. Depending on the size of the roundabout, the raised *central*

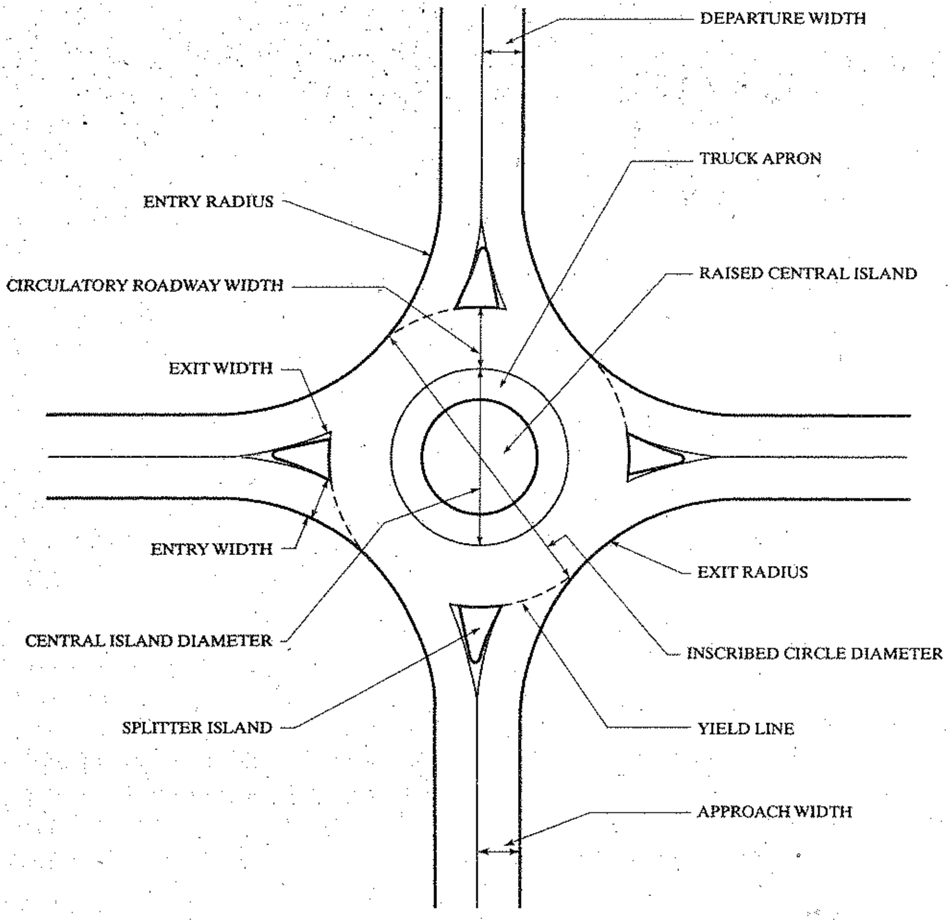


Figure 2.4.27 Geometric elements of a roundabout [2.13].

island may be surrounded by an *apron* to provide ample space for trucks, buses, emergency vehicles and other large vehicles to negotiate the roundabout. To discourage passenger vehicles from encroaching it, the apron is often delineated by a contrasting pavement color.

As of 1999 no standard guide for the design of roundabouts had emerged in the United States. At least two states, Maryland [2.14] and Florida [2.15], had issued their own guidelines. In 1971 the British Ministry of Transport issued its first design guide and several revisions followed between 1971 and 1993 [2.16]. French, German, and Australian design guides also existed at that time. The decision of whether to implement a roundabout treatment rather than channelized signalized intersections, two-way stop control (TWSC) intersections or all-way stop control (AWSC) intersections must be based on a careful study of the safety, capacity, space requirements, and cost of each alternative. Some computer software (see Chapter 15) can perform simulation and capacity analysis of roundabouts. The software includes NETSIM, which can be used to approximately model a roundabout as a series of yield signs, SIDRA (a product of the Australian Research Board), and ARCADY (a product of the United Kingdom's Transport Research Laboratory).

2.4.15 Traffic Calming Devices

The choice of *design speed* is critical to the geometric design of roadways. The design speed, for example, affects the choice of curve radius, the rate of superelevation, and the required safe sight distances. Once selected, the design speed becomes an intrinsic attribute of the roadway. Thus if a horizontal curve is properly designed for 60 mi/h, it can be traversed safely by vehicles traveling at that speed under design conditions.

The *posted speed* (or speed limit) on the other hand is typically set below the design speed. Traditional practice uses the measured 85th percentile speed of observed free-flowing vehicles as a first approximation of the proper speed limit, subject to factors such as the design speed, accident experience, conflicts with pedestrians, and parking maneuvers [2.17].

The posted speed is set below the design speed to minimize the frequency of unsafe conditions encountered by those drivers that, for whatever reason, choose to exceed it. Other reasons are based on considerations beyond kinematics. For example, many municipalities establish relatively low speed limits in the vicinity of schools or in residential neighborhoods out of concern for the safety of children, pedestrians, and bicyclists. In addition to excessive speeds, ITE identified "unwanted" through traffic and curb parking on neighborhood streets by people whose actual destinations are outside the neighborhood as sources of "a basic discrepancy between . . . vehicular traffic and the tranquility of a residential street" [2.18].

Because of the existing geometry of many local streets, posted speed limits have been neither successful in inducing drivers to reduce speeds nor have they discouraged drivers from using them for through movements: Wide, straight roadways with long sight distances and smooth pavement surfaces send drivers a contrary message. To counter these driver tendencies, geometric treatments or *traffic calming* devices have been finding increasing application in the United States since 1980. It is generally accepted that this trend had its modern beginnings in the Dutch city of Delft in 1970. In 1976 the Netherlands Ministry of Transport and Public Works adopted official standards for residential precincts (or *woonerven*) that incorporated these concepts [2.19].

The term *traffic calming* itself is a translation of the German word *verkehrsberuhigung* which was associated during the 1970s with a planning philosophy that residential neighborhoods should be designed to give preference to residents rather than to the automobile. The concepts of *livable communities* and *neotraditional* urban design are often used as the larger context for this objective (e.g., [2.20, 2.21].)

Traffic calming strategies can encompass a wide variety of options that include simple traffic control actions, such as the use of stop signs, striping and turning restrictions, automobile-free zones including pedestrian and transit malls, as well as regulatory policies, enforcement strategies, parking regulations, and community design principles. ITE defines traffic calming in a more restricted sense as

> the combination of mainly physical measures that reduce the negative effects of motor vehicle use, alter driver behavior and improve conditions for non-motorized street users. [2.22]

Two classes of measures are generally available: those that aim mainly at physically discouraging spillover traffic from nearby congested arterials from using local streets as bypasses and those that aim principally at speed reduction. Both types of actions may be applied either to retrofit existing areas or to be part of the design of new areas. They are applicable not only to residential neighborhoods but also to activity centers and rural locations.

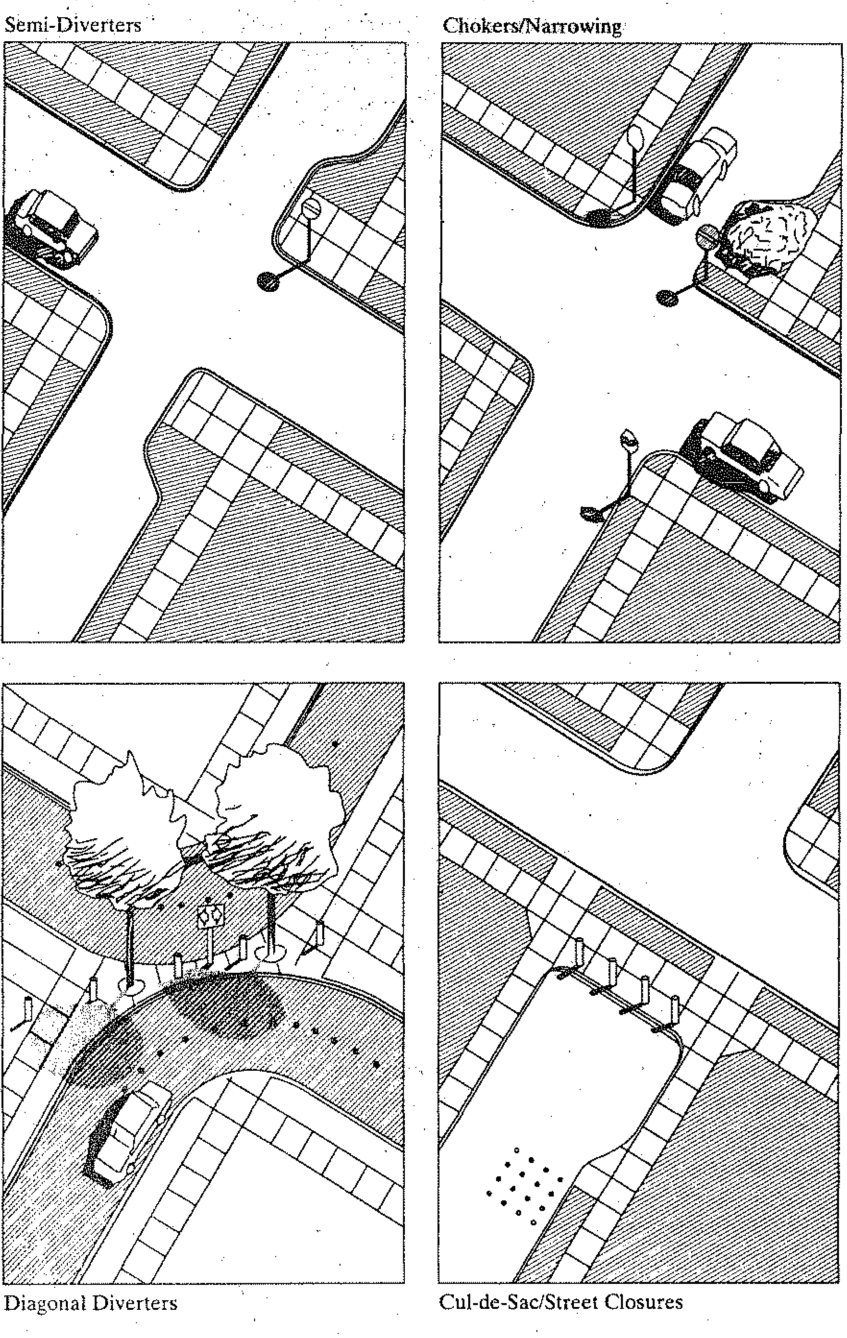
One way to discourage "unwanted" traffic from entering a traffic-calmed area is to eliminate unimpeded straight paths that are often present in the typical gridiron pattern of neighborhood street design. This may be accomplished by a number of devices, including *median barriers* such as those illustrated in Fig. 2.4.24. Other means include the introduction of *cul-de-sacs, diverters,* and *chicanes* (e.g., Fig. 2.4.28). Cul-de-sacs can be applied at existing intersections or midblock. *Diagonal diverters* are implemented at existing intersections to divide them into two independent circuitous paths, whereas *semidiverters* (or half-closures) are used to restrict entrance to otherwise two-way streets at the far side of an intersection. *Chicanes* transform straight segments of streets into S-shaped paths by introducing alternating obstructions on either side of the roadway.

Methods aiming primarily at speed reduction may be classified into *mini-roundabouts* or Seattle-type traffic circles, *street narrowings* of various kinds, and *vertical undulations.* Mini-roundabouts employ the yield-at-entry and diversion principles described in Section 2.4.14, even though in some instances nonconforming features, such as requiring minor street traffic to stop, may be introduced. Street narrowings (Fig. 2.4.28) include *bulbouts* or *neckdowns* that typically narrow approaches to intersections and *chokers* that reduce the width of streets midblock. In addition to inducing drivers to reduce their speeds, these devices also reduce pedestrian crossing distances.

Vertical undulations include *rumble strips, speed humps,* and *speed tables.* Rumble strips introduce a variable texture or roughness to segments of the pavement to let drivers know that they should slow down. Speed humps are by far the most popular traffic calming devices in the United States [2.23]. They differ from *speed bumps* in that they can be negotiated at selected design speeds, usually around 20 mi/h. Speed bumps, on the other hand, are very short and abrupt in the direction of travel (typically less than 3 ft wide), causing drivers to slow down to almost a stop (see Fig. 2.4.29). Humps are anywhere from 12 to 20 ft wide and rise to heights of 3 to 4 ins. To be visible to approaching vehicles, they may be striped with diagonal white lines. Their cross section (in the direction of travel) can be sinusoidal, circular, or flat-topped. *Cushions* are similar to humps except that they do not extend across the entire cross section of the travel way. This allows wide-bodied vehicles, such as buses, and bicycles to pass unimpeded but cause automobiles to slow down. Speed tables (or *plateaus*) are similar to flat-topped humps but are much wider. When placed at intersections, they result in *raised junctions* rising to the height of the surrounding sidewalks. They are often constructed with contrasting pavement textures and are delineated with *bollards* to prevent vehicles from encroaching on the sidewalks.

A traffic calming design for a particular area is usually custom-made to meet problems that are specific to the area. It must consider the appropriate combination of devices, appropriate spacing between them, and accompanying warning signs and lighting. Many municipalities have established procedures that place high emphasis on resident participation and approval. Design guidelines by both national standards entities and municipalities are emerging.

Despite their potential to discourage cut-through traffic and to reduce speeds, as illustrated in Fig. 2.4.30, traffic-calming actions also cause certain negative impacts that must be considered. Among these are adverse effects on emergency vehicle response times



Semi-Diverters

Chokers/Narrowing

Diagonal Diverters

Cul-de-Sac/Street Closures

Figure 2.4.28 Various neighborhood traffic control measures. (From National Cooperative Highway Research Program, Synthesis of Highway Practice 139, *Pedestrians and Traffic Control Measures,* Transportation Research Board, National Research Council, 1988. [2.24])

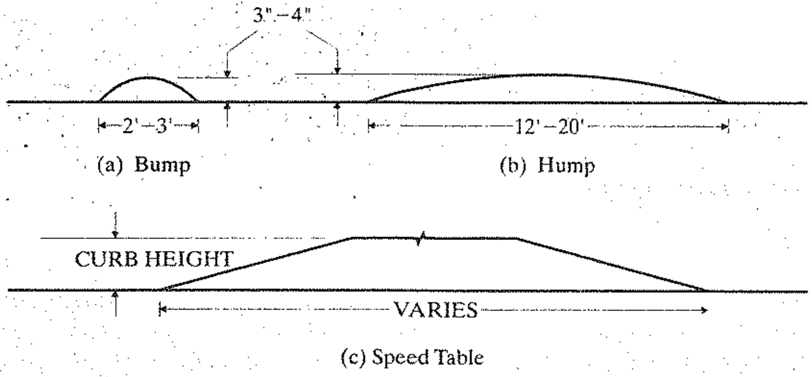


Figure 2.4.29 Vertical undulations.

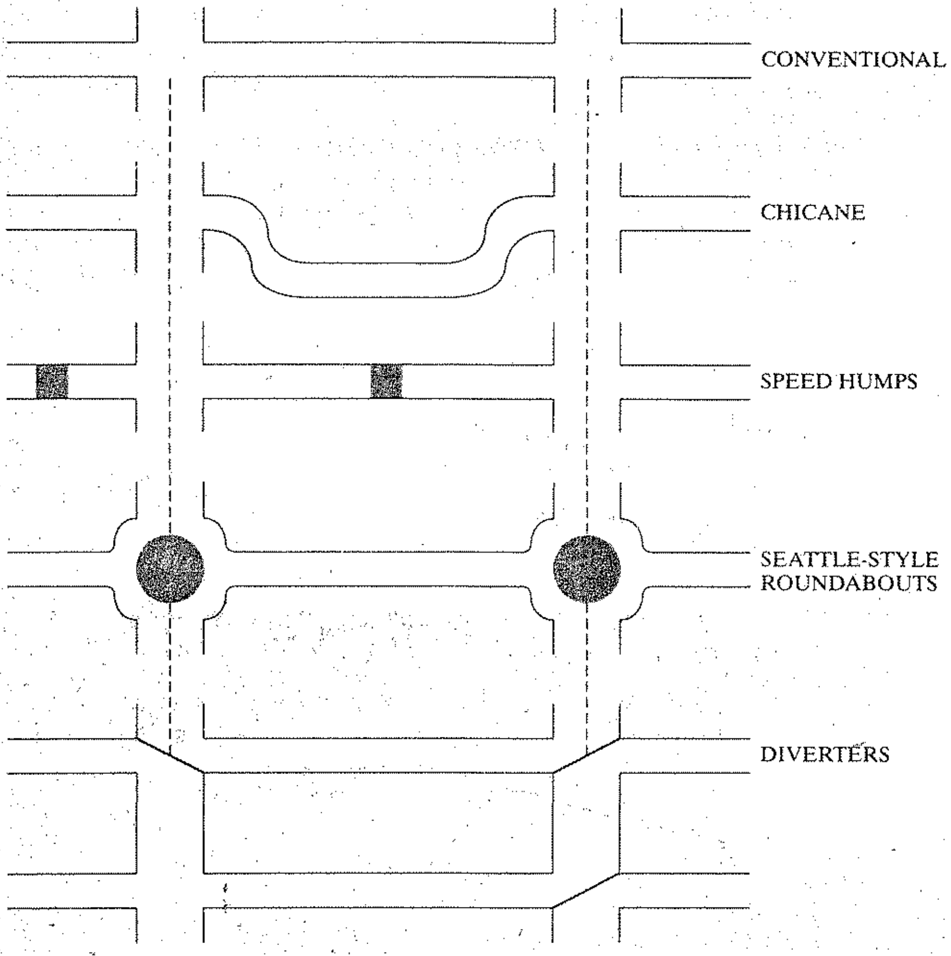


Figure 2.4.30 Basic roadway calming designs.

(e.g., [2.25]), removal of parking spaces, a proliferation of warning and advisory signage, increases in traffic noise, restricted bicycle movements, and inconvenience to residents of the area. In several cases traffic calming devices demanded by neighborhood residents have been subsequently removed because of such adverse impacts.

2.5 PAVEMENT STRUCTURES

2.5.1 Background

Pavements serve structural, functional, and safety purposes. They are necessary not only for roadways but also for parking lots, airports (i.e., runways, taxiways, aprons, and service roads), industrial sites, ports, and so forth. The structural performance of a pavement is aimed at distributing the loads under the wheels of vehicles over larger areas to prevent stressing, beyond its load-bearing capacity, the native soil (or subgrade) on which the pavement system is constructed. Figure 2.5.1 illustrates this situation: The load at the interface between the wheel and the pavement surface is applied over a relatively small area, causing high stresses at that point, but these stresses decrease with depth as the load is spread over larger areas. The degree of load distribution decreases from the top to the bottom of the pavement structure.

The functional performance of pavements is related to the users' requirement for smooth and comfortable riding conditions. The quality of riding comfort is typically measured by the Present Serviceability Index (PSI), which was developed in 1957 by the American Association of State Highway Officials (AASHO). The PSI is based primarily on measurements of pavement roughness. This is accomplished by a variety of available equip-

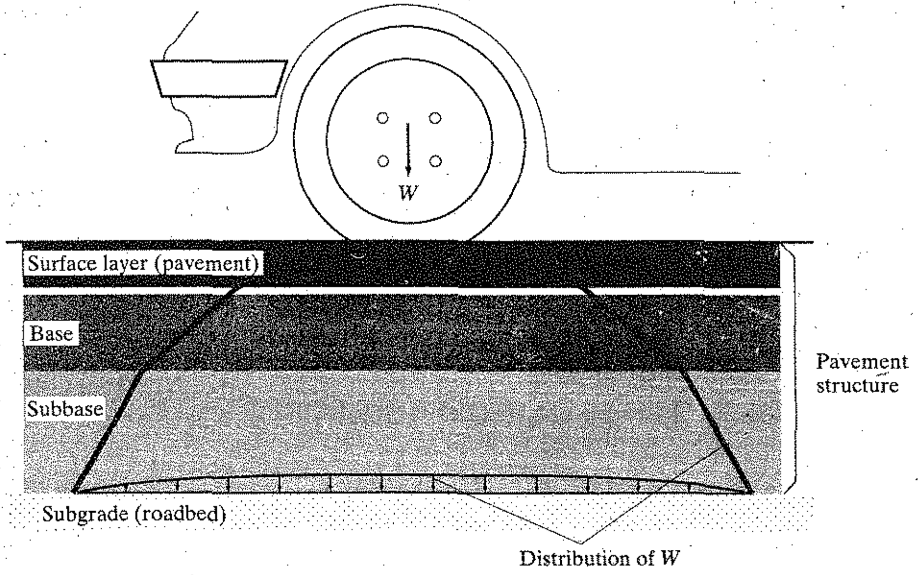


Figure 2.5.1 Distribution of weight of wheel from the contact area to the native soil.

ment that essentially measure the profile of the pavement along the traveled way. The PSI deteriorates with usage and pavement age and is one of several criteria employed to aid decisions relating to maintenance, rehabilitation, or reconstruction of the pavement. Typical symptoms of pavement distresses are longitudinal and transverse cracking, breaking, swelling, and heaving. They affect the structural integrity of a pavement and the level of service to the users [2.26].

The safety performance of pavements is mainly related to the skid resistance developed at the pavement–tire interface. This friction or skid resistance can be enhanced by the choice of materials and the various treatments, such as texturing the pavement surface. Another characteristic that is related to safety is the light reflectance of the pavement surface.

2.5.2 Pavement Materials and Types

Throughout the ages compaction of the native soil by repetitive use was the common way in which roadbeds were maintained. Some exceptions were the heavy stone roadways constructed by some advanced ancient civilizations, including the Romans, whose famous Appian Way is extant to the present day. Any attempt to replicate the works of these ancient societies given today's material and labor costs would be prohibitively expensive. A derivative of this method of roadbuilding, known as the French method, was used into the nineteenth century, and the practice of overlaying a prepared roadbed with natural or artificial stones (e.g., cobblestones) continued until fairly recently. Other roadbed treatments include artificial compaction and stabilization of unpaved roads, typically found in rural areas, on farms, and at construction sites; surface treatments of the native soil; the use of asphalt-based pavements; and the construction of portland-cement concrete pavements. The last two are the most common types of pavements in use today.

The Asphalt Institute [2.27] describes asphalt as a "strong cement, readily adhesive, highly waterproof, and durable. It is a plastic substance which imparts controllable flexibility to mixtures of mineral aggregates with which it is usually combined. Although a solid or semisolid at ordinary atmospheric temperatures, asphalt may be readily liquefied by the application of heat." Asphalt is classified as a "bituminous" cement, a term which refers to the fact that it consists of hydrocarbons.

The combination of asphalt with graded mineral aggregates is known as asphalt concrete. It is in this manner that asphalt is usually applied to pavement design. Asphalt concrete mixtures are sometimes discussed in terms of the gradation of their component mineral aggregates and fillers into categories such as open-graded, coarse-graded, or fine-graded. Open-graded aggregate contains little mineral filler material, and consequently it is characterized by relatively large void areas between aggregate particles in a compacted mix. Coarse-graded aggregates exhibit a continuous grading of sizes but show a predominance of coarse sizes, whereas fine-graded aggregates have a predominance of fine sizes (i.e., those passing the No. 8 sieve). A coarse aggregate of uniform size, known as a macadam aggregate, received its name from the Scottish engineer John McAdam, who first used it in an asphalt mix. Sheet asphalt refers to a special type of mix that contains a well-controlled combination of asphalt, sand (i.e., fine aggregate), and mineral filler.

When the asphalt concrete is placed, spread, and compacted at atmospheric temperature, it is referred to as a cold-laid mixture, in contrast to hot-laid mixtures, which involve elevated temperatures. The combinations of material characteristics and proportions on one hand and mixing and placing conditions on the other can lead to asphalt concretes of differing characteristics in terms of their stability, durability, and flexibility to suit a variety of application requirements [2.27].

It is interesting to note that natural deposits of asphalt exist. For example, skeletons of prehistoric animals have been preserved in such deposits at the La Brea pit near Los Angeles, CA. Asphalt material was used in Mesopotamia for roadway construction and waterproofing purposes. According to the Asphalt Institute, imported rock asphalt was used in Philadelphia, PA in 1838, and the first asphalt pavement was built in Newark, NJ in 1870. Today asphalt is recovered from petroleum in the process of separation and refinement of constituents.

Portland cement is mainly a calcium aluminum silicate that is produced by fusing limestone and clay in a rotary kiln to form a clinker material, which is then ground into a fine powder. First used by an Englishman, Joseph Aspdin, who patented the substance in 1824, portland cement derives its name from its ability to react with water and, through hydration, to produce an artificial stone that resembles the limestone deposits found on the Isle of Portland in England. It is a "nonbituminous" cement and is classified as a "hydraulic" cement because it solidifies under water. In various combinations with water, graded mineral aggregates, and other admixtures and additives, portland cement is used to produce numerous construction materials, such as grout, plaster, mortar, and portland-cement concrete. By controlling the combinations of the constituents of the mix, a wide variety of desirable characteristics (e.g., strength, durability, and workability) can be obtained to suit particular applications. The most notable characteristic of portland-cement concrete is its compressive strength: It far exceeds its tensile strength, which is only about 10% of its compressive strength. For this reason portland-cement concrete pavements are typically designed to resist compressive forces only. The first reported use of portland-cement concrete for pavement construction in the United States occurred in 1891 in Bellefontaine, OH.

In case prominent tensile forces are present under particularly heavy loads continuous reinforcement with steel bars is applied to withstand the applied tension. Also, prestressing the structural element is another way for taking full advantage of the compressive strength. This involves the application, via tendons embedded in the concrete, of a compressive load on the structure prior to applying the service loads. The compressive prestress counterbalances tensile stress produced by the loads in such a way so that the concrete remains under compression. Continuously reinforced or prestressed concrete pavements are commonly used for airport runways and aprons.

A fundamental difference between asphalt and portland-cement concrete pavements lies in the fact that the former is characterized by flexibility, whereas the latter provides rigidity. For this reason the two types of pavement are classified, respectively, as flexible and rigid pavements. The following are the corresponding AASHTO definitions [2.26].

Flexible pavement. A pavement structure which maintains intimate contact with and distributes loads to the subgrade and depends on aggregate interlock, particle friction, and cohesion for stability

Rigid pavement. A pavement structure which distributes loads to the subgrade, having as one course a portland-cement concrete slab of relatively high bending resistance

Rigid pavements are further subdivided according to the method of reinforcement into plain (unreinforced) with or without dowels, conventionally reinforced, continuously reinforced, and prestressed. Dowels are steel rods that connect individual pavement slabs to facilitate the transfer of loads between them. In the case of continuously reinforced portland-cement concrete pavements the reinforcing steel serves this function.

Both types of pavement tend to deform under the applied loads as illustrated in Fig. 2.5.2, with the top fibers of the pavement in compression and the bottom fibers in tension. Because of the rigidity and stiffness expected from portland-cement concrete pavements, they require special structural engineering attention. This includes the provision of contraction joints at recurrent intervals to control transverse cracking, expansion joints, load transfer devices, and joint sealants to prevent water and incompressible debris from entering the joint reservoirs. Figure 2.5.3 shows an undoweled and a doweled contraction joint as well as an example of doweled contraction joints and transverse joints with distributed steel.

Composite pavements consisting of an asphalt surface overlay on a portland-cement concrete slab are common. Moreover, a recent practice of overlaying a portland-cement concrete layer on old asphalt pavements, called "whitetopping" appears to be gaining in popularity.

2.5.3 Pavement Structure

A pavement structure consists of a series of layers, beginning with the native soil that constitutes the prepared roadbed (or subgrade), which is typically overlaid by the subbase and base layers. The strength of these layers increases from the bottom up to conform with the

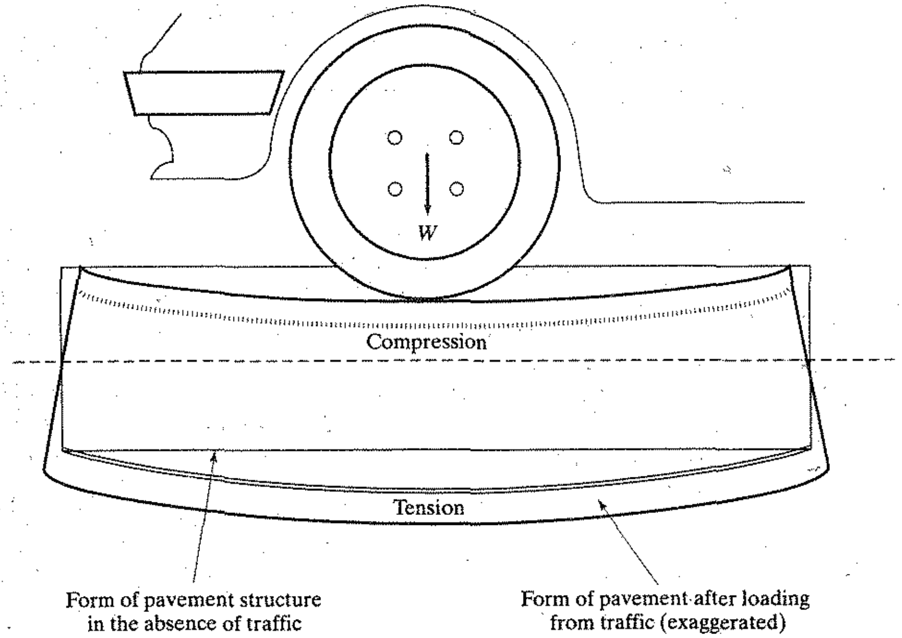

W
Compression
Tension
Form of pavement structure in the absence of traffic
Form of pavement after loading from traffic (exaggerated)

Figure 2.5.2 States of pavement structure with and without traffic load.

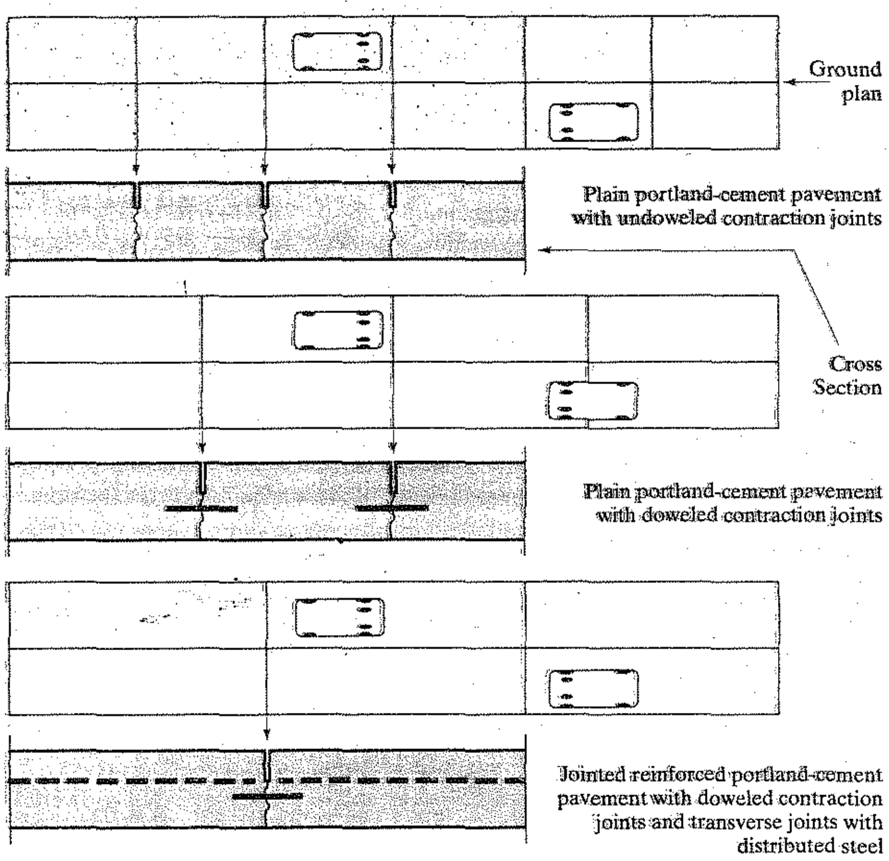

Ground plan
Plain portland-cement pavement with undoweled contraction joints
Cross Section
Plain portland-cement pavement with doweled contraction joints
Jointed reinforced portland-cement pavement with doweled contraction joints and transverse joints with distributed steel

Figure 2.5.3 Basic types of jointing for portland-cement concrete pavement slabs.

increasing requirements of stress and load distribution (Fig. 2.5.1). This practice contributes to economical and efficient use of materials by avoiding "overdesigning" the lower layers. Some of these layers may be omitted, depending on the strength of native soil and material availability. An asphalt pavement that is placed directly on the subgrade is known as a full-depth asphalt pavement. In this case the thickness of the base and subbase is simply replaced by appropriate "layers" of asphalt concrete to form a thicker layer than would otherwise be required. In the case of rigid pavements the wheel loads are distributed over larger areas than those in the case of asphalt pavements, and this reduces the strength requirements of the base and subbase. Figure 2.5.4 presents examples of flexible and rigid pavements. A flexible pavement may include the following five layers (from bottom up): (1) prepared roadbed; (2) subbase course, typically from compacted granular material; (3) base course, which provides structural support to the pavement, made from aggregates such as crushed gravel, which may be treated with fly ash, cement, or asphalt; (4) drainage layer, which is part of the base made with select aggregates or fabrics with sufficient permeability so that water can be quickly removed from the structure; and (5) surface course mix of bituminous materials and aggregates. The proportion of bituminous material and aggregates as well as

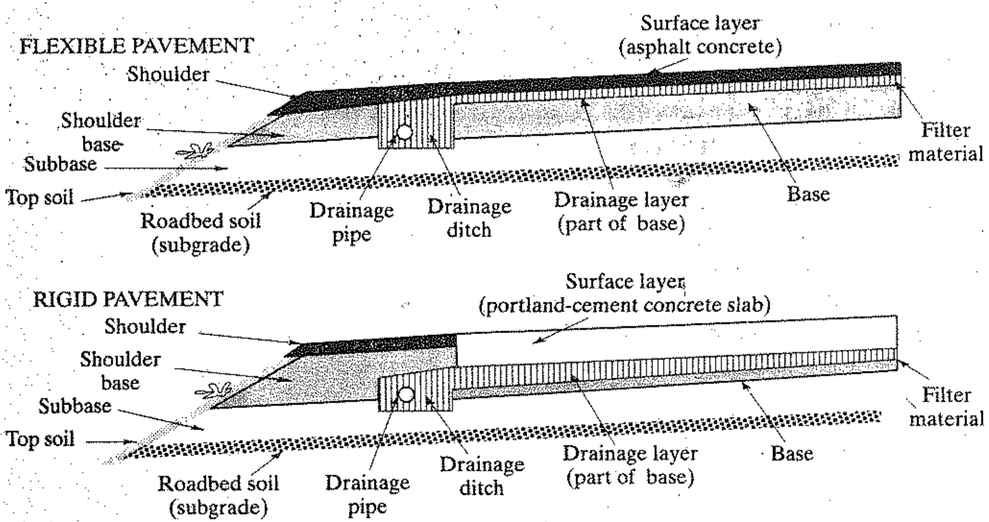


Figure 2.5.4 Cross-section examples of flexible and rigid pavements.

the gradation, strength, abrasion, and other characteristics of the aggregates determine the resistance to cracking and the supplied skid resistance. The surface mix must be properly compacted.

The rigid pavements usually consist of four major layers: (1) prepared roadbed (as for flexible pavements); (2) subbase course, with characteristics similar to those for flexible pavements with the exception that often lean concrete (econocrete) subbases are constructed to reduce erosion of the bottom side of the slabs, the joints between slabs, and the edges of the slabs; (3) base course, which may contain a drainage layer (often the base and subbase are combined into one supportive layer, with or without a drainage layer); and (4) pavement slab, which is a concrete mix with portland cement, aggregates, and various optional admixtures. The interlocking of aggregates provides the mechanism with which loads are transferred in portland-cement concrete pavements (aggregate interlock).

The proportion of portland cement and aggregates as well as the gradation, strength, abrasion, and other characteristics of the aggregates determine the resistance to cracking and the supplied skid resistance. The slabs may be reinforced with steel; they may have steel reinforced joints (for the proper transfer of loads between slabs; see Fig. 2.5.3) and joint sealants. Continuously reinforced concrete pavements consist of one continuous slab without joints.

Drainage is a paramount element and the rule of thumb is avoidance of "bathtub" design, whereby rainfall or groundwater accumulates in the layers of the pavement structure. The drainage of water should be rapid; thus certain gradations of materials that supply a high proportion of voids and no fines as well as careful design should be implemented (Fig. 2.5.4). The incorporation of drainage in the pavement structure (often neglected) increases the initial construction cost, but it can prolong the life of the structure substantially, thereby offsetting the higher initial cost.

2.5.4 Pavement Design

Pavement design includes, among other elements, selection of a pavement type (i.e., flexible, rigid, or composite), the design of the concrete mix (i.e., asphalt or portland cement, gradation of aggregates, admixtures, etc.), the selection of materials for the soil layers below the pavement, and the thickness of the soil and pavement layers. This process is complex and hard to define unequivocally because of the multiplicity of engineering and nonengineering factors involved. According to AASHTO [2.26], "currently, the most realistic pavement type selection process can result by obtaining five to ten most nearly optimal cost solutions for each pavement type being considered and examining these options qualitatively in the light of [these] factors." For this reason AASHTO recommends designing pavements with the use of computers. Another reason is that manual application based on tables and nomographs is bound to result in multiple approximations that may compromise final accuracy. Major factors affecting pavement design include: (1) traffic load, (2) soils, (3) environment, and (4) reliability. These are discussed next.

Traffic load is the most important factor in pavement design because it largely determines the thickness of the pavement structure. The measurement of traffic load varies according to the method utilized. In general, pavements are designed according to the heavy vehicle traffic that they are expected to receive. The AASHTO [2.26] method utilizes 18-kip ESALs (equivalent single-axle load). All traffic is transformed in equivalent 18,000-lb single-axle loads. Large loaded trucks with trailers are estimated to cause loads the equivalent of which exceeds 1,000 passenger cars. For example, the ESAL factor for passenger cars is 0.0008 and for tractor-semitrailers with five or more axles it is 2.3719.

Other methods utilize the annual average daily traffic (AADT) and the percentage of heavy vehicles per category (i.e., single truck, semitrailer, etc.) in the AADT. The major difficulty with regard to traffic loads is that they need to be forecast over the design life of the roadway (i.e., 30 years). Forecast data are partly based on a time series of highway traffic data and truck weight station logs. Then growth rates are estimated or assumed to derive the future expected traffic.

Soils on which the roadway structure will be founded are of critical importance. Weak and/or unstable soil beds may require extensive soil improvement efforts before placement of the pavement structure. Expansive and frost-heave susceptible soils are most problematic for the placement of a pavement structure. The major load-bearing property of the soilbed considered is the resilient modulus. The resilient modulus of the compacted roadbed soil represents the amount of the recoverable deformation at given stress levels. Standardized tests are conducted to determine this property [2.26, 2.28].

Environment affects pavements mainly through rainfall and temperature. Rainfall penetrating the structure alters the properties of layers and makes the pavement vulnerable to loads. Temperature also affects the properties of pavements by generating stresses, contraction, and expansion. The combined effect of the presence of water in the pavement layers and low temperatures (i.e., below the freezing point) creates frost heaving (expansion), and in thawing periods the bearing capacity of a pavement may be greatly reduced (e.g., several rural roads in northern states limit the weight per axle during spring to avoid excessive pavement damage). Even with good drainage, thawing from the surface downward is most destructive since water from melted ice is trapped by the frozen layers underneath. Soils containing fine particles are most susceptible to frost heaving. Thus selection of properly graded soils and aggregates can ameliorate most of the frost heaving problem.

Reliability analysis of pavement structures accounts for a number of uncertainties over time. A pavement structure designed with average values has a 50% probability of fulfilling its required performance life. Thus the uncertainty in major design factors such as pavement structure factors, which include traffic loads, roadbed soil factors, climatic factors, and pavement condition (i.e., serviceability) factors must be taken into account. Proper adjustments can be made to the design factors to achieve a desired reliability level (i.e., 80 to 99% for most major streets and highways).

2.5.5 Design Methods

All or most of the preceding factors are incorporated in the two major pavement thickness analysis methodologies which apply to both flexible and rigid pavements: the traditional experimental/statistical method of AASHTO [2.26] and the mechanistic/empirical method advocated by the Asphalt Institute [2.27] and the Portland Cement Association [2.28]. The former is based on the 2-year extensive real-world pavement testings conducted in the late 1950s by AASHO (the then acronym of AASHTO) in Ottawa, IL, and the continuous collection of pavement performance data from the nationwide highway system. The method is a statistical modeling of the behavior of a large variety of pavements given their specifications (construction characteristics such as type, materials, thickness, base and subbase, etc.), the number of loadings until observable failure, and soil and environmental factors. The latter method is based on theory (i.e., it assumes that the pavement is a multilayered elastic structure on an elastic foundation) and is calibrated to conform to empirical observations of performance for a variety of prevailing conditions.

The selection of the type of pavement (rigid or flexible) is decided on the basis of a number of factors, such as economic (i.e., initial and life-cycle costs, availability of funds, etc.), local supply and availability of materials, past experience with various pavement types, construction considerations, continuity of pavement type, recycling opportunities, safety considerations (i.e., friction, contrast, reflectance), competition among industries, and local government preference.

2.5.6 Life-Cycle Economic Analysis

Life-cycle costs analysis accounts for several cost components, their time of occurrence, the design life of the structure, and its salvage value at the end of its design life. The main cost components are (1) the initial cost, which includes the land acquisition and all construction costs; (2) annual maintenance costs; (3) the designed rehabilitation cost (i.e., resurfacing at the tenth and twentieth years of a pavement with a design life of 30 years); and (4) user costs, which include vehicle operating costs (i.e., fuel consumption, tire wear, maintenance, etc., depend partly on the type and condition of the pavement), the user travel-time cost (which becomes a major component during rehabilitation stages when lanes or a segment of a roadway facility is closed to traffic), and the traffic accident cost (i.e., accidents attributed to the pavement's condition, or during construction and rehabilitation periods).

There are two widely used evaluation methods that aid in the choice of pavement design alternatives: the equivalent uniform annual cost method and the present worth method. In the former method all costs are translated into an equivalent annual cost for each year of the life of the structure. Thus annual costs remain unchanged and fixed costs are spread over the life of the structure. In the latter method all costs are collapsed to present time, thereby reflecting the total present time worth of the structure given its design life. Chapter 12 presents various evaluation methods in detail.

The critical element that allows either allocation of costs (uniform spread over all time points or collapse at a single time point) is the discount rate, which is the difference between the prime interest rate and the prevailing inflation rate. In other words, the discount rate reflects the true cost of borrowing money. The fluctuations of interest and inflation rates tend to counterbalance each other. As a result, the discount rate tends to remain fairly stable [2.28]. Consider the following example.

Paving a road with portland cement or asphalt concrete in the mid-1990s was estimated to cost the following amounts per mile:

- *Portland-cement concrete:* Initial cost = \$400,000, annual maintenance = \$500, 30-year life
- *Asphalt concrete:* Initial cost = \$325,000, resurfacing at the fifteenth year = \$100,000 (present time value), annual maintenance = \$1,000, 30-year life

If the prime interest and the inflation rate were 7 and 4%, respectively, the corresponding present cost per mile is portland cement = \$410,235 and asphalt = \$412,627 (i.e., first inflate by 4% and then discount by 7%). If the prime interest and the inflation rate is 14 and 11%, respectively, the corresponding present time cost per mile is portland cement = \$410,134 and asphalt = \$414,627. (*Note:* The present worth formulas are given in Chapter 12.)

Thus despite the large difference between the interest and inflation rates in the hypothetical examples, the stable (3%) discount rate resulted in nearly identical cost estimates. The salvage (remaining) value of the road at the end of its design life (i.e., worth as recyclable material or as a secondary facility) should be subtracted from the total cost estimates.

2.5.7 Pavement Management Systems

With nearly all of the interstate highway and roadway system complete in the United States, the emphasis is being switched from construction to maintenance. The development of a pavement management system that originated near the turn of the century has grown steadily due to the increased need for pavement maintenance.

A pavement management system incorporates the coordination of activities associated with the design, planning, construction, maintenance, research, and evaluation of pavements. Most of these activities are focused on existing pavements. The system consists of three essential elements: (1) surveys related to pavement condition and serviceability and compilation of a continuously updated data base; (2) prioritization of needs, alternative repair options, evaluation, and decision for action; and (3) implementation procedures.

The condition and serviceability of a pavement as perceived by the user is represented by its roughness, which is a measure of the irregularities in the pavement surface, causing discomfort to the users. Longitudinal, transverse, and horizontal components of roughness affect the comfort and safety of users. Roughness is assessed with various mechanical devices called profilometers. Advanced techniques utilize video imaging, radar, sonics, and infrared technologies for assessing surface and structural pavement damage [2.29].

The FHWA as well as state agencies are largely responsible for the development of strategies for the prioritization of needs, evaluation, and decision for action. Implementation involves the so-called 4R procedures (i.e., resurfacing, restoration, rehabilitation, and reconstruction). *Resurfacing* is self-explanatory; both asphalt layer on a rigid pavement or portland-cement layer on a flexible pavement are feasible resurfacing options, in addition to asphalt on asphalt and portland cement on portland cement. When resurfacing asphalt concrete pavements, a layer 1 to 3 in. thick is usually removed before the new surface layer is applied. *Restoration* includes the removal and replacement of portland-cement slabs, the patching of potholes, the sealing of cracks, the retrofitting of edge support, and various other localized repairs. *Rehabilitation* is large-scale restoration. It includes elements, such as replacement of bridge decks, resurfacing of a substantial segment of a roadway facility, recycling of materials, and minor subgrade work incidental to other repairs. *Reconstruction* is the complete removal of the pavement structure to the base layer and the replacement with virgin or recycled materials. Recycling includes the removal and crushing of portland-cement slabs and the use of the product as coarse aggregates or stockpile material for bases. In the case of flexible pavements, removed asphalt material is used as base material or it is recycled at a hot-mix plant. The overall network priorities and scheduling for 4R constitute a major part of pavement management systems (PMS) [2.30]. Most modern PMS are implemented as geographic information system (GIS) applications (see Chapter 15) [2.31].

2.5.8 High Performance Concrete, Superpave, and LTPP

Traditionally pavement design specifications have been based on empirical properties (such as "percent of air voids") which had been correlated to pavement performance. A recent trend has been the identification of *performance-based* properties that can be used directly to predict pavement performance.

Performance-based specifications have been at the center of a major initiative of the Federal Highway Administration that was authorized by the Surface Transportation and Relocation Act of 1987. Known as the *Strategic Highway Research Program* (SHRP), this 5-year $150 million endeavor consisted of four major components: (1) portland-cement concrete (PCC) and structures, (2) asphalt, (3) long-term pavement performance (LTPP), and (4) highway operations. Thus three of the four program elements were directly or indirectly associated with pavements. Upon completion of SHRP, FHWA proceeded with the implementation phase in cooperation with industrial partners, several lead states, and university-based centers. The implementation phase continues to yield modifications and improvements in the earlier research-prescribed methods and practices.

High performance PCC. The pavement portion of the PCC component of SHRP emphasized *high performance* PCC [2.32]. The term "high performance" in this context does not necessarily indicate "high strength." It is, instead, allowed to take a meaning that is applicable to the intended use. Examples of possible applicable criteria include durability, rapid setting for high early strength, low permeability, and low life-cycle costs. Rapid setting is accomplished by the use of low water-cement ratios and various additives. In many applications rapid setting is considered highly desirable because it allows the opening of rehabilitated highway segments to the traffic soon after placement.

Superpave. The most important element of the asphalt component of SHRP was the development of a hot-mix *superior performing asphalt pavement* known as *Superpave* [2.33]. Three mix design levels requiring increasingly more elaborate procedures have been developed. *Level 1* applies to low traffic volumes (less than 10^6 ESALs) and is similar to traditional volumetric design methods based on empirical performance-related properties [2.34, 2.35, 2.36]. *Level 2* applies to intermediate traffic loads (between 10^6 and 10^7 ESALs) and builds upon the *level 1* design by requiring additional performance-based tests and software analyses aimed at predicting pavement performance in terms of predicting fatigue cracking, low temperature cracking, and permanent deformation versus time. *Level 3* design applies to high traffic loads (more than 10^7 ESALs) and involves more comprehensive performance prediction models [2.33].

Level 1 design (which is the starting point for the other two levels) basically consists of the following steps:

1. Selection of materials (i.e., asphalt binder and aggregates)
2. Selection of the design aggregate structure
3. Evaluation of trial mixes in terms of their volumetric properties
4. Selection of design mix

Superpave asphalt binders are referred to as performance grade (PG) binders and are designated as ($PG\ T_{HIGH} - T_{LOW}$) where T_{HIGH} is the average seven-consecutive day maximum temperature and T_{LOW} is the lowest temperature prevailing at the location where the pavement is to be constructed. The high temperature is measured at a 20-mm depth, whereas the low temperature is measured at the surface of the pavement layer. The temperature ratings of PG binders are specified in increments of 6°C and are further adjusted to account for expected traffic conditions (e.g., slow traffic with frequent stops warrants incrementing the high temperature rating) and traffic levels (in terms of ESALs). A variety of binder tests are required to ensure conformance with the selected PG specification [2.37].

Aggregate selection is based on characteristics, such as normal maximum size (depending on pavement depth), coarse and fine aggregate angularity, toughness, soundness, clay and deleterious material content, dust proportion, and percent of thin elongated particles. Aggregate gradation is restricted by specified control points (i.e., percent passing specified sieve sizes) and excludes a "restricted zone" containing a portion of fine aggregates thought to cause mix instability leading to pavement rutting. Superpave aggregate specifications were established by consensus of expert opinion rather than by research. Subsequent field experience has led to certain disagreements, including the efficacy of the restricted zone.

Trial mixes are compacted with either a specially designed *Superpave gyratory compactor* or another gyratory compactor meeting certain criteria [2.36]. The mixes are evaluated in accordance to volumetric properties, such as air voids (4% recommended for all level 1 designs), voids in the mineral aggregate (VMA), asphalt volume absorbed into the aggregate, asphalt content, voids filled with asphalt (VFA), and density. The number of gyrations applied by the compactor depends on the anticipated traffic loads and the maximum temperature rating. Test procedures for the moisture susceptibility of the trial mixes are included as part of the design mix selection process.

As stated earlier, level 2 and 3 designs involve additional tests to provide inputs to performance prediction software. A specially designed *Superpave shear test device* is used to subject the samples to loads simulating the compression and shear forces applied by vehicle tires to the pavement, whereas an indirect tension creep tester is used to measure low temperature cracking. A setback to the advanced level procedures occurred during the late 1990s when certain critical flaws were discovered in the then existing performance tests and predictive computer models [2.38].

Long-term pavement performance. The LTPP component of SHRP is an extensive effort to extend the understanding of pavement performance and to support research by constructing a comprehensive database of field testing and monitoring more than 2400 asphalt and PCC test sections throughout the United States and Canada [2.39]. Key features for each test location include pavement characteristics, construction method, test results, maintenance practices, and so forth. Updated versions of the database, along with data exploration and extraction utilities, are periodically issued on CD-ROM (e.g., [2.40]).

2.6 SUMMARY

In this chapter we reviewed the fundamental kinematic and kinetic equations of particle motion and developed the formulas that govern the rectilinear and curvilinear motion of single vehicles. Basic models of human factors were then introduced to illustrate how driver responses to stimuli in the driving environment can be incorporated into highway design. The factors examined included driver perception-reaction, visual acuity, and responses to laterally placed objects. Perception-reaction was shown to affect the total distance covered by a stopping vehicle and the presence of "dilemma zones" at signalized intersections. Laterally located objects create a tendency for the driver to slow down and to steer the vehicle away from them even when the vehicle is not on a collision course with the objects.

The basic aspects of geometric design, that is, the proportioning of the visible elements of fixed facilities, were covered next. The elements described included cross-section design, horizontal alignment, superelevation, vertical alignment, and channelization. To ensure safe operation, driver needs were shown to be met through the provision of adequate stopping and passing sight distances and by the proper selection and placement of channelization treatments, such as pavement markings, elongated and triangular islands, and medians as well as traffic calming. Vehicle characteristics are reflected in the selection of appropriate design vehicles. Human and design factors are summarized in Table 2.6.1.

Fundamental principles of pavement structures and design, including Superpave, as well as elements of economic analysis for flexible and rigid pavements were presented.

TABLE 2.6.1 Summary of Human and Design Factors

Human factors		
Cause/Fact	**Effect**	**Elements of interest**
Perception reaction time	Useful time is lost in critical situations	Delayed undertaking of corrective action
Ambiguity in decision making	Driver in dilemma zone	Any action will compromise safety (proper selection of duration of amber)
Acuity of vision	Potential inability to perceive conditions correctly	May result in poor judgment (wrong information fed to brain)
Instinctive fear or insecurity	Reduction of speed or positioning at a longer distance from hazard or vehicle ahead	Lateral displacement; car-following behavior
Fatigue, intoxication, or other impairment	Inability to perceive conditions correctly and react in a timely manner	May result in poor judgment (body or brain incapable of functioning properly)

Design factors			
Cause/Fact	**Effect**	**Element of interest**	**Critical unit**
Must stop	Discomfort if abrupt or danger if not successful	Stopping distance	Bus with standing passengers plus wet conditions
Must be able to see ahead	Safe navigation of vehicle	Sight distance	Sight distance ≥ stopping distance
Driver and vehicle must be able to follow roadway comfortably	Vehicle under driver's control; comfortable and safe ride when speed limits are observed	Maximum slopes; minimum radii; superelevation; design speed; speed limit	Heavy vehicle; wet conditions; conservative height of eyes and headlights
Lack of proper channelization	Confusion and failure to follow proper paths	Inefficient service process; accident occurrence	Potential conflicts; volumes; available space and resources

EXERCISES

1. Given the acceleration pattern shown in Fig. E2.1, (a) derive and plot the relationship between speed and time and (b) calculate the total distance traveled during the 20-s interval. At $t = 0$ the vehicle was traveling at 12 mi/h.
2. The driver of a car traveling up a 2% grade at an initial speed V_0 applied the brakes abruptly and the vehicle slid to a complete stop at an average deceleration of 8 ft/s^2. Was the pavement wet or dry?
3. At time $t = 0$ two persons entered the elevator of the tower shown in Fig. E2.3. The first person rode to the restaurant level. The second person went to the observation deck. Plot the time–elevation and the velocity–elevation diagrams for each of the two persons considering that the elevator started up 3 s after they entered, made no intermediate stops between the ground and the restaurant levels, and stayed for 6 s at the restaurant level. The elevator manufacturer's brochure provides the following technological specifications: acceleration is 5 ft/s^2; deceleration is 4 ft/s^2; and maximum cruising velocity is 20 ft/s.

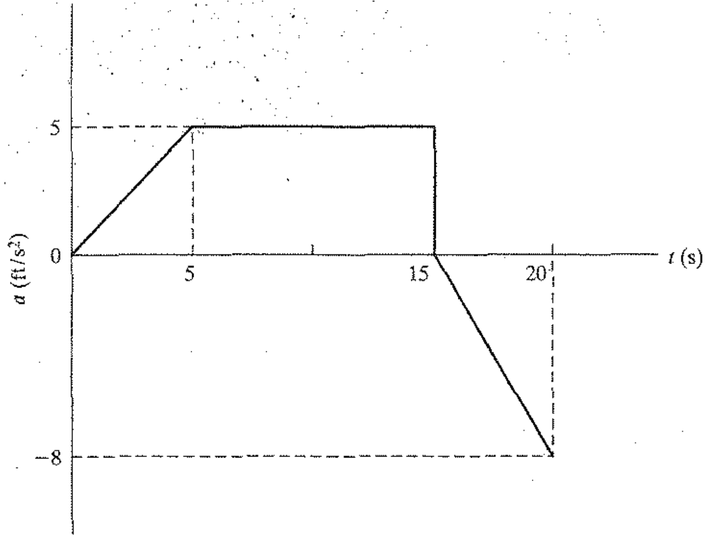


Figure E2.1

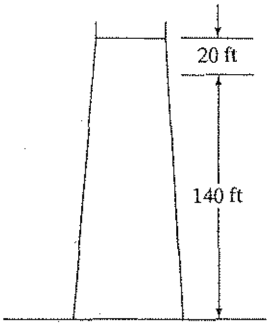


Figure E2.3

4. A rapid-transit system uses a tubular guideway in which a close-fitting vehicle is propelled by a pressure difference between the front and rear cross-sectional areas [Fig. E2.4(a)]. At departure time t_0 the pressure difference $P_2 - P_1$ is instantaneously raised from zero to some initial level and the vehicle accelerates forward. The pressure difference is then decreased until it reaches zero and the vehicle attains its cruising velocity, which is sustained until the vehicle begins its deceleration toward the next station [Fig. E2.4(b)]. Neglecting friction, (a) express acceleration, velocity, and distance traveled as functions of time and (b) sketch the relationship between velocity and time for the movement between two stations 1 mi apart. Use the following data:

A = cross-section area of the vehicle = 100 ft^2

W = vehicle weight = 40,000 lb

α = 100 lb/ft^2

β = 3.33 lb/ft^2-s

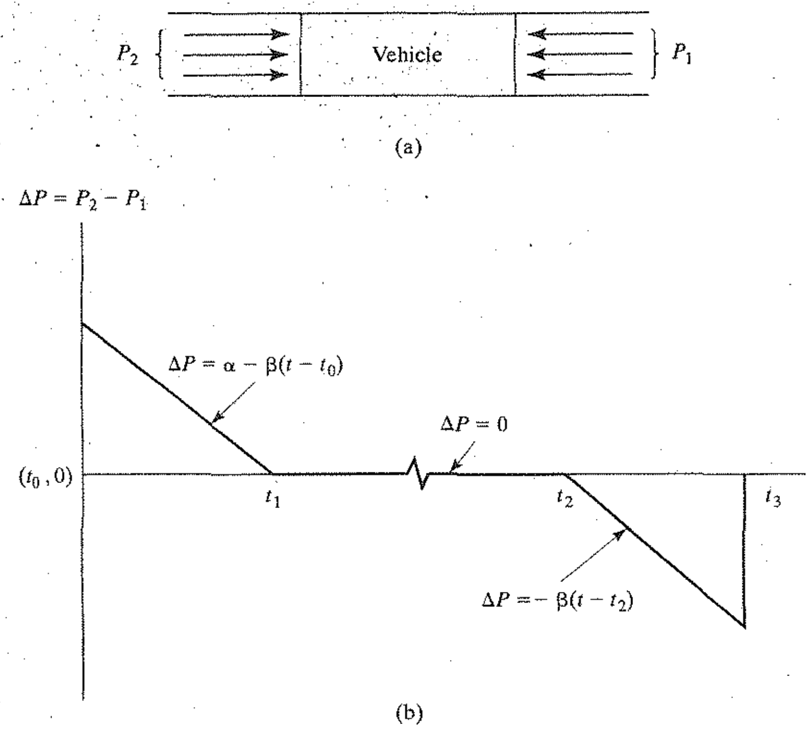


Figure E2.4

5. A car collided with a telephone pole and left 20-ft skid marks on the dry pavement. Based on the damages sustained, an engineer estimated that the speed at collision was 15 mi/h. If the roadway had a +3% grade, calculate the speed of the car at the onset of skidding.
6. A large rock became visible to a driver at a distance of 175 ft. Assuming a perception-reaction time of 0.8 s, an initial speed of 42 mi/h, a coefficient of friction equal to 0.5, and a level roadway, calculate the speed at impact.
7. Plot the relationship between the approach speed v and the length of the dilemma zone for the following data: $a_2 = 0.5g$, $\delta_2 = 1.0$ s, $w = 65$ ft, $L = 15$ ft, and $\tau = 4.5$ s. To help you interpret this plot, draw another diagram in which the v versus x_c and the v versus x_0 relationships are superposed.
8. Assuming a comfortable deceleration of 8 ft/s^2, an intersection width of 42 ft, a perception-reaction time of 0.9 s, and a vehicle length of 18 ft, plot $\tau_{\min}$ versus approach speed v_0. At what speed does $\tau_{\min}$ attain a minimum? What is that minimum?
9. Prepare a computer program for the general case of Exercise 8. Run your program several times and plot the results.
10. A driver with 20/40 vision and a sixth-grade education needs 2 s to read a directional sign. The letter size is such that the sign can be read by a person with 20/20 vision from a distance of 200 ft. Does the subject driver have enough time to read the sign at a speed of 30 mi/h?
11. The street name sizes on the signs at a certain location can be discerned by a person with 20/40 vision from a distance of 300 ft. How much larger should the street names be in order to be legible to a person with 20/50 vision from a distance of 450 ft?

12. A witness with 20/60 vision supplied to the investigating officer the license plate number of a vehicle involved in a hit-and-run accident. If under the conditions that prevailed at the time of the accident a license plate can be read by a person with 20/20 vision from a distance of 180 ft, what is the maximum distance for which the witness's testimony can be relied upon?

13. Vehicles A and B are traveling toward each other in opposing lanes on a straight segment of a two-lane highway at 35 and 40 mi/h, respectively. If the critical rates of angular change of the two drivers are 0.0065 and 0.0055 rad/s, determine (a) which driver will be the first to displace laterally and (b) the longitudinal distance between vehicles when the displacement will occur. Assume that the lateral separation between the two vehicles is 6 ft.

14. Inside a tunnel the distance between the drivers and the curb is 10 ft. Assuming that the drivers fix their eyes on the curb at an angle of 2°, calculate the appropriate speed limit for a critical angular rate of change of 0.005 rad/s.

15. How far to the side of the drivers of Exercise 14 should the curb be to allow a speed of 45 mi/h?

16. What is the maximum allowable degree of curve (arc definition) for a two-lane highway if $e_{\max} = 0.08, f_s = 0.12$, and the design speed is 50 mi/h?

17. A simple highway curve is planned to connect two horizontal tangents that intersect at sta. 2500 + 00.00 at an external angle of 52°. For a design speed of 60 mi/h and a curve radius of about 1.25 times the minimum allowable, calculate (a) the design rate of superelevation and (b) the required length of superelevation runoff. Assume a four-lane undivided highway and 10-ft lanes. Clearly state any other assumptions that you think are needed.

18. Sketch the plan view and the longitudinal profile of your curve design (Exercise 17) assuming a normal crown of 0.02 ft/ft and pavement rotation about the centerline.

19. A sight obstruction is located at a distance of 20 ft from the center of the inside lane of a highway that prescribes a circular curve (Fig. 2.4.14). If the degree of curve for the centerline of the inside lane is 15°, calculate (a) the curve's radius and (b) the available horizontal sight distance. Does the computed sight distance meet the AASHTO stopping criterion for a speed of 35 mi/h?

20. A 2000-ft vertical curve connects a +3% grade to a −5% grade. If the vertical tangents intersect at sta. 52 + 60.55 and elevation 877.62 ft, calculate the elevations at the VPC, VPT, high point, and sta. 54 + 00.

21. A −4% grade and a 0% grade meeting at sta. 24 + 40.00 and elevation 2421.54 ft are joined by an 800-ft vertical curve. The curve passes under an overpass at sta. 25 + 00.00. If the lowest elevation of the overpass is 2439.93 ft, calculate the available clearance.

22. A +2% grade meets a +6% grade at sta. 10 + 30.00 and elevation 168.21 ft. For a design speed of 55 mi/h, find the minimum length of vertical curve that satisfies the 1984 AASHTO stopping sight-distance criterion. Also, calculate the elevation of the middle point of this curve.

23. Calculate the available passing sight distance on a 2000-ft vertical curve with $G_1 = +5\%$ and $G_2 = -2\%$ for (a) the pre-1984 and (b) the post-1984 oncoming vehicle and driver eye heights.

24. Prepare a computer program that, given the station location and elevation of the intersection of two vertical grades and the required sight distance, calculates (a) the minimum length of vertical curve, (b) the station location and elevation of the VPC and the VPT, and (c) the curve elevation of any specified intermediate point.

REFERENCES

2.1 Drew, D. R., *Traffic Flow Theory and Control,* McGraw-Hill, New York, 1968.

2.2 American Association of State Highway and Transportation Officials, *A Policy on Geometric Design of Highways and Streets,* AASHTO, Washington, DC, 1990.

2.3 JOHANNSON, G., and K. RUMAR, "Drivers' Brake Reaction Times," *Human Factors*, 13, 1 (1971): 23–27.

2.4 GAZIS, D., R. HERMAN, and A. MARADUDIN, "The Problem of the Amber Signal in Traffic Flow," *Operations Research*, 8 (1960): 112–132.

2.5 INSTITUTE OF TRANSPORTATION ENGINEERS, *Determining Vehicle Signal Change and Clearance Intervals*, Report IR-073, ITE Technical Council Committee 4TF-1, 1994.

2.6 PAPACOSTAS, C. S., and N. H. KASAMOTO, "The Intergreen Interval Controversy: Toward a Common Framework," *Transportation Research Record 1324*, National Research Council, Washington, DC, 1991, pp. 21–32.

2.7 GRAHAM, J. R., A. FAZAL, and L. E. KING, "Minimum Luminance of Highway Signs Required by Older Drivers," *Transportation Research Record 1593*, National Research Council, Washington, DC, 1997, pp. 91–98.

2.8 KANTOWITZ, B. H., and R. D. SORKIN, *Human Factors: Understanding People–System Relationships*, John Wiley, New York, 1983.

2.9 TARAGIN, A., "Driver Behavior as Affected by Objects on Highway Shoulders," *Highway Research Board Proceedings*, 34 (1955): 453–472.

2.10 MICHAELS, R. M., and L. W. GOZAN, *Perceptual and Field Factors Causing Lateral Displacement*, Highway Research Record 25, Highway Research Board, National Research Council, Washington, DC, 1963.

2.11 TRANSPORTATION RESEARCH BOARD, *Intersection Channelization Design Guide*, National Cooperative Highway Research Program Report 279, National Research Council, Washington, DC, 1985.

2.12 FEDERAL HIGHWAY ADMINISTRATION, *Manual on Uniform Traffic Control Devices*, U.S. Department of Transportation, Washington, DC, 1991.

2.13 TRANSPORTATION RESEARCH BOARD, Modern Roundabout Practice in the United States, NCHRP Synthesis 264, National Research Council, Washington, DC, 1998.

2.14 STATE OF MARYLAND DEPARTMENT OF TRANSPORTATION, *Roundabout Design Guidelines*, Hanover, 1995.

2.15 FLORIDA DEPARTMENT OF TRANSPORTATION, *Florida Roundabout Guide*, Tallahassee, 1996.

2.16 BRITISH MINISTRY OF TRANSPORT, *Geometric Design of Roundabouts*, TD/93, United Kingdom, 1993.

2.17 INSTITUTE OF TRANSPORTATION ENGINEERS, *Traffic Engineering Handbook*, 4th ed. Prentice-Hall, Englewood Cliffs, NJ, 1992.

2.18 INSTITUTE OF TRANSPORTATION ENGINEERS, *Residential Street Design and Traffic Control*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

2.19 SCHLABBACH, K., "Traffic Calming in Europe," *ITE Journal*, Institute of Transportation Engineers, July 1997, pp. 38–40.

2.20 TRANSIT COOPERATIVE RESEARCH PROGRAM, *Transit-Friendly Streets: Design and Traffic Management Strategies to Support Livable Communities*, TCRP Report 33, Transportation Research Board, National Research Council, Washington, DC, 1998.

2.21 FRIEDMAN, B., S. B. GORDON, and J. B. PEERS, "Effect of Neotraditional Neighborhood Design on Travel Characteristics," *Transportation Research Record 1466*, National Research Council, Washington, DC, 1997, pp. 63–70.

2.22 LOCKWOOD, I. M., "ITE Traffic Calming Definition," *ITE Journal*, Institute of Transportation Engineers, July 1997, pp. 22–24.

2.23 Institute of Transportation Engineers, *Guidelines for the Design and Application of Speed Humps: A Recommended Practice,* Washington, DC, 1997.

2.24 National Cooperative Highway Research Program, *Pedestrians and Traffic Control Measures,* Synthesis of Highway Practice 139, Transportation Research Board, National Research Council, Washington, DC, 1988.

2.25 Fire and Rescue Commission, The Effects of Speed Humps and Traffic Circles on Responding Fire-Rescue Apparatus, Montgomery County, MD, 1997.

2.26 American Association of State Highway and Transportation Officials, *AASHTO Guide for Design of Pavement Structures 1986,* AASHTO, Washington, DC, 1986.

2.27 The Asphalt Institute, *Thickness Design: Asphalt Pavements for Highways and Streets,* Manual Series No. 1, AI, 1981.

2.28 Portland Cement Association, *Thickness Design for Concrete Highway and Street Pavements,* PCA, Skokie, IL, 1984.

2.29 Bomar, L. C., W. F. Horne, D. R. Brown, and J. L. Smart, *Determining Deteriorated Areas in Portland Cement Concrete Pavements Using Radar and Video Imaging,* National Cooperative Highway Research Program Report 304, Transportation Research Board, National Research Council, Washington, DC, 1988.

2.30 Voight, G. F., and M. J. Knutson, *Development and Selection of the Preferred 4R Strategy,* American Concrete Pavement Association, Arlington Heights, IL, 1989.

2.31 National Highway Institute, *Pavement Management Systems,* CD-ROM, Federal Highway Administration, U.S. Department of Transportation, Washington, DC, 1998.

2.32 Zia, P., M. L. Leming, and S. H. Ahmad, *High Performance Concretes: A State-of-the-Art Report,* SHRP-C/FR-91-103, Strategic Highway Research Program, National Research Council, Washington, DC, 1991.

2.33 Kennedy, T. W. et al., *Superior Performing Asphalt Pavements (Superpave): The Product of the SHRP Asphalt Research Program,* SHRP-A-410, Strategic Highway Research Program, National Research Council, Washington, DC, 1994.

2.34 Cominsky, R., R. B. Leahy, and E. T. Harrigan, *Level One Mix Design: Materials Selection, Compaction and Conditioning,* SHRP-A-408, Strategic Highway Research Program, National Research Council, Washington, DC, 1994.

2.35 The Asphalt Institute, *Superpave Level 1 Mix Design,* SP2, Lexington, KY, 1995.

2.36 Cominski, R. et al., *The Superpave Mix Design Manual for New Construction and Overlays,* SHRP-A-407, Strategic Highway Research program, National Research Council, Washington, DC, 1994.

2.37 The Asphalt Institute, *Performance Graded Asphalt Binder Specification and Testing,* SP1, Lexington, KY, 1995.

2.38 Anderson, R. M., J. R. Buckowski, and P. A. Turner, "Evaluating Asphalt Mixtures Using Superpave Performance Tests," Paper presented at the 78th Annual Meeting of the Transportation Research Board, National Research Council, Washington, DC, 1999.

2.39 Hadley, W. O., *SHRP-LTPP Overview: Five Year Report,* SHRP-P-416, Strategic Highway Research Program, National Research Council, Washington, DC, 1994.

2.40 Federal Highway Administration, *DataPave97: LTPP Information Management System,* CD-ROM, Strategic Highway Research Program, National Research Council, Washington, DC, 1998.

3

Traffic Stream Flow Models

3.1 INTRODUCTION

Chapter 2 was concerned primarily with the motion of a single vehicle. From the resulting equations of motion the basic geometric design formulas for highways were derived and their application was illustrated. Occasionally single vehicles traverse the transportation facilities without significant interference from other vehicles. But the same facilities also experience simultaneous usage by streams of vehicles. The resulting traffic conditions range from almost free flow when only a few relatively unconstrained vehicles occupy a roadway to highly congested conditions when the roadway is jammed with slow-moving vehicles. In this chapter we examine the consequences of vehicular interactions. The equations developed in Chapter 2 are used to formulate a general model of a vehicular stream for the simple case of identically scheduled vehicles on an exclusive right-of-way. This model is then extended to the case of highway traffic, where considerable variability prevails. The determinant of these traffic-flow models is the car-following rule adopted by drivers in an attempt to maximize their speeds while maintaining an acceptable level of safety. They accomplish this by adjusting the distance between vehicles, depending on their speed. The basic variables that describe the prevailing conditions within a vehicular stream (i.e., flow, concentration, and mean speed) are introduced, and the fundamental relationship between the three stream variables is postulated and applied to several traffic phenomena, including the propagation of shock waves in traffic.

3.2 VEHICULAR STREAM MODELS

3.2.1 Vehicular Following

Consider the case of vehicles following each other on a long stretch of roadway or guideway. Furthermore, assume that these vehicles are not required to interrupt their motion for reasons that are external to the traffic stream, such as traffic lights, transit stations, and the like. In this case of *uninterrupted flow* the only interference that a single vehicle experiences is caused by other vehicles on the roadway. Figure 3.2.1 shows two typical stream vehicles traveling at a speed v and a *spacing* s between the front of the leading vehicle to the front of the following vehicle. As a general rule the spacing between vehicles should be such that if a sudden deceleration becomes necessary for a leading vehicle, the following vehicle has ample time and distance to perceive the situation, react to it, and be able to decelerate safely without colliding with the stopping, leading vehicle. A similar rule was applied in Chapter 2 to compute the necessary safe stopping distance that served as a criterion for the proper geometric design of roadways.

Figure 3.2.1(a) shows the locations of the leading and following vehicles described earlier at the moment when the leading vehicle begins to decelerate, and Fig. 3.2.1(b) shows the limiting acceptable conditions at the end of the stopping maneuver of the following vehicle. Parenthetically, the term *vehicle* may be taken to mean a vehicular train consisting of a number of articulated vehicles rather than a single vehicle. Using the following notation, a relationship among spacing, speed, and deceleration (assumed constant) can be developed:

v = initial speed of the two vehicles

d_l = deceleration rate of the leading vehicle

d_f = deceleration rate of the following vehicle

δ = perception-reaction time of the following vehicle

x_o = safety margin after stop

L = length of vehicle

N = number of vehicles in a train (if applicable)

Under constant deceleration the braking distance of the leading vehicle is

$$x_l = \frac{v^2}{2d_l} \tag{3.2.1}$$

Including perception-reaction, the total distance that would be covered by the responding following vehicle is

$$x_f = v\delta + \frac{v^2}{2d_f} \tag{3.2.2}$$

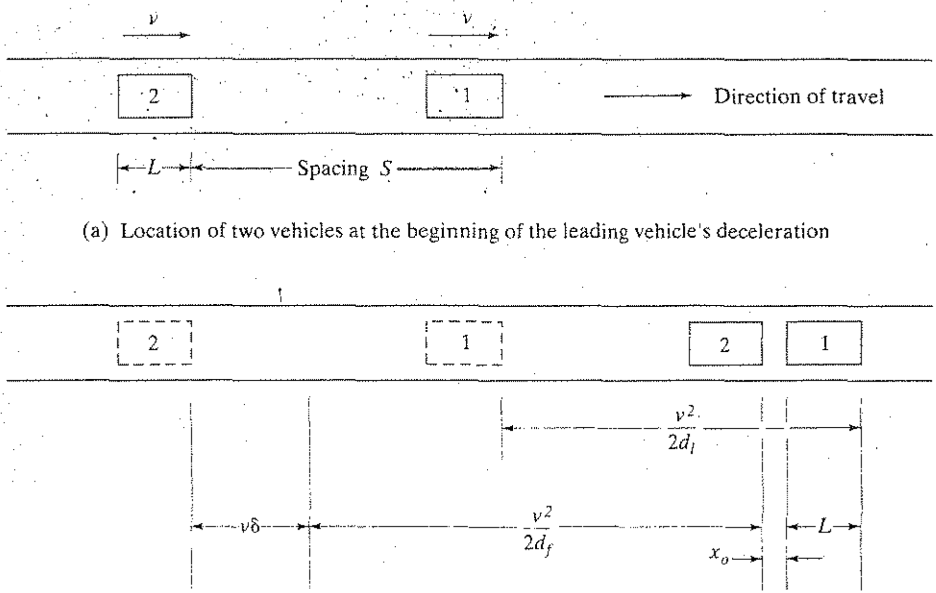
In terms of the initial spacing s, the length of the vehicular unit (NL), and the safety margin x_o,

$$x_f = s + x_l - NL - x_o \tag{3.2.3}$$

Substituting Eqs. 3.2.1 and 3.2.2 in Eq. 3.2.3 and solving for s gives

$$s = v\delta + \frac{v^2}{2d_f} - \frac{v^2}{2d_l} + NL + x_o \tag{3.2.4}$$

Thus, given the speed of normal operation of the system and the other performance parameters, it is possible to compute the necessary spacing so that the following vehicle will just



(a) Location of two vehicles at the beginning of the leading vehicle's deceleration

(b) Distances traveled

Figure 3.2.1 Vehicle following concept.

be able to avoid a collision by anticipating a potential stopping maneuver by the vehicle ahead. The application of this equation to a specific system requires the specification of the *anticipated* deceleration of the leading vehicle and the *desired* deceleration of the following vehicle. The combined choice of particular values for these variables has some important implications with respect to the level of safety provided by the system's operation.

3.2.2 Safety Considerations

Three particular values of deceleration are relevant to the operation's safety level [3.1]:

d_n = normal or comfortable deceleration

d_e = emergency deceleration

∞ = "instantaneous" or "stonewall" stop

Normal deceleration is related to passenger comfort as discussed in Chapter 2. The instantaneous stop condition may arise when an accident or a stalled vehicle or other obstruction suddenly comes within the perception field of the subject vehicle.

The safest level of operation occurs when the spacing between vehicles is such that the following vehicle can safely stop by applying *normal* deceleration even when the leading vehicle comes to a stonewall stop. A lower level of safety results when the spacing is selected so that the following vehicle would have to apply an *emergency* brake rather than normal deceleration in order to avoid a collision. The combinations of leading

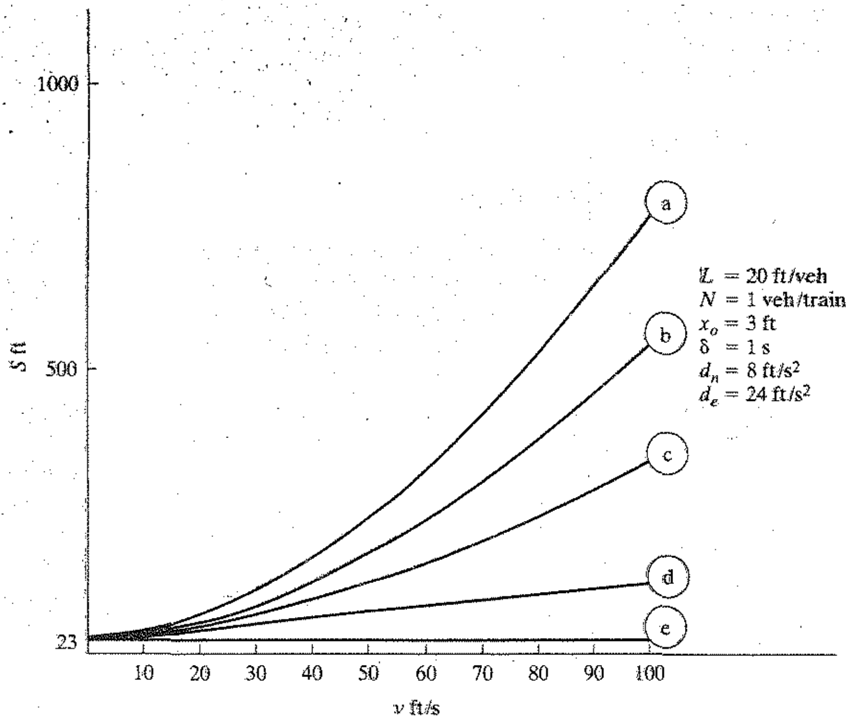


Figure 3.2.2 Spacing versus speed.

and following vehicle decelerations that designate various *safety regimes* are shown in Table 3.2.1.

TABLE 3.2.1 Safety Regime Definitions

Regime	Deceleration of leading vehicle	Deceleration of following vehicle
a	∞	d_n
b	d_e	d_n
c	∞	d_e
d	$d_l = d_f$	
e	(no braking)	

Note: For $d_e < 2d_n$, regime c is safer than regime b.
Source: Vuchic [3.1].

Figure 3.2.2 plots spacing versus speed (Eq. 3.2.4) for the four safety regimes corresponding to the values inserted in the figure. Also included in Fig. 3.2.2 is the limiting case of a hypothetical continuous train, which is assumed to operate at any constant speed without ever having to decelerate [3.1]. The figure clearly shows that the higher the level of safety is, the higher the required spacing will be just to avoid a collision. On this basis alone it would seem reasonable to choose the safest level of operation. However, by increasing the level of safety, the capacity of the system (i.e., the maximum number of vehicles or passengers that can be accommodated during a given period of time) suffers. Consequently a *trade-off* between safety and capacity exists.

3.3 STREAM VARIABLES

3.3.1 Spacing and Concentration

Consider the uniform operation described in the preceding section and assume that a single photograph of a roadway segment is taken at an instant of time. The photograph would show a number of equally spaced vehicles along the roadway segment. The ratio of the number of vehicles appearing on the photograph to the length of the roadway segment is defined as the *concentration* k of the vehicular stream. This is an instantaneous measurement taken at the instant when the photograph was taken. Since the operation of the system described here is uniform (i.e., constant spacing and operating speed), the numerical value of concentration obtained at any instant of time on any segment of roadway will be the same. However, if the spacings and speeds of the vehicles that make up the stream are not equal, as is the case with the typical operation of highways, the value of concentration can vary with time and also differ from one location to another at the same time. The dimensions of concentration (which is often referred to as *density*) are given in terms of vehicles per length of roadway, for example, vehicles per mile (or veh/mi). The relationship between spacing (or average spacing when not constant) and concentration is

$$s = \frac{1}{k} \tag{3.3.1}$$

3.3.2 Headway and Flow

Consider a stationary observer next to the roadway. Vehicles pass the observer's location one after another at intervals of time defined as the *headways* between vehicles and denoted by the letter h. In the simple example described earlier the headway between vehicles is constant and can be computed by dividing the constant spacing by the constant speed of system operation. It is not too difficult, however, to imagine a situation, such as highway traffic, where the measured headway between subsequent vehicles varies. In either case, during a period of observation T, the observer would count a number of headways, each corresponding to an individual vehicle in relation to its leader, the sum of which equals the total time of observation T. The number of vehicles counted at the point of observation divided by the total observation time is defined as the stream *flow* $\bar{q}$ (sometimes referred to as *volume* V) and measured in vehicles per unit time, for example, vehicles per hour (veh/h). Flow is a *measurement at a point* on the roadway over time. The relationship between headway (or average headway when not constant) and flow is

$$h = \frac{1}{q} \tag{3.3.2}$$

3.3.3 Average or Mean Speed

The third basic measurement of traffic is that of *average*, or *mean, speed*. In the case of the uniform vehicular stream described previously, all vehicles were assumed to operate at the same speed v. Therefore the average speed of any group of vehicles in the stream is also equal to v. This is not always the case, however. In a typical highway situation, for example, vehicles are traveling at different speeds, which they adjust as they traverse the highway. The problem of when, where, and how to take speed measurements that are representative of the traffic stream is not trivial [3.2]. For example, the speeds of successive vehicles may be taken at a single point of the roadway over a long period of time. These speeds are also known as *spot speeds*. Alternatively, the speeds of all the vehicles occupying a length of highway may be taken at the same instant. Also, by taking two aerial photographs of the highway separated by a small interval of time, the speed of each vehicle may be calculated by dividing the distance traveled by that time interval. The method by which the speed measurements are taken and the way in which their average is computed affect the results and interpretation of this quantity. Two common ways of computing the average, or mean, speed are the *time mean speed* and the *space mean speed*. The time mean speed u_t is the *arithmetic average* of the spot speeds just defined; that is,

$$u_t = \frac{1}{N}\sum_{1}^{N} v_i \tag{3.3.3}$$

where N is the number of observed vehicles and v_i is the spot speed of the ith vehicle.

The space mean speed is calculated on the basis of the average travel time it takes N vehicles to traverse a length of roadway D. The ith vehicle traveling at speed v_i will take

$$t_i = \frac{D}{v_i} \tag{3.3.4}$$

seconds to cover the distance D. Thus the average travel time for N vehicles will be

$$t_{\text{ave}} = \frac{1}{N}\sum_{1}^{N} \frac{D}{v_i} \tag{3.3.5}$$

and the average speed based on the average travel time (i.e., the space mean speed) is the *harmonic average* of the spot speeds, or

$$u_s = \frac{1}{\dfrac{1}{N}\displaystyle\sum_{1}^{N}\frac{1}{v_i}} \tag{3.3.6}$$

The two average speeds may be calculated alternatively by

$$u_t = \frac{\displaystyle\sum_{1}^{N} \Delta x_i}{N\,\Delta t} \tag{3.3.7}$$

and

$$u_s = \frac{N\,\Delta x}{\displaystyle\sum_{1}^{N} \Delta t_i} \tag{3.3.8}$$

where

Δx_i = distance traveled by the ith vehicle during a fixed time interval Δt

and Δt_i = time taken by the ith vehicle to cover fixed distance Δx [3.3, 3.4]

There are, of course, many other ways to take speed measurements and averages. However, for the purposes of this book it suffices to state that the space mean speed (and not the time mean speed) is the proper stream speed average needed in this chapter's mathematical models.

Example 3.1

The spot speeds of four vehicles were observed to be 30, 40, 50, and 60 ft/s, respectively. Compute the time mean speed and the space mean speed.

Solution The time mean speed is the arithmetic average of the spot speeds, or

$$\frac{30 + 40 + 50 + 60}{4} = 45 \text{ ft/s}$$

On the other hand, the space mean speed is the harmonic average. Equation 3.3.6 yields

$$u_s = 42.1 \text{ ft/s}$$

Discussion The same results may be obtained by applying Eq. 3.3.7 (with, say, $\Delta t = 1$ s) and Eq. 3.3.8 (with $\Delta x = 1$ ft). The time mean speed is greater than the space mean speed. This is always the case because of the relative contribution to each average of slow- and fast-moving vehicles.

3.3.4 Time-Distance Diagrams of Flow

The vehicular variables (e.g., spacing, headway, and vehicle speed) and stream variables (e.g., flow, concentration, and mean speed) just described can be clearly illustrated via a *time–distance diagram* of the trajectories of the vehicles constituting a traffic stream. Figure 3.3.1 is such a diagram for the simple case of uniformly operated vehicles represented as particles. Since in this case the speed of the vehicles is constant, the time-distance plot for each vehicle is simply a straight line, the slope of which, *dx*/*dt*, equals the speed, *v*. A point on a plot represents the location of the subject vehicle at the corresponding instant of time. A horizontal line (e.g., line *AA*) intersects a number of time-distance lines and the (time) difference between pairs of vehicles along the horizontal line is the headway between those vehicles. Also, this horizontal line represents a stationary observer whose location does not change with time. The number of vehicles that the observer would be able to count over a period of observation *T* is equal to the number of times the horizontal line *AA* intersects a vehicle time-distance line: The higher the number of vehicles that are counted during time *T*, the higher the stream flow will be.

A vertical line (*BB*) represents the conditions prevailing at a given instant. The difference between subsequent vehicles is the spacing between vehicles. Also, line *BB* represents an aerial photograph of the stream at that instant: The number of time-distance lines that are intersected by line *BB* corresponds to the number of vehicles that would appear on a photograph of the roadway segment shown. The smaller the number of such vehicles is, the lower the stream concentration will be.

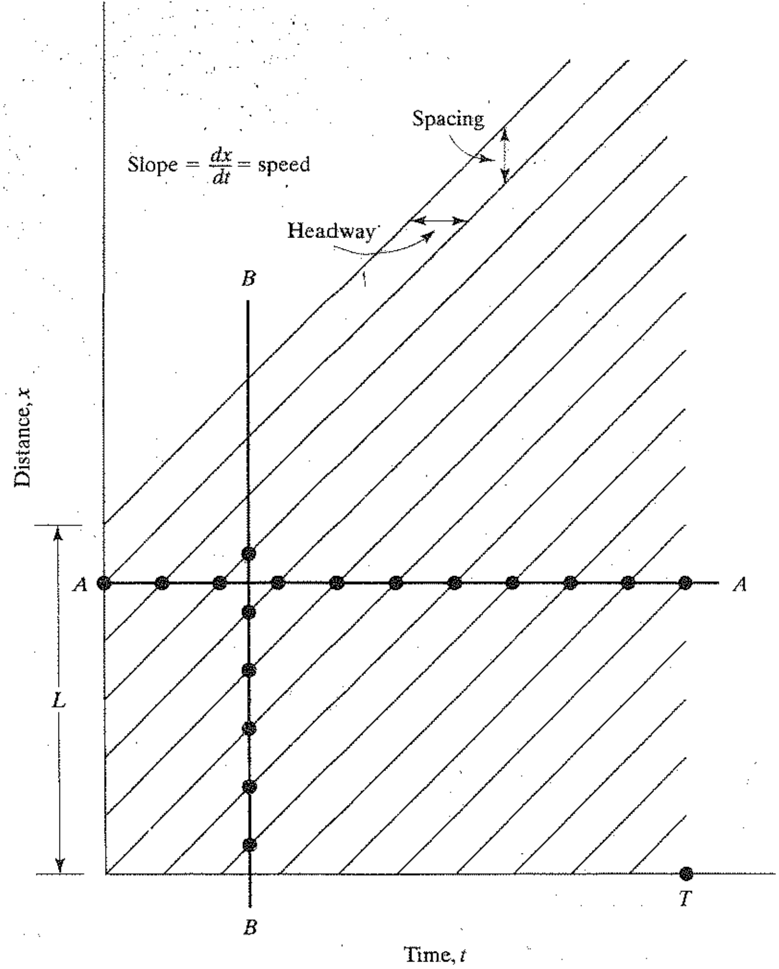


Figure 3.3.1 Time–distance diagram: uniform flow.

The problem of determining a representative measure of mean speed becomes clearer when viewing the time-distance diagram of the stream. One way to average the speeds of vehicles in the stream is to measure their speeds as they pass a given location (i.e., the slopes of the time-distance lines as they cross line *AA*), the speeds of all vehicles in the stream at an instant (i.e., the slopes of time-distance lines as they cross line *BB*), the speeds computed for a small interval of time (Δt) over the length of the highway, the speeds computed for a small interval of distance (Δx) over a long period of time, or even by computing the average speed for each vehicle over the entire length of roadway and averaging that. This may seem irrelevant in the case of uniformly scheduled vehicles shown in Fig. 3.3.1 because all the vehicles are assumed to maintain the same speed throughout their movement. But in cases of nonuniform operations the problem becomes clear. The reader is encouraged to consider the differences obtained by computing the alternate speed averages just described for the

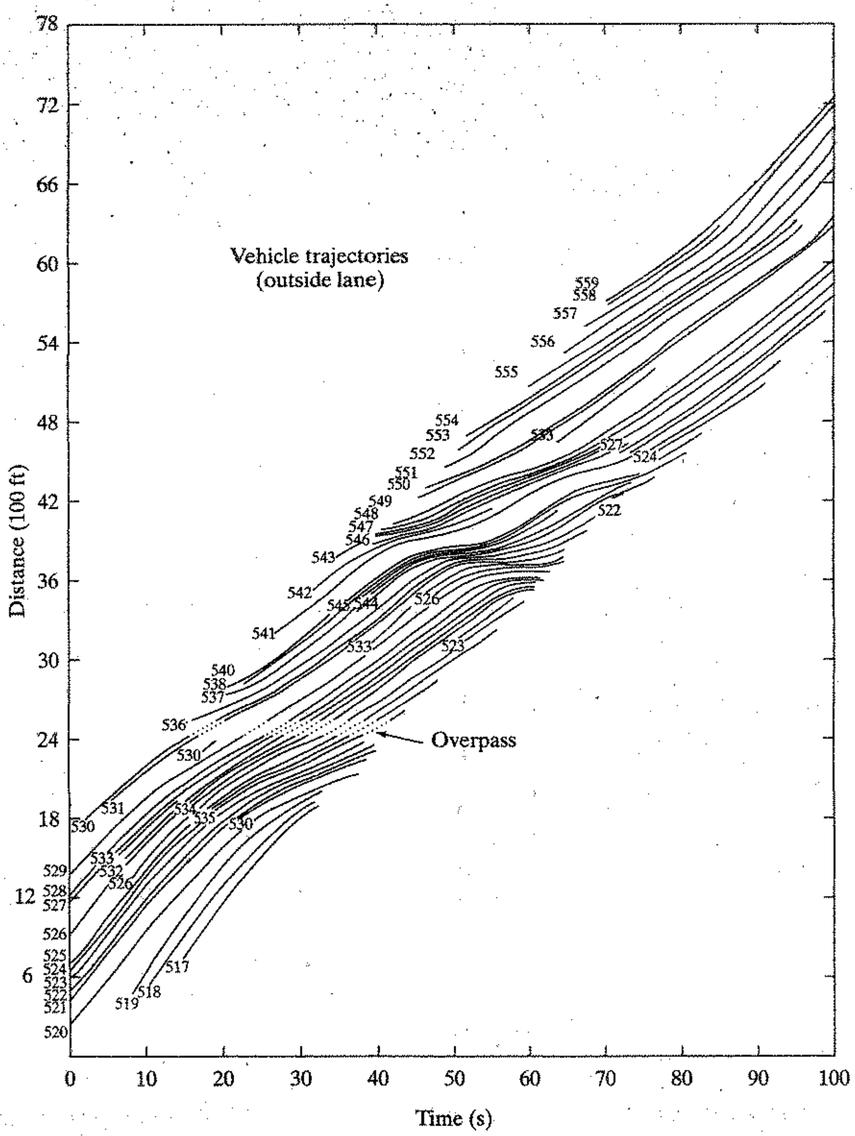


Figure 3.3.2 Highway flow.
(From Rockwell and Treiterer [3.5].)

typical highway traffic time–distance diagram illustrated by Fig. 3.3.2, which represents a stream of vehicles on the curb lane of a highway in Columbus, OH [3.5]. The data shown were collected by aerial photogrammetry from a helicopter flying above the highway. The extent of any vertical line *BB* on the diagram is the range within the view of the camera at the corresponding instant. Any such line shows a great amount of variability in spacings at and between instants and at various locations. The slopes of the vehicle time–distance diagrams change, indicating speed changes, and the concentration of vehicles is seen to exhibit great variability, higher when and where the lines are densely packed and lower when and where they are sparsely packed. It is interesting to note that concentration is highest at points where the speeds are the lowest. This phenomenon is eminently reasonable: When the speeds are low, the safe spacing selected by individual drivers is shorter leading to higher concentrations. The effect of this relationship on stream flow is not as obvious.

3.4 VEHICULAR STREAM EQUATIONS AND DIAGRAMS

3.4.1 The Fundamental Equation of a Vehicular Stream

If two vehicles are traveling at a spacing s and speed u, the headway between them is simply $h = s/u$. Substituting Eqs. 3.3.1 and 3.3.2 in this relationship leads to the fundamental equation describing a traffic stream:

$$q = uk \tag{3.4.1}$$

Note that the units balance to vehicles per hour on both sides of this equation, which represents a three-dimensional relationship between the basic vehicular stream variables: flow, mean speed, and concentration. It is of the utmost importance to realize that the three variables vary simultaneously. Consequently it would generally be incorrect to attempt to compute the value of one of the three variables by varying another while holding the third constant. As shown earlier, when speed is increased, the safe spacing between vehicles also increases, causing the concentration to decrease. According to Eq. 3.4.1, the resulting flow is given by the product of a *higher speed* times a *lower concentration*. Hence the flow may increase, decrease, or remain the same, depending on the relative magnitudes of these two opposing effects.

To gain a clearer understanding of this phenomenon, consider the two-dimensional *projections* of Eq. 3.4.1 on the u–k, u–q, and q–k planes, first for the simple case of uniform flow and then for the more complex case of highway traffic.

3.4.2 The Case of Uniform Flow

Substituting Eq. 3.3.1 into Eq. 3.2.4, solving for k in terms of u, and adjusting the units of concentration to vehicles per mile leads to

$$k = f(u) = \frac{1}{u\delta + \dfrac{u^2}{2d_f} - \dfrac{u^2}{2d_l} + NL + x_o} \tag{3.4.2}$$

This equation is plotted in Fig. 3.4.1 with k on the abscissa and u on the ordinate for the values that are inserted in the figure and for four of the five safety regimens discussed earlier, including the limiting case of the hypothetical continuous train. Excepting the hypothetical case, the relationship between speed and concentration is seen to be monotonically decreasing as should be expected: The higher the speed is, the longer the required spacing is, and consequently the lower the concentration will be. The conditions around very low concentration and very high speed are referred to as *free-flow conditions* and the maximum speed at zero concentration is known as *free-flow speed* u_f. Although Eq. 3.4.2 shows speed

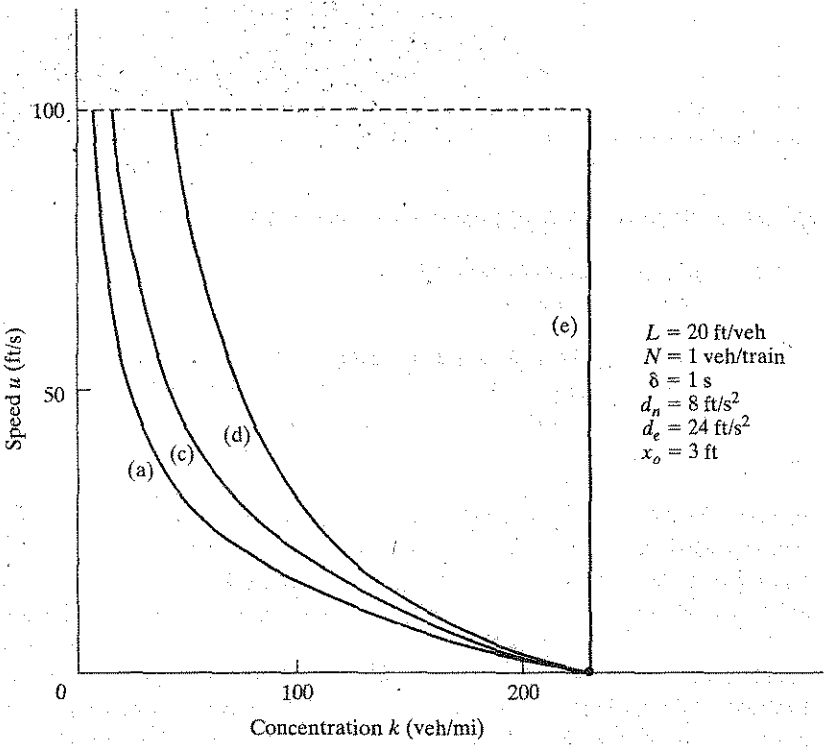


Figure 3.4.1 Speed–concentration curves.

to approach infinity asymptotically as concentration approaches zero, for all practical purposes there exists a maximum speed (see dashed line), which depends on the technological characteristics of the system.

In view of Eq. 3.4.1 multiplication of both sides of Eq. 3.4.2 by the mean speed u leads to

$$q = \frac{u}{u\delta + \dfrac{u^2}{2d_f} - \dfrac{u^2}{2d_l} + NL + x_o} \tag{3.4.3}$$

Figure 3.4.2 shows the plots of this relationship for each of the four safety regimes. The units of q have been converted to vehicles per hour. For each value of u, each of the curves shown represents the flow that is attainable if the spacing is kept *just long enough* to avoid a collision in accordance with the corresponding safety regime. Stream operations between the curves, that is, at safety levels along the continuum from one safety regime cutoff point to another, are quite possible. In other words, if viewed alone, each of the four curves outlines an area of operation in terms of q and u that offers a level of safety *equal to or better* than the safety regime represented.

Each curve indicates zero flow at zero speed, meaning that since no vehicle is moving, zero vehicles *per unit time* flow by a point on the facility. At the high-speed end

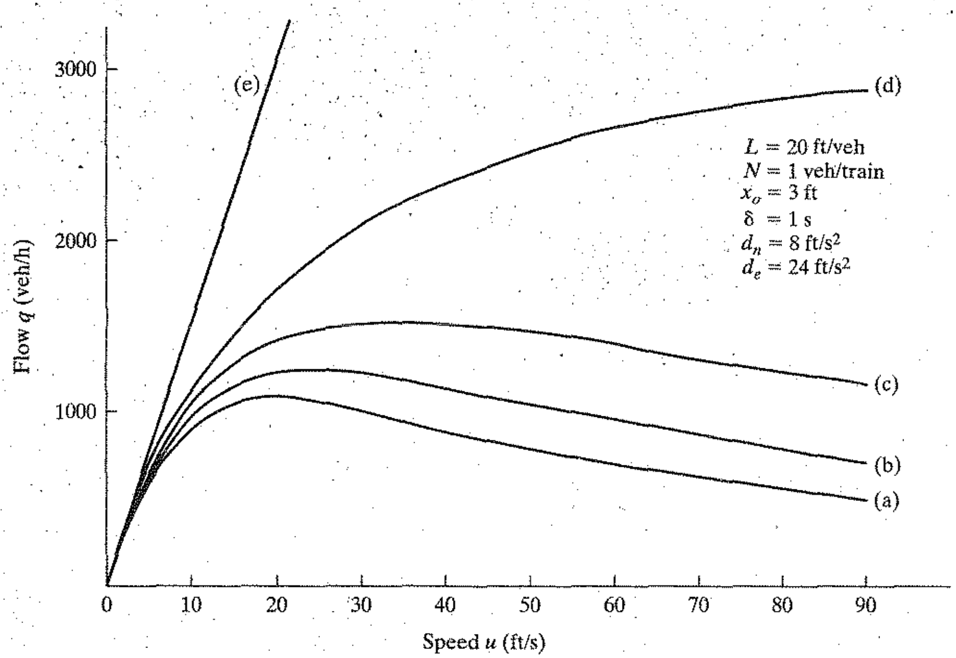


Figure 3.4.2 Speed–flow relationships.

the flow exhibits a decline because of the increasingly longer spacing requirements for safe operation. The *maximum flow* ($q_{\max}$) shown on each curve is the *capacity* of the roadway or guideway at the specified safety regime. The units of capacity are the same as the units of flow, that is, vehicles per unit time and *not* simply vehicles. Capacity occurs at an intermediate speed u_m and not at maximum (i.e., free) flow: Up to u_m, increasing speed corresponds to increasing flow; beyond u_m, increasing speed is associated with decreasing flow. Hence in this range a trade-off exists between speed and flow: Higher speeds can be attained only by sacrificing the throughput capability of the highway or guideway.

Finally, the relationship between flow and concentration can be examined by solving Eq. 3.4.2 for u in terms of k and multiplying both sides by k to obtain

$$q = k\,u(k) \tag{3.4.4}$$

Figure 3.4.3 shows a typical plot of this relationship for safety regime b as described before and, for the sake of discussion, the hypothetical train as well. The free-flow end of Fig. 3.4.3 (i.e., low flow and low concentration) corresponds to the high-speed end of Figs. 3.4.1 and 3.4.2. At the other end of the diagram concentration attains its maximum value, the flow is zero, and the speed is also zero. In other words, the roadway or guideway is occupied by as many vehicles as it can hold, but no vehicle is moving. Hence no flow is developed. These conditions correspond to a traffic jam, where maximum "packing" of stationary vehicles occurs. The value of concentration at that end is denoted by the *jam concentration* k_j. Again, maximum flow or capacity occurs at intermediate values of speed u_m and concentration k_m.

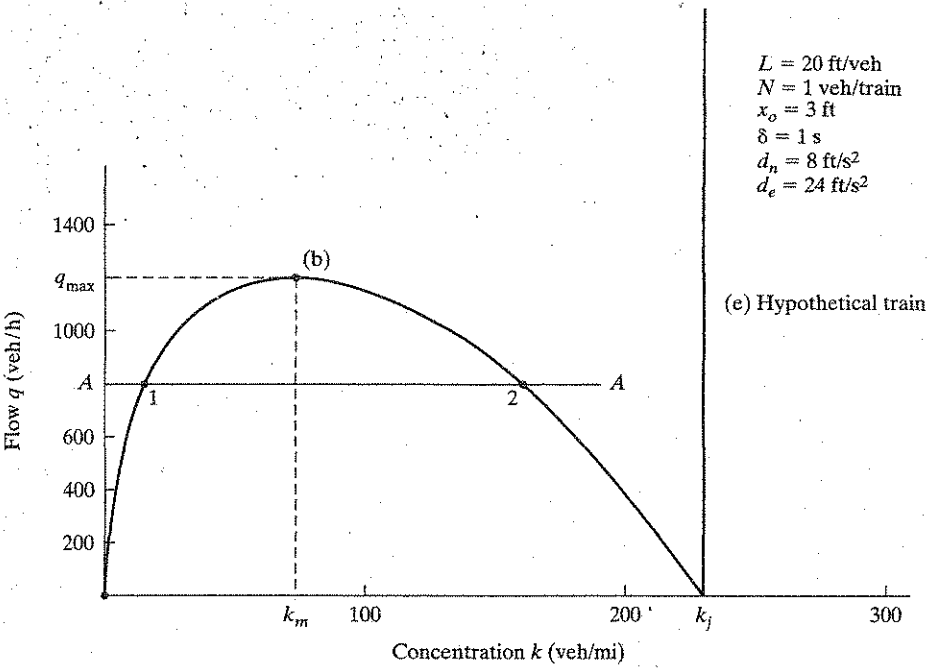


Figure 3.4.3 Flow–concentration curve.

The horizontal line AA in Fig. 3.4.3 intersects the q–k curve at *two* points. Although the flow is the same at these points, the concentration is different. Also, the speeds corresponding to these two points are different (see Figs. 3.4.1 and 3.4.2). Point 1 represents conditions that are closer to free flow, whereas point 2 represents conditions that are more congested. If a straight line is drawn from the origin to a point on the q–k curve, the *slope* of this line is simply equal to q/k, which according to Eq. 3.4.1, is equal to the mean speed u. Therefore it is possible to specify the numerical values of the three basic stream variables (q, k, and u) by using only one of the three diagrams. It is customary (especially in highway traffic analysis) to use the q–k diagram for this purpose.

Figure 3.4.4 shows a q–k curve for some safety regime, say b. As discussed earlier, operating conditions that do not lie exactly on the curve are quite possible. For example, all points associated with safety regime d lie above the regime b curve shown. Thus it is the desired level of safety that fixes the q, k, and u points on a particular curve and not the physical capabilities of the system. Consider, for example, the limiting hypothetical case of a continuous train that operates on a closed loop at a constant speed u and that is never required to decelerate [3.1]. In this case considerations of safe stopping are not relevant. Theoretically the concentration of the continuous train can be kept at jammed conditions on the track. Point C represents a stationary train at jam concentration, zero flow, and zero speed. If the train is operated at some constant speed u, its concentration remains at jam concentration, but the flow becomes finite (see point D, Fig. 3.4.4). If the operating speed is higher, as exemplified by the slope of line BE, the conditions associated with point E

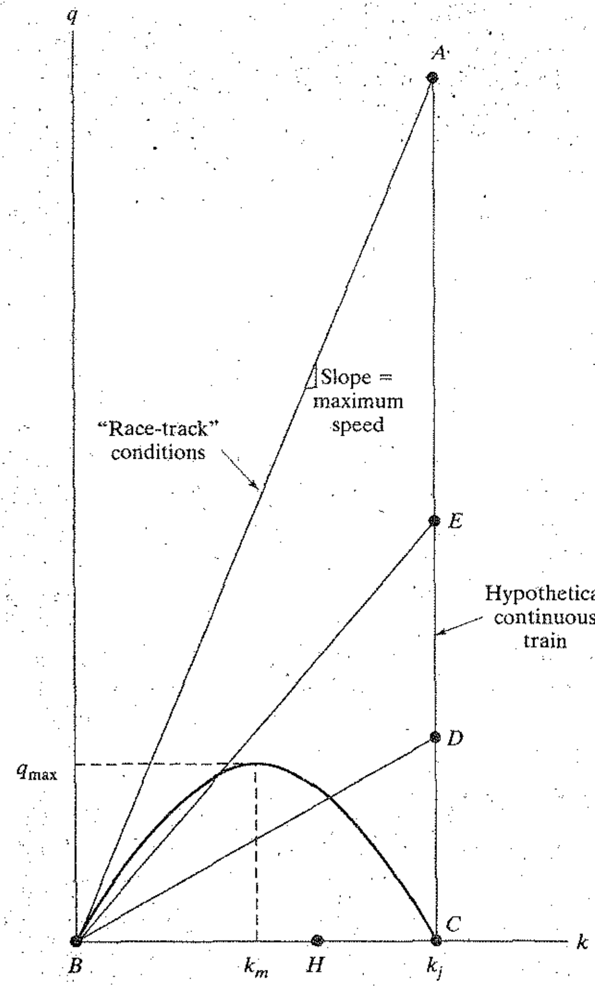


Figure 3.4.4 Flow interpretations.

result. Thus the vertical line at jam concentration represents this hypothetical case. Clearly it is physically impossible to operate the system at any of the conditions shown to the right of this line.

Now consider line *AB*. The slope of this line represents a high speed. This line would represent situations where the same high speed can be maintained at all values of concentration, a situation that is approached in the case of car racing: Irrespective of concentration, the speeds that race-track drivers sustain are very high. Of course, in the case of car racing the predominant consideration is not safety but speed.

Points below this race-track line are also attainable. Thus for a given transportation technology, triangle *ABC* encloses the area on the q–k plane within which it is physically possible to operate the system. Within this triangle, the conditions described by points below as well as above the q–k curve shown are physically possible. Even points lying on the k-axis may be given a physical interpretation. For example, point H may represent a sparsely occupied parking lane at concentration below jammed conditions, zero speed, and zero flow. But in this case speed (i.e., getting to a destination as quickly as possible) is not important!

What gives rise to the q–k curve shown within the region ABC in a typical travel situation is the trade-off between a desire to get to one's destination as quickly as possible (i.e., maximize speed) on one hand and getting there safely (as reflected by the preferred safety regime) on the other.

3.4.3 The Case of Highway Traffic Flow

The case of uniform flow considered earlier approximates the operation of a uniformly scheduled transit service on an exclusive right-of-way, where the decisions relating to the trade-off between safety and speed that give rise to the typical q–k diagram (see Fig. 3.4.3) are made explicitly by the operator of the system. In the case of highway traffic, drivers make their own decisions regarding this trade-off. Some drivers keep close to the car in front of them and try to increase their speeds when possible, whereas others keep unusually long spacings by stressing safety more than speed [3.6]. In addition, highway vehicles are not identical but exhibit a great amount of variability in size and technological attributes. The upshot of all these individual differences is a statistical *clustering* of points representing the stream conditions around a curve similar in shape to that shown on Fig. 3.4.3, and the stream diagrams and equations are typically estimated by statistical methods. This difference not withstanding, the flow diagrams of highway traffic exhibit the same general form and are subject to the same kind of interpretation as those developed for the simple case of uniformly scheduled rapid transit.

Figure 3.4.5 illustrates the general form of the u–k, u–q, and q–k diagrams corresponding to highway flow. The u–k relationship is monotonically decreasing, reflecting the rule that drivers follow on the average as they follow one another. The rule of the road suggested by many city traffic ordinances of keeping a distance of one car length for each 10-mi/h increment of speed is but one such *car-following rule.* The q–u and q–k relationships are "backward bending" as before, with maximum flow occurring at an intermediate speed u_m and concentration k_m. Typically, given the u–k relationship, it is possible to estimate the other two relationships by following the procedure that was applied earlier for the case of uniformly scheduled transit.

Example 3.2

Assume that drivers in fact follow the rule of the road of keeping a gap of one car length (L) for each 10-mi/h increment of speed. Assuming a car length of 20 ft, develop the equations of stream flow and draw the u–k, q–u, and q–k diagrams.

Solution According to the rule of the road, the safe spacing is a function of speed, or

$$s = L + \left(\frac{u}{10}\right)L = \frac{20 + 2u}{5280} \text{ mi/veh}$$

Applying Eq. 3.3.1 to find the implied k–u relationship yields

$$k = \frac{1}{s} = \frac{2640}{10 + u} \text{ veh/mi}$$

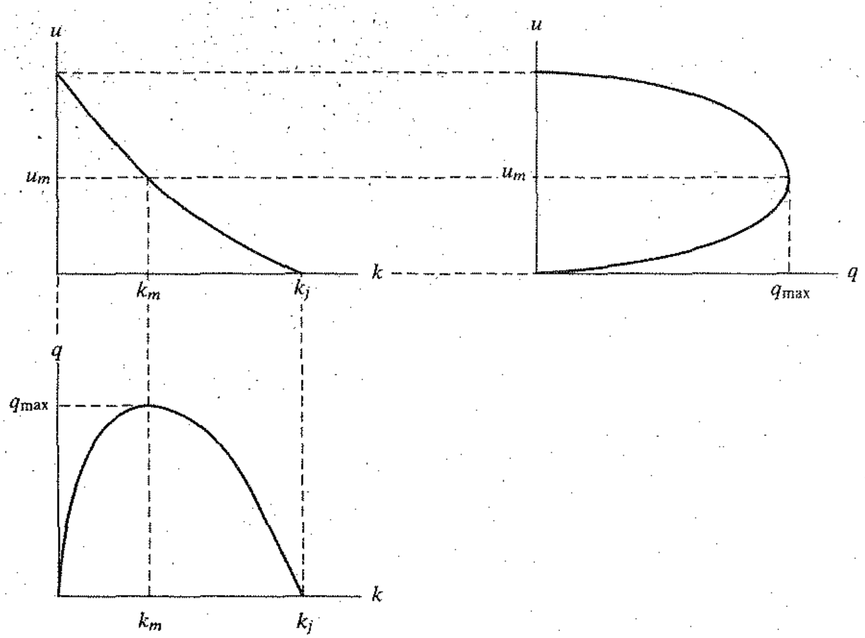


Figure 3.4.5 Flow curves.

or

$$10k + uk = 2640$$

But according to Eq. 3.4.1, $q = uk$. Thus the relationship between q and k becomes

$$q = 2640 - 10k \text{ veh/h}$$

Finally, expressing this equation in terms of u rather than k yields

$$q = 2640 - \frac{26{,}400}{10 + u}$$

The three diagrams of flow are shown in Fig. 3.4.6.

Discussion The u–k diagram has the expected general shape, showing a monotonically decreasing function. Note that the shaded area equals the product $q = uk$. The q–u and q–k diagrams, however, seem to deviate from the expected "backward bending" shape illustrated by the dashed line on the q–k diagram. In view of the observed conditions on actual facilities this implies that the rule of the road becomes unrealistic for low concentrations and high speeds. In fact, the equations just developed allow vehicles to travel at very high speeds, which is unrealistic. In other words, if the dashed line represents realistic conditions, it may be said that the rule of the road is a linear approximation of the stream conditions at the upper range of concentration. Moreover, the capacity of the roadway is not to be found at zero concentration, as the extrapolating of the straight line beyond its proper range would indicate.

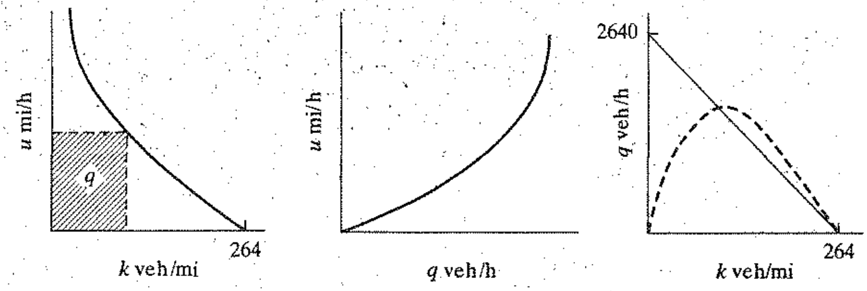


Figure 3.4.6 Numerical example of flow curves.

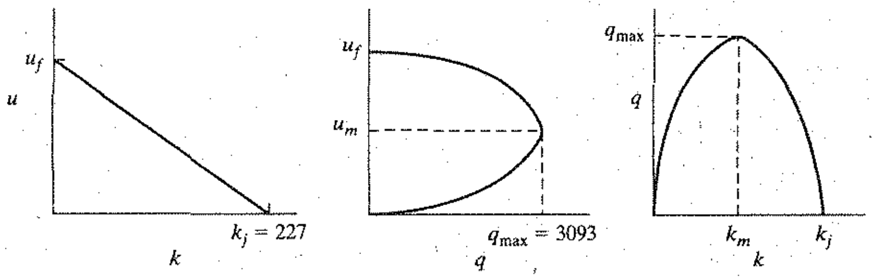


Figure 3.4.7 Derivation of flow properties from flow curves.

Example 3.3

Given that the relationship between speed and concentration obtained from actual data is $u = 54.5 - 0.24k$, repeat the steps of Example 3.2 to estimate $q_{\max}$, u_m, and k_j.

Solution To find the relationship between q and k, multiply both sides of the given u–k relationship by k and substitute the fundamental Eq. 3.4.1 in the result:

$$q = uk = 54.5k - 0.24k^2$$

To find the q–u relationship, solve the given equation for k and multiply both sides of the result by u:

$$k = 227 - 4.17u$$

and

$$q = uk = 227u - 4.17u^2$$

The plots of the three flow relationships are shown in Fig. 3.4.7.

To find k_j, we evaluate the given equation at $u = 0$. Thus $k_j = 227$ veh/mi. The free-flow speed u_f occurs at $k = 0$ and equals 54.5 mi/h. The capacity of the highway is $q_{\max} = u_m k_m = 3093$ veh/h. The reader is asked to verify these results by applying the calculus to maximize q using either the q–k or the q–u relationship.

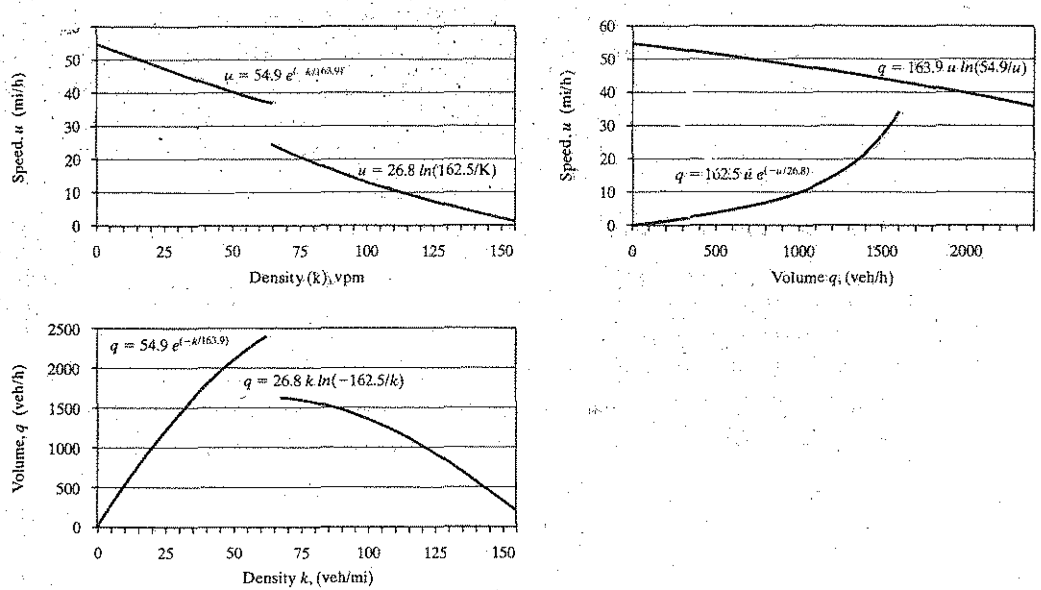


Figure 3.4.8 Example of realistic q–u–k relationships.
(From Drake, J. S., J. L. Schofer, and A. D. May, "A Statistical Analysis of Speed–Density Hypotheses." Highway Research Record 154, (1967): 53–87 [3.8].

Discussion The earlier results indicate that the general mathematical form of a *linear u–k* relationship is

$$u = u_f\left(1 - \frac{k}{k_j}\right)$$

This relationship was first postulated by Greenshields [3.7]. If the *u–k* relationship is linear, the *q–u* and *q–k* relationships are both parabolic. In that case capacity, or $q_{\max}$, occurs at $u_m = u_f/2$ and $k_m = k_j/2$. Figure 3.4.8 illustrates the best of 21 *q–u–k* relationships obtained through calibration using data from the middle lane of the Eisenhower Expressway in Chicago [3.8].

3.5 STREAM MEASUREMENTS: THE MOVING-OBSERVER METHOD

3.5.1 Background

The method of least squares can be used to determine the relationship between two or more variables based on a set of experimental observations. Examples of calibrating *u–k* relationships are presented in Chapter 13. The data used in curve fitting are obtained from an appropriate experiment or experimental observation session. Many vehicular stream-measurement techniques are available for collecting the necessary data [3.2, 3.3]. Because flow, speed, and concentration are interrelated, a proper measurement technique must take simultaneous measurements on two of the three variables; the third variable can be computed by applying Eq. 3.4.1. Taking measurements of only one of the three variables cannot describe the prevailing vehicular stream conditions. For example, the stream flow can be measured as the ratio of the number of vehicles crossing a pneumatic tube recorder that is stretched across a highway at a given location divided by the total time of measurement. Recall, however, that the same value of q is found at two points on the q–k diagram (Fig. 3.4.3), one closer to free flow and the other toward the jammed-flow end of the diagram. This means that in order to distinguish between these two conditions, the values of k and u must also be known.

3.5.2 The Moving-Observer Method

The moving-observer method of traffic stream measurement has been developed to provide simultaneous measurements of stream variables. It involves an observer, who is taking certain measurements while moving in relation to the traffic stream being measured. Referring to the two-way street operation illustrated by Fig. 3.5.1, consider the problem of measuring the stream conditions prevailing in the northbound direction. To develop the appropriate equation of the moving-observer method, consider two cases, cor-

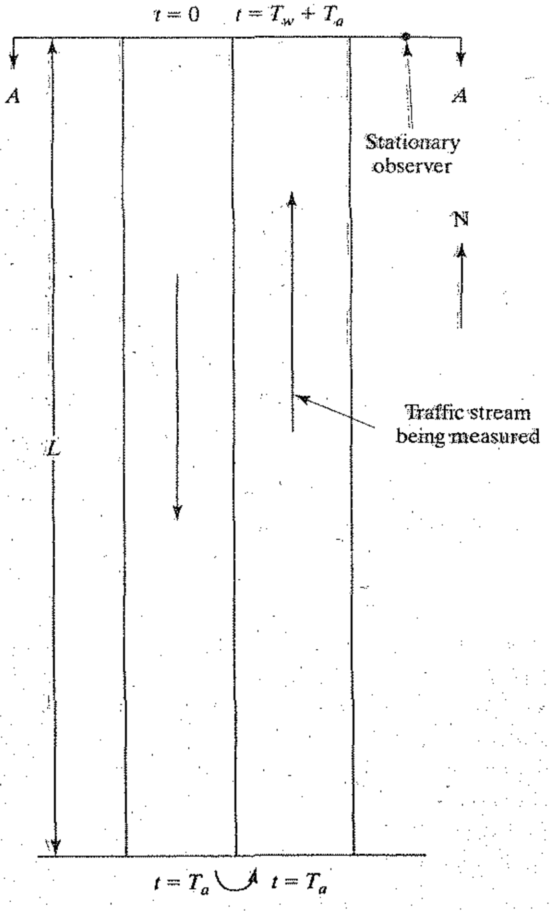


Figure 3.5.1 Moving–observer method.

responding to the *relative motion* between the observer and the vehicular stream being measured.

The first case considers the traffic stream relative to the observer; that is, it assumes a stationary observer and a moving vehicular stream. If N_o vehicles overtake the observer during a period of observation T, the observed flow is simply equal to

$$q = \frac{N_o}{T} \qquad \text{or} \qquad N_o = qT \tag{3.5.1}$$

The second case (i.e., the movement of the observer relative to the stream) assumes that only the observer is moving and the rest of the traffic is stationary. By traveling a distance L, the observer would overtake a number of vehicles N_p. Thus the concentration of the stream being measured may be computed as

$$k = \frac{N_p}{L} \qquad \text{or} \qquad N_p = kVT \tag{3.5.2}$$

where V is the observer's speed and T is the time it takes the observer to traverse distance L.

Now consider that the observer is actually moving within the traffic stream being measured. In that case some vehicles M_o will overtake the observer, and some vehicles M_p will be overtaken by the observer in a test vehicle. The magnitudes of the two counts will depend on the relative speeds between the test vehicle and the rest of the traffic: If the test vehicle is traveling faster than average, it will overtake more vehicles than will overtake it, and vice versa. This case is the combined effect of the "relative" counts described for the two cases of the previous paragraph. Denoting the difference between the two counts as M gives us

$$M = M_o - M_p = qT - kVT \tag{3.5.3}$$

and dividing both sides of Eq. 3.5.3 by T yields

$$\frac{M}{T} = q - kV \tag{3.5.4}$$

This is the basic equation of the moving-observer method, which relates the stream variables q and k to the counts M, T, and V that can be obtained by the test vehicle. The test vehicle's speed V should not be confused with the unknown mean speed u of the stream.

The values of M, T, and V taken on any particular test run are substituted in Eq. 3.5.4, leaving the two unknown stream variables q and k. To solve for these unknowns, we need two independent equations. A second test run at a different test vehicle speed to ensure independence can provide the second equation. Normally one test run is performed *with traffic* (i.e., moving in the direction of the stream being measured) and the other is performed *against traffic* (i.e., moving in the opposite direction). In both cases, however, the test vehicle counts the M_o and M_p vehicles in the vehicular stream whose conditions are being measured.

When the test vehicle is moving against traffic, it will only be overtaken (in a relative sense) by vehicles in the stream; it will overtake no vehicles. In this case then M is simply equal to the number of vehicles in the northbound stream that the test vehicle encounters while traveling south.

Substituting the measurements taken during the two test runs into Eq. 3.5.4 and using subscripts w and a to refer to the test runs "with" and "against" traffic, respectively, we obtain

$$\frac{M_w}{T_w} = q - kV_w \tag{3.5.5a}$$

$$\frac{M_a}{t_a} = q + kV_a \tag{3.5.5b}$$

The plus sign in the second equation reflects the fact that the test vehicle travels in the negative direction.

The simultaneous solution of these equations yields

$$q = \frac{M_w + M_a}{T_w + T_a} \tag{3.5.6}$$

The units of q are vehicles per unit time, which is consistent with the definition of stream flow. However, whether this value is in fact the unknown stream flow is a legitimate question. Recalling that flow is a point measurement, the answer to this question would be affirmative if a point along the roadway length L can be found where a flow measurement during the total observation time $(T_w + T_a)$ yields the stream flow obtained by Eq. 3.5.6. Point A on Fig. 3.5.1 is such a point.

To prove this claim, consider the following situation: Assume that the test vehicle begins its run against traffic at time zero. At the same time an independent observer located at point A begins to count the vehicles passing that point and continues to do so until the test vehicle crosses the same point going north. The test vehicle reaches the end of the run against traffic T_a units of time after the start of the test. It then turns around instantaneously and begins the test run with traffic, which takes T_w units of time. The total number of vehicles that would cross line A during the total time $(T_a + T_w)$ is equal to the number of vehicles M_a that the test vehicle encounters during its run against traffic *plus* the number of vehicles that overtake the test vehicle during its run with traffic *minus* any vehicles that the test vehicle overtakes during its run with traffic. The difference between the latter two counts taken during the run in the direction of the stream is simply equal to M_w, as defined before. The sum of M_a and M_w is exactly the number of vehicles that the independent observer at point A will be able to count during the time $(T_w + T_a)$. Consequently the computation of Eq. 3.5.6 yields the required stream flow q.

To calculate the space mean speed u for the vehicular stream, Eq. 3.5.5a is rewritten as

$$\frac{M_w}{T_w} = q - \frac{q}{u}\left(\frac{L}{T_w}\right) \tag{3.5.7}$$

The quantity (L/u) is the time T_{ave} that it takes the average vehicle in the stream to traverse the length L. This average time can be computed from Eq. 3.5.7:

$$T_{\text{ave}} = T_w - \frac{M_w}{q} \tag{3.5.8}$$

where

T_w = travel time of the test vehicle in the direction of the stream being measured

M_w = count taken during that run

q = flow computed by Eq. 3.5.6

Equation 3.5.8 relates the travel time of the test vehicle to the average travel time of the vehicles in the stream. If the test vehicle is traveling faster than average, it will overtake more vehicles than those that will overtake it, and M_w will be negative. Consequently the average stream travel time will be greater than the test vehicle's travel time. If the test vehicle is slower than the rest of the traffic, M_w will be positive, and the average stream travel time (Eq. 3.5.8) will be less than that of the test vehicle. Finally, if the test vehicle is traveling at the average stream speed, it will (on the average) overtake as many vehicles as will overtake it, and Eq. 3.5.8 will reflect this fact. Once the average stream travel time is computed from Eq. 3.5.8, the average stream speed can be obtained from

$$u = \frac{L}{T_{\text{ave}}} \tag{3.5.9}$$

This speed is the *space mean speed* because it is computed on the basis of travel time as described in Section 3.3. The calculation of the stream concentration is a matter of substitution of the flow computed from Eq. 3.5.6 and the mean speed computed from Eq. 3.5.9 into Eq. 3.4.1.

To ensure statistical reliability, the test is run a number of times (usually, five or six) and the average results are employed in the final calculations.

Example 3.4

A bicycle racer practices every day at different times. Her route includes a ride along a 0.5-mi bikeway and back, as shown in Fig. 3.5.2. Since she is a traffic engineer, she has made it a habit to count the number of cars in lane A that she meets while riding southward (M_s), the number of cars in lane A that overtake her while riding northward (M_o), and the number of

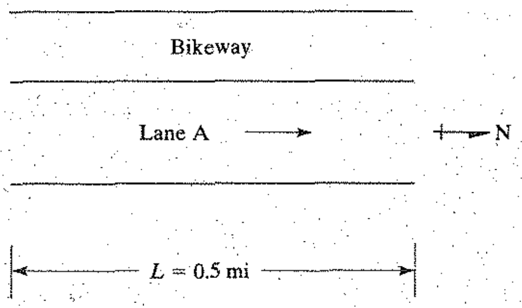


Figure 3.5.2 Example of a moving observer.

cars in lane A that she overtakes while riding northward (M_p). The table summarizes the average measurements that she obtained for each period of the day.

Time of day	M_s	M_o	M_p
8:00–9:00 A.M.	107	10	74
9:00–10:00	113	25	41
10:00–11:00	30	15	5
11:00–12:00	79	18	9

Given that the bicyclist travels at a constant speed of 20 mi/h, (a) find the traffic stream conditions for each period of the day, (b) calibrate $u = a + bk$ and plot the q–k relationship, and (c) estimate the capacity of lane A.

Solution (a) It takes the bicyclist 0.5 mi/20 mi/h = 0.025 h to traverse the half-mile distance. Hence $T_a = T_w = 0.025$ h. For each period of the day the flow, average travel time, mean speed, and concentration of lane A are computed as illustrated for the 8:00 to 9:00 A.M. period:

$$q_1 = \frac{107 + 10 - 74}{0.025 + 0.025} = 860 \text{ veh/h}$$

$$u_1 = \frac{0.5}{0.025 - \dfrac{10 - 74}{860}} = 5 \text{ mi/h}$$

$$k_1 = q_1/u_1 = 172 \text{ veh/mi}$$

The results for the other periods of the day are

$$q_2 = 1940 \qquad q_3 = 800 \qquad q_4 = 1760$$

$$u_2 = 15 \qquad u_3 = 40 \qquad u_4 = 25$$

$$k_2 = 129 \qquad k_3 = 20 \qquad k_4 = 70$$

(b) To find the speed–concentration relationship of the form $u = a + bk$, apply simple linear regression to the pairs (k, u), with u as the dependent variable. The result is

$$u = 42.76 - 0.22k$$

Multiply both sides of this equation by k and substitute $q = uk$:

$$q = 42.76k - 0.22k^2$$

which is a parabola similar to that of Example 3.3. The plots of the last two equations are shown in Fig. 3.5.3, along with the original data points.

(c) As in Example 3.3, $q_{\max}$ occurs at $u_m = u_f/2$ and $k_m = k_j/2$. Moreover, $u_f = 42.76$ mi/h at $k = 0$, and $k_j = 194$ veh/mi at $u = 0$. Hence

$$q_{\max} = 2074 \text{ veh/h}$$

Discussion This example illustrates the use of the moving-observer method to take traffic stream measurements, it applies the method of least squares to calibrate the relationship between u and k, and it applies the fundamental characteristics of traffic streams to find the q–k curve.

The first step assumed that the traffic conditions prevailing during each period of the day remain relatively stable from day to day since the observations were averaged by time of day.

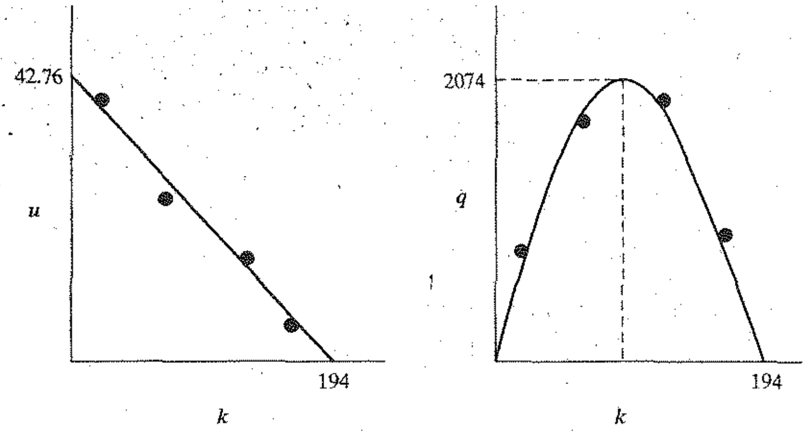


Figure 3.5.3 Flow curves from moving–observer data.

The results show that during the morning peak hour between 8:00 and 9:00 A.M. lane A is very congested with very low speed ($u = 5$ mi/h) and high concentration ($k = 172$ veh/mi). Between 9:00 and 10:00 A.M., the traffic eases and the prevailing conditions move closer to free flow. Finally, between 11:00 A.M. and 12:00 noon concentration increases and speed decreases once more, perhaps due to the lunchtime crowd.

3.6 SHOCK WAVES IN TRAFFIC

3.6.1 Background

Suppose that a traffic stream is moving on a roadway at a given flow, speed, and concentration as illustrated by point 1 on the q–k diagram of Fig. 3.6.1. Based on the calibrated diagram shown, point 1 corresponds to a flow of 1000 veh/h, a concentration of 25 veh/mi, and a mean speed (i.e., the slope of chord 0–1) of 40 mi/h. The spacing between vehicles may be computed by Eq. 3.3.1 to be about 212 ft. Now assume that a truck in the stream decides to slow down to 10 mi/h. If passing is not permitted, the following vehicles will also have to slow down to match the truck's speed. With time, a moving platoon of vehicles traveling at 10 mi/h will grow behind the truck. At any instant the last vehicle to join the platoon will be traveling at 10 mi/h, but farther upstream vehicles would continue to approach the platoon at the original conditions. Since the vehicles within the platoon are traveling slower than before they joined, they will tend to adjust their spacing to a shorter safe spacing than before. The resulting stream conditions for vehicles within the platoon are represented by point 2 on Fig. 3.6.1, where the slope of chord 0–2 is the platoon speed. In this example the values of platoon flow and concentration are shown to be 1200 veh/h and 120 veh/mi, respectively.

Thus at any time after the truck slowed down a stationary observer will see a platoon defined by the truck at its front and the last vehicle to join at its rear. The platoon, consisting of slow-moving vehicles at relatively high concentrations, is moving with a speed of 10 mi/h and grows in length as more vehicles join it. After some time the traffic conditions in front of the truck are at free flow (i.e., zero concentration). Behind the last vehicle

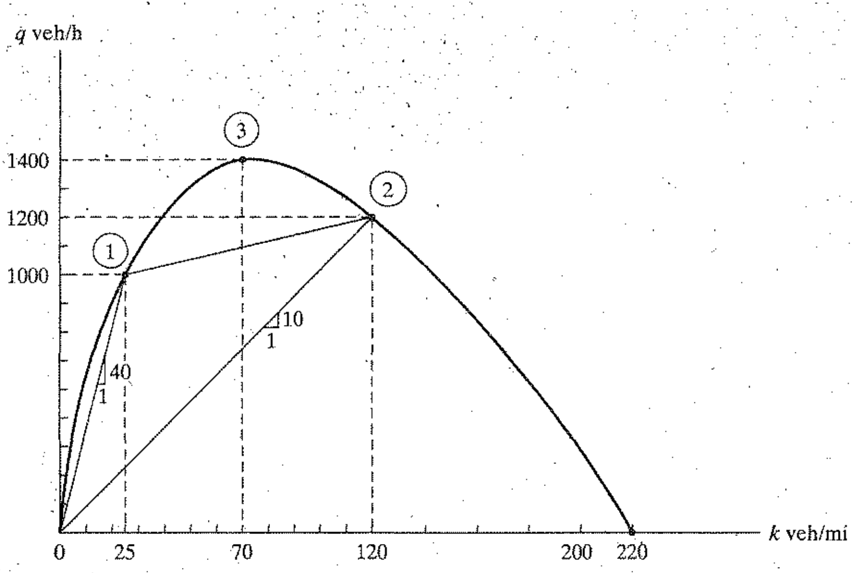


Figure 3.6.1 Shock wave description.

to join the platoon the stream conditions are at 40 mi/h and a concentration of 25 veh/mi. Figure 3.6.2 illustrates the dynamics of platoon formation described earlier: At time $t = 0$ the vehicles in the stream are shown to travel at the average stream conditions corresponding to point 1 of Fig. 3.6.1. The truck slows down to 10 mi/h at this instant. A short time later the truck has displaced somewhat and a following vehicle is shown to have matched its speed. The platoon now contains two vehicles [Fig. 3.6.2(b)]. Later on additional vehicles join the moving platoon, as illustrated. Figure 3.6.2(d) shows a clear roadway in front of the truck, the high concentration platoon behind the truck, and approaching vehicles at the original stream conditions farther upstream. The same situation is described by the time-distance diagram of Fig. 3.6.3.

Using a hydrodynamic analogy [3.9], a *shock wave* is said to exist whenever traffic streams of varying stream conditions meet. In the preceding example there are two such lines of demarcation, or shock waves. One is seen between the platoon conditions and the free-flow conditions in front of the platoon (line *AA*, Fig. 3.6.2). The other is seen between the approach conditions and the platoon conditions (line *BB*, Fig. 3.6.2). The shock wave at the front of the platoon is defined by the truck, whereas the shock wave at the rear of the platoon is defined by the last vehicle to join the platoon. Figures 3.6.2 and 3.6.3 show that the shock waves *AA* and *BB* displace with time in relation to the roadway. The rate at which the platoon grows is related to the relative speeds of the two shock waves *AA* and *BB*. Moreover, given the platoon concentration (120 veh/mi in this case) and the length of the platoon, the number of vehicles within the platoon can easily be computed.

If after a time the truck driver decides either to accelerate or to exit the highway, the vehicles stuck behind the truck will be free to increase their speeds, and another shock wave will begin between the release conditions and the platoon conditions. The next section shows that the fundamental diagram of stream flow can be used to explain the shock wave phenomenon.

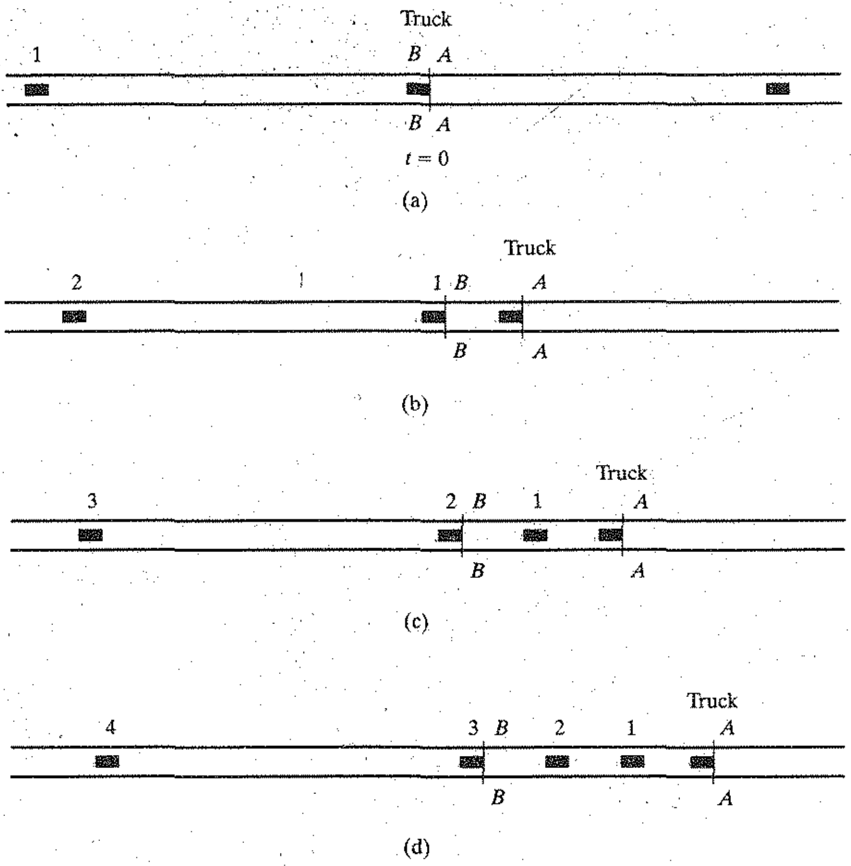


Figure 3.6.2 Platoon formation.

3.6.2 The Shock Wave Equation

It has been shown [3.9] that the speed of a traffic stream shock wave is given by the slope of the chord connecting the two stream conditions that define the shock wave (e.g., points 1 and 2, Fig. 3.6.1). Labeling the two conditions as *a* and *b in the direction of traffic movement*, the *magnitude and direction* of the speed of the shock wave between the two conditions are given by

$$u_{sw} = \frac{q_b - q_a}{k_b - k_a} \tag{3.6.1}$$

If the sign of the shock wave speed is positive, the shock wave is traveling in the direction of stream flow; if it is zero, the shock wave is stationary with respect to the roadway; if it is negative, the shock wave moves in the upstream direction. The example illustrated in Figs. 3.6.2 and 3.6.3 shows two shock waves, both traveling in the direction of the stream of vehicles. Situations arise where the shock wave travels in the opposite

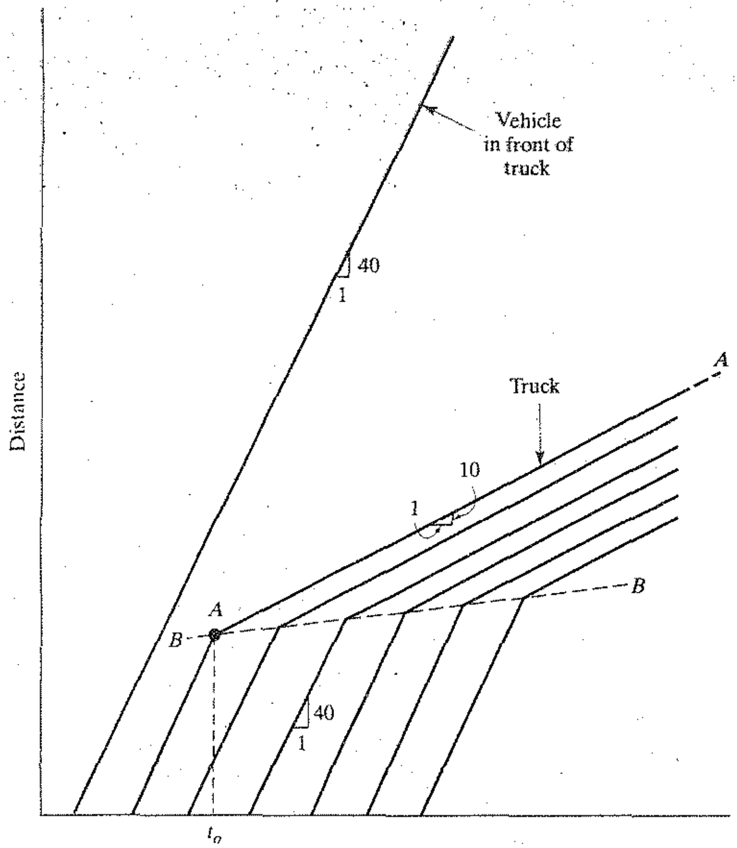


Figure 3.6.3 Time–distance diagram of platoon formation.

direction. For example, consider the case where a vehicular stream is interrupted by a traffic signal. Vehicles stopped by the red light are packed at jam concentration (i.e., zero speed and flow). Upstream of the stationary platoon, vehicles approach the platoon at the approach conditions. The shock wave between the approaching vehicles and the jammed vehicles defined by the last vehicle to be stopped is, in fact, moving in the negative (i.e., upstream) direction.

Example 3.5

For the illustration of Fig. 3.6.1, determine the magnitude and direction of the speeds of the two shock waves *AA* and *BB* and determine the rate at which the platoon is growing behind the truck.

Solution The conditions that define the shock wave at the front of the platoon are (1) the platoon conditions (i.e., q_a = 1200 veh/h and k_a = 120 veh/mi) and (2) the free-flow conditions in front of the truck (i.e., $q_b = 0$ and $k_b = 0$). Equation 3.6.1 yields the speed of the shock wave AA:

$$u_{sw}(AA) = \frac{0 - 1200}{0 - 120} = +10 \text{ mi/h}$$

which happens to be the speed of the truck, as expected in this situation. The speed of the shock wave at the rear of the platoon defined by (1) the approach conditions (i.e., $q_a = 1000$ veh/h and $k_a = 25$ veh/mi) and (2) the platoon conditions is

$$u_{sw}(BB) = \frac{1200 - 1000}{120 - 25} = +2.1 \text{ mi/h}$$

The front of the platoon moves at 10 mi/h forward relative to the roadway and the rear of the platoon travels at 2.1 mi/h in the same direction. The rate of growth of the platoon is given by the relative speed between the two, or 10.0 − 2.1 = 7.9 mi/h. The platoon grows at this rate as it travels forward.

Discussion The speed of the front of the platoon is the same as the speed of the truck only because the conditions in front of the truck are at free flow (see Fig. 3.6.3). In the general case, however, the speed of the shock wave should not be confused with the speed of any of the vehicles in the stream, as the speed of shock wave *BB* in this example clarifies. Platoon vehicles are traveling at 10 mi/h, approaching vehicles are traveling at 40 mi/h (see slope of line 0–1, Fig. 3.6.1), but the shock wave between the two travels at 2.1 mi/h. It is of interest to note that although the truck forced the traffic to slow down, the flow increased from 1000 to 1200 veh/h, a situation of which the frustrated drivers would be unaware. In certain cases slowing the traffic via the traffic control system may be a good way of increasing the flow. But this consequence goes totally unnoticed by the drivers.

Example 3.6

For Example 3.5 assume that the truck exited the traffic stream 10 min after slowing down. Vehicles at the front of the platoon were then released to a speed of 20 mi/h and a concentration of 70 veh/mi. Compute the amount of time it took the 10-mi/h platoon to disappear.

Solution The release conditions imply a flow of (20 mi/h)(70 veh/mi) = 1400 veh/h (i.e., point 3 on Fig. 3.6.1). At the end of 10 min (or $\frac{1}{6}$ h) the platoon had grown to a length of

$$L = (7.9 \text{ mi/h})(\tfrac{1}{6} \text{ h}) = 1.3 \text{ mi}$$

Incidentally, at that instant the 120-veh/mi platoon contained (1.3)(120) = 156 vehicles. After the truck exited the traffic stream a shock wave between (1) the platoon conditions and (2) the release conditions developed. The speed of this shock wave is

$$u_{sw} = \frac{1400 - 1200}{70 - 120} = -4.0 \text{ mi/h}$$

relative to the roadway. Thus the shock wave at the front of the platoon moved upstream at 4.0 mi/h, whereas the shock wave at the rear of the platoon continued to move downstream at 2.1 mi/h. The relative speed of the two waves was 4.0 + 2.1 = 6.1 mi/h. Since the platoon was 1.3 mi long to begin with, it took (1.3)/(6.1) = 0.21 h or 12.6 min after the truck's departure for the platoon to dissipate totally.

Discussion The time–distance diagram in Fig. 3.6.4 plots the location of the front and the rear of the platoon from the moment the truck slowed down to the moment when the last

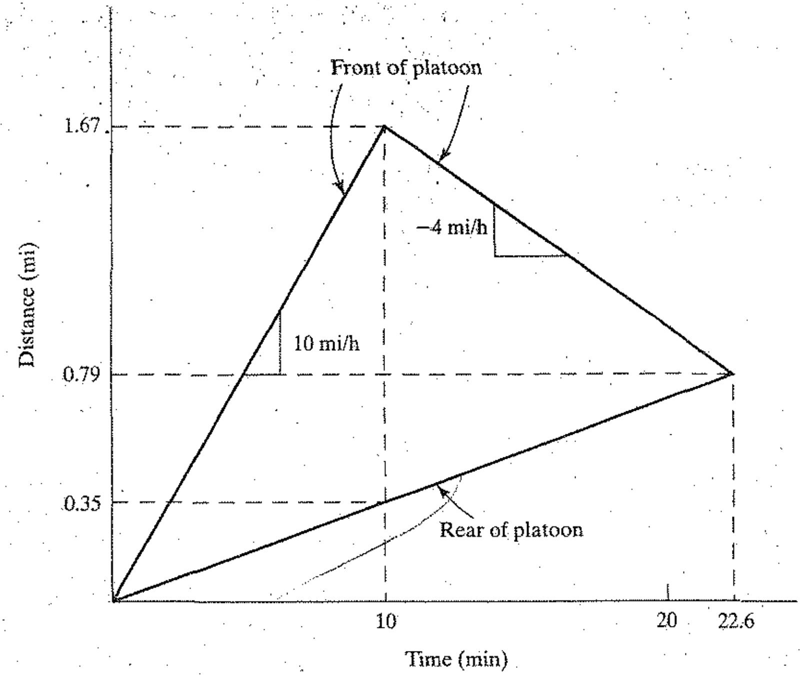


Figure 3.6.4 Time–distance diagram of platoon positions.

vehicle caught in the platoon was released. At any instant the difference between the two represents the length of the platoon, which is seen to grow from 0 to 1.3 mi during the first 10 min and then to shrink back to 0 approximately 12.6 min after the truck's exit. During this second phase the front and rear of the platoon were defined by different vehicles at different times as vehicles at the front were sequentially released and additional vehicles joined at the rear.

The point where the platoon disappeared was 0.79 mi from the initial point, even though the front of the platoon had been as far as 1.67 mi ahead of the location where the truck slowed down.

Typically the congestion relief time is longer than the duration of the flow disruption (i.e., 10 and 12.6 min in this example). This is why freeway accidents during peak periods tend to create long-lasting jams.

Example 3.7

For Example 3.6, determine the speed of the shock wave that commenced at the instant when the 10-mi/h platoon was totally eliminated.

Solution After the last platoon vehicle was released, a shock wave commenced between (1) the approach conditions behind this last vehicle and (2) the release conditions in front of it. The speed of this new shock wave was

$$u_{\text{sw}} = \frac{1400 - 1000}{70 - 25} = +8.9 \text{ mi/h forward}$$

Discussion The appearance of this, perhaps unexpected, shock wave illustrates the complexity of the dynamics of traffic flow. These accordionlike movements occur back and forth as vehicles slow down and accelerate in response to various stimuli, including other vehicles, traffic controls, sight-distance restrictions, sharp horizontal curves, and so forth. The 8.9 mi/h shock wave presumably continued as long as the approach and release conditions were sustained. In reality, these conditions change over time, as a comparison between the traffic at midnight vis-à-vis the morning rush hour would attest. Moreover, even under identical *average* conditions, there is enough variability in individual vehicle conditions to cause the continuous commencement and dissipation of such shock waves.

3.7 SUMMARY

In this chapter we extended the equations of single-vehicle motion to vehicular interactions. A general relationship between safe spacing and speed was developed and shown to affect the capacity of highways and transit ways. Based on the single-vehicle variables of speed, spacing, and headway, the fundamental variables of vehicular streams (i.e., average speed, concentration, and flow) were defined, and the fundamental relationship between the stream variables was examined. A method of measuring stream conditions (the moving-observer method) was also presented. Finally, the phenomenon of shock waves in traffic streams was illustrated.

EXERCISES

1. A rapid-transit system employing single vehicles is scheduled at constant headways. For safety regime b (Table 3.2.1), plot the relationship between spacing in feet (ft) and speed in feet per second (ft/s) using the following data: perception-reaction time = 1.5 s, normal deceleration = 8 ft/s^2, emergency deceleration = 32 ft/s^2, vehicle length = 40 ft, and safety clearance x_o = 4 ft.
2. Repeat the solution to Exercise 1 for safety regime a.
3. For Exercises 1 and 2, calculate the maximum flows in vehicles per hour (veh/h) and the corresponding speeds.
4. Given $s = 0.30/(60 - u)$, where s is the spacing in miles (mi) and u is the speed in miles per hour (mi/h), derive the relationships u–k, u–q, and q–k. Also, estimate the capacity (i.e., $q_{\max}$) of the roadway.
5. For the data of Exercise 4, plot spacing in ft versus headway in seconds (s).
6. Prepare a computer program, which, given the necessary inputs and a particular safety regime, calculates the spacing for increasing values of speed.
7. A study of the traffic using a tunnel showed that the following speed-concentration relationship applies:

$$u = 17.2 \ln(228/k) \text{ mi/h}$$

   Find (a) the capacity of the tunnel, (b) the values of speed and concentration at capacity, and (c) the jam concentration.

8. The u–k relationship for a particular freeway lane was found to be

$$u + 2.6 = 0.001(k - 240)^2 .$$

   Given that the speed is in mi/h and the concentration is in veh/mi, find (a) the free-flow speed, (b) the jam concentration, (c) the lane capacity, and (d) the speed at capacity.

9. The following relationship applies to a particular urban highway:

$$q = 273u - 70u \ln u$$

   Calculate $q_{\max}$, u_m, k_m, and u_f.

10. According to the General Motors Research Laboratory, the fuel consumption rate of passenger cars is of the form

$$F = K_1 + \frac{K_2}{V}$$

where

F = fuel consumption in gallons per mile (gal/mi)
V = space mean speed
K_1 and K_2 = calibration parameters

The following data were obtained by an experiment using a typical mix of passenger cars.

F	0.40	0.10	0.12	0.07	0.06
V	2.50	10.20	14.08	25.12	52.00

Calibrate, sketch, and interpret this model.

11. The following data were taken on a highway:

u	52	34	36	22	21	mi/h
k	8	41	48	70	105	veh/mi

(a) Estimate the free-flow speed assuming that $q = AkB^k$. (b) Calculate the capacity of the highway. (c) Find u_m and k_m.

12. A moving observer conducted two test runs on a 5-mi stretch of roadway. Both tests were in the direction of traffic. Given the following measurements, calculate the flow, concentration, space mean speed, average spacing, and average headway of the traffic stream.

Test run	Test vehicle speed (mi/h)	$M_o - M_p$ vehicles
1	10	100
2	20	−150

13. While taking measurements by the moving-observer method, a test vehicle covered a 1-mi section in 1.5 min going against traffic and 2.5 min going with traffic. Given that the traffic flow was 800 veh/h and that the test vehicle passed 10 more vehicles than passed it when going with traffic, find (a) the number of vehicles encountered by the test vehicle while moving against traffic, (b) the speed of the traffic being measured, (c) the concentration of the traffic stream, and (d) whether on its run with traffic the test vehicle was traveling faster or slower than the traffic stream.

14. Given the following u–k relationship,

$$u = 30 \ln\left(\frac{300}{k}\right) \text{ mi/h}$$

find the jam concentration, the capacity of the roadway, and the speed of the shock wave between conditions $u_a = 60$ mi/h and $u_b = 40$ mi/h.

15. A line of traffic moving at a speed of 30 mi/h and a concentration of 50 veh/mi is stopped for 30 s at a red light. Calculate (a) the velocity and direction of the stopping wave, (b) the length of the line of cars stopped during the 30 s of red, and (c) the number of cars stopped during the 30 s of red. Assume a jam concentration of 250 veh/mi.

16. A vehicular stream at $q_a = 1200$ veh/h and $k_a = 100$ veh/mi is interrupted by a flag-person for 5 min beginning at time $t = t_0$. At time $t = t_0 + 5$ min vehicles at the front of the stationary platoon begin to be released at $q_b = 1600$ veh/h and $u_b = 20$ mi/h. Assuming that $k_j = 240$ veh/mi, (a) plot the location of the front of the platoon versus time and the location of the rear of the platoon versus time and (b) plot the length of the growing platoon versus time.

17. A 15-mi/h school zone is in effect from 7:30 to 9:00 A.M. Traffic measurements taken on October 10, 1985, showed that at precisely 9:00 A.M., the conditions presented in Fig. E3.17 prevailed. How long did it take for the 3-mi platoon to disappear, and what was the speed of the shock wave that commenced at the moment when the platoon dissipated completely?

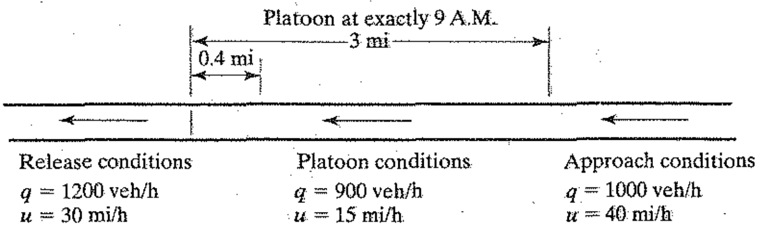


Figure E3.17

18. You were the driver of the *sixth* car to be stopped by a red light. *Ten* seconds elapsed after the onset of the following green before you were able to start again. Given that the release flow was 1200 veh/h, calculate the release concentration and the release mean speed. Assume an average car length $L = 18$ ft and a safety margin $x_o = 2$ ft.

19. Prepare a computer program that calculates the speed of the shock wave between two conditions specified by input values for k_a and k_b. Assume that $u = C - Dk$, where C and D are parametric values to be specified by the user of your program. Your program should be constrained by

$$40 \leq u_f \leq 70 \text{ mi/h}$$

$$200 \leq k_j \leq 300 \text{ veh/mi}$$

Run your program several times and interpret the results.

REFERENCES

3.1 VUCHIC, VUKAN R., *Urban Public Transportation Systems and Technology*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

3.2 GERLOUGH, D. L., AND M. J. HUBER, *Traffic Flow Theory, A Monograph*, Special Report 165, Transportation Research Board, National Research Council, Washington, DC, 1975.

3.3 TRANSPORTATION RESEARCH BOARD, *Traffic Flow Theory: A State-of-the-Art Report,* National Research Council, Washington, DC, 1997.

3.4 INSTITUTE OF TRANSPORTATION ENGINEERS, *Transportation and Traffic Engineering Handbook,* Prentice-Hall, Englewood Cliffs, NJ, 1976.

3.5 ROCKWELL, T. H., and J. TREITERER, *Sensing and Communication between Vehicles,* National Cooperative Highway Research Program Report 51, Highway Research Board, National Research Council, Washington, DC, 1968.

3.6 SYNODINOS, N. E., and C. S. PAPACOSTAS, "Driving Habits and Behaviour Patterns of University Students," *International Review of Applied Psychology,* 34 (1985): 241–257.

3.7 GREENSHIELDS, B. D., "A Study of Traffic Capacity," *Highway Research Board Proceedings,* 14 (1935): 448–477.

3.8 DRAKE, J. S., J. L. SCHOFER, and A. D. MAY, "A Statistical Analysis of Speed-Density Hypotheses," *Highway Research Record 154,* Highway Research Board, National Research Council, Washington, DC (1967): 53–87.

3.9 LIGHTHILL, M. J., and G. B. WHITMAN, "On Kinematic Waves: II. A Theory of Traffic Flow on Long Crowded Roads," *Proceedings of the Royal Society* (London), Series A, 229, 1178 (1955): 317–345.

4

Capacity and Level of Service Analysis

4.1 INTRODUCTION

This chapter is concerned with the capacity and performance analysis of actual transportation facilities and systems. Transportation facilities and systems work under uninterrupted or under interrupted flow conditions. A freeway section without on- and off-ramps and a transit guideway between two stations are good examples of uninterrupted flow. A signalized intersection and a rail-transit station are good examples of interrupted flow. Uninterrupted flow can often be approximated by fluid dynamics analogies or other continuous mathematical formulations. Interrupted flow is usually more complex and involves more interacting elements and probabilities for event occurrence. The mathematical formulation usually yields capacity in units per hour [e.g., vehicles per hour (veh/h)] and one or more measures of effectiveness such as speed, density, or delay. The Highway Capacity Manual (HCM) includes specific definitions of the level of service (LOS) for each type of facility. LOS ranges from A (the best) to F (the worst) and is defined based on ranges of values for a specific measure of effectiveness (e.g., density for freeways, delay for intersections.)

The concepts and analytical procedures in this chapter largely reflect methodologies presented in published volumes of the HCM (1994 and 1997 editions) as well as draft materials of HCM 2000. Substantial amounts of information found in the HCM has not been included here. Some procedures have been simplified and others have been expanded.

The HCM is a "living" document subject to frequent updates. As a result, the material in this chapter is appropriate for education and learning, but it should not be used for conducting real-world analyses. Instead, the HCM itself or locally approved procedures should be used.

This chapter begins with the capacity and performance analysis of pedestrian and bicycle facilities. Then it examines transit facilities, separately for uninterrupted and interrupted flow conditions. The capacity and performance of highways is presented next, again separately for uninterrupted and interrupted flow. Interrupted highway flow is given special attention because it represents the urban traffic intersection systems. Extensive sections on the capacity and performance of signalized and unsignalized intersections follow along with a section on traffic data collection methods.

4.2 PEDESTRIAN AND BICYCLE FACILITIES

4.2.1 Background

Pedestrian-flow models have been developed that bear a close resemblance to the concepts discussed in connection with vehicular streams. The speed of a pedestrian regime is, naturally, measured in units of distance divided by time, for example, feet per second. Flow is given in terms of pedestrians per unit width of a walkway per unit time. It is thus a point measurement in the same way as highway flow, where the point at which flow is observed stretches across a number of lanes. Pedestrians, of course, are not normally obliged to follow strictly any type of lane assignment, but pedestrian flow per linear foot of walkway width is a tangible measure. Density is specified as the number of pedestrians per unit area, for example, pedestrians per square foot. The reciprocal of pedestrian density is called *space* and has units of surface area per pedestrian (e.g., square feet per pedestrian). Its vehicular stream equivalent is spacing. The fundamental relationship $q = uk$ has been found to apply in the case of pedestrians under generally uninterrupted conditions.

4.2.2 Pedestrian-Flow Models

Figure 4.2.1 presents the calibrated pedestrian speed-density relationships obtained by three researchers. These diagrams conform to the general shape observed in the case of vehicular flow; that is, they are monotonically decreasing from free-flow speed at zero density to zero speed at jammed density. Figure 4.2.2 presents the speed-flow relationships corresponding to the aforementioned calibrated u–k curves. Obviously the vehicular stream parallel extends to this diagram as well: The maximum pedestrian flow (i.e., capacity) occurs at an intermediate point between free-flow and jammed conditions. Figure 4.2.3 plots the flow versus space, and Fig. 4.2.4 plots the relationship between speed and space.

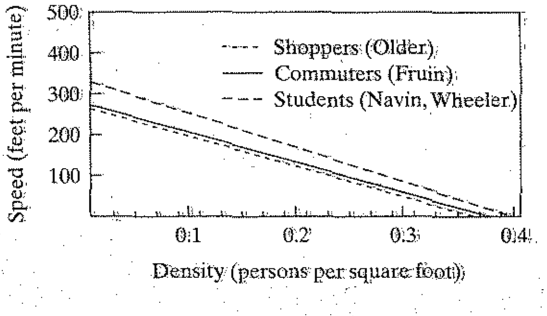


Figure 4.2.1 Pedestrian speed and density. (From Transportation Research Board [4.1].)

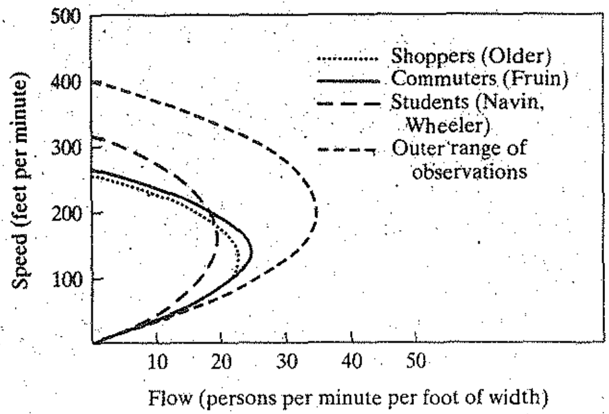


Figure 4.2.2 Pedestrian speed and flow. (From Transportation Research Board [4.1].)

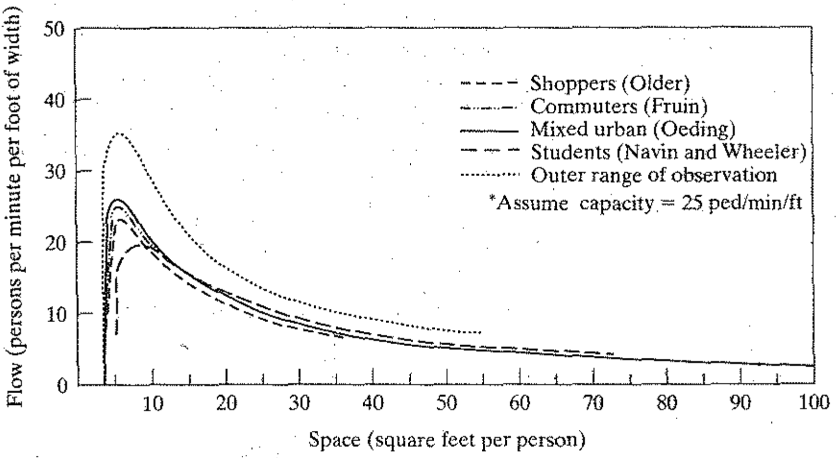


Figure 4.2.3 Pedestrian flow and space. (From Transportation Research Board [4.2].)

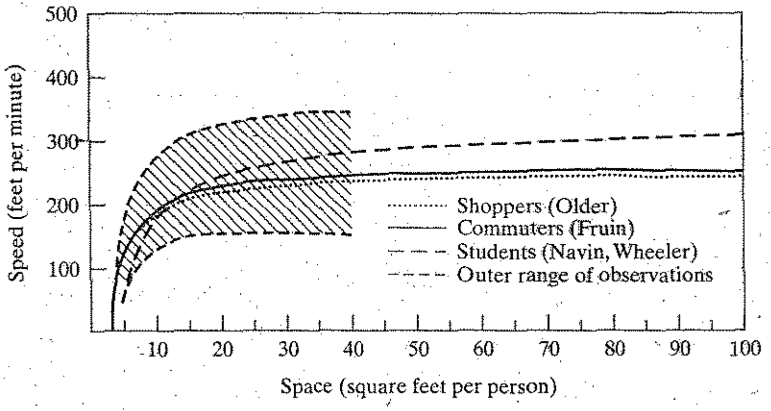


Figure 4.2.4 Pedestrian speed and space. (From Transportation Research Board [4.2].)

4.2.3 Pedestrian Level of Service

Table 4.2.1 presents the recommended ranges of pedestrian levels of service in terms of the space for walkways and sidewalks and queuing areas. The LOS at signalized intersections also is given. For signalized intersections a qualitative assessment of noncompliance is given as well. It reflects the higher propensity to walk against a "Don't Walk" signal when delays increase. Figure 4.2.5 illustrates the typical levels of density that are encountered within each level of service. The differences between the 1985 HCM (Fig. 4.2.5) and the HCM 2000 (Table 4.2.1) are apparent.

4.2.4 Bicycle Level of Service

HCM 2000 [4.3] has adopted the concept of *hindrance* as a measure for estimating the level of service (LOS) on exclusive or shared bicycle facilities. Hindrance is a rather nebulous term that is usually operationalized by using the number of events as a surrogate measure. HCM 2000 has adopted the concept of a method introduced by Botma but has not adopted specific formulae. This text summarizes Botma's proposed method for LOS estimation for exclusive and shared bicycle facilities. The main determinant of LOS is factor MP, defined as the total number of events (meeting and passing events) per hour that the average bicyclist experiences. Events include passing bikers in the same direction and meeting bikers traveling in the opposing direction. MP is estimated as follows.

$$\mathrm{MP}_{\text{Exclusive}} = V_o + 0.118 V_s \qquad (4.2.1)$$

$$\mathrm{MP}_{\text{Shared}} = 2.5\, V_{po} + V_{bo} + 3\, V_{ps} + 0.118\, V_{bs} \qquad (4.2.2)$$

where

TABLE 4.2.1 Level of Service for Pedestrian Flow

Walkways and sidewalks[a,d]	Transportation terminals[b,d]	Queuing areas[a,d]	Level of service (LOS)	Signalized intersections[c,d] Delay	Noncompliance[c,d]
	Space (ft²/person)			(s/pedestrian)	(likelihood)
> 62	> 26	> 13	A	< 10	Low
> 41–62	> 14–26	> 10–13	B	≥ 10–20	
> 24–41	> 11–14	> 7–10	C	> 20–30	Moderate
> 16–24	> 9–11	> 3–7	D	> 30–40	
> 8–16	> 8–9	> 2–3	E	> 40–60	High
< 8	≤ 8	≤ 2	F	> 60	Very High

[a] *Source:* Transportation Research Board, *Draft materials on HCM 2000.*

[b] *Source:* Davis, D. and J. Braaksma, "Levels of Service for Platooning Pedestrians in Transportation Terminals," *ITE Journal,* April 1987.

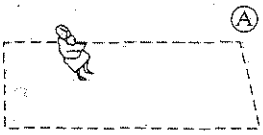
[c] *Source:* Dunn, R. and R. Pretty, "Mid-block Pedestrian Crossings: An Examination of Delay," *Proceedings of the 12th Annual AARB Meeting,* Tasmania, August 1984.

[d] Check the current version of HCM for the validity of correspondence with LOS.

LEVEL OF SERVICE A

Pedestrian Space: ≥ 130 sq ft/ped Flow Rate: ≤ 2 ped/min/ft

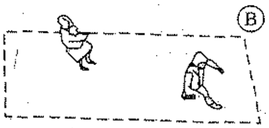
At walkway LOS A, pedestrians basically move in desired paths without altering their movements in response to other pedestrians. Walking speeds are freely selected, and conflicts between pedestrians are unlikely.



LEVEL OF SERVICE B

Pedestrian Space: ≥ 40 sq ft/ped Flow Rate: ≤ 7 ped/min/ft

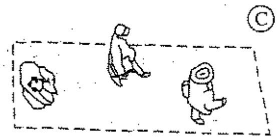
At LOS B, sufficient area is provided to allow pedestrians to freely select walking speeds, to bypass other pedestrians, and to avoid crossing conflicts with others. At this level, pedestrians begin to be aware of other pedestrians, and to respond to their presence in the selection of walking path.



LEVEL OF SERVICE C

Pedestrian Space: ≥ 24 sq ft / ped Flow Rate: ≤ 10 ped/min/ft

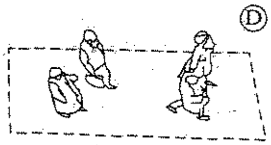
At LOS C, sufficient space is available to select normal walking speeds, and to bypass other pedestrians in primarily unidirectional streams where reverse-direction or crossing movements exist, minor conflicts will occur, and speeds and volume will be somewhat lower.



LEVEL OF SERVICE D

Pedestrian Space: ≥ 15 sq ft/ped Flow Rate: ≤ 15 ped/min/ft

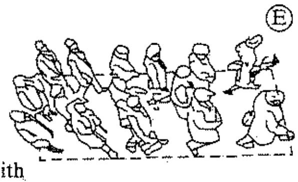
At LOS D, freedom to select individual walking speed and to bypass other pedestrians is restricted. Where crossing or reverse-flow movements exist, the probability of conflict is high, and its avoidance requires frequent changes in speed and position. The LOS provides reasonably fluid flow; however, considerable friction and interaction between pedestrians is likely to occur.



LEVEL OF SERVICE E

Pedestrian Space: ≥ 6 sq ft/ped Flow Rate: ≤ 25 ped/min/ft

At LOS E, virtually all pedestrians would have their normal walking speed restricted, requiring frequent adjustment of gait. At the lower range of this LOS, forward movement is possible only by "shuffling." Insufficient space is provided for passing of slower pedestrians. Cross- or reverse-flow movements are possible only with extreme difficulties. Design volumes approach the limit of walkway capacity, with resulting stoppages and interruptions to flow.



LEVEL OF SERVICE F

Pedestrian Space: ≤ 6 sq ft/ped Flow Rate: variable

At LOS F, all walking speeds are severely restricted and forward progress is made only by "shuffling." There is frequent unavoidable contact with other pedestrians. Cross- and reverse-flow movements are virtually impossible. Flow is sporadic and unstable. Space is more characteristic of queued pedestrians than of moving pedestrian streams.

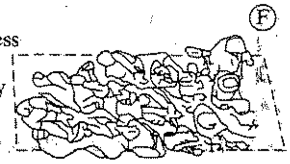


Figure 4.2.5 Levels of service on walkways. (From Transportation Research Board [4.2].)

MP = total meeting and passing maneuvers in one hour

V_o = flow rate of bicycles on the opposing direction

V_s = flow rate of bicycles on the subject direction

V_{po} = flow rate of pedestrians on the opposite direction

V_{bo} = flow rate of bicycles on the opposing direction

V_{ps} = flow rate of pedestrians on the subject direction

V_{bs} = flow rate of bicycles on the subject direction

Using MP, one then enters Table 4.2.2 to determine the prevailing LOS. In addition, the LOS on bicycle paths at signalized intersections can be assessed by estimating the average control delay as described in Section 4.7. It is highly unusual that an on-street bike path is saturated with bicycle traffic; thus, typically only the delay component d_1 in Eq. 4.7.1e (page 188) is used. Capacity is estimated by multiplying the green split (green-to-cycle ratio) by 2000 bicycles per hour.

TABLE 4.2.2 Level of Service for Bicycle Flow

Frequency of events (MP)	Control delay		
Exclusive bicycle or shared bicycle-pedestrian off-street paths[a,c] (2-way, 2-lane facilities)	Signalized intersections[b,c]	Hindrance (%)[b,c]	Level of service[b,c]
⩽ 38	< 5	⩽ 10	A
> 38–60	⩾ 5–10	> 10–20	B
> 60–102	> 10–20	> 20–40	C
> 102–144	> 20–30	> 40–70	D
> 144–180	> 30–45	> 70–100	E
> 180	> 45		F

[a] *Source:* Botma, H., "Method to Determine Level of Service for Bicycle Paths and Pedestrian-Bicycle Paths," *Transportation Research Record 1503:* 38–44, Transportation Research Board, Washington, D.C., 1995.

[b] *Source:* Transportation Research Board, *Draft Materials on HCM 2000.*

[c] Check the current version of HCM for the validity of correspondence with LOS.

4.3 TRANSIT SYSTEMS: UNINTERRUPTED FLOW

4.3.1 Background

Major urban mass transportation systems operating on roadways or exclusive guideways are defined next. Their characteristics are summarized in Table 4.3.1. The functions of these systems in an urban context are described in Section 6.4.1.

Light rail transit (LRT) is essentially a modernized electric streetcar with possibilities of articulation and capable of being operated both in mixed traffic and on exclusive rights-of-way. The PCC car is a very successful electric streetcar design that resulted from the cooperative efforts of the Electric Railway Presidents' Conference Committee during the 1930s.

TABLE 4.3.1 Characteristics of Typical Transit Systems[a]

Item[b]	Unit	Standard bus	Articulated bus	Single four-axle LRT vehicle	Personal rapid transit	Two-car AGT	Two-car-articulated LRT	Eight-car AGT	Six-car RRT	Ten-car RRT/RGR
n	veh/TU	1	1	1	1	2	2	8	6	10
l'	m	12.0	17.0	14.0	2.3	6.5	24.0	10.7	18.0	21.0
C_v	sps/veh	53[c]	73[c]	110	4[c]	40	189	70	145	175
S_0	m	1	1	2	1	1	2	1	2	2
t_r	s	1	1	1	0	0	1	0	0	0
b_n	m/s^2	1.4	1.4	1.2	1.6	1.4	1.2	1.4	1.1	1.1
b_e	m/s^2	4.0	4.0	3.0	5.0	4.0	3.0	4.0	1.8	1.8
$V_{\max}$	km/h	90	80	80	70	50	90	80	100	120
$v_{\max}$	m/s	25.0	22.2	22.2	19.4	13.9	25.0	22.2	27.7	33.3
Operating safety regime		c	c	b	c	a	a	a	a	a
Typical model		GMC's "New Look"	M.A.N.	PCC	Aramis	Airtrans	DÜWAG U-2	Skybus	Munich U-Bahn	San Francisco BART

[a] Data given are typical for the selected models. They are taken from the models given, with the exception of a few rounded or modified values to eliminate nontypical features (e.g., BART's A and C cars have different lengths; Airtrans' $V_{\max} = 27$ km/h). Acceleration rate $a = 0.8$ m/s^2 and station standing time $t_s = 20$ s are assumed for all modes.

[b] TU, transit unit (train); l', effective vehicle length; C_v, passenger spaces/vehicle; t_r, perception-reaction time; b_n, normal deceleration; b_e, emergency deceleration.

[c] Assumes seating only.

Source: Vuchic [4.4].

Personal rapid transit (PRT) refers to system designs that generally employ small vehicles operating on a network of exclusive pathways and characterized by route and scheduling flexibility.

Automated guideway transit (AGT) is designed to operate on exclusive guideways without intervention from an onboard driver in an attempt to minimize labor costs and to increase productivity; it may or may not be a PRT system.

Rail rapid transit (RRT) generally refers to high-performance, electrically propelled, multicar systems that operate on grade-separated rail facilities with few stops.

Regional rail (RGR) or commuter rail (CR) denotes multicar diesel or electric trains with very few stops that had its origins as an extension of intercity railroads.

An articulated bus is a long urban motor bus consisting of two sections, which are connected by a flexible joint to allow for short turning radii and increased passenger-holding capacity vis-à-vis the standard 40-ft urban bus.

4.3.2 Uninterrupted Speed–Flow Relationships

Figures 4.3.1 and 4.3.2 present the q–u relationships derived by Vuchic [4.4] and specify the typical operational safety regimes of the technological systems listed in Table 4.3.1. Vuchic uses the term *way capacity* to refer to the flow on the curve corresponding to a specific safety regime. This is to capture the fact that each q–u curve represents the maximum

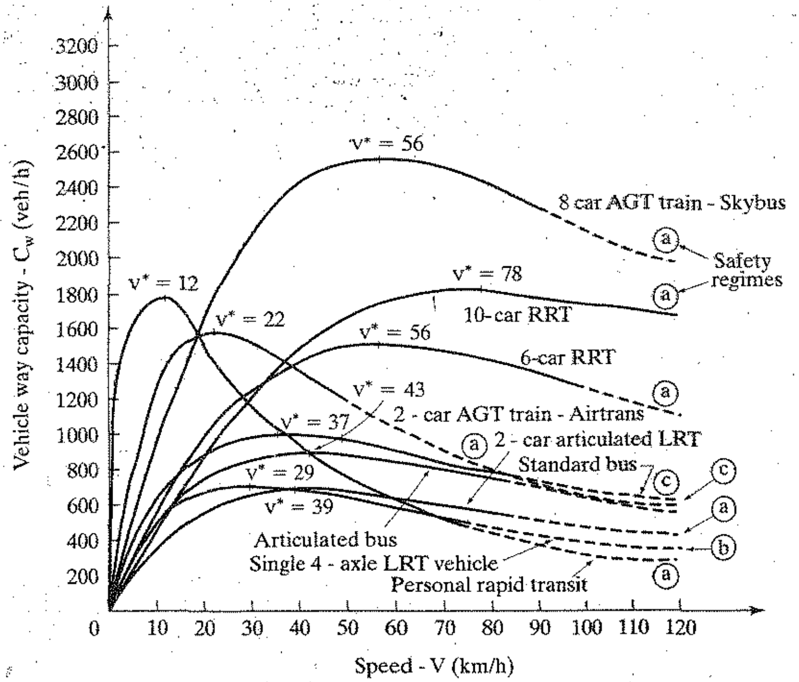


Figure 4.3.1 Actual system speed flow; vehicles per hour. (From Vuchic [4.4].)

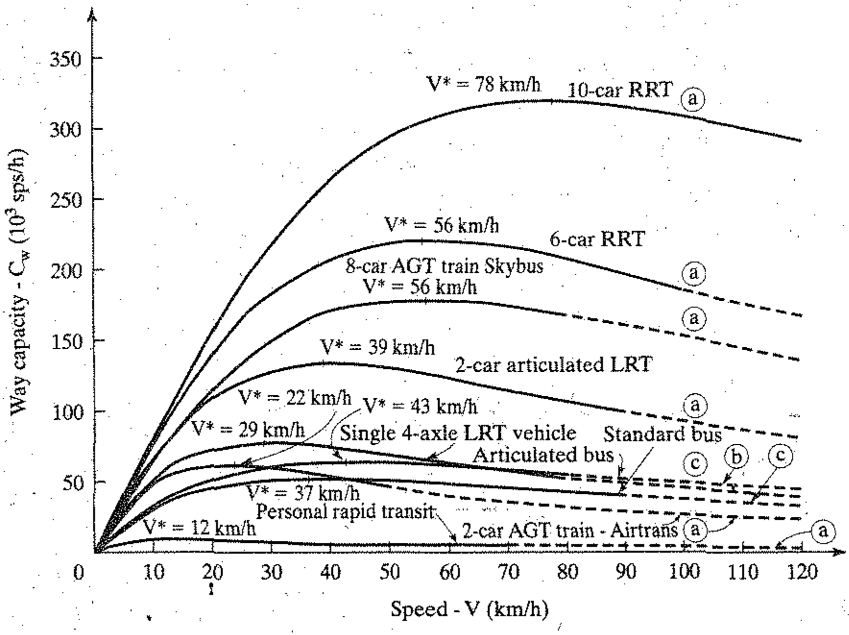


Figure 4.3.2 Actual system speed flow; spaces per hour. (From Vuchic [4.4].)

flow possible for a given speed and a specific safety regime. Note, however, that for each curve, there exists a speed for which the flow (or way capacity) is maximum. In this book this is referred to as the capacity of the system, $q_{\max}$.

Figure 4.3.1 presents flow in terms of *vehicles per hour.* Also, note that certain of the systems plotted permit *transit units* (TUs) or trains, each consisting of a number of such vehicles. Each train may be thought of as a series of vehicles having a jam spacing in contrast to the much longer spacing between transit units. Thus the coupling of vehicles into transit units in effect organizes the stream into a series of small spacings (and headways) followed by a larger spacing (and headway). This arrangement results in a higher vehicular flow than if single vehicles were scheduled individually at the constant spacing required by the same safety regime.

Figure 4.3.2 presents the flow characteristics of the same technological systems, not in terms of vehicles per hour but in terms of what Vuchic calls *spaces per hour.* Consequently these curves represents the passenger-carrying capability of each system.

Two important points relating to Figs. 4.3.1 and 4.3.2 must be stressed here. First, system comparisons of passenger flow on one hand and vehicular flow and capacity on the other do not lead to the same conclusion with respect to dominance of one system over another. It should also be noted that just because a system is capable of carrying a certain flow of passengers, it does not necessarily mean that it actually carries that maximum. To illustrate this point, consider the private automobile: The passenger-holding capacity of a typical automobile is about five to six passengers. But actual usage shows an average loading of 1.5 to 2.0 passengers per automobile, depending on the time of day, the trip purpose, and other factors.

The second point worthy of mention in connection with Figs. 4.3.1 and 4.3.2 is the fact that irrespective of how flow is considered (i.e., passenger or vehicular), there are ranges where one system dominates another, but there are also ranges where the reverse is true. Thus in many cases it is not possible to state unequivocally that one system is always superior to another. Differences in safety regime, capital and operating costs, and system flexibility are among the variables that enter the calculus of choosing between systems.

4.3.3 Fleet Size

The *number* of vehicles needed to sustain a transit line *flow* of q vehicles per hour for a time period T is affected by the fact that some vehicles may be able to traverse the line more than one time during T. A vehicular count over the time period T will yield

$$N = qT \quad \text{vehicles} \tag{4.3.1}$$

some of which will be counted more than one time. Assuming that the round-trip time of a single vehicle is T_{rt}, this vehicle will, on the average, traverse the line approximately T/T_{rt} times. Hence, to provide N vehicle departures during T, a fleet F of

$$F = N\left(\frac{T_{rt}}{T}\right) = qT_{rt} \tag{4.3.2}$$

is needed.

The average round-trip travel time is the sum of the average travel times between stops plus the average dwell time at each stop multiplied by the number of stops. The dwell time varies between 20 and 90 s, depending on the geometry of the access gates (i.e., number, width of doors, existence of steps, etc.), the number and split of passengers processed (i.e., boarding and alighting), and the degree of passenger "packing."

Example 4.1

A transit line employing nonarticulated vehicles is expected to carry 10,000 passengers during the 2-h morning peak period. Given a round-trip time of 30 min and an average vehicle occupancy of 75 passengers, calculate the hourly flow q and the number of vehicles F required to provide this flow.

Solution The number of vehicular departures needed to carry the given demand is

$$N = \frac{10{,}000}{75} = 133.3 \qquad \text{or} \qquad 134 \text{ departures in 2 h}$$

The hourly flow is

$$q = \frac{134}{2} = 67 \text{ veh/h}$$

Assuming that this flow is attainable, the number of vehicles needed is

$$F = (67 \text{ veh/h})(0.5 \text{ h}) = 33.5 \qquad \text{or} \qquad 34 \text{ vehicles}$$

4.3.4 Transit Network Fleet Size

In the preceding subsection we developed a simple formula that can be used to estimate the number of vehicles necessary to accommodate a known passenger demand on a single line. The calculation of the fleet size needed to provide services on a large transit network consisting of many transit lines is complicated by the fact that the travel desires of people vary by time of day and spatial orientation. Moreover, the selection of appropriate transit lines is a difficult task, which frequently results in overlapping lines and transferring between lines. As a simple illustration of line overlapping, consider the two-line network of Fig. 4.3.3.

The vehicle flow requirements for each link of the network are shown in parentheses for the inbound movement (i.e., toward node 3) and in brackets for the outbound movement (i.e., away from node 3). The two transit lines shown overlap on the link 2–3; hence passengers on this link can choose either line. To accommodate the known demand, the line requirements must satisfy the following conditions:

$$N_{\text{I}} \geqslant 30 \text{ departures} \tag{4.3.3a}$$

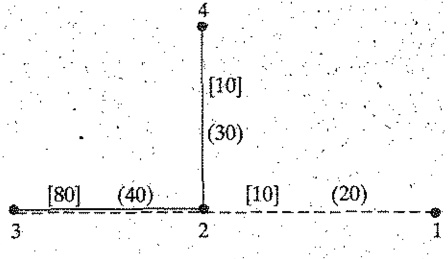
$$N_{\text{II}} \geqslant 20 \tag{4.3.3b}$$

and

$$N_{\text{I}} + N_{\text{II}} \geqslant 80 \tag{4.3.3c}$$

The problem then is to distribute the extra flow on link 2–3 (i.e., 80 − 30 − 20 = 30) between the two lines. This may be accomplished by minimizing the fleet size, for example,

$$\min F = \Sigma \left(\frac{N_i T_{rt,i}}{T} \right) \tag{4.3.4}$$



Line	Service
1	1:1, 2, 3 2:3, 2, 1
2	3:4, 2, 3 4:3, 2, 4

Figure 4.3.3 Transit line overlap. (From Papacostas [4.5].)

or because this minimization would generally favor shorter lines, by some other rule, such as distributing the extra flow to the overlapping lines in proportion to the minimum needs according to constraints 4.3.3a and 4.3.3b [4.5]. Another option involves the introduction of a shuttle line between points 2 and 3. Computer-based methods of transit network analysis are available in the technical literature [4.6].

The level of service (LOS) definition for mass transit is complex. It incorporates a multitude of factors, such as geographic coverage, on-time performance, scheduling and frequency of service, speed, comfort, safety, and security. The 1994/1997 *Highway Capacity Manual* [4.7] includes several tables on mass transit LOS. The approach of the manual is much simplified since LOS is determined only on the basis of square footage available per passenger. Table 4.3.2 presents two such tables from HCM.

TABLE 4.3.2 Examples of Mass Transit Level of Service Assessment

Peak-hour level of service	Passengers	Approx. ft^2/pass.	Pass./seat (approx.)
A	0–26	13.1 or more	0.00–0.50
B	27–40	13.0–8.5	0.51–0.75
C	41–53	8.4–6.4	0.76–1.00
D	54–66	6.3–5.2	1.01–1.25
E (maximum scheduled load)	67–80	5.1–4.3	1.26–1.50
F (crush load)	81–85	< 4.3	1.51–1.60

Peak-hour level of service	Approx. ft^2/pass.	Approx. pass./seat
A	15.4 or more	0.00–0.65
B	15.2–10.0	0.66–1.00
C	9.9–7.5	1.01–1.50
D	6.6–5.0	1.51–2.00
E-1	4.9–4.0	2.01–2.50
E-2 (maximum scheduled load)	3.9–3.3	2.51–3.00
F (crush load)	3.2–2.6[a]	3.01–3.80

[a] The maximum crush load can be realized in a single car but not in every car on the train.

Source: Transportation Research Board [4.7].

4.4 TRANSIT SYSTEMS: INTERRUPTED FLOW

4.4.1 Background

The movement of a transit vehicle *between* stations can be described by the equations of motion covered in Chapter 2. A typical vehicle or train leaving a station will accelerate to a *cruising speed* and maintain that speed until the point when it must begin to decelerate to a complete stop at the next station. The distance and time over which the cruising speed is maintained depend on the distance between the two stations, that is, the *station spacing*. Following a *dwelling time* at a station, the vehicle again enters its acceleration phase.

Figure 4.4.1 illustrates the time-distance diagram of uniformly scheduled arrivals and departures at a typical station. For simplicity, individual transit units are shown as points. The identical decelerating, dwelling, and accelerating phases of two consecutive units are shown in Fig. 4.4.1(a) for the situation where only one unit is permitted to occupy the station at any time. The headway between two units, measured horizontally, is constant as long as it is measured between points when the two vehicles are, respectively, in the same state (e.g., entering the station, leaving the station, halfway in their dwelling phase). On the other hand, the spacing between units, measured vertically, varies with time. Figure 4.4.1(b) shows two consecutive units that are permitted to dwell at the station simultaneously. The time period shown begins during the dwelling phase of unit 1 and ends with the departure of unit 3. Unit 4 is shown to have arrived prior to the departure of unit 3. The way in which headway and spacing are measured in this situation is identical to that of Fig. 4.4.1(a).

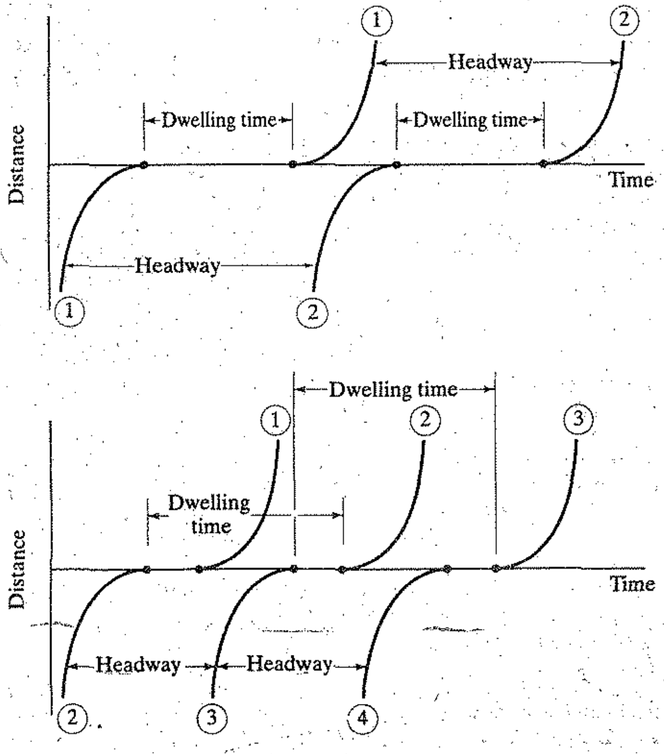


Figure 4.4.1 Transit station operations.

The minimum headway at which units enter and exit a station is affected by the combined effect of the dwelling time and the number of units that can be accommodated at the station simultaneously. This minimum station headway is most often longer than the minimum headway that is technologically attainable under uninterrupted conditions.

4.4.2 Transit Stations

The basic element of a transit station is the *platform*, where vehicles or trains stop to take on and drop off passengers. Although differing in geometric design and also in the terminology employed, all passenger-serving stations, including simple bus stops, rapid transit stations, harbors, and airport terminals, share this basic element. A *terminal* is a large station that accommodates high volumes of entering, leaving, and transferring passengers or freight. The physical facilities preceding and following the platform, on which vehicles decelerate and accelerate, respectively, can be considered a part of the station. In addition, some station designs include vehicle-holding areas on either side of the platform to be occupied by vehicles awaiting clearance to enter or exit the station.

The maximum number of vehicles (and passengers) that a station can process in a given period of time, that is, the station's capacity, depends on the number and type of platforms available, the desired level of safety, and related rules of operation. The length of each platform limits the maximum number of vehicles that can be accommodated simultaneously either singly or in trains.

Station platforms can be either *off-line* or *on-line*. Off-line platforms branch out of the mainline so that when a local vehicle or train is at the station, another unit that is not scheduled to serve the station can proceed on the mainline. Because overtaking is not possible at on-line platform locations, vehicles that are not scheduled to serve a station may be delayed behind local vehicles serving the station.

4.4.3 Single-Platform Capacity

The simplest type of transit station consists of a single on-line platform capable of accommodating a train of N vehicles of length L and allowing only one train to occupy the platform at any given time. Figure 4.4.2 shows the operation of two consecutive trains of length (NL) from the moment when the first train begins its deceleration into the station to the moment when the second train comes to a complete stop at the station, the location of which is shown on the distance axis. Each train is represented by the two parallel trajectories of

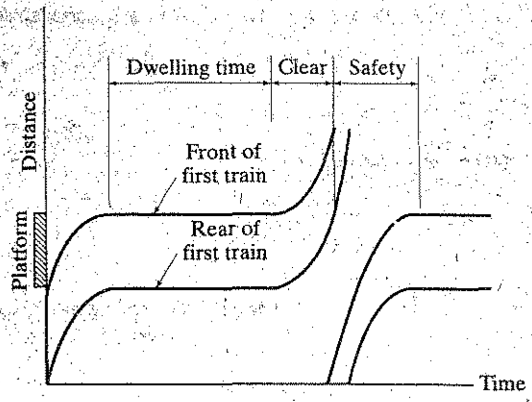


Figure 4.4.2 Transit station times.

its front and rear. The headway between the two trains consists of three parts: the dwelling time, the time it takes the first train to clear the platform, and a "safety" clearance interval. A clearance interval of zero represents the limiting situation when the second train reaches the front of the platform at the moment when the first vehicle clears it. The length of the clearance interval is related to the level of safety associated with the operation. The technical literature (e.g., [4.4]) discusses the details of safety-regime analysis of station operations for various technologies. For the purposes of this book, the minimum headway under uninterrupted conditions for a desired safety regime (given by the reciprocal of Eq. 3.4.3) is used as an approximation for the length of the clearance interval. The minimum station headway then becomes

$$h_s(\min) = T_{\text{dwell}} + \left(\frac{2NL}{a_n}\right)^{1/2} + \left(\delta + \frac{u}{2d_f} - \frac{u}{2d_1} + \frac{NL + x_o}{u}\right) \qquad (4.4.1)$$

where the second term represents the time it takes a train of N vehicles to clear the platform while accelerating at normal acceleration a_n and the third to the last terms give the minimum headway under uninterrupted conditions.

Example 4.2

Plot the relationship between cruising speed and station vehicle flow for the system described in connection with Fig. 3.4.2 for safety regime *a*. Assume a dwelling time of 10 s and a normal acceleration of 8 ft/s^2.

Solution For safety regime *a* the following vehicle maintains a spacing such that it can stop at normal deceleration in the eventuality that the leading vehicle stops "instantaneously." Substituting in Eq. 4.4.1 gives

$$h_s(\min) = 10 + (5.0)^{1/2} + 1 + \frac{u}{16} - 0 + \frac{20 + 3}{u}$$

$$= 13.24 + \frac{u}{16} + \frac{23}{u}$$

$$q = (3600)\left(13.24 + \frac{u}{16} + \frac{23}{u}\right)^{-1} \text{ veh/h}$$

This equation is plotted in Fig. 4.4.3.

Discussion The maximum flow for safety regime *a* occurs at $u = 19.18$ ft/s and equals 230 veh/h, which is significantly below the capacity corresponding to uninterrupted conditions (see Fig. 3.4.2).

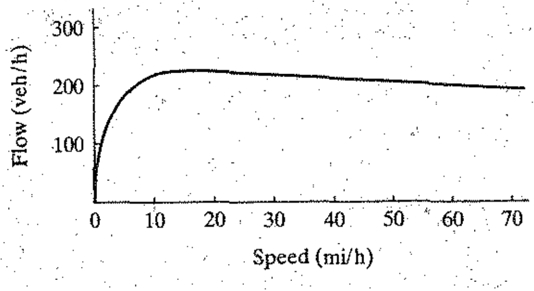
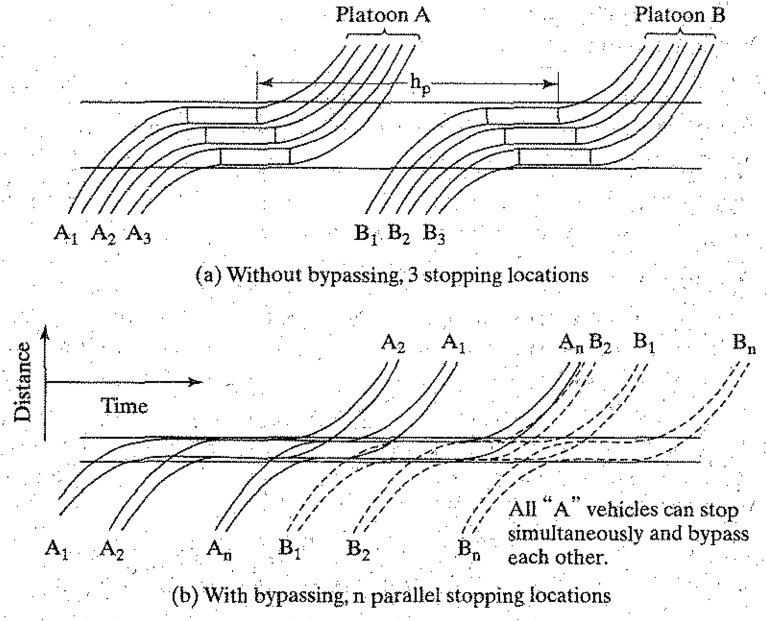


Figure 4.4.3 Transit vehicle flow and speed.

4.4.4 Other Designs

The capacity of a station can be enhanced by providing multiple parallel platforms, simultaneous dwelling, off-line platforms, and possible combinations of the three. Figure 4.4.4 illustrates two possibilities. The first involves on-line simultaneous dwelling by platoons of three units, whereas the second shows the case of *n* parallel platforms. It is noteworthy that the pattern of headways becomes more complex; in the case depicted by Fig. 4.4.4(a), the headways within platoons are much shorter than the headways between platoons.



(a) Without bypassing, 3 stopping locations

(b) With bypassing, n parallel stopping locations

Figure 4.4.4 Operation of stations with simultaneous standing. (From Vuchic [4.4].)

4.5 HIGHWAYS: UNINTERRUPTED FLOW

4.5.1 Background

The general relationships among highway flow, concentration, and speed were derived and interpreted in Chapter 3, which also illustrated the moving-observer method of measuring these variables.

From the practical point of view it is often difficult to devote the resources needed for data collection and curve fitting for each specific highway segment under investigation. Moreover, taking measurements on a facility that is under design is impossible before it is actually built. Yet the designer must anticipate the operational characteristics of the facility in order to make prudent geometric design decisions. Such estimates are based on observations of existing facilities of similar types, and practical methods of analysis and design that utilize these estimates have evolved and been codified over the years [4.1–4.3, 4.8, 4.9]. This section summarizes the basic method by which the capacity of long segments of highway facilities can be assessed. Freeway analysis based on the Highway Capacity Manual (HCM) requires the examination of three elements separately: basic segments (as defined earlier), weaving areas, and ramp junctions. This section focuses on basic segments as well as on extended segments. It presents the procedure for an overall assessment of the level of service of a basic or an extended freeway segment, the latter having variable vertical alignment features and a number of interchanges. Computer simulation with data in short intervals (e.g., 1 to 15 min) covering an entire peak period (e.g., 6 to 10 A.M.) is recommended for a detailed look at a long section (e.g., two or more miles) of a freeway given that any bottleneck causes upstream propagation of congestion, which cannot be represented in the analysis of an isolated pipeline, weaving area, or ramp [4.3].

4.5.2 Level of Service

The fundamental diagrams of vehicular streams, q–k or u–q, enclose a region that subtends the stream conditions meeting a certain safety level. It was also shown that the counteracting incentives of safety versus speed tend to cause actual stream conditions to cluster around the curve. Moreover, since the q–k and u–q curves were "backward bending," each level of flow q was shown to correspond to two distinct stream conditions on either side of capacity: one closer to free flow and the other toward traffic jam conditions. To capture this difference, the 1965 *Highway Capacity Manual* [4.8] introduced the concept of *level of service,* as illustrated by Fig. 4.5.1. The overall shape of this conceptualized relationship is the familiar u–q curve, except that the abscissa plots the normalized flow (or volume), that is, flow divided by the capacity of the roadway. The resulting volume-to-capacity (v/c) ratio ranges from 0 to 1. The area encompassed by the normalized u–q curve is divided into six subareas denoted by the letters A to F, each designating a specific level of service.

The qualitative descriptions of the conditions that correspond to each level of service can be found in the HCM 2000 [4.3].

Level-of-service A describes free-flow operations. Free-flow speeds prevail. Vehicles are almost completely unimpeded in their ability to maneuver within the traffic stream. Even at

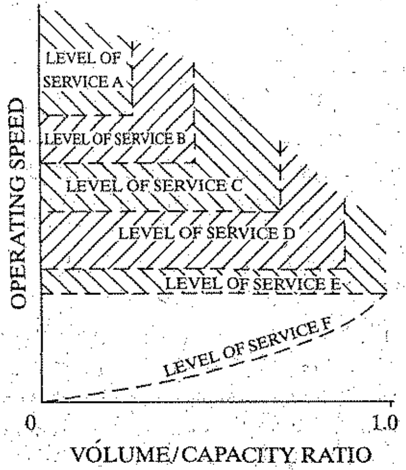


Figure 4.5.1 Levels of service. (From Highway Research Board [4.8].)

the maximum density for LOS A, the average spacing between vehicles is about 167 m [550 ft], or 27 car lengths, which affords the motorist a high level of physical and psychological comfort. The effects of incidents or point breakdowns are easily absorbed at this level.

Level-of-service B represents reasonably free flow, and free-flow speeds are maintained. The lowest average spacing between vehicles is about 100 m [330 ft], or 16 car lengths. The ability to maneuver within the traffic stream is only slightly restricted, and the general level of physical and psychological comfort provided to drivers is still high. The effects of minor incidents and point breakdowns are still easily absorbed.

Level-of-service C provides for flow with speeds at or near the free-flow speed of the freeway. Freedom to maneuver within the traffic stream is noticeably restricted and lane changes require more care and vigilance on the part of the driver. Minimum average spacings are in the range of 67 m [220 ft], or 11 car lengths. Minor incidents may still be absorbed, but the local deterioration in service will be substantial. Queues may be expected to form behind any significant blockage.

Level-of-service D is the level in which speeds can begin to decline slightly with increasing flows and density begins to increase somewhat more quickly. Freedom to maneuver within the traffic stream is more noticeably limited, and the driver experiences reduced physical and psychological comfort levels. Even minor incidents can be expected to create queuing, as the traffic stream has little space to absorb disruptions. At the limit, vehicles are spaced at about 50 m [160 ft] or 8 car lengths.

Level-of-service E describes operation at capacity. Operations in this level are volatile, as there are virtually no usable gaps in the traffic stream. Vehicles are spaced at approximately 6 car lengths, leaving little room to maneuver within the traffic stream at speeds which are still over 80 km/h [50 mi/h]. Any disruption to the traffic stream, such as vehicles entering from a ramp or a vehicle changing lanes, can establish a shock wave that propagates throughout the upstream traffic flow. At capacity the traffic stream has no ability to dissipate even the most minor disruptions, and any incident can be expected to produce a serious breakdown with extensive queuing. Maneuverability within the traffic stream is extremely limited, and the level of physical and psychological comfort afforded the driver is poor.

Level-of-service F describes breakdowns in vehicular flow. Such conditions generally exist within queues forming behind breakdown points. Breakdowns occur for a number of reasons:

- Traffic incidents can cause a temporary reduction in the capacity of a segment, such that the number of vehicles arriving at the point is greater than the number of vehicles that can move through it.
- Points of recurring congestion, such as merging or weaving areas and lane drops, experience very high demand in which the number of vehicles arriving is greater than the number of vehicles discharged.
- In forecasting situations the projected peak-hour (or other) flow rate can exceed the estimated capacity of the location.

The foregoing description of the six levels of service reveals several interesting facts. First, level-of-service E around maximum flow or capacity does not correspond to acceptably comfortable and convenient conditions from the point of view of the driver. Second, as the description of level-of-service A implies, the specific *u–q* curve and the actual capacity of a roadway depend on its physical and operating characteristics. The former include items such as grades and sight distances, as described in Chapter 2, and the latter include factors such as posted speed limits and vehicle mix. Third, along the u–q curve, each level of service constitutes a range of speeds and flows demarcated by upper and lower limits in the values of concentration, as illustrated in Fig. 4.5.2.

4.5.3 Freeway Base Conditions

This subsection introduces the procedure recommended by the HCM 2000 [4.3] for the calculation of the level of service of a geometrically uniform section of a freeway. The procedure estimates the density of a freeway in terms of passenger-car equivalents under the following "base conditions," and then modifies this estimate to capture the effect of any deviations from the base conditions that are present at the freeway section under study:

- 12-ft minimum lane width
- 6-ft minimum right side lateral clearance between the edge of the travel lane and the nearest object that influences driving behavior
- 2-ft minimum lateral clearance from the left-side median
- All passenger-car traffic composition
- Five or more lanes per direction (urban freeways only)
- Access spacing of 2 mi or greater
- Level terrain (grades no greater than 2%)
- Driver population consisting mostly of regular users of the facility (commuters)

Adjustments to these conditions yield the prevailing free-flow speed as explained in the following subsection. Prior to this, however, the definitions of volume, flow rate, and peak-hour factor are given next. Volume is defined as

> the number of vehicles passing a point on a highway or highway lane during one hour, expressed as vehicles per hour [4.8].

whereas rate of flow is defined as

> the number of vehicles passing a point on a highway or highway lane during some period of time *less* than one hour, expressed as *an equivalent rate* in vehicles per hour [4.9] (emphasis added by the authors).

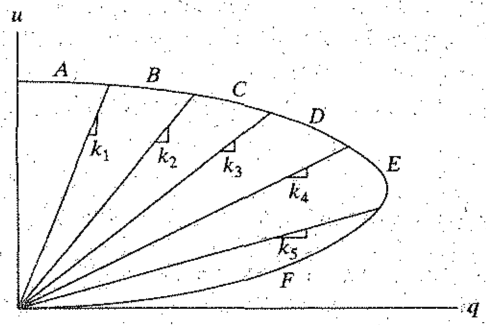


Figure 4.5.2 Level of service, speed, flow, and density.

Thus if 100 vehicles were counted during a 5-min period, the rate of flow would be 100 vehicles per 5-min period *times* twelve 5-min periods per hour equals 1200 veh/h. The actual volume V counted during the entire hour to which the above 5-min period belongs may or may not equal 1200 veh/h. The consecutive 60 min (1-h period) of the day when a highway experiences the highest volume as just defined is known as the *peak hour.* The ratio of the peak-hour volume to the *maximum* rate of flow computed on the basis of an interval t within the peak hour is known as the *peak-hour factor* (PHF). Thus, given a volume V and a maximum rate of flow q based on a set interval t less than 1 h (e.g., 5 min), the peak-hour factor is

$$\text{PHF} = \frac{V}{q} = \frac{V}{N_t(60/t)} \tag{4.5.1}$$

where N_t is the maximum number of vehicles counted during any interval t within the hour. The PHF is a measure of demand uniformity or demand peaking, as the following examples illustrate.

Example 4.3 Uniform Demand

Assume that 50 vehicles were counted during each of all possible 5-min intervals during the peak hour. Compute the PHF.

Solution The total number of vehicles that were counted during the entire hour was 600 vehicles. Thus

$$V = 600 \text{ veh/h}$$

The rate of flow based on the maximum number of vehicles observed during *any* 5-min period was, according to the denominator of Eq. 4.5.1,

$$q = 50\left(\frac{60}{5}\right) = 600 \text{ veh/h}$$

Hence the PHF is

$$\text{PHF} = \frac{600}{600} = 1.00$$

Example 4.4 Extremely Peaked Demand

Consider the extreme case where 250 vehicles were counted during a 15-min interval and no vehicles were observed during the rest of the hour.

Solution The counted volume was 250 veh/h. On the other hand, the rate of flow based on the 15-min interval was

$$q = 250\left(\frac{60}{15}\right) = 1000 \text{ veh/h}$$

Hence the PHF is

$$\text{PHF} = \frac{250}{1000} = 0.25$$

Discussion The two examples calculated the PHF for two (unrealistically) extreme conditions and found that for an absolutely uniform demand the PHF is equal to unity, whereas for an absolutely peaked demand (i.e., the entire hourly volume observed during a 15-min period), it was 0.25. In realistic conditions the PHF would lie between these limits: The closer it is to unity, the more uniform the demand and, conversely, the closer the PHF is to zero, the more peaked the demand will be. It can easily be shown that the theoretical lower bound of the PHF is 0.25 when t is taken to be 15 min. In the discussion that follows, the term *flow* denotes the rate of flow. Consequently when hourly volumes are given, they must be converted to flows by

$$q = \frac{V}{\text{PHF}}$$

HCM 2000 notes that on freeways typical PHF values range between 0.80 and 0.95.

4.5.4 Freeway Capacity and Level of Service

The technical literature contains highly refined procedures for incorporating the above (and other) factors in the study of freeway segments. Only a rough outline of these procedures is included here to familiarize the reader with the rationale of the method. For the sake of clarity the notation used in the following paragraphs is partly the authors'.

Figure 4.5.3 presents the speed-flow relationship per freeway lane for several freeway designs, defined by the design speed and the number of lanes available under base conditions. The design speed is the speed used to design the facility (see Chapter 3) and not the posted speed limit. The figure shows that under base conditions, the capacity of a freeway

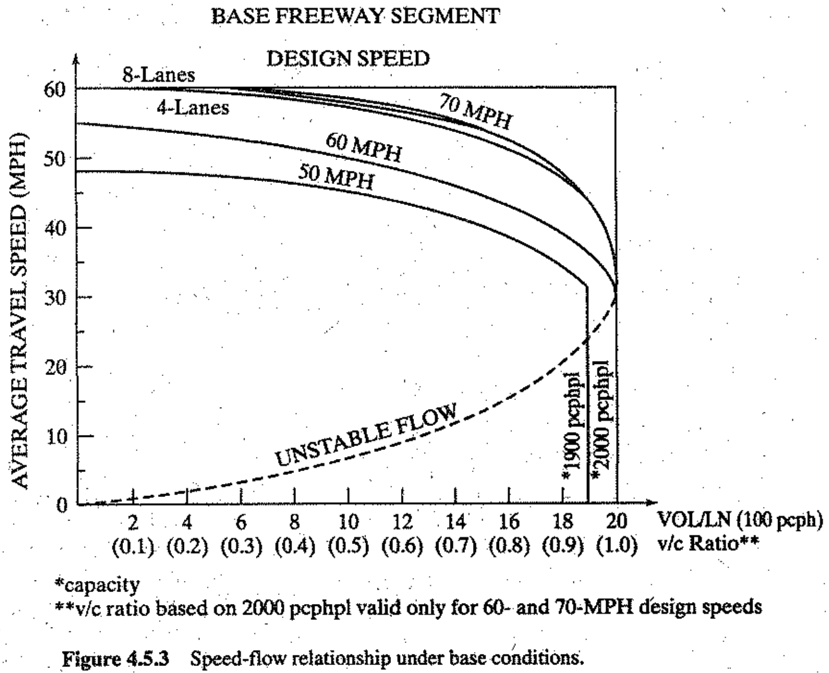


Figure 4.5.3 Speed-flow relationship under base conditions. (From Transportation Research Board [4.1].)

lane is about 2000 passenger cars per hour (pc/h) for design speeds of 60 and 70 mi/h and about 1900 pc/h for design speeds of 50 min/h.

Table 4.5.1 summarizes the speed, flow, saturation, and density levels corresponding to the six levels of service for the same freeway designs. In each case capacity corresponds to the limiting flow for level-of-service E. With appropriate adjustments, the values found in the table can be used either to analyze the conditions prevailing on a particular facility, that is, to determine the operating level of service when the number of lanes and the actual freeway volume are known, or to design a facility, that is, to determine the number of lanes needed to accommodate a given volume under a desired level of service.

As mentioned earlier, density is the measure of effectiveness that primarily determines the level of service of basic freeway segments (pipelines) as well as extended freeway segments. In the later applications, approximations of terrain and interruptions by interchanges (on- and off-ramps) are made. Once density is known, the LOS is determined based on the ranges in Table 4.5.1. Density is estimated as follows:

$$D = \frac{V_p}{S} \tag{4.5.2}$$

where

V_p = flow rate in pc/h/ln and is estimated with Eq. 4.5.3

S = free-flow speed, estimated with Eq. 4.5.6

$$V_P = \frac{V}{\text{PHF} \cdot N \cdot f_{\text{HV}} \cdot f_{dp}} \tag{4.5.3}$$

where

V = volume in veh/h

PHF = peak-hour factor

N = number of lanes

TABLE 4.5.1 Freeway Segment LOS

		Level of service				
Free-flow speed (mi/h)	Criteria	A	B	C	D	E
55	Density range (pc/mi/ln)	0–11.3	11.3–17.7	17.7–25.8	25.8–35.4	35.4–45.1
	Minimum speed (mi/h)	56	56	56	55	50
	Maximum saturation (V/c)	0.28	0.44	0.64	0.87	1
	Maximum flow rate (pc/h/ln)	630	990	1440	1955	2250
70	Minimum speed (mi/h)	68	68	67	60	52
	Maximum saturation (V/c)	0.33	0.51	0.74	0.91	1
	Maximum flow rate (pc/h/ln)	770	1210	1740	2135	2350

Source: Transportation Research Board [4.3].

f_{HV} = heavy vehicle factor and estimated with Eq. 4.5.4. Note that unlike earlier versions, HCM 2000 combines trucks and buses because there was evidence that they do not perform differently in freeway traffic ($E_T = E_B$).

f_{dp} = driver population factor; it ranges from 1, which corresponds to all-commuter motorists on the freeway (high familiarity) to 0.85 in the presence of many unfamiliar motorists such as tourists and travelers.

$$f_{HV} = \frac{1}{1 + P_T(E_T - 1) + P_R(E_R - 1)} \tag{4.5.4}$$

where

P_T = proportion of trucks and buses in traffic

P_R = proportion of recreational vehicles in traffic

E_T = passenger-car equivalent for trucks and buses

E_R = passenger-car equivalent for recreational vehicles

For specific basic (pipeline) segments one should refer to HCM 2000 tables for deriving passenger-car equivalencies based on length of segment, uphill or downhill grade, and percentage of heavy vehicle traffic. For extended freeway segments the HCM includes the following simple passenger-car equivalencies:

Factor	Level terrain	Rolling terrain	Mountainous terrain
E_T	1.5	2.5	4.5
E_R	1.2	2.0	4.0

At this point V_p can be estimated for subsequent use in Eq. 4.5.2. Next step is the estimation of the free-flow speed, which also can be estimated in the field by collecting speed samples under free-flow conditions (i.e., LOS A or B.) One should be careful not to bias the speed data collection process by causing rubbernecking (e.g., highly visible crew on an overpass), lateral displacement (e.g., vehicle or observer on the shoulder), or use of radar/laser speed-measuring equipment (which can be detected by in-vehicle devices.) All these introduce a downward bias to the speed sample.

Alternatively, the free-flow speed can be estimated as follows:

$$\text{FFS} = \text{BFFS} - f_{LW} - f_{LC} - f_N - f_{ID} \tag{4.5.5}$$

where

FFS = free-flow speed in mi/h (shown as S in Eq. 4.5.2)

BFFS = base free-flow speed, typically 70 mi/h (110 km/h)

f_{LW} = adjustment for lane width shorter than 12 ft

f_{LC} = adjustment for lateral clearance shorter than 6 ft on the right side

f_N = adjustment for less than five lanes per direction on urban freeways

f_{ID} = adjustment for the density of access points per mile (interchange density)

Factors $f_{\text{LW}}, f_{\text{LC}}, f_{\text{N}}$, and f_{ID} can be estimated with the use of tables in HCM 2000. Alternatively, these factors can be described by simple equations that, if substituted in Eq. 4.5.5, result in the following comprehensive equation:

$$\begin{aligned} S = {} & \text{BFFS} \\ & - 3.1 \times (12 - \text{W})^{1.77} \\ & - (2.4 - 0.4 \times \text{LC}) \\ & - (7.5 - 1.5 \times \text{N}) \\ & + 4.4 - 8.45 \times \text{ACCESS} \end{aligned} \tag{4.5.6}$$

where

W = lane width (10 or 11 ft)

LC = lateral clearance (from 0 to 5.9 ft)

N = number of lanes per direction (2, 3, or 4)

ACCESS = number of interchanges per mile (from 0 to 1.9)

Example 4.5: LOS Estimation

An extended freeway segment with largely level terrain has an observed free-flow speed of approximately 110 km/h, three lanes per direction, a 3-ft lateral clearance, and about one interchange per mile. It has an observed volume of 3080 veh/h with corresponding PHF = 0.88 and 154 trucks and buses, and no recreational vehicles. An all-commuter motorist composition may be assumed. Estimate the LOS for this set of conditions.

Solution Given the level terrain and 5% trucks and buses, first we apply Eq. 4.5.4:

$$f_{\text{HV}} = \frac{1}{1 + 0.05(1.5 - 1)} = 0.976$$

Then we apply Eq. 4.5.3 to estimate V_P:

$$V_P = \frac{3080}{0.88 \times 3 \times 0.976 \times 1} = 1195 \text{ veh/h}$$

Estimation of the free-flow speed S is the next step; we employ Eq. 4.5.6:

$$S = 70 - 5.5 - 3.1 \times (12 - 11)^{1.77} + 0.4 \times 3 + 1.5 \times 3 - 8.45 \times 1 = 58.7 \text{ mi/h}$$

LOS is determined by estimating density with Eq. 4.5.2:

$$D = 1195 \div 58.7 = 20.4 \text{ pc/h/l}$$

Referring to Table 4.5.1, we conclude that under this set of conditions the LOS is *C*.

4.5.5. Freeway Congestion Quantification

Although the level of service may be seen as a representation of congestion, other measures are available for assessing the level of congestion of a freeway or street. Sample metrics of congestion include the following:

- Travel rate in minutes per mile
- Delay = [actual travel time] − [acceptable travel time]

- Relative delay = [actual travel time] ÷ [acceptable travel time]
- Total delay in vehicle hours
- Corridor mobility index = [passenger volume] × [average speed in mi/h] ÷ [normalizer]*
- Accessibility = Σ {[objective ability to reach opportunities "o" (jobs, retail, entertainment, etc.)] with [actual travel time]$_o$ < [acceptable travel time]$_o$}

  Note that acceptable travel time varies by the type of opportunity; for example, a woman may be willing to travel up to 90 min to her job but less than 20 min to a movie theater.

Congestion of a freeway may be quantified by estimating the prevailing or operating speed. The Manual on Uniform Traffic Control Devices [4.11] specifies that:

> A good measure of recurring freeway congestion is freeway operating speed. An early indication of a developing congestion pattern would be freeway operating speeds less than 50 mph, occurring regularly for a period of half an hour. Freeway operating speeds of less than 30 mph for a half-hour period would be an indication of severe congestion.

The average freeway operating speed on a typical day can be estimated as follows [4.10]:

$$S_{\text{PH}} = 91.4 - 2.0\ \text{ADT} - 2.85\ \text{ACCESS} \tag{4.5.7}$$

where

S_{PH} = peak-hour speed, in mi/h

ADT = measure of annual daily traffic per lane, in thousands

ACCESS = number of access points per mile

For example, the westbound I-294 freeway has four lanes, ADT = 82,000, and six on- and off-ramps on a 2.5-mi section. Given this set of conditions, the peak-hour speed is estimated to be

$$91.4 - 2 \times 82 \div 4 - 2.85 \times 6 \div 2.5 = 43.56 \text{ mi/h}$$

This suggests that the freeway segment is congested throughout the peak hour, but not severely.

General discussion on the subject of urban traffic congestion can be found in Section 6.4.2. Quantification of arterial street congestion is covered in Section 4.7.5.

4.5.6 Capacity Restrictions

The preceding subsection dealt with the uninterrupted flow of vehicles on long freeway sections. Under these circumstances variations in flow conditions and stoppages are possible as a result of *nonrecurring* random events or incidents, such as accidents, spilled loads, disabled and slow-moving vehicles, and other extraordinary events. In addition, geometric restrictions give rise to high concentrations, or congestion, along the stream channel. These include reductions of capacity at lane drops (i.e., at points where the number of lanes

*Normalizer is equal to 25,000 for streets and 125,000 for freeways [4.10].

decreases), at points of abrupt alignment changes, and at locations where two streams come together, such as merging areas. When the approaching volume (i.e., the demand) exceeds the capacity of these locations, conditions of high density begin to appear upstream, and shock waves develop between these conditions and the approaching flow farther upstream. The duration and the severity of the congested flow depend on the degree of capacity reduction and on the pattern of demand over time. The high-density platoon may spill into similar conditions at adjacent capacity restrictions. Because such restrictions are permanent in space and the demand for travel exhibits a daily regularity, these effects are *recurring* events that take place during the regular periods of high demand [4.10].

4.6 HIGHWAYS: INTERRUPTED FLOW

4.6.1 Background

The most common interruption of highway flow, especially in urban areas, is the at-grade intersection where a common space is shared by several traffic streams. The conflicts between streams may be reduced by either separating them in space (i.e., by constructing overpasses) or separating them in time (i.e., by interrupting each stream via signal controls). To maintain a smooth progression of traffic through intersectional areas, the geometric design of intersections frequently includes the addition of regular and special turning lanes.

4.6.2 Types of Signals

The typical traffic signal controlling an intersection provides a sequential display of the green, yellow, red, and special indications, such as single or combined turning arrows, to each approach. One complete sequence of the signal displays constitutes the signal cycle, the duration, or *length,* of which is equal to the sum of the durations of its components.

Traffic signals are *pretimed* or *demand-actuated* [4.12–4.14]. Pretimed signals repeat a preset constant cycle. Demand-actuated signals have the capability to respond to the presence of vehicles or pedestrians at the intersection. They are equipped with detectors and the necessary control logic to respond to the demands placed on them. Ensuring a proper clearance interval between the green and the red phases (see Section 2.3,2) is part of this logic. *Semiactuated* signal controls are implemented at intersections of a major and a minor street, with the detectors placed only on the minor street approaches to the intersection. The heavily used major street is given a guaranteed green display, which is interrupted only when either vehicles are detected on the lightly used minor street or when pedestrians press the push button to cross the major street.

Fully actuated signals employ detectors on all legs of the intersection and are applicable to intersections of streets that carry about equal but fluctuating flows. *Volume-density* controllers are capable of sensing more detailed demand information. Complex signal control systems employ a central computer to control the flows on large highway networks. Special types of demand-actuated signals recognize and give priority to particular classes of vehicles, such as city buses or emergency vehicles.

Arterial street intersection signalization includes *isolated intersection* control, *arterial system* (also known as *open network*) control, and *network system* (or *closed network*) control.

Generally speaking, isolated intersections are located more than about half a mile from other intersections. As a result, vehicles arrive at the various approaches to the intersection randomly, and this pattern of arrivals is best suited for demand-actuated control. At high-speed (i.e., above 35 mph) isolated intersections, volume-density control (described later) is most appropriate.

An arterial system consists of a series of intersections, usually along a major street, that require time coordination to improve the efficiency of flow. Depending on the relative volumes between the arterial and the cross streets, either pretimed or demand-actuated control may be appropriate.

A network system typically takes the form of closely spaced intersections in a grid pattern such as that found in central business districts. Most of the intersections on the grid require signal control. Because of the considerable interactions between the intersections, pretimed signal control is most prevalent. Semiactuated control is sometimes employed at midblock pedestrian crossings and alley exits.

4.6.3 Signal Detectors and Controllers

In 1976 the National Electrical Manufacturers Association (NEMA) promulgated a Standards Publication relating to the various components of traffic control equipment and functional specifications. A revised NEMA standard, issued in 1983 [4.15], covers vehicle detector systems, basic and advanced signal controller units, interface (i.e., input and output) standards, solid-state flashers, and other signalization devices.

The most common type of vehicle detector used in the United States is the inductive loop detector, which employs a wire sensor loop embedded in the roadway pavement. Figure 4.6.1 illustrates two such loops, one for each of two traffic lanes. A vehicle within the detection zone of the sensor affects the magnetic field of the loop by causing a decrease in its inductance. A loop detector unit, which energizes and monitors the loop, responds to a preset decrease in inductance and sends an output signal to the controller unit. The sensitivity of the sensor can be adjusted by selecting the magnitude of the inductance drop caused by a vehicle that would generate an output signal indicating the presence or the passage of a vehicle. NEMA specifications require a sensitivity to detect small and large motorcycles and automobiles causing a signal reduction of 0.13, 0.32, and 3.2%, respectively. Thus vehicles occupying the detection zone may be classified according to the magnitude of the inductance drop they cause.

Vehicle detectors can be used to accomplish several functions, the two most basic being passage detection and presence detection. Passage detection is accomplished with a small loop that is occupied only briefly by a moving vehicle. In this case a short-duration pulse is generated to signal the vehicle's passage. Presence detection is accomplished via a

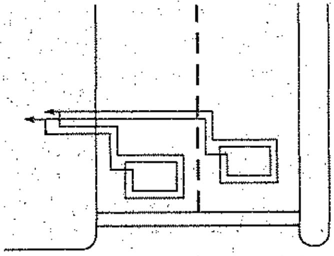


Figure 4.6.1 Example of inductive loop detectors.

long loop or a series of interconnected short loops as shown in Fig. 4.6.2. The figure also shows that a combination of short loops spaced at an appropriate distance apart may be used to respond to vehicles approaching at various speeds.

NEMA [4.15] uses the term "detector mode" to describe the duration and conditions of the channel output of a detector. In the case of presence detectors four modes are specified: *Pulse mode* refers to the case when the detector produces a short-duration pulse when vehicle detection occurs; *controlled output* refers to the case when a set-duration pulse is produced, irrespective of the length of time over which a vehicle occupies the detection zone; *continuous-presence mode* refers to the operation when the detector output continues as long as at least one vehicle occupies the zone of detection; and *limited-presence mode* corresponds to the operation when the output continues for some limited period if vehicles remain within the detection zone. Among the many features of a standard NEMA detector system is the ability, when selected, to delay its output for a certain period of time and to inhibit the output if the actuating vehicle leaves before this time expires. This feature is useful in situations where right turns on red are permitted because it helps to avoid the situation of changing a signal phase for a vehicle that has already departed. A feature allowing the extension of the detector output for a set time after the vehicle's departure is useful in permitting sluggish vehicles (e.g., slow trucks) to clear the intersection prior to a signal phase change.

The controller unit is the "brain" of a traffic controller system: It receives "calls" from the detectors and interfaces with the signal display equipment to provide for the sequencing and timing of the traffic signal displays. NEMA provides physical and functional standards for basic and advanced units to ensure compatibility between the products of various manufacturers. Some manufacturers, however, offer certain features beyond the NEMA standard. In addition to the most common NEMA controller assembly, a system known as Type 170 was developed jointly by the states of California and New York. This and updated versions (Types 179 and 2070) involve the specifications for a general purpose microprocessor

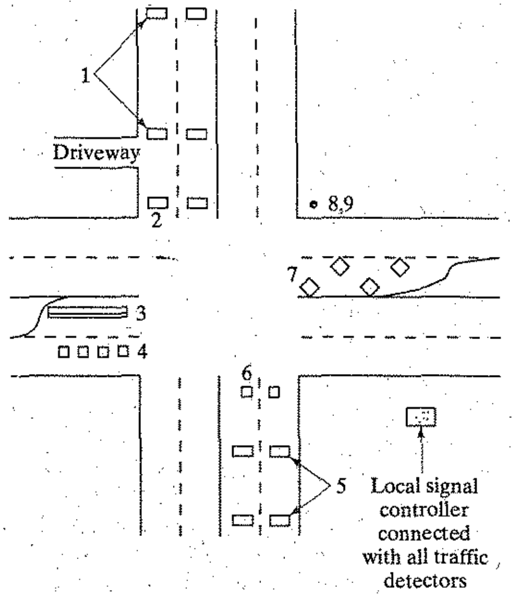

1
Driveway
2
8,9
7
3
4
6
5
Local signal controller connected with all traffic detectors

Figure 4.6.2 Examples of forms and uses of loop detectors: (1) two sets of loops, 60 ft apart for 30- to 35-mph speeds; (2) two 6- × 8-ft loops to detect driveway activity; (3) long loop for presence detection; (4) four 6- × 6-ft loops over 54-ft length for presence detection; (5) two sets of loops 80 ft apart for 40-mi/h speeds; (6) two 6 × 6-ft loops for pulse operation; (7) diamond loops for presence detection; (8, 9) pedestrian and bike crossing button.

controller [4.16]. The functionality of the system is implemented through software rather than through the specific switch-setting options provided by the NEMA controller. Both types of controller units are capable of implementing a variety of phasing and timing strategies, including pretimed (fixed) and complex actuated control schemes.

Two fundamental concepts that aid the understanding of the operation of traffic signal controllers are the definition of the terms "phase" and "ring." A *phase* is defined as consisting of the green interval, the yellow interval, and where applicable, the subsequent short red (clearance) interval that are associated with a combination of movements which are always given the right-of-way simultaneously. As will be explained later, modern actuated controllers are capable of displaying two phases at the same time. It is thus easy for a casual observer of a traffic signal to think of the two simultaneously displayed phases as a single phase, even though under alternate traffic demand conditions the controller has the ability to display the two phases independent of each other. A phase is said to be active if any of its three component parts is being displayed; otherwise the phase is inactive (red).

A *ring* is defined as a sequence of phases in the order in which they would be displayed if demand existed for all of them. A single ring may contain from two to four such phases. By convention the phases in a ring are designated by the Greek capital letter phi (for phase) followed by a phase number (e.g., Φ1). Figure 4.6.3 illustrates the phase designations for a two-, three-, and four-phase ring. The wraparound arrow shows the direction of the phase sequence. If the signal were to operate in a fixed timing pattern, each phase would be displayed in sequence and would be of the pretimed duration. If the signal were to operate in the actuated

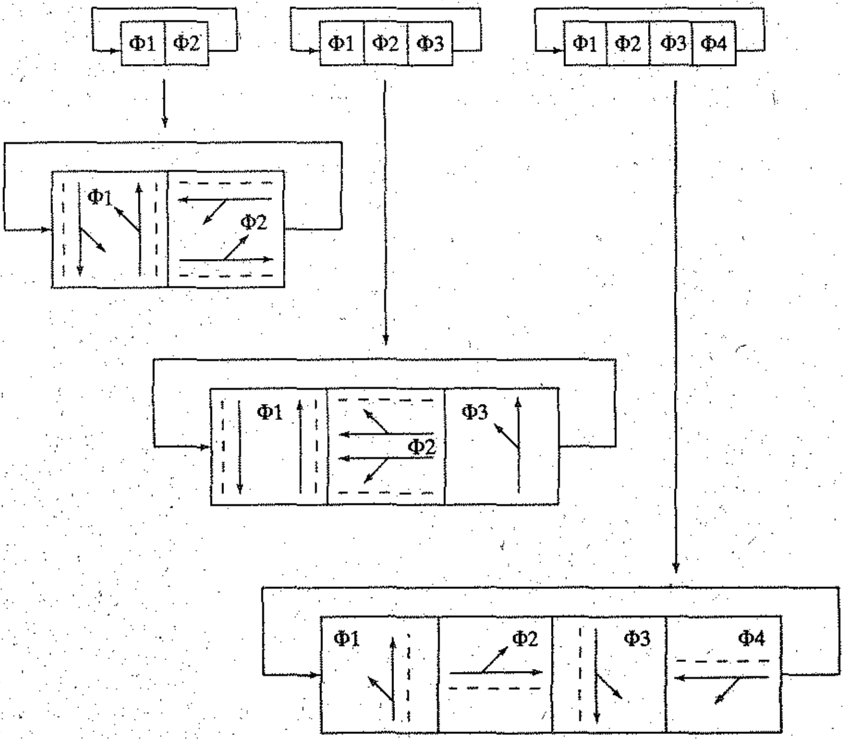


Figure 4.6.3 Examples of single-ring operation.

mode, the duration of each phase would depend on the traffic demand placed on it (see below). For each of the three rings shown in Fig. 4.6.3 a typical phasing example is presented. Each phase shows its associated movements. The dashed straight-line segments that appear in connection with some of the illustrated phases indicate the pedestrian movements that are allowed during each of those phases.

For more complex phasing patterns dual-ring controllers are used. These consist of two parallel four-phase rings as shown in Fig. 4.6.4, so that a maximum of eight phases can be defined. The two rings allow for the possibility of displaying two phases concurrently, one from each ring. However, the two rings are interlocked at two reference points (also known as *barriers*). The two rings are forced to cross these barriers simultaneously. In other words phases 2 and 6 on one hand and 4 and 8 on the other must terminate concurrently. This is required to avoid conflicting movements that are typically found on opposite sides of the barriers.

Obviously it is critical when defining the movements associated with the various phases to avoid conflicting movements between the phases defined for the two rings that lie on the same side of a barrier. Figure 4.6.5 illustrates a typical two-ring phasing scheme known as a "quad left" operation. A mode of operation that requires that one phase from each ring must always be active is known as the *dual-entry mode* of operation. By contrast, a *single-entry mode* allows for only a phase belonging to one ring to be active, whereas all phases of the other ring could be inactive in the absence of calls for phases in the other ring.

Figure 4.6.6 illustrates the possible phase combinations that would result under dual-entry operations. Of note is the fact that based on the traffic demand placed on the various phases on the same side of a barrier, alternate phase display combinations are possible. For example, if the demand for left-turn phase 5 is larger than for phase 1, phase 5 would continue to be active in combination with phase 2 after the termination of phase 1. If on the other hand the demand for phase 1 is greater than for phase 5, phase 1 will be displayed concurrently with phase 6 after the termination of phase 5. Finally, if the demand for phases 1 and 5 is equal, the next combination to be displayed would consist of phases 2 and 6. In this arrangement phases 2 and 6 must terminate concurrently irrespective of their relative demands because of the barrier constraint; otherwise unacceptable conflicting phases (such as 2 and 7 or 6 and 3) could result. A useful way of visualizing the relationship between the dual-ring phasing diagram (such as Fig. 4.6.5) and the resulting phasing combinations has been suggested by McShane and Roess [4.17] as illustrated in Fig. 4.6.7. In that figure the

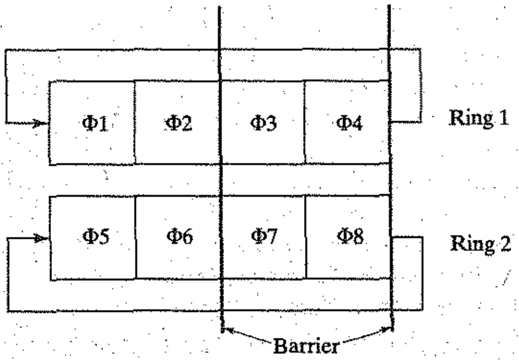


Figure 4.6.4 Example of double-ring operation.

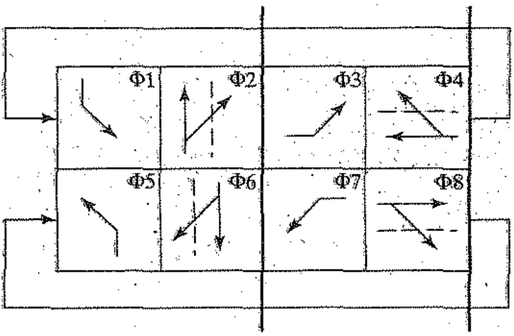


Figure 4.6.5 Example of eight-phase "quad-left" operation.

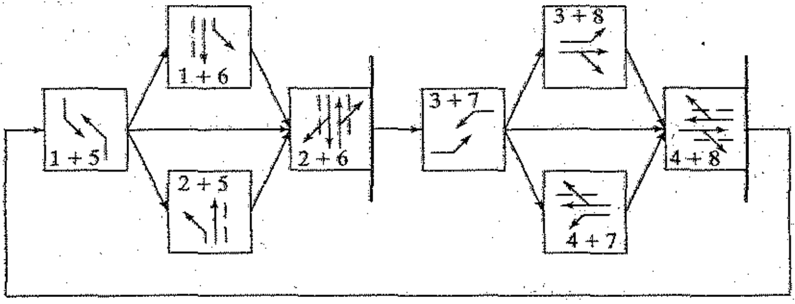


Figure 4.6.6 Phase combinations in eight-phase operation.

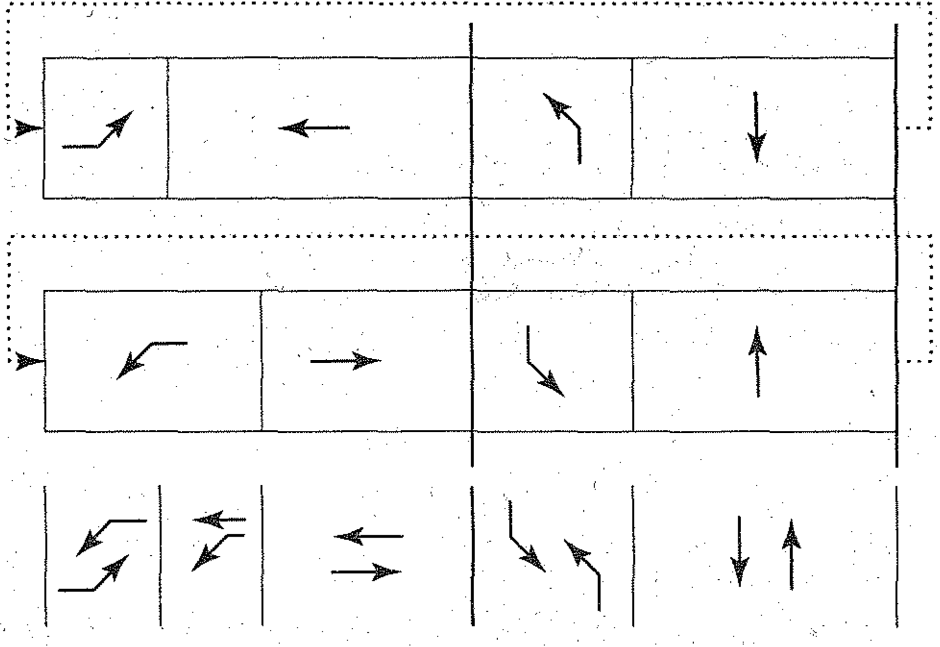


Figure 4.6.7 The ring concept applied to an eight-phase roller.
(From *Traffic Engineering* by McShane/Roess, ©1990. Reprinted by permission of Prentice-Hall, Inc., Upper Saddle River, NJ.)

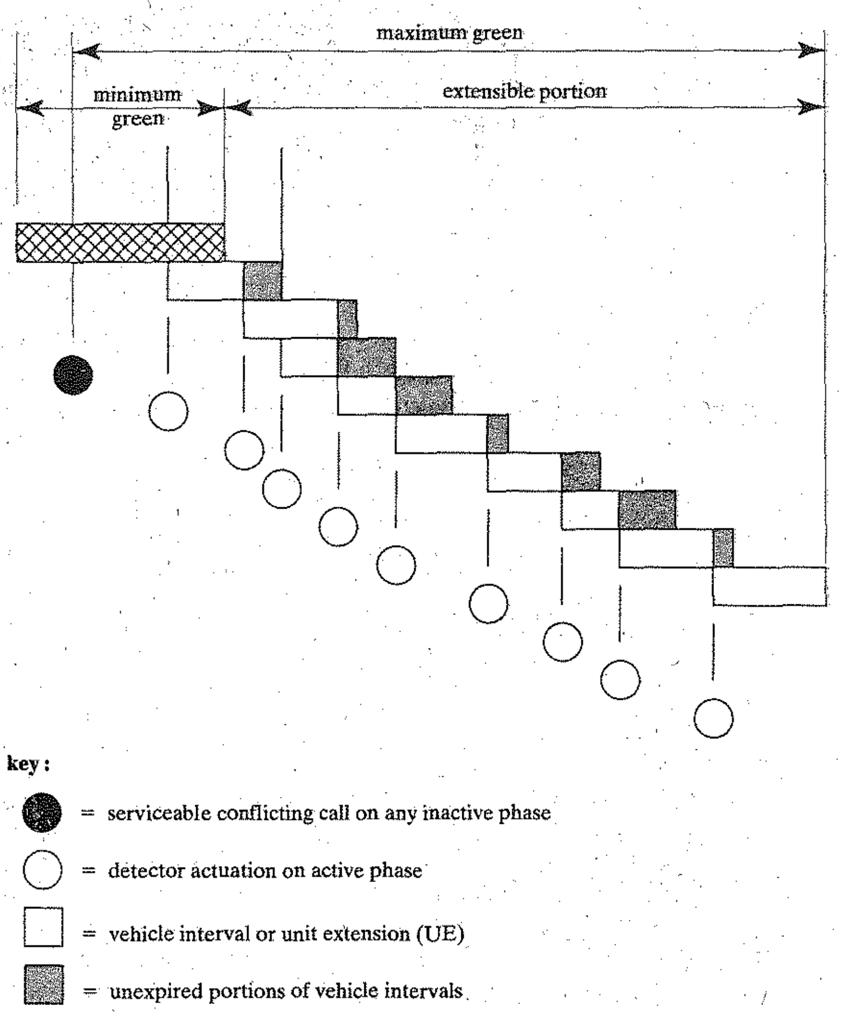

maximum green
minimum green
extensible portion
key:
● = serviceable conflicting call on any inactive phase
○ = detector actuation on active phase
□ = vehicle interval or unit extension (UE)
▨ = unexpired portions of vehicle intervals

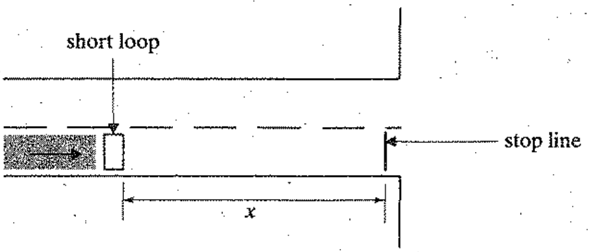
Figure 4.6.8 Basic parameters for green timing under actuated control.

"length" of each phase is proportioned according to a possible demand pattern subject to the barrier constraints; the corresponding phase display combinations are clearly seen.

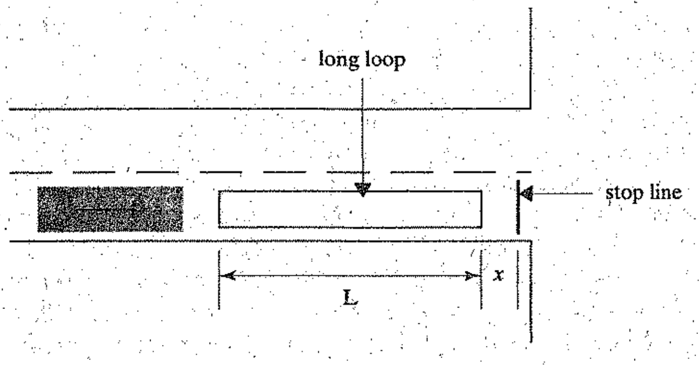
In the preceding discussion reference has been made to the ability of actuated controllers to adjust the duration of phases based on the corresponding traffic demand. An exhaustive coverage of how this is accomplished is beyond the scope of this book and only the basic principles are presented next. In the simplest case the duration of the green component of a given phase may be thought of as consisting of a minimum green and an extensible portion as illustrated in Fig. 4.6.8. When the phase becomes active, the preset minimum green is displayed, and depending on the gaps between detector actuations for that phase, green is extended by the passage time, also known as the vehicle extension. This interval represents the time required by a vehicle to travel the distance from the detector to the stop line.

This is illustrated in Fig. 4.6.8, where the abscissa represents time and the bullets designate the times of vehicle detection. The maximum green limits the length of the extensible portion. In the case of NEMA controllers the maximum green begins to time when a serviceable conflicting call is received rather than from the onset of green. In the case shown continuous demand is present and the green "maxes out." If any one of the extensible portions were to expire without a vehicle actuation, the phase would have "gapped out." In the case just described the minimum green was assumed to be fixed at a preset duration. NEMA, however, provides for a more complex possibility where the initial green (prior to the extensible portions) is variable and consists of a minimum initial interval plus an additional interval that depends on the number of vehicles that are queued prior to the onset of green.

The two most common methods used to extend the green in response to vehicle demand are point detection using short loops (known as *conventional detection*) and long loop presence detection (also referred to as *loop-occupancy detection*).



In the case of point detection a short loop is commonly placed at a distance x upstream of the stop line that would be traveled by a vehicle at the selected approach speed in 2 to 5 s as shown in the preceding diagram. This time interval is known as the *passage time* and is equal to the *vehicle extension* (or *unit extension*) interval. The basic idea is to extend the green by this amount for each actuation during the green phase so as to allow the detected vehicle to reach the intersection. In other words, when the green phase for this movement is active, a vehicle that actuates the detector will need this amount of time to reach the stop line. This is the reason that an actuation resets the timing of the vehicle extension interval as shown in Fig. 4.6.8. A gap-out would occur when the passage time elapses without an additional actuation and a call for service is waiting on an inactive (i.e., red) conflicting phase.



Extension of the green interval with presence detection using a long loop (or a series of interconnected short loops to obtain the same effect) is illustrated. A long loop of length L is placed with its trailing edge at a short distance x from the stop line. Thus the vehicle interval required to travel the distance x is relatively short (typically 0 to 1.5 s). In this mode of operation the vehicle interval is *held* at its beginning (i.e., the green indication is sustained) as long as a vehicle occupies the loop's detection area. The required length of the loop is determined by the formula

$$L = 1.47\, V(G - V_i) - L_v$$

where

L = loop length, in ft

V = approach speed, in mi/h

G = desired allowable gap, 2 to 5 s

V_i = vehicle interval to travel distance x, in s

L_v = average vehicle length (usually 20 ft)

1.47 = conversion factor from mi/h to ft/s

In other words the desired gap is exceeded when no vehicle occupies the loop in which the controller unit can service a call waiting on a conflicting phase.

As a general rule, the duration of the *minimum green* interval is related to the distance x from the trailing edge of the detector to the stop line. The minimum green interval is selected so as to allow vehicles that were stored within this distance to enter the intersection, allowing for a start-up delay after the onset of green. A commonly used equation for estimating the duration of the minimum green is $(4 + 2n)$ seconds, where n is the number of vehicles that can be potentially stored within the length x. This equation assumes a startup delay of 4 s and 2 s discharge headway between vehicles. In the case of long loop presence detection, the minimum green can be close to zero if the loop is placed near the stop line. However, in cases where pedestrians are permitted to cross the intersection concurrently with a green display the duration of the minimum green is controlled by pedestrian crossing time (see Section 4.6.4).

The maximum green interval can be expressed in terms of a limit to vehicle extensions as shown in Fig. 4.6.8. It is typically set between 30 and 60 s based on analysis. Some agencies specify the maximum green to be approximately 1.5 times the required green as computed for pretimed signals (see Section 4.6.4). One reason that a maximum green interval is specified is to avoid excessive delays and queues to vehicles calling for service on conflicting phases.

As described earlier, the duration of the minimum green interval is set to allow vehicles queued between the detector and the stop line to be serviced. In the case of high-speed approaches (i.e., above 35 mi/h) with point detection, this distance x can potentially store a large number of vehicles, implying a lengthy minimum green. Consequently green time would be wasted whenever a smaller number of vehicles actually accumulate during red. In such cases a *volume-density* mode of actuation is usually employed. This mode has two special features. First, it employs detectors that are capable of counting the number of vehicles arriving during red and, optionally, yellow. Second, the volume-density mode employs an advanced NEMA feature known as the *gap reduction* function.

In the case illustrated in Fig. 4.6.8 the duration of the gap between vehicles that would cause a gap-out is fixed. By contrast, the gap reduction feature provides for a reduced gap between vehicle actuations. At the start of a green phase the gap that can trigger a gap-out is set at its maximum value. After a specified delay following a call for a conflicting phase, this gap begins to decrease with time (see Fig. 4.6.9). The gap is not allowed to decrease below a minimum value. Depending on vehicle demand, a gap-out can occur at the maximum, the minimum, or any of the intermediate levels.

Other controller functions include the provision for pedestrian "Walk" and clearance (i.e., flashing "Don't Walk") displays and several options relating to storing and recalling information relating to detector calls for future use. A feature known as detector (or locking) memory allows the detector to remember a call placed during red even if the actuating vehicle has left the approach during red as it often happens when right-turn-on-red (RTOR) is permitted. Locking memory (sometimes called *memory on*) is employed with point detection because the system has no knowledge of the movements of vehicles after they enter the space between the detector and the stop line. Consequently the controller would provide a green indication at the next opportunity in order to serve any vehicles occupying that space.

Long-loop presence detection can be operated with either locking or nonlocking memory. A common application of long-loop presence detection with *memory off* (i.e., nonlocking memory), is found on exclusive left-turn bays with permitted turns against opposing through traffic during the circular green followed by an exclusive left-turn green arrow (protected turns). With memory off, vehicles that left the detection area during circular green will not be remembered. Thus the exclusive green indication will be displayed only if left-turning vehicles remain at the end of the circular green. Some other common options available in modern controllers include the following:

Vehicle permit. Allows a full vehicle interval (unit extension) after max-out or gap-out to allow the last vehicle to pass.

(Min or Max) *Vehicle recall.* Displays the specified green (min or max) even when there is no demand for it.

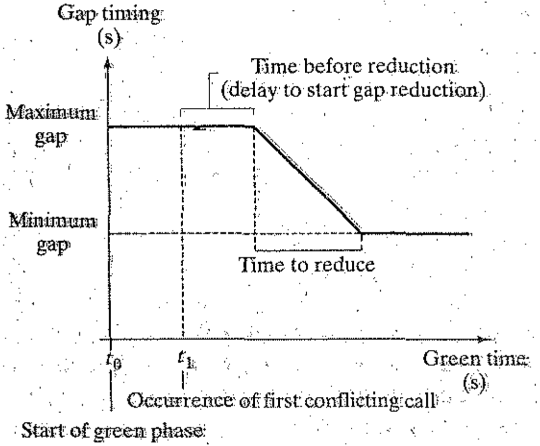


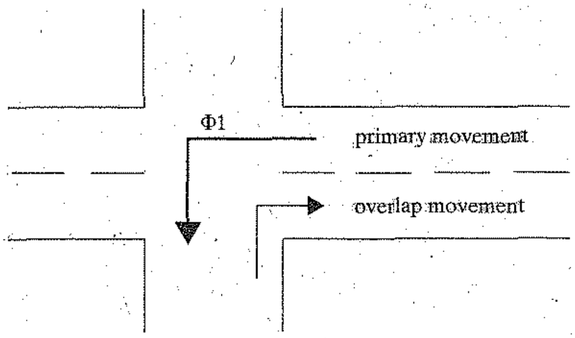
Figure 4.6.9 Example of timing with gap reduction.

Pedestrian recall. Activates the pedestrian phase even when the push button has not been pressed.

Red rest. Instructs the controller to display red (rest in red) if there is no demand for the green indication. This option (or *flag*) cannot be used if one of the vehicle recall options has been set.

Simultaneous gap-out. In a dual-ring controller the two phases (one from each ring) will not terminate unless both either gap-out or max-out. This applies to the two phases that immediately precede the barriers (see Fig. 4.6.4).

Overlaps. NEMA provides for up to four overlapping phases, allowing a specified green phase to be displayed simultaneously with phases across the barrier. The diagram here shows an example where the overlap phase is permitted whenever Φ1 is displayed. The overlap phase shown is also known as a *shadowed movement.*



Finally, it should be noted that the two detection techniques described here are very basic. For special situations (particularly to obviate the dilemma-zone problem at high-speed approaches) more complex detection schemes are needed.

4.6.4 Signal Timings

Signal timings describe the set of parameters defining the operation of a signalized intersection (i.e., the sequence and duration of the signal indications for each intersection approach). From an analytical perspective it involves the identification of the sequence by which the various movements at an intersection are served as well as the time duration of service (i.e., green time) for each movement.

The process of identifying the sequence of service is called phasing and it precedes all other signal timing steps. Then the cycle length is estimated and green times are allocated to each phase according to the relative magnitude of traffic flows served in each phase. The latter part also includes the allocation of phase change intervals (yellow and all red). Certain constraints must be checked to ensure the safe and efficient processing of vehicles and pedestrians at an intersection.

Signal phasing. Phasing is the sequence by which the various movements of both vehicles and pedestrians are being served at a signalized intersection. In traffic engineering the definition of phasing is slightly different from the one encountered in the discussion of signal controllers in the previous section. For example, the equivalent of the eight-phase operation in Fig. 4.6.6 is a four-phase scheme (Fig. 4.6.11) because controller phases (usually denoted by numbers, e.g., Φ1, Φ2, etc.), which are executed simultaneously, are taken as one phase (usually denoted by letters, e.g., ΦA, ΦB, etc.) in traffic analysis. This difference should be better understood after reading the rest of this section.

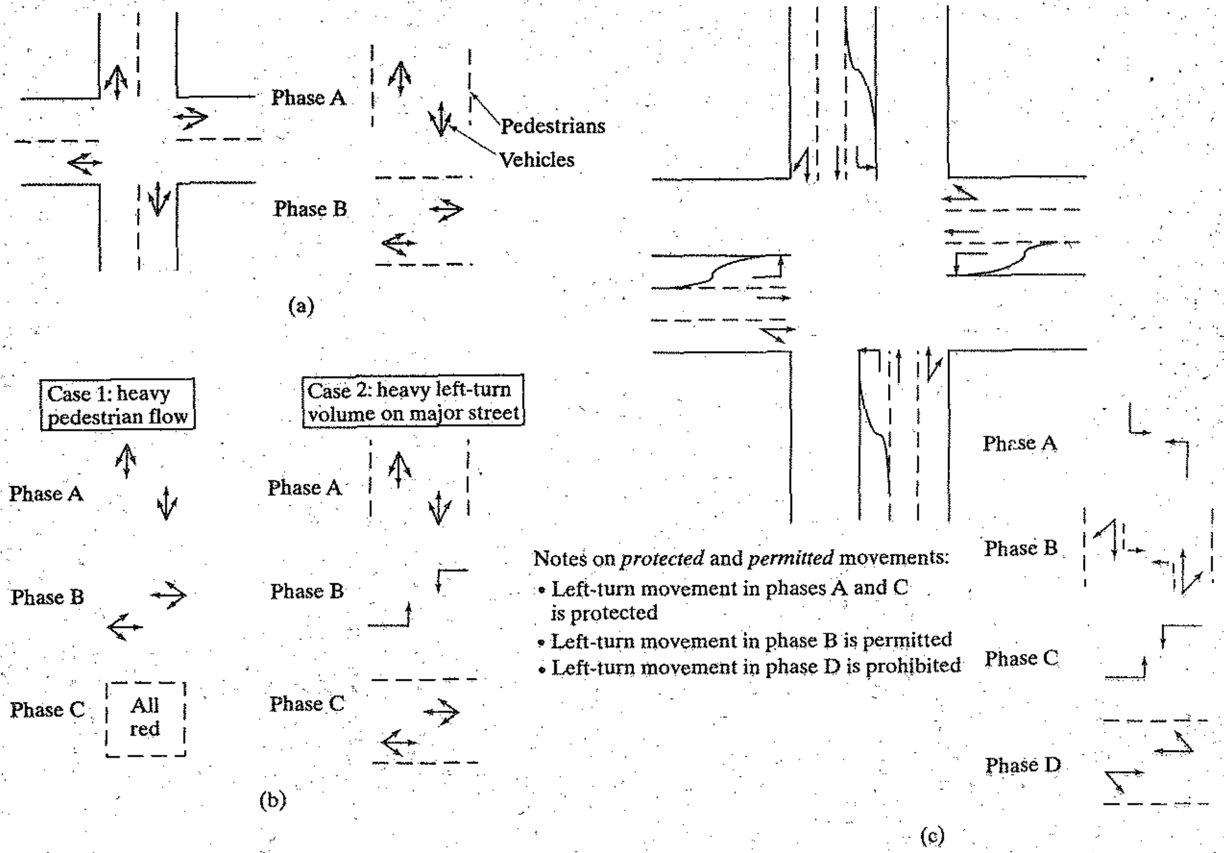
The objective of phasing is the minimization of the potential hazards arising from the conflicts of vehicular and pedestrian movements, while maintaining the efficiency of flow through the intersection. A large number of phases may be required if all conflicts are to be eliminated. Typical conflicts are (1) left-turning vehicles conflict with opposing through traffic as well as with pedestrians and (2) right-turning vehicles conflict with pedestrians [see Fig. 4.6.10(a)]. Increasing the number of phases promotes safety but hinders efficiency because it results in increasing delays. Delays increase because (1) start-up lost times increase (i.e., the time between the display of green and the discharge of the first vehicle in queue), (2) phase change intervals increase (i.e., the number of yellow and red clearance intervals required for transition from one phase to the next increase), and (3) minimum phase duration requirements have to be met. These requirements are based on minimum pedestrian crossing times; they are discussed later in this section.

Three common phasing schemes are presented in Fig. 4.6.10. The simplest phasing scheme is a two-phase operation [Fig. 4.6.10(a)]. This operation is appropriate at intersections with low pedestrian volumes, low-to-moderate turning volumes, and vehicle arrivals with an adequate number of sufficiently long gaps that permit left-turning vehicles to be served within the green time allotted to the phase. Right-turning vehicles conflict with pedestrians, whereas left-turning vehicles conflict with both opposing through traffic and pedestrians.

A three-phase operation is appropriate when one of the conditions under a two-phase operation is violated:

1. *High volume of pedestrians* [Fig. 4.6.10(b), case 1]. In this case pedestrians are prohibited to cross when vehicles are served (phases A and B), and an exclusive phase is provided to serve pedestrians (phase C). Phase C is called the all-red phase because all vehicle approaches have a red signal indication.
2. *High left-turning volume on one of the two intersecting streets* [Fig. 4.6.10(b), case 2]. In this case a specific phase (phase B) is allocated to serve left turns on one of the two streets. The left turns served in this fashion are protected (*protected movement*) because they have no conflicts with either vehicles or pedestrians. Left turns may be allowed or disallowed in the next phase (phase C). If they are allowed, they are a *permitted movement*; they can be served only if conditions permit (i.e., if there are long enough gaps in the opposing vehicular and pedestrian flow). The case where left turns are not allowed (after the end of the phase serving the left turns) is usually implemented with an exclusive set of traffic signals with arrows pointing to the left and a sign notifying that "left turn on arrow only" is allowed. High left-turning volumes varying by time (i.e., different directional left-turn overloads during morning and evening peak periods) can be treated with leading or lagging left-turn green allocations. Their discussion is beyond the scope of this document.

If heavy left-turning volumes are present on both intersecting streets, a four-phase operation is preferred [Fig. 4.6.10(c)]. This signalization scheme is most effective when



Notes on *protected* and *permitted* movements:

- Left-turn movement in phases A and C is protected
- Left-turn movement in phase B is permitted
- Left-turn movement in phase D is prohibited

Figure 4.6.10 Examples of (a) two-, (b) three-, and (c) four-phase signal operation. (From Berry, [4.18].)

coupled with left-turning bays or exclusive left-turn lanes and with actuated signal controllers. Left-turn bays or exclusive lanes make the operation of the intersection more efficient by reducing interference with the through movements on each approach. The traffic-actuated controller gives the ability to skip or elongate left-turning phases depending on the presence of low or high demand for left turns in each approach.

Figure 4.6.11 illustrates the operation of a four-phase scheme under actuated control. In the beginning of a new cycle the controller assesses the demand for left turns. If there is sufficient demand in both the east- and westbound directions, it selects the top box for phase A. If there is only eastbound left-turning demand, it selects the box second from the top. Thus, along with the eastbound left turns, it releases the through and right-turning eastbound movements. If there is only westbound left-turning demand, it selects the box second from the bottom. If there is no left-turning demand, the controller skips phase A altogether and proceeds to phase B. Similar decisions are made for the north–south left-turning traffic (phase C). Comparison of Figs. 4.6.6 and 4.6.11 reveals that ΦA is $\Phi 1 + \Phi 5$ or $\Phi 1 + \Phi 6$ or $\Phi 2 + \Phi 5$, ΦB is $\Phi 2 + \Phi 6$, and so on.

Actuated controllers are also able to modify the cycle length as well as the durations of green to better serve the actual demand. In light traffic green durations are kept to a minimum, resulting in a short cycle length. The opposite happens when traffic is heavy. Minimum and maximum phase durations are prespecified by the traffic analyst prior to implementation. Minimum greens are required for safe processing of pedestrians. Maximum greens are required so that the movements which receive correspondingly long reds will not accumulate more vehicles than the length of the block can handle (i.e., queue backups spilling over to adjacent intersections). The flexibility of these systems results in more efficient service of traffic and in the minimization of delays. Actuated controllers perform best at isolated locations (i.e., intersection not a part of a coordinated signal network system) and during off-peak times (i.e., drastic reduction of unnecessary delays).

Cycle-length selection. Cycle length is a complete sequence of signal indications; it is the duration of time in which the whole set of phases at a signalized intersection takes place once. The length of cycle should not be set arbitrarily. Unnecessarily long cycle lengths cause substantial delays, whereas too short cycle lengths may cause congestion or endanger the processing of pedestrians through an intersection.

The appropriate cycle length can be estimated by Webster's formula; it results in the optimal cycle length:

$$C_o = \frac{1.5L + 5}{1 - \text{CS}} \tag{4.6.1}$$

where

C_o = optimal cycle length, in s

L = total lost time during a cycle, which consists of the startup delay minus the portion of yellow utilized by drivers (see Fig. 4.7.3); 3 to 4 s per phase is a good approximation

CS = sum of the flow ratios of critical movements (discussion follows)

Before estimating the cycle length, the phasing at the intersection must be set. There is no technique or computer algorithm that can produce an optimal phasing scheme other than a

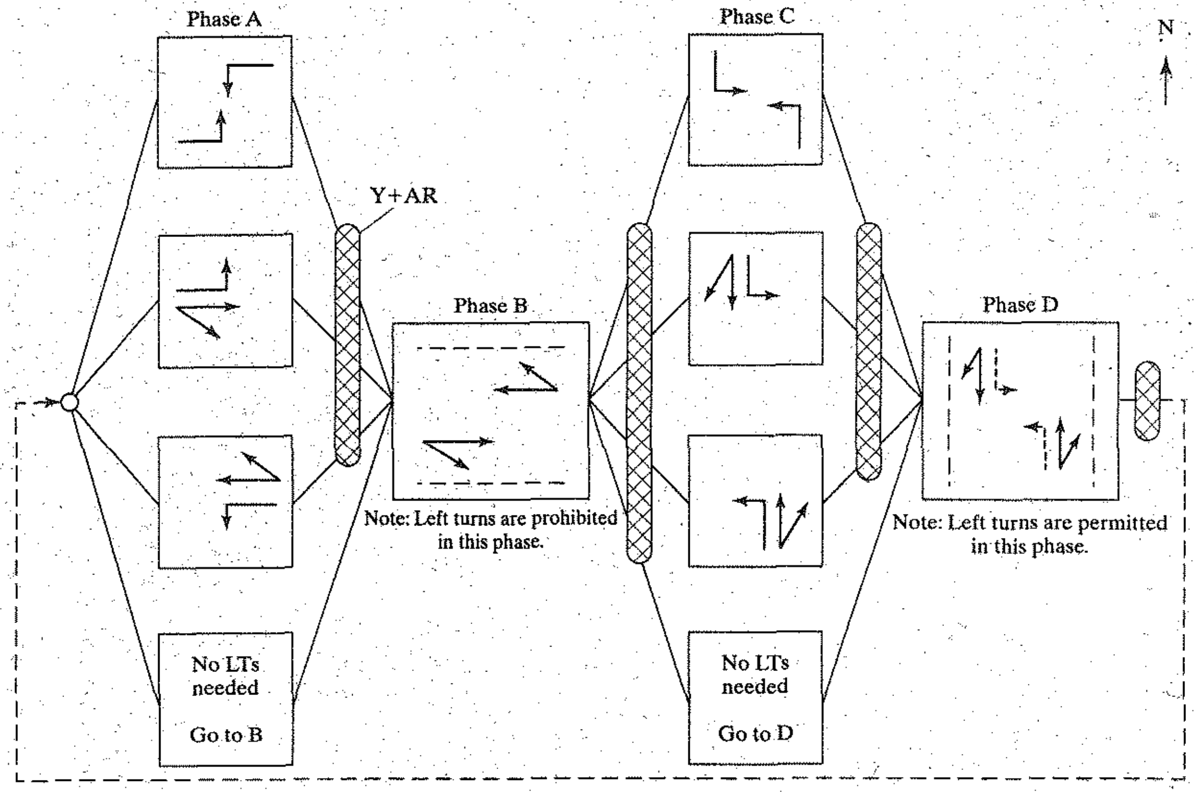
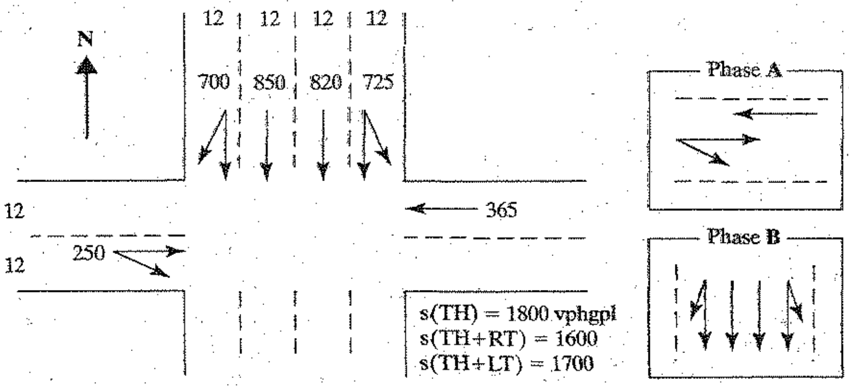

Phase A
Phase B
Phase C
Phase D
N
Y+AR
Note: Left turns are prohibited in this phase.
Note: Left turns are permitted in this phase.
No LTs needed
Go to B
No LTs needed
Go to D

Figure 4.6.11 Typical four-phase operation at an intersection with actuated signal controller.

tedious analysis of many combinations of phasing schemes and lane channelization options on each approach. Usually reliance on common sense, experience, and trial and error are the tools for identifying a phasing scheme. Ultimately the best phasing scheme is the one that coupled with an optimal cycle length results in the shortest delays for the vehicles using the intersection.

Flow ratio is the demand over the servicing rate, or in traffic engineering terminology, the volume over the saturation flow (flow ratio = *v/s*). The saturation flow is discussed in detail in Section 4.6.2. Meanwhile, it will be treated as a given property.

Critical movements are discussed in the following example of signal timings estimation. Figure 4.6.12 displays an intersection that is assumed to be unsignalized. Due to the high peak-hour volumes, signalized control is warranted. The north-south direction is one way (southbound only). It has four lanes. The rightmost and leftmost lanes operate in a shared-use fashion (i.e., through as well as turning traffic share the use of those lanes). The



ESTIMATE THE SUM OF CRITICAL FLOW RATIOS

$$\text{Phase A: } \max\left\{\frac{250}{1600}, \frac{365}{1800}\right\} = \max\{0.147, 0.203\} = 0.203$$

$$\text{Phase B: } \max\left\{\frac{700}{1600}, \frac{850+820}{2\times 1800}, \frac{725}{1700}\right\} = \max\{0.438, 0.464, 0.426\} = 0.464$$

$$\text{CS} = 0.203 + 0.464 = 0.667$$

LOST TIME

Assuming lost time is 4 s per phase, the total lost time per cycle is: L = 2 × 4 = 8 s

OPTIMAL CYCLE LENGTH

$$C_o = \frac{1.5 \times 8 + 5}{1 - 0.667} = 51 \text{ s}$$

Figure 4.6.12 Estimation of optimal cycle length.

east-west traffic can go both ways. This is a small street with only one lane per direction and low traffic volumes. All lanes are 12 ft wide.

First the phasing scheme is established; a two-phase operation seems appropriate; dashed lines represent the simultaneous serving of pedestrians. Then the critical movement in each phase is identified. It corresponds to the lane or groups of lanes with the highest flow ratio. For phase A the critical movement is the westbound traffic, whereas for phase B the critical movement is the through-southbound movement (both lanes are combined since they serve the same movement). Factor CS results by simply adding up the maximum flow ratios. For a two-phase operation two maximum flow ratios should be identified and added up (one per phase). For a three-phase operation, three maximum flow ratios should be identified and added up, and so forth.

The cycle length is estimated to be equal to 51 s. Empirical results show that cycle lengths within a ±30% from the optimal cycle-length estimate from Webster's formula are still performing nearly optimally [4.18, 4.19]. Therefore all cycle lengths between 35 and 65 s are likely to perform satisfactorily at this intersection. According to the Highway Capacity Manual, cycle lengths usually vary between 60 and 120s; under very congested conditions they may reach 150s.

Green allocation. After the cycle length has been estimated green and clearance times must be allocated to each phase. The proper way to identify the duration of clearance is to go through the dilemma-zone calculations (Section 2.3.2). The dilemma-zone analysis may yield different durations of clearance due to higher speeds or different crossing lengths (i.e., in the intersection examined, the east-west traffic needs to cross four lanes to clear the intersection, whereas the southbound traffic needs to cross only two lanes). Variable clearance durations can easily be accommodated in the green allocation procedure. Clearance includes the time for both yellow and clearance red (see case study 1 later in this chapter).

The top table in Fig. 4.6.13 presents the green time allocation. Green time is allocated proportionally to the critical flow ratios for each phase. Note that the total clearance

Phase	Cycle length	Y+AR	Available cycle length	Flow ratio	Critical sum	Allocation	Green	G_p	Ped. check
A	51	5	42	0.203	0.667	30.4%	12.8	14.0	Not ok
B		4		0.464		69.6%	29.2	9.0	ok

↓

Increase the cycle length in steps of 1 s until the pedestrian requirement is met.

↓

Phase	Cycle length	Y+AR	Available cycle length	Flow ratio	Critical sum	Allocation	Green	G_p	Ped. check
A	55	5	46	0.203	0.667	30.4%	14.0	14.0	ok
B		4		0.464		69.6%	32.0	9.0	ok

Figure 4.6.13 Green time allocation.

duration is subtracted from the cycle length for the allocation. The total of green times and clearance times should be equal to the cycle length. For this example we find that yellow (Y) is 3/s for each phase and clearance red (AR) is 2s for ΦA and 1s for ΦB.

Before accepting these timings as final, a check of whether pedestrians can be safely served by the allotted times is necessary. Pedestrians served during phase A must cross four lanes. According to 1997 and earlier editions of the HCM the minimum time required* is

$$G_p = 7 + \frac{W}{4} - Y = 7 + \frac{4 \times 12}{4} - 5 = 14 \text{ s} \tag{4.6.2}$$

where

W = width of the crossing, in ft

Y = total clearance interval time, in s

The green time allotted to phase A cannot pass the minimum pedestrian time requirement, whereas the green time allotted to phase B is sufficient for safe pedestrian crossing: A minimum of 9 s is needed and 29.2 s are available.

Increasing the cycle length to 55 s (which is still in the near-optimal cycle length range) results in the green time allocation presented in the bottom table of Fig. 4.6.13. The final green times are now acceptable. Figure 4.6.14 presents the signalization stripes corresponding to this example application.

Another constraint imposed on cycle length selection is the networkwide cycle length. According to the Manual on Uniform Traffic Control Devices (MUTCD), if the studied intersection is 0.5 mi away from the nearest signalized intersection, it may be considered as isolated (i.e., not in a network); therefore the preceding constraint does not apply. If the distance between neighboring intersections is shorter than 0.5 mi, the studied intersection belongs to a network, and for efficiency reasons progression of platoons of vehicles should be maintained (arterial progression is discussed in Section 4.6.6). Effective progression cannot be achieved unless all intersections in the network operate under the same cycle length. This may impose a severe constraint in selecting an optimal cycle length for a specific intersection. Some flexibility exists, however. Progression can be achieved if some intersections operate in half-cycle or double-cycle. Thus if the network cycle length is 80 s, some intersections may operate in a 40- or 160-s cycle length.

*A different formula for G_p has been proposed for HCM 2000. For pedestrian crossings with a width that does not exceed 18 ft, the minimum green time for a given phase is:

$$G_p = 3.2 + \frac{L}{4} + 0.27\,N_{\text{ped}}$$

L = length of crosswalk in ft, equal to the W in Eq. 4.6.2

N_{ped} = number of pedestrians using the crosswalk during the phase

3.2 = pedestrian start-up time

This formula is less helpful in planning and design applications; it does not consider the safety buffer afforded by Y+AR and introduces a start-up time that may vary widely among locales (busy downtown vs. rural town), and users (hurried vs. inattentive individuals).

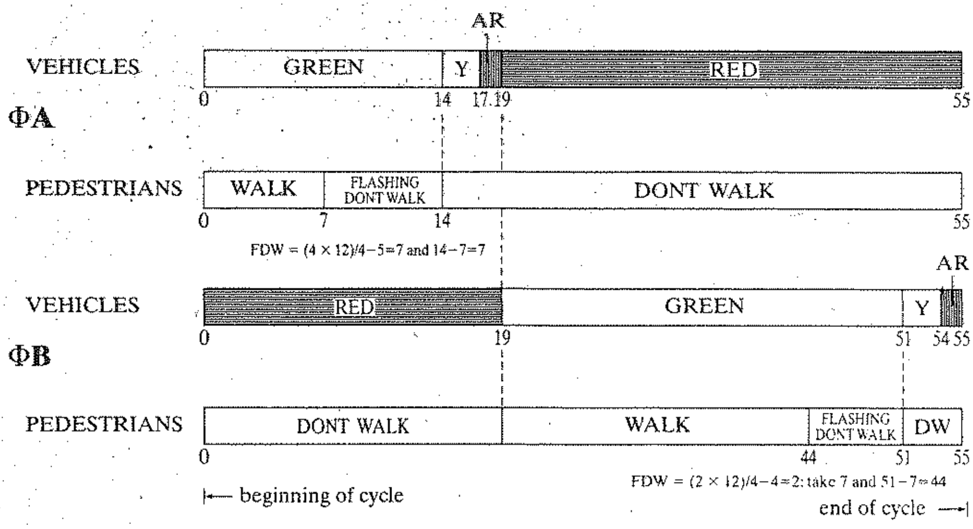


Figure 4.6.14 Example of signalization stripes for vehicular and pedestrian traffic.

Network analysis packages such as TRANSYT [4.13] can help to identify an optimal networkwide cycle length given the phasing, traffic loads, and saturation flows at each intersection as well as other network characteristics, such as distances between intersections and speeds.

A previously unsignalized intersection may have been intimidating (perceived as unsafe) or inconvenient (long delays incurred) for some drivers. The emplacement of signals may alter these perceptions, and therefore new traffic may divert to the newly signalized facility. Hence a few months after the installation new traffic counts must be obtained and new timings should be estimated if the net volumes or the directional distribution of flows has changed. This may not be necessary if actuated signalization has been installed because of its inherent ability to accommodate fluctuating demands, provided that minimum and maximum phase durations have been set appropriately.

4.6.5 Time-Distance Diagram of Interrupted Flow

Figure 4.6.15 shows an idealized time-distance diagram for an interrupted traffic stream. The signal is stationary and its display changes over time. A total of 12 vehicles is shown. At time $t = 0$ vehicle 1 is stopped by the red light. Vehicles 2 to 8 consecutively join the stopped platoon. Line AB represents the shock wave between the approach conditions and the stationary-platoon conditions (see Section 3.6). After an *initial* (or *startup*) *delay* due mostly to the first driver's perception-reaction following the onset of green, the platoon leader moves through the intersection. Subsequent vehicles follow at a shorter release headway. Line CB represents the shock wave at the front of the platoon between the jam and the release conditions. When the two shock waves meet, the stationary platoon is totally dissipated. In the case shown this event occurs before the onset of the following red, so that vehicles 9, 10, and 11 are able to clear the intersection without interruption. Finally, vehicle

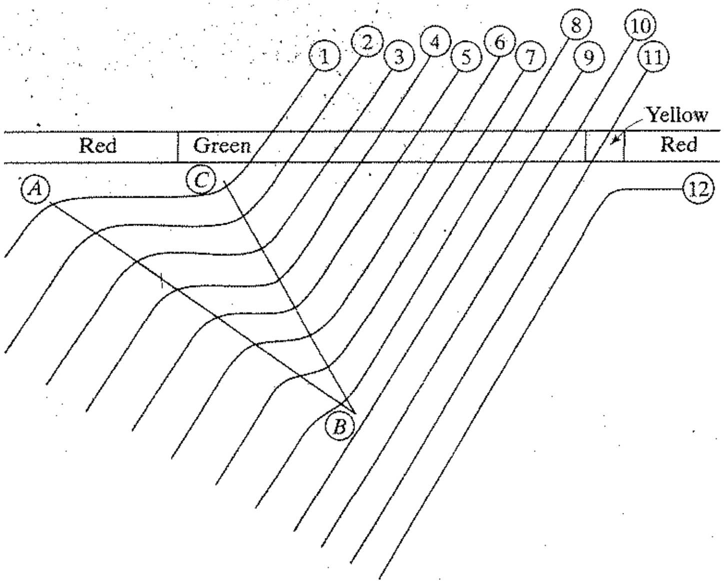


Figure 4.6.15 Traffic interruption.

12 is obliged to stop for the next red display. As explained in Chapter 3, a third shock wave between the approach and the release conditions may begin when the stationary platoon disappears, but, for simplicity, this shock wave is not shown in Fig. 4.6.15.

4.6.6 Pretimed Signal Coordination

The fact that certain vehicles can avoid stopping at an intersection presents the opportunity to coordinate a series of signals to allow platoons of vehicles to clear all the signals without interruption [4.20]. This scheme works when the signals being coordinated have the same cycle length or multiples of a common cycle length but not necessarily the same distribution of green, yellow, and red within the common cycle. Figure 4.6.16 shows a system of four intersections, three of which are signalized. The relative timing of each signal is specified by its *offset,* which is the time difference between a reference time and the beginning of the *first complete* green phase thereafter. The two pairs of parallel lines drawn on the figure represent the constant speed trajectories of the first and last vehicles in each direction that can clear all intersections without stopping. The time difference between the parallel trajectories in each direction of movement is known as the *through band* for the direction. Dividing the through band by the average vehicular headway gives the number of vehicles constituting the uninterrupted platoon. The *width of the through band,* measured in seconds, may be adjusted by "sliding" each signal diagram horizontally. A *balanced design* refers to the case when the through bands in the two directions of travel are equal. A balanced design, however, does not always represent the best design. For instance, a *preferential design* may be more appropriate during the morning or evening peak periods on streets with unbalanced directional flows.

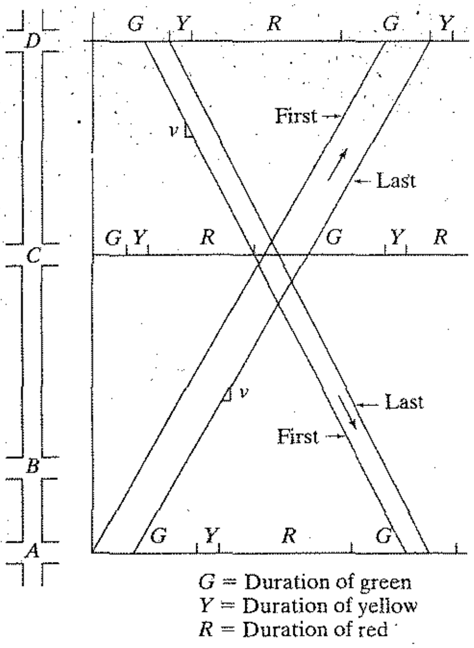


Figure 4.6.16 Pretimed signal coordination.

The solution to a signal coordination problem may be accomplished graphically, analytically, or by computer, using several simple equations. For example, the time it takes a vehicle to travel between intersections at a constant speed equals the distance traveled divided by the speed. Also, the following equation may be used to discern the status of a signal at any time $t = T$ after the reference time $t = 0$:

$$\text{Time into cycle} = \text{remainder of } \frac{\{T - \text{offset}\}}{C} \tag{4.6.3}$$

Knowledge of the duration of the green, yellow, and red displays can pinpoint the exact status of the signal at $t = T$.

Example 4.6

A signal has an offset of 10 s, a green, $G = 50$ s, a yellow, $Y = 5$ s, and a red, $R = 65$ s. Find the status of the signal at times (a) $t = 45$ s, (b) 150 s, (c) 720 s, and (d) 782 s.

Solution The signal cycle $C = (G + Y + R) = 120$ s. Apply Eq. 4.6.3.

(a) For $t = 45$: $(45 - 10)/120 = 0$ remainder 35; 35 s into the $(0 + 1)$ (i.e., first) cycle Since $35 < G$, the display is green.

(b) For $t = 150$: $(150 - 10)/120 = 1$ remainder 20; 20 s into the second cycle. Since $20 < G$, the display is also green.

(c) For $t = 720$: $(720 - 10)/120 = 5$ remainder 110: 110 s into sixth cycle. Since $(G + Y) < 110 < C$, the display is red.

(d) For $t = 782$: $(782 - 10)/120 = 6$ remainder 52; 52 s into the seventh cycle. Since $G < 52 < (G + Y)$, the display is yellow.

Example 4.7

The signals at the intersections of the one-way street have been pretimed and coordinated as follows:

Intersection	Green	Yellow	Red	Offset	Distance from A
A	40 s	5 s	35 s	0 s	—
B	50 s	5 s	25 s	40 s	2000 ft
C	35 s	5 s	40 s	10 s	5000 ft

Given a design speed of 30 mi/h, determine the width of the resulting through band.

Solution The three signals have equal cycles of 80 s. Therefore signal coordination is possible. A vehicle clearing intersection A at time $t = 0$ will arrive at B (2000 ft)/(44 ft/s) = 45.5 s later. At this instant the display at B is 5.5 s into the first green display. Hence the vehicle can proceed toward C without interruption. It will reach C at $t = 45.5 + (3000/44) = 113.7$ s. At this time the signal at C is 23.7 s into the second cycle and is green. Therefore, this vehicle can clear all three intersections.

To find the last vehicle that can do the same, the remaining green duration after the passage of the first vehicle is calculated (see Fig. 4.6.17):

At A: $40 - 0 = 40$ s

At B: $90 - 45.5 = 44.5$ s or $50 - 5.5 = 44.5$ s

At C: $125 - 113.7 = 11.3$ s or $35 - 23.7 = 11.3$ s

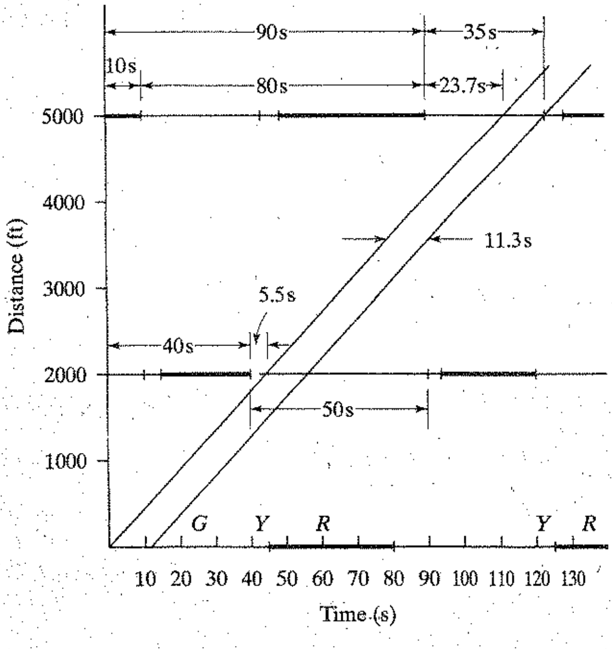


Figure 4.6.17 Example of bandwidth derivation.

The minimum of these values defines the width of the through band (i.e., 11.3 s) and fixes the trajectory of the last vehicle as shown in the figure.

Discussion Often the width of the through band is taken to include the yellow; that is, the last vehiclc is allowed to clear an intersection on yellow. In this event the width of the through band would be reported as 16.3 s. Note that in this example the width of the through band can be increased by increasing the offset of intersection C. As a general rule, allowing a few seconds of green to elapse before the first vehicle in the main platoon reaches a signal is considered good practice because it allows any main- or side-street vehicles caught by the preceding red phase to clear the way before the platoon's arrival.

4.6.7 Actuated Signal Coordination

Signal coordination is also possible with demand-actuated controls. In this case it is important to ensure that a through band is maintained, that is, not destroyed by the response to demand from the side streets. This is accomplished by allowing side-street traffic to be served with green only during limited *permissive periods* within the cycle. In addition, the requirement of cycle lengths that are equal to whole number multiples of a base cycle length must be adhered to. The requirement is accomplished by terminating noncoordinated phases at specific points within the selected cycle length. Such terminations are known as *force-offs.*

4.7 CAPACITY OF SIGNALIZED INTERSECTIONS

4.7.1 Background

Under uninterrupted conditions the definition of flow is simply the number of vehicles that pass a point during a specified time interval. At intersections a unique point where flow measurements can be taken does not exist. Figure 4.7.1 shows the variety of movement desires that a typical four-leg intersection is expected to accommodate and the resulting points of conflict between these movements. The ability of a signalized intersection to process the approaching flows is affected by the magnitudes and vehicular composition of these volumes, their movement desires, the geometric design of the intersection, and the characteristics of the signal. The presence of bus stops in the immediate vicinity of the intersection also affects its operating conditions. Several practical methods of signalized intersection analysis are found in the technical literature. Most of these methods are empirical and approximate. Hence their ability to explain the many subtleties encountered at intersections is, accordingly, limited.

The HCM 2000 explicitly recognizes two limitations of the methodology:

- Inability to account for congestion effects propagating from locations downstream the approaches of the subject intersection
- Inability to account for congestion effects on through lanes caused by the overfilling of left-turn lanes or bays

Given that truly isolated signalized intersections are rare, traffic projects focused on a single intersection are atypical, and computer-based analysis is affordable and expedient instead of individual intersection capacity analysis, a network analysis of two or more signalized

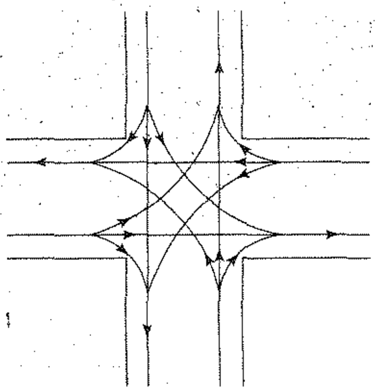


Figure 4.7.1 Intersection movement desires.

intersections using one or several traffic simulation models (some of which incorporate the HCM's delay models) may be a more prudent and cost-effective course of action. Several traffic simulation models are presented in Chapter 15.

4.7.2 Capacity and Performance Analysis

The operating characteristics of signalized intersections can be estimated and evaluated with a procedure of capacity and performance analysis. The capacity of an intersection represents the throughput of the facility (i.e., the maximum number of vehicles that can be served in 1 h). An important outcome of the capacity analysis is the volume-to-capacity ratio (*V/c* ratio), which is also called the degree of saturation (*X*). This ratio indicates the proportion of the capacity (supply) utilized by the existing traffic volume (demand).

The performance of an intersection is based on estimates of average delay for each vehicle utilizing the facility. Short delays result in a good level of service (LOS), whereas long delays result in poor LOS (e.g., average delay equal to 5 s per vehicle corresponds to LOS A, whereas 40 s per vehicle corresponds to LOS D). Facilities performing at D or worse may need to be upgraded (i.e., improvements in signal timings and progression, rechannelization or widening of the road space). Intersection performance is discussed in detail toward the end of this section.

The 2000 *Highway Capacity Manual* (HCM) [4.3] procedures for analysis of signalized intersections are widely accepted in traffic engineering practice. Its predecessors (HCM-1965, TRB Circular 212, HCM-1985, HCM-1994, and 1997 updates) are still used by some. Among other differences, the pre–1985 procedures do not require the estimation of delays. Levels of service are derived directly from V/C ratios.

A summary of the HCM 2000 procedures for analysis of signalized intersections is presented next. A substantial number of inputs is required for the application of these procedures. The inputs may be classified in five categories: (1) traffic characteristics, such as volumes by direction (i.e., through, right and left for each intersection approach, also referred to as *turning movements*); (2) traffic composition, such as proportion of heavy vehicles in traffic by lane or approach; (3) geometric characteristics, such as the number of lanes, lane widths, approach grades; (4) signal timing characteristics, such as type of control: pretimed, actuated, cycle length, duration of greens, phase-change intervals; and (5) other operating characteristics, such as arterial progression, existence of parking and frequency of parking maneuvers, and bus stop blockage. Right-turns-on-red (RTOR) volumes may be excluded from the subsequent analysis if they have been collected in the field.

After all inputs have been gathered, *traffic volumes are adjusted* to reflect peak-period conditions. This is done by multiplying peak hourly volumes by the peak-hour factor (PHF), which is defined as follows:

$$\text{PHF} = \frac{\text{peak-hour volume}}{4(\text{peak 15-min volume})} \tag{4.7.1}$$

Usually traffic volumes are recorded every 15 min. Based on the counts, first the peak hour is identified: For example, traffic volume counts taken between 4:00 and 6:00 P.M. may indicate that the peak hour is between 4:30 and 5:30; this is the peak-hour volume to be used in Eq. 4.7.1. Then the highest peak 15-min count is selected from the peak hour; this is the peak 15-min volume to be used in Eq. 4.7.1. Each intersection approach may have different peak characteristics (i.e., each may reach its peak at a different time of day).

The final step in the adjustment of volumes is the grouping of directional flows in lane groups based on the utilization of each lane (i.e., types of movements utilizing a lane) and the phasing scheme. For example, a two-lane approach on a two- by two-way street intersection (Fig. 4.7.2) may be analyzed in one of three possible ways: (1) one-lane group serving all three movements (usually selected when turning flows are low); (2) two-lane

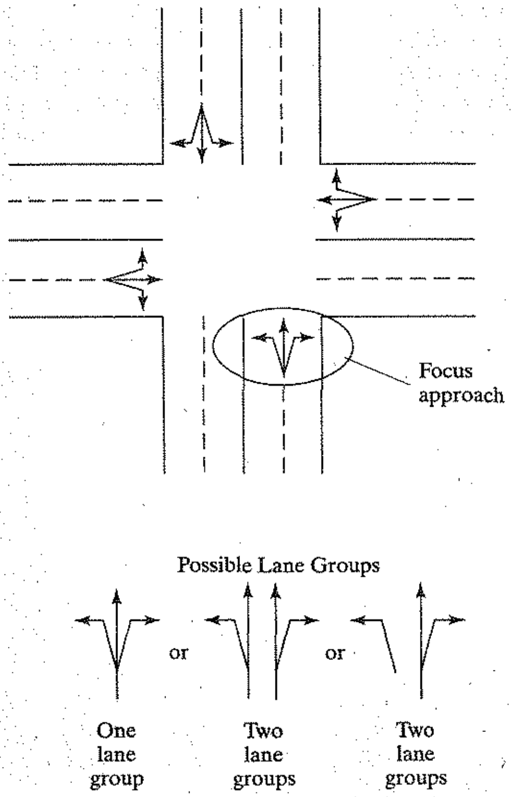


Figure 4.7.2 Example of possible groupings of movements into lane groups.

groups: one serving left-turning and through movements and the other serving right-turning and through movements (usually selected when moderately high turning volumes prevail); and (3) two-lane groups: one lane serving the left-turn movement only and one serving the through and right-turning movements; this happens when left-turn volume is high, so that left-turning traffic essentially occupies the left lane at all times (de facto left-turn lane) or when there exists an exclusive left-turn phase in the signalization plan.

The second step in the analysis is the *estimation of prevailing saturation flows* for each lane group. The saturation flow describes the (behavioral) way of driver discharge from an intersection stop line. The saturation flow is essentially the service rate: It represents the maximum number of vehicles that can be served in 1 h, assuming a continuous display of green and a continuous queue of vehicles. The saturation flow is expressed in vphg (vehicles per hour of green).

Figure 4.7.3 presents the concept of saturation flow. Assume an intersection approach with one lane that has an infinite number of cars waiting in queue. An x–y plot is employed to represent the discharge pattern of drivers. The signal is red. At time t_0 green is displayed. There is a typical reaction and action time delay (i.e., react to signal and then shift into gear and press the gas pedal), and at time t_1 the first car in queue crosses the stop line. Each car crossing the stop line is considered discharged.

Approximately after the fourth car the discharge rate becomes rapid and fairly uniform: Cars pass by at a fairly constant rate and the headways between them are almost equal. The saturation point has been reached. This saturation discharge rate is the highest attainable under normal conditions; this is the saturation flow. Normally the discharge process may not be as uniform, but the plot in Fig. 4.7.3 depicts reality reasonably well.

At the end of green, yellow is displayed (at time t_2). Some drivers proceed to clear the intersection and others stop. There are occasions when t_3 falls after t_4, which denotes

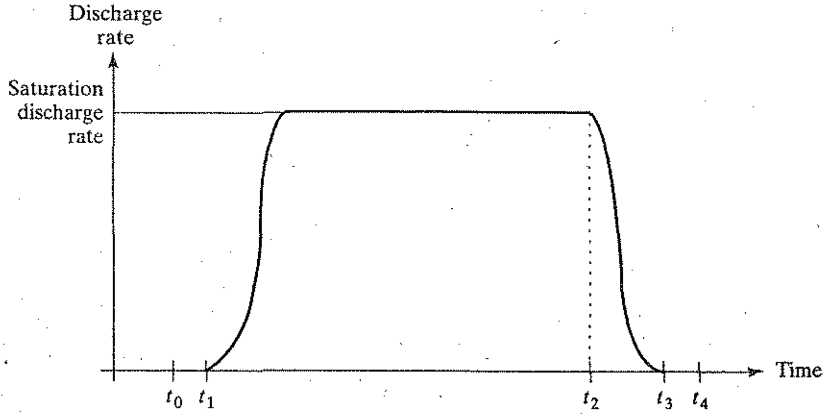


Figure 4.7.3 Concept of saturation flow, t_0, beginning of green; t_1, first vehicle crosses stop line; t_1–t_0, startup delay; t_2, beginning of yellow (end of green); t_3, last vehicle to discharge during this cycle; t_3–t_2, yellow utilization; t_4, end of yellow (beginning of red); t_2–t_1, signal green time; t_3–t_1, effective green time.

the beginning of red. This means that some drivers utilized a small part of the red phase, often mistakenly. These drivers are called "sneakers" in traffic engineering jargon. Substantial presence of sneakers calls for both enforcement of traffic laws and reevaluation of the duration of the yellow and red clearance interval (dilemma-zone problem).

The major determinant of the saturation flow is the average headway between vehicles discharging from an intersection. The saturation flow (s) is defined as

$$s = \frac{3600}{\bar{h}} \tag{4.7.2}$$

where $\bar{h}$ is the average headway in seconds.

Since s describes driver behavior, vehicle characteristics such as size and acceleration characteristics, traffic conditions, and environmental factors as well as driving habits affect the saturation flow. Thus the HCM recommends the local data collection and derivation of ideal saturation flow levels. Studies have shown that saturation flows are higher in suburban areas [4.21], and lower in small urban areas [4.22] or under adverse weather conditions.

For intersection capacity analysis a base saturation flow is selected first; s_0 is usually equal to 1,900 pcphgpl (passenger cars per hour of green per lane). The base saturation flow represents driver behavior in large U.S. urban areas on facilities with specific geometric and operational characteristics. Then s_0 is adjusted to reflect actual conditions. Adjustments are not made for environmental conditions; daylight and dry pavement conditions are always assumed. The prevailing saturation flow for a specific lane group is estimated as follows:

$$s = s_0 \cdot N \cdot f_w \cdot f_{\mathrm{HV}} \cdot f_g \cdot f_P \cdot f_{bb} \cdot f_a \cdot f_{\mathrm{LU}} \cdot f_{\mathrm{RT}} \cdot f_{\mathrm{LT}} \cdot f_{pb} \tag{4.7.3}$$

where

Number of lanes (N) is the number of lanes serving the lane group.

Lane width (f_w), in ft. Drivers tend to feel more comfortable on wider lanes (i.e., less interference with vehicles in adjacent lanes and less lateral displacement, as in Section 2.3.4) resulting in higher saturation flows. The value for this factor is 1 for 12-ft wide lanes (base condition). Equation 4.7.4 produces estimates for f_w for other lane widths. W is the average lane width in feet. It is suggested that two lanes are considered for widths exceeding 16 ft.

$$f_w = 1 + \frac{W - 12}{30} \tag{4.7.4}$$

Heavy vehicles (f_{HV}). Heavy vehicles typically accelerate slowly, which slows the discharging process. As a result, headways are elongated and the saturation flow decreases. In the absence of heavy vehicles this factor is equal to 1 (base condition). Equation 4.7.5 produces estimates for f_{HV} for the prevailing share of heavy vehicles. In it % HV is the percent of heavy vehicles and $E_T = 2.0$ is the passenger-car equivalency of the average heavy vehicle.

$$f_{\mathrm{HV}} = \frac{100}{100 + \%\,\mathrm{HV}(E_T - 1)} \tag{4.7.5}$$

Grade (f_g). Uphill grade tends to decrease acceleration, thus headways elongate and saturation flow decreases. The opposite is true for downhill grades. On level terrain this factor is equal to 1 (base condition). Equation 4.7.6 produces estimates for f_g for the prevailing slope. In it $\%\,G$ is the percent grade:

$$f_g = 1 - \frac{\%\,G}{200} \tag{4.7.6}$$

Parking (f_P). Parking adjacent to traffic lanes tends to interfere with the flow of traffic and parking maneuvers disrupt the normal discharge process. The impact of parking is greater when fewer lanes are available. The number of parking maneuvers also is related to saturation flow: the more parking maneuvers per hour, the lower the saturation flow. In the absence of curb parking this factor is equal to 1 (base condition). Equation 4.7.7 produces estimates for f_P based on prevailing conditions. In it N_m is the number of parking maneuvers per hour.

$$f_P = \frac{N - 0.1 - \dfrac{18 \cdot N_m}{3600}}{N} \tag{4.7.7}$$

Bus blockage (f_{bb}). Transit buses often stop at intersection corners to serve passengers. This usually disrupts intersection operations. One lane may be temporarily blocked, during green, or following vehicles may have to slow down and maneuver around the stopped bus, which caused a temporary decrease of the saturation flow. The impact of bus blockage is greater when fewer lanes are available. In the absence of bus stops this factor is equal to 1 (base condition). Equation 4.7.8 produces estimates for f_{bb} based on the number of buses stopping in 1 h. In it N_B is the number of buses stopping per hour.

$$f_{bb} = \frac{N - \dfrac{144 \cdot N_B}{3600}}{N} \tag{4.7.8}$$

Area type (f_a). The type of the area surrounding the intersection, CBD or non-CBD, has an impact on driving behavior and consequently on the saturation flow. CBD-like locales exhibit very high interference due to pedestrians, parking, delivery vehicles, and so on. Besides city centers, CBD-like conditions can be found in campuses, at urban beachfronts, on a few blocks of a suburban main street, and so on. At those locations $f_a = 0.9$. For all other locales $f_a = 1$ (base condition).

Lane utilization (f_{LU}). On occasion the distribution of traffic on multilane approaches is uneven (e.g., a subset of lanes leads to the freeway or to a large activity center). A downward adjustment of the saturation flow is required because space on some lanes is underutilized. In the absence of unequal traffic distribution this factor is equal to 1 (base condition). Equation 4.7.9 produces estimates for f_{LU} based on the distribution of demand on the busiest lane for the lane group. In it, V_g is the unadjusted demand volume for the lane group and V_{gl} is the unadjusted demand volume for the single lane that carries the highest volume.

$$f_{\text{LU}} = \frac{V_g}{N \cdot V_{gl}} \tag{4.7.9}$$

Right turns (f_{RT}). This adjustment reflects the required slowing of a vehicle in order to negotiate the right-turn curve. In the absence of right turns this factor is equal to 1 (base condition). Equations 4.7.10, 4.7.11, and 4.7.12 produce estimates for f_{RT} based on prevailing operating conditions and the proportion of right turns in the lane group.

$$\text{Exclusive lane:} \quad f_{RT} = 0.85 \tag{4.7.10}$$

$$\text{Shared lane:} \quad f_{RT} = 1 - 0.15 \cdot P_{RT} \tag{4.7.11}$$

$$\text{Single lane:} \quad f_{RT} = 0.9 - 0.135 \cdot P_{RT} \tag{4.7.12}$$

where P_{RT} is the proportion of right turns in the lane group.

Left-turn movement (f_{LT}). This adjustment reflects the required slowing of a vehicle in order to negotiate the left-turn curve. In the absence of left turns this factor is equal to 1 (base condition). Equations 4.7.13 and 4.7.14 produce estimates for f_{LT} based on prevailing operating conditions and the proportion of left turns in the lane group. These equations apply to protected phasing only. A complex analysis is needed for permitted and protected-permitted operations; it can be found in an appendix of HCM 2000.

$$\text{Exclusive lane:} \quad f_{LT} = 0.95 \tag{4.7.13}$$

$$\text{Shared lane:} \quad F_{LT} = 1/(1 + 0.05 \cdot P_{LT}) \tag{4.7.14}$$

where P_{LT} is the proportion of left turns in the lane group.

Pedestrian and bicycle (f_{pb}). Both left- and right-turn movements may conflict with pedestrians and bicyclists. If such conflicts are absent, then this factor is equal to 1 (base condition). Otherwise it is derived by Eqs. 4.7.15 and 4.7.16 for right and left turns, respectively.

$$f_{Rpb} = 1.0 - P_{RT}(1 - A_{pbT})(1 - P_{RTA}) \tag{4.7.15}$$

$$f_{Lpb} = 1.0 - P_{LT}(1 - A_{pbT})(1 - P_{LTA}) \tag{4.7.16}$$

where

P_{RTA} (P_{LTA}) = proportion of right (left) turns under protected green

A_{pbT} = permitted phase adjustment. In the case of pedestrian-only conflicts, $A_{pbT} = 1 - fV_{ped} \div 2000$ with $f = 0.6$ or 1. If the number of turning lanes is the same as the receiving lanes, then $f = 0.6$; if the number of turning lanes is smaller than the number of receiving lanes, then $f = 1$.

The adjusted volumes and the derived saturation flows for each lane group are combined in the *capacity analysis*. The capacity for each lane group is estimated by

$$c_i = s_i \cdot \frac{g_i}{C} \tag{4.7.17}$$

where

c_i = capacity of lane group i, in vehicles per lane (veh/l)

s_i = prevailing saturation flow of lane group i

g_i = green time allotted to lane group i

C = cycle length

The degree of saturation is estimated as follows:

$$X_i = \frac{V_i}{c_i} \tag{4.7.18}$$

where

X_i = degree of saturation of lane group i

V_i = adjusted peak volume of lane group i

c_i = capacity of lane group i

In order to be able to derive a degree of saturation for the entire intersection, critical movements must be identified for each phase. If more than one lane group is served in one phase, then the lane group with the highest flow ratio $(V/s)_i$ is the critical one. The process of critical movement selection is identical to the one presented in the cycle-length estimation process. The critical degree of saturation (X_c) for the entire intersection is estimated as follows (recall the critical sum, CS, from Eq. 4.6.1):

$$X_c = \Sigma\,(V/s)_{\text{critical}} \cdot C\ /\ (C - L) = \text{CS} \cdot C\ /\ (C - L) \tag{4.7.19}$$

where L is the total lost time during a cycle; it includes the startup lost time and the unutilized portion of the phase change interval (Y + AR). $L = 3s$ per phase is usually taken.

During busy periods X_c is useful in assessing signal timings problems at intersections, as follows:

If all $X_i \leq 1$ and $X_c \leq 1$, then the signal timings are adequate (but not necessarily good or optimal).

If one or more $X_i > 1$ and $X_c \leq 1$, then the cycle length is adequate, but the green allocation is incorrect.

If regardless of X_i $X_c > 1$, then the cycle length is too short.

If several $X_i > 1$ and $X_c > 1$, then another (simpler) phasing may yield improvement. If this fails, and $C > 150$ s, then the ability of the signal to handle the traffic demand has been practically exhausted. Measures such as lane addition and left-turn movement deletion are needed.

HCM 2000 suggests that for planning applications X_{cm} is used instead of X_c. X_{cm} is estimated with Eq. 4.7.19 except that the maximum cycle length is used instead of the prevailing cycle length. Planning analysis is described in the next section.

The last step in the intersection capacity analysis process is the *performance evaluation* of the facility. The performance evaluation is based on the delay incurred by all vehicles utilizing the facility. There are several definitions of delay. Two of the most common ones are travel delay and stopped delay. *Travel delay* for an individual vehicle is the difference between the time a vehicle passes a point downstream of the intersection where it has regained normal speed and the time it would have passed that point had it been able to continue at its approach speed [4.18]. *Stopped delay* for an individual vehicle is the time duration of "substantially standing still" while waiting in queue at a signalized intersection approach. Substantially standing still is usually taken equal to 3 mi/h or less. Empirical results show that division of the total delay by 1.3 results in the stopped delay. The 1985 and 1994 editions of the HCM utilize stopped delay. The 1997 and 2000 editions of the HCM utilize control delay, which is similar to the travel delay.

Exhibit 4.7.1 shows a part of the evolution in delay estimation formulas, including the fundamental equation of Webster and a 1970s version of the Australian Road Research Board (ARRB) delay equation followed by the 1985 and 1994 HCM equations. The HCM 2000 equation is shown at the bottom of the exhibit (Eq. 4.7.1e); it is the most complex to date and consists of three components:

d_1 = uniform delay, which is attributed to the signal; that is, part of the time the signal is red. This delay increases or decreases depending on the quality of progression.

d_2 = overflow delay caused by influxes of demand that cannot clear the intersection in one green. Such delay may also be caused by the signal controller giving priority to special classes of vehicles (e.g., emergency, light rail, etc.).

d_3 = delay due to a queue that exists prior to the period of analysis studied (e.g., residual demand from earlier cycle failures)

Most factors in the delay formulas of Exhibit 4.7.1 have already been explained (e.g., c, C, g, s, V, X). The remaining ones are explained here.

T = time period of analysis; usually 0.25 or 1 h

k = factor depending on the controller setting of seconds of unit extension (UE). Specifically

$$k = (1 - 2 \cdot k_{\min})(X - 0.5) + k_{\min} \tag{4.7.20}$$

For example $k_{\min} = 0.11$ for UE = 3.0 s. Note that for $X = 1$, $k = 0.5$. The same value of k applies to pretimed intersections. Refer to the HCM for other values.

I = the effect of metered arrivals due to an upstream restrictive signalized intersection

Q_b = unmet demand at the beginning of period T

t = portion of T during which demand exceeded capacity

DF = delay factor accounting for the type of signal controller and the quality of progression (see table in 1994 HCM for values)

PF = progression factor representing the quality of signal coordination. It is estimated as follows:

$$PF = \frac{(1 - P)f_{\text{PA}}}{1 - \dfrac{g}{C}} \tag{4.7.21}$$

where

P = proportion of vehicles arriving during green

f_{PA} = supplemental adjustment factor; it is equal to 1 for random arrivals. For other conditions consult the HCM for the appropriate value.

The amount of estimated delay defines the level of service of a lane group, an approach, and the intersection as a whole. Approach delays result by weighting lane group delays with the respective lane group volumes. The intersection delay results by weighting approach delays with the respective approach volumes (weighted average). The following correspondence between level of service (LOS) and control delay is specified in the HCM 2000:

Delay (s/veh)	Level of service
≤ 10	A
> 10–20	B
> 20–35	C
> 35–55	D
> 55–80	E
> 80	F

Section 4.7.4 presents two comprehensive case studies of signalized intersections. They include signal timings derivation as well as capacity and performance analysis. Another example of intersection analysis is given as part of a traffic impact study in Section 9.2.6.

4.7.3 Planning Analysis

The planning analysis is usually applied in the case of an intersection that does not exist (e.g., a connection to a planned subdivision) or of an unsignalized facility close to a future development site. Given the absence of detailed and precise data under these circumstances, a streamlined process that can produce an adequate geometric and signal design is used. The key output of the planning analysis is the X_{cm}. If the planning analysis of the proposed intersection design yields $X_{cm} \leq 0.85$, the design is deemed "under capacity," and thus adequate. If X_{cm} is between 0.85 and 0.95, the design is deemed "near capacity"; between 0.95 and 1.00 is deemed "at capacity" and above 1.00 is deemed "over capacity." Planning analysis based on HCM 1997 is performed in four sequential steps, as follows:

Step 1: Determination of the volumes for each movement. This is usually done through a trip generation and distribution process described in Chapters 8 and 9. After the volumes are known they need to be assigned on each lane. Adjustments for curb parking as well as exclusive or shared left turns are provided for.

Step 2: Decision on the type of left-turn operation: protected, permitted, or combination of both.

Step 3: Selection of a phasing plan from six basic plans.

Step 4: Estimation of the critical flow for each phase and the X_{cm} for the intersection.

Optional steps include the estimation of a basic signal timing plan and the estimation of left-turn bay lengths, if applicable. Case study 3 in the next section provides a comprehensive numerical example of a planning analysis.

$$d_{\mathrm{WEBSTER}} = \frac{C \cdot \left(1 - \frac{g}{C}\right)^2}{2 \cdot \left(1 - \frac{V}{s}\right)} + \frac{X^2}{2 \cdot V \cdot (1 - X)} - 0.65 \sqrt[3]{\frac{c}{V^2}} \cdot X^{2+5\,(g/C)} \tag{4.7.1a}$$

$$d_{\mathrm{AARB}} = c\,\frac{T}{4}X\left[(X - 1) + \sqrt{(X - 1)^2 + 12\,\frac{X - X_0}{cT}}\,\right] \text{ with } X_0 = 0.67 + \frac{sg}{600} \tag{4.7.1b}$$

$$d = 0.38 \cdot \frac{C \cdot \left(1 - \frac{g}{C}\right)^2}{1 - X \cdot \frac{g}{C}} + 173 \cdot X^2 \cdot \left[(X - 1) + \sqrt{(X - 1)^2 + 16\,\frac{X}{c}}\,\right]$$

$$d_{1985}^{HCM} = \mathrm{PF} \cdot d \tag{4.7.1c}$$

$$d_1 = 0.38 \cdot \frac{C \cdot \left(1 - \frac{g}{C}\right)^2}{1 - \frac{g}{C} \cdot \min\{X, 1.0\}}$$

$$d_2 = 173 \cdot X^2 \cdot \left[(X - 1) + \sqrt{(X - 1)^2 + m \cdot \frac{X}{c}}\,\right]$$

$$d_{1994}^{HCM} = d_1 \cdot \mathrm{DF} + d_2 \tag{4.7.1d}$$

$$d_{2000}^{HCM} = d_1 \cdot \mathrm{PF} + d_2 + d_3$$

$$d_1 = 0.5\,\frac{C \cdot \left(1 - \frac{g}{C}\right)^2}{1 - \frac{g}{C} \cdot \min\{X, 1.0\}}$$

$$d_2 = 900\,T\left[(X - 1) + \sqrt{(X - 1)^2 + \frac{8kIX}{cT}}\,\right]$$

$$d_3 = \frac{1800\,Q_b\,(1 + u)t}{cT} \quad \text{with} \quad u = 1 - \frac{cT}{Q_b}\,(1 - \min\{X, 1.0\})$$

$$\text{for } t \geqslant T, \text{ else } u = 0 \tag{4.7.1e}$$

Exhibit 4.7.1

4.7.4 Case Studies

This section presents three single-intersection analysis case studies in considerable detail. The first case (in Section 4.7.4.1) is a typical major-minor street intersection in a downtown area. Narrow lanes, bus stops, parking, and left-turn prohibitions are present. The objective is the derivation of a detailed (pretimed) signal timing plan and estimation of capacity and performance of the facility. The second case (in Section 4.7.4.2) presents a complex intersection with a five-phase signal timing containing both overlapping phases and approach-exclusive phasing. The objective is the estimation of an optimal timing plan and the expected performance of the facility. The third case (in Section 4.7.4.3) presents an application of a planning-level analysis of an intersection that presently does not exist.

4.7.4.1 Simple Signalized Intersection

Figure 4.7.4.1 presents the intersection of Date Street and Dole Street at the downtown location of a city with about 150,000 inhabitants. Date Street is a narrow arterial street

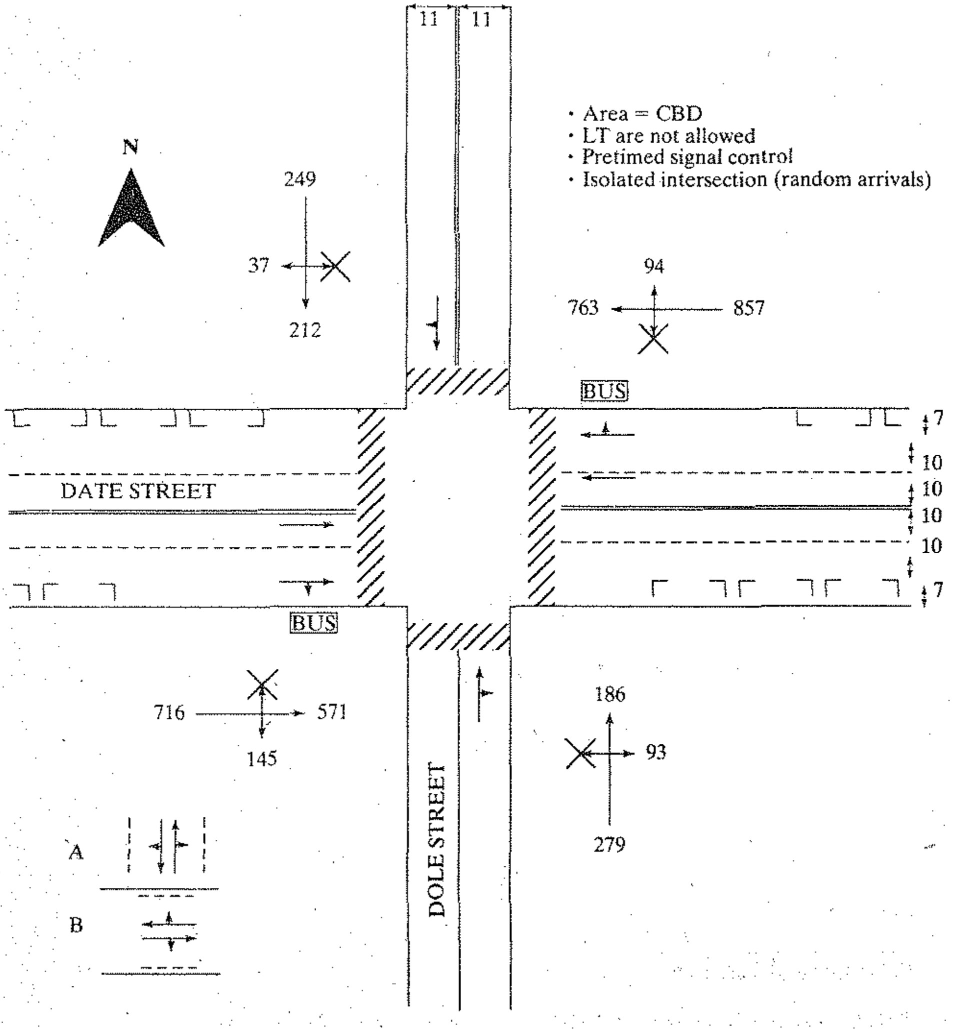


Figure 4.7.4.1 Sample intersection for basic signal timings and capacity analysis.

and Dole Street is a secondary street with considerable volume during the peak periods. Table 4.7.4.1(a) summarizes the field data. Volumes per direction are given along with the proportion of right-turning traffic. Left turns are banned from all directions at this intersection. Some notable conditions include a 3.6% slope in the north-south direction (northbound is uphill), the presence of parking and bus stops on the arterial street, as well as a considerable volume of pedestrians.

Given these conditions, the prevailing saturation flows are estimated in Table 4.7.4.1(b). The table simply replicates all the components of Eq. 4.7.3. The last column is the product of all columns from column (s_o) to column (pb). Since both V and s are now known, the flow ratio (V/s) can be estimated. The critical ratios can be selected by comparing those movements executed in the same phase and selecting the larger one. Specifically both NB and SB traffic moves during phase A. The NB traffic has the larger flow ratio (0.211 versus 0.176); 0.211 is selected. The same occurs for phase B.

Prior to the estimation of the cycle length and the green allocation, the proper duration of the yellow and overlapping red (all red) must be estimated. The *Traffic Engineering Handbook* [4.20] of the Institute of Transportation Engineers utilized Eq. 2.3.7 for this purpose. The equation consists of three components. The first two, when added up, constitute the amount of yellow time and the third component constitutes the all red.* The generalized denominator of the second component may be written as $2a \pm 2Gg$, which accounts for the grade of the approach. The handbook makes the following assumptions for the estimation of Y + AR: $\delta = 1.0$, $a = 10$, $G = 32$, and $L = 20$. W is defined as the distance from the stop line to the far edge of the conflicting traffic. For this case study the distance was determined in the field as 55 ft and 30 ft for the north-south and east-west direction, respectively.

At the particular location a speed study was conducted and the 85th percentile of prevailing speeds is shown in Table 4.7.4.1(c), in mi/h. Speed was converted into feet per second (ft/s) for use in Eq. 2.3.7. At this point all required data are available and Y and AR can be estimated separately and for each individual approach. Since both north and south movements have the right-of-way simultaneously, the end of their right-of-way should also occur simultaneously. As a result, the largest Y (3.8 s in this case) and AR (1.7 s in this case) for the NB and SB approaches are selected to form the final Y + AR for the north-south phase change interval (5.5 s in this case). A similar process applies to the east-west direction.

The estimation of the cycle length is done next assuming a lost time $L = 4$ s per phase. Webster's formula (Eq. 4.6.1) yields a cycle length equal to 37.9 s. Initial greens are obtained by simple proportioning of the available green (cycle = 37.9 minus total Y + AR = 10.4 s gives 27.5 s) based on the critical flow ratios. In this case we observe that the resultant green times do not satisfy the pedestrian green time requirement (G_p), so the cycle is manually increased in 1-s intervals until the requirement is satisfied. This occurs when cycle length reaches 51 s.

At this point all inputs for capacity analysis are available and Table 4.7.4.1(d) is developed using the HCM 2000 default value of 0.9 for the peak-hour factor. All approaches are undersaturated and all PHF · X products are below 1. This is a requirement for valid delay estimates.

*These definitions for Y and AR are not universally accepted. Section 2.3.2 presents the underlying principle of dilemma-zone avoidance in the determination of the yellow and clearance intervals, as well as a discussion on common allocations of time between Y and AR.

TABLE 4.7.4.1(a) Field Data

	Approach	Movmt.	V	% HV	Width	Slope	Park	Bus/hr	Peds.	% turn
1	NB	TH + RT	279	2	11	3.6%	N	N	140	33
2	SB	TH + RT	249	3.5	11	−3.6%	N	N	120	15
3	EB	TH + RT	716	8	10	0%	30	25	75	20
4	WB	TH + RT	857	5	10	0%	30	18	90	17

Note: Pedestrians: 140 on the north side crosswalk, 120 on the south side crosswalk, and so on.

TABLE 4.7.4.1(b) Saturation Flows and Flow Ratios

	Approach	Movmt.	s_0	N	w	a	HV	g	P	bb	RT	LT	pb	s	V/s	Crit?
1	NB	TH + RT	1900	1	0.967	0.9	0.980	0.982	1	1	0.855	1	0.993	1351	0.206	1
2	SB	TH + RT	1900	1	0.967	0.9	0.966	1.018	1	1	0.880	1	0.992	1419	0.175	0
3	EB	TH + RT	1900	2	0.933	0.9	0.926	1	0.875	0.950	0.970	1	0.982	2339	0.306	0
4	WB	TH + RT	1900	2	0.933	0.9	0.952	1	0.875	0.964	0.967	1	0.979	2428	0.353	1

TABLE 4.7.4.1(c) Signal Timings

							Speed							Final		
	Approach	Movmt.	δ	a	G	g	mph	ft/s	Y	W	L	AR	Y + AR	Y + AR	Y	AR
1	NB	TH + RT	1.0	10	32	0.036	30.1	44.1	3.0	55.0	20	1.7	4.7	5.5	3.8	1.7
2	SB	TH + RT	1.0	10	32	−0.036	33.4	49.0	3.8	55.0	20	1.5	5.3	5.5	3.8	1.7
3	EB	TH + RT	1.0	10	32	0	39.7	58.2	3.9	30.0	20	0.9	4.8	4.9	4.0	0.9
4	WB	TH + RT	1.0	10	32	0	41.0	60.1	4.0	30.0	20	0.8	4.8	4.9	4.0	0.9

Phase	V/s crit.	Avail. time	Initial g	G_p	Check	Next g	Check	Final g
A	0.206	28.2	10.4	15.0	not	15.0	ok	15.0
B	0.353	28.2	17.8	7.6	ok	25.6	ok	25.6
CS =	0.559							
C_0 =	38.6		38			51		51.0

TABLE 4.7.4.1(d) Capacity Analysis

	Approach	Movmt.	V	PHF	V_a	g	C	g/C	c	$X = V/c$	PHF·X
1	NB	TH + RT	279	0.90	310	15.0	51	0.29	397	0.78	0.70
2	SB	TH + RT	249	0.90	277	15.0	51	0.29	417	0.66	0.60
3	EB	TH + RT	716	0.90	796	25.6	51	0.50	1175	0.68	0.61
4	WB	TH + RT	857	0.90	952	25.6	51	0.50	1220	0.78	0.70

for $L = 8$ ⟶ $X_c = 66.3\%$

TABLE 4.7.4.1(e) Level of Service

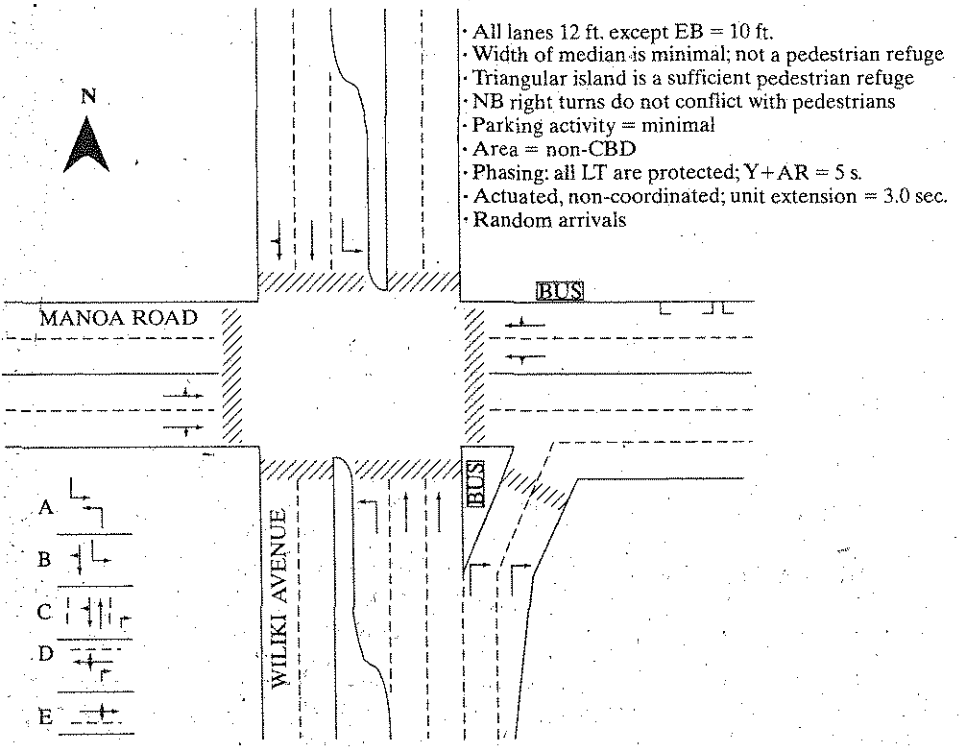
									Approach		Intersection	
	Approach	Movmt.	D_1	k	D_2	PF	Delay	LOS	Delay	LOS	Delay	LOS
1	NB	TH + RT	16.5	0.50	14.1	1.0	30.6	C	30.6	C		
2	SB	TH + RT	15.8	0.50	8.1	1.0	23.9	C	23.9	C	17.5	B
3	EB	TH + RT	9.6	0.50	3.1	1.0	12.7	B	12.7	B		
4	WB	TH + RT	10.4	0.50	5.0	1.0	15.4	B	15.4	B		

Delay and level-of-service (LOS) estimates are derived in Table 4.7.4.1(e). Default values of $k = 0.50$ (pretimed controller) and PF = 1.0 (isolated intersection less the absence of signal coordination) are used. Overall the intersection is operating at LOS B.

4.7.4.2 Complex Signalized Intersection

Figure 4.7.4.2 presents the intersection of Wiliki Avenue and Manoa Road at a suburban location of a city with about 750,000 inhabitants. Wiliki Avenue is a major regional arterial street and Manoa Road is a busy collector street which leads to a college campus with more than 12,000 students. All movements are allowed at this intersection and left-turn lanes are present on the north-south arterial. Table 4.7.4.2(a) summarizes the field data. Volumes per direction and movement (as applicable) are given for each 15-min period during the morning peak hour. The proportion of turning traffic and heavy vehicles also are given. Notable conditions include a 3% slope in the north-south direction (northbound is uphill), the presence of parking on one approach and bus stops on two approaches as well as a heavy volume of pedestrians. The intersection was heavily congested during the time of the data collection, but flow was not obstructed by any nearby bottlenecks.

Given these conditions, the prevailing saturation flows are estimated in Table 4.7.4.2(b). The table simply replicates all the components of formula 4.7.3. The last column is the product of all columns from column (s_0) to column (pb). Since both V and s are now known,



- All lanes 12 ft. except EB = 10 ft.
- Width of median is minimal; not a pedestrian refuge
- Triangular island is a sufficient pedestrian refuge
- NB right turns do not conflict with pedestrians
- Parking activity = minimal
- Area = non-CBD
- Phasing: all LT are protected; Y+AR = 5 s.
- Actuated, non-coordinated; unit extension = 3.0 sec.
- Random arrivals

Figure 4.7.4.2 Sample intersection for complex signal timings and capacity analysis.

TABLE 4.7.4.2(a) Field Data

	Approach	Movmt.	15-min volumes Q_1	Q_2	Q_3	Q_4	% HV	Slope	Park	Bus/h	Peds.	% turn	V
1		LT	38	42	39	37	0					100	156
2	NB	TH	124	137	141	133	4	3%	N	10	225	0	535
3		RT	223	208	187	195	2.5					100	813
4	SB	LT	55	58	47	49	0.5	−3%	N	N	153	100	209
5		TH + RT	188	192	166	147	2					4	693
6	EB	TH + RT	66	73	68	70	2	0%	N	N	57	85	277
7		TH + LT	37	33	51	26	3					75	147
8	WB	TH + RT	74	78	69	66	1	0%	Y	10	144	92	287
9		TH + LT	81	97	84	75	0					27	337

Note: Pedestrians: 225 on the north side crosswalk, 153 on the south side crosswalk, and so on.

TABLE 4.7.4.2(b) Saturation Flows and Flow Ratios

	Approach	Movmt.	s_0	N	w	a	HV	g	P	bb	RT	LT	pb	s	V/s	Crit?
1		LT	1900	1	1	1	1	0.985	1	1	1	0.950	1	1778	0.088	1
2	NB	TH	1900	2	1	1	0.962	0.985	1	0.980	1	1	1	3527	0.152	0
3		RT	1900	2	1	1	0.976	0.985	1	1	0.743	1	1	2713	0.300	0
4	SB	LT	1900	1	1	1	0.995	1.015	1	1	1	0.950	1	1823	0.115	1
5		TH + RT	1900	2	1	1	0.980	1.015	1	1	0.991	1	0.987	3699	0.187	1
6	EB	TH + RT	1900	1	0.933	1	0.980	1	1	1	0.849	1	0.986	1456	0.190	1
7		TH + LT	1900	1	0.933	1	0.971	1	1	1	1	0.950	1	1636	0.090	0
8	WB	TH + RT	1900	1	1	1	0.990	1	0.900	0.960	0.799	1	0.980	1272	0.226	1
9		TH + LT	1900	1	1	1	1	1	0.900	1	1	0.950	1	1625	0.207	0

TABLE 4.7.4.2(c) Signal Timings

	V/s							
Phase	Initial	Adj. 1	Adj. 2	Avail. time	Initial g	G_p	Check	Final g
A	0.088	0.088	0.088	72	9.9	—	N.A.	9.9
B	0.115	0.027	−0.017	72	−1.9	—	N.A.	−1.9
C	0.187	0.160	0.152	72	17.0	14.0	ok	17.0
D	0.226	0.226	0.226	72	25.3	20.0	ok	25.3
E	0.190	0.190	0.190	72	21.4	20.0	ok	21.4
CS =	0.806	0.691	0.638					
C_0 =	180	113	97					97

the flow ratio (V/s) can be estimated. The critical ratios need to be selected carefully because the "select the largest V/c ratio" rule does not apply to phases that serve movements in an overlapping fashion, that is, movements that are served in more than one phase. This is an outcome of the actuated signal controller, which enables the extension of green on approaches having heavier volumes. Let us work out this selection phase-by-phase.

Phase A. It serves the NB and SB left turns. Observe that the SB left turns also are served in phase B. Therefore phase A must serve the NB left turns because they are not served in any other phase. Thus the NB left-turn movement is critical.

Phase B. It is an extension of phase A having an objective to serve the remaining volume of SB left turns. The SB-through movement does not conflict with the SB left turns, so it is

TABLE 4.7.4.2(d) Capacity Analysis

	Approach	Movmt.	V_r	PHF	V_a	g	C	g/C	c	$X = V/c$	PHF·X
1		LT	156	0.93	168	9.9	96.7	0.10	181	0.93	0.86
2	NB	TH	535	0.95	564	17.0	96.7	0.18	622	0.91	0.86
3		RT	813	0.91	892	47.4	96.7	0.49	1330	0.67	0.61
4	SB	LT	209	0.90	232	12.9	96.7	0.13	244	0.95	0.86
5		TH + RT	693	0.90	768	20.1	96.7	0.21	769	1.00	0.90
6	EB	TH + RT	277	0.95	292	21.4	96.7	0.22	322	0.91	0.86
7		TH + LT	147	0.72	204	21.4	96.7	0.22	362	0.56	0.41
8	WB	TH + RT	287	0.92	312	25.3	96.7	0.26	334	0.94	0.86
9		TH + LT	337	0.87	388	25.3	96.7	0.26	426	0.91	0.79

for L = 20 ——————→ $X_c = 80.4\%$

TABLE 4.7.4.2(e) Level of Service

	Approach	Movmt.	D_1	k	D_2	PF	Delay	LOS	Approach delay	LOS	Inters. delay
1		LT	43.0	0.44	46.3	1.0	89.3	F			
2	NB	TH	39.0	0.43	17.1	1.0	56.2	E	40.1	D	
3		RT	18.7	0.24	1.3	1.0	20.0	C			
4	SB	LT	41.6	0.46	44.5	1.0	86.0	F			54.8
5		TH + RT	38.3	0.50	32.0	1.0	70.3	E	73.9	E	delay
6	EB	TH + RT	36.7	0.43	28.0	1.0	64.7	E			
7		TH + LT	33.5	0.16	2.0	1.0	35.6	D	54.6	D	
8	WB	TH + RT	35.7	0.45	32.9	1.0	68.6	E			
9		TH + LT	34.6	0.43	23.5	1.0	58.1	E	62.9	E	

also given green, but it is also served in phase C. For this phase the SB left turn is the critical movement (with a proper adjustment as shown later).

Phase C. It serves the NB and SB-through traffic. The rule "select the largest *V/c* ratio" applies here, but an adjustment is needed, as shown below. The SB-through movement is the critical one.

Phases D and E. They do not contain any overlapping movements, thus the select the largest *V/c* ratio rule applies. It results in the WB-through and right turn as critical for phase D and the EB-through and right turn as critical for phase E.

This intersection has been signalized for several years. The overseeing authority utilizes Y + AR = 5 s for all approaches, which is taken as given for the purposes of this case study. Next the cycle length and green allocation is done in Table 4.7.4.2(c) taking $L = 4$ s per phase.

In Table 4.7.4.2(c), the critical flow ratios, as identified at the last column of Table 4.7.4.2(b), are copied next to phases A through E. Using Webster's formula, an initial estimation of the cycle length is made. The estimate is a huge 180 s. This will be reduced significantly because of two necessary adjustments.

Flow ratio adjustment 1 (adj. 1) pertains to overlapping movements. Specifically:

- SB, LT is served in both phases A and B and is critical for phase B. The magnitude for phase B needs to be reduced by that included in phase A: 0.115 − 0.088 = 0.027.
- SB, TH + RT is served in both phases B and C. The magnitude for phase C needs to be reduced by that included in phase B: 0.187 − 0.027 = 0.160.

The estimate of cycle length after this adjustment is 113 s, which is below the upper bound of large cycle lengths suggested in HCM 2000 (120 to 150 s for severely congested locations.)

Flow ratio adjustment 2 (adj. 2) also pertains to overlapping movements and accounts for the fact that while a movement is expiring and receives Y + AR, the overlapping movement receives a continuous green. Observe this in the phasing diagram in Fig. 4.7.4.2. During the transition from phase A to phase B the NB left turns receive 5 s of Y + AR; at the same time the SB left turns receive 5 s of green, which needs to be subtracted from the allocation. The specific adjustment for phase B is

$$\text{adj. } 2 = 0.027 - 5 \div 113 = -0.017$$

where 5 s is the Y + AR time and 113 s is the cycle length estimate from the previous adjustment. The estimate for adjustment 2 may be positive, equal to zero, or even negative, as in this case. A negative estimate in reality translates to this: In the transition from phase A to phase B, NB left turns receive 5 s of Y + AR. During a few (e.g., 3.5 s) of the 5 s the SB left turns receive green, then the signal turns yellow for the balance (e.g., 1.5 s). So phase B may have a value between any positive number and −4.9 s.

The same adjustment for phase C is more challenging:

$$\text{adj. } 2 = 0.160 - 5 \div 113 = 0.116$$

which is less that the flow ratio for NB, TH (0.152), which in turn becomes critical and is used for the signal timings estimation. The reason is that NB, TH is served exclusively in phase C and requires a proportion of green, which is greater than or equal to 0.152/CS.

The final cycle length is a reasonable 97 s and is sufficient for fulfilling the pedestrian crossing requirements. For example, for phase E

$$\text{Minimum } G_p = 7 + (6 \times 12) \div 4 - 5 = 20.0 < 21.4 \text{ s}$$

Note that left turns and pedestrian crossings never coincide. The final greens are given in the last column of Table 4.7.4.2(c). (*Caution:* These are the greens for each phase.) The greens of overlapping movements must be estimated separately. Specifically the green times for the two overlapping movements mentioned earlier are

$$\begin{aligned} \text{SB, LT} &= 9.9 + 5.0 - 1.9 = 12.9 \text{ s} \\ \text{SB, TH + RT} &= -1.9 + 5.0 + 17.0 = 20.1 \text{ s} \end{aligned}$$

Also, the NB right-turn movement operates in two adjacent phases; its total green time is 17.0 + 5.0 + 25.3 = 47.4 s.

At this point all inputs for capacity analysis are available and Table 4.7.4.2(d) is developed using the peak-hour factor estimates derived by using the four-quarter-hour measurements in Table 4.7.4.2(a). Several approaches are nearly saturated but all PHF · X products are ≤1. This is a requirement for valid delay estimates.

Delay and level-of-service estimates are derived in Table 4.7.4.2(e). Values for k are derived using Eq. 4.7.10 and $k_{\min} = 0.11$. PF = 1.0 is used because this is an isolated intersection (absence of signal coordination). The delay for each approach is estimated as a weighted average. For example, the delay for the NB approach is

$$(89.3 \cdot 156 + 56.2 \cdot 535 + 20.0 \cdot 813) \div (156 + 535 + 813) = 40.1 \text{ s/veh}$$

Overall the intersection is estimated to operate at LOS D (marginally below E), which is consistent with the level of congestion observed in the field.

4.7.4.3 Planning a Signalized Intersection

Figure 4.7.4.3 presents the intersection of New Avenue and Bonsai Road. New Avenue will serve as a collector road for a proposed retail and residential subdivision at a suburban location of a city with about two million inhabitants. Bonsai Road is a secondary arterial street with low-to-moderate traffic. All movements will be allowed at this intersection and left-turn lanes are planned on the proposed north-south collector road. Figure 4.7.4.3 also presents the available data along with typical default values from HCM 1997.

Table 4.7.4.3(a) is a spreadsheet synthesized from the HCM 1997 planning procedure. Step 1(e) presents the HCM-based rule of thumb for selecting protected or permitted

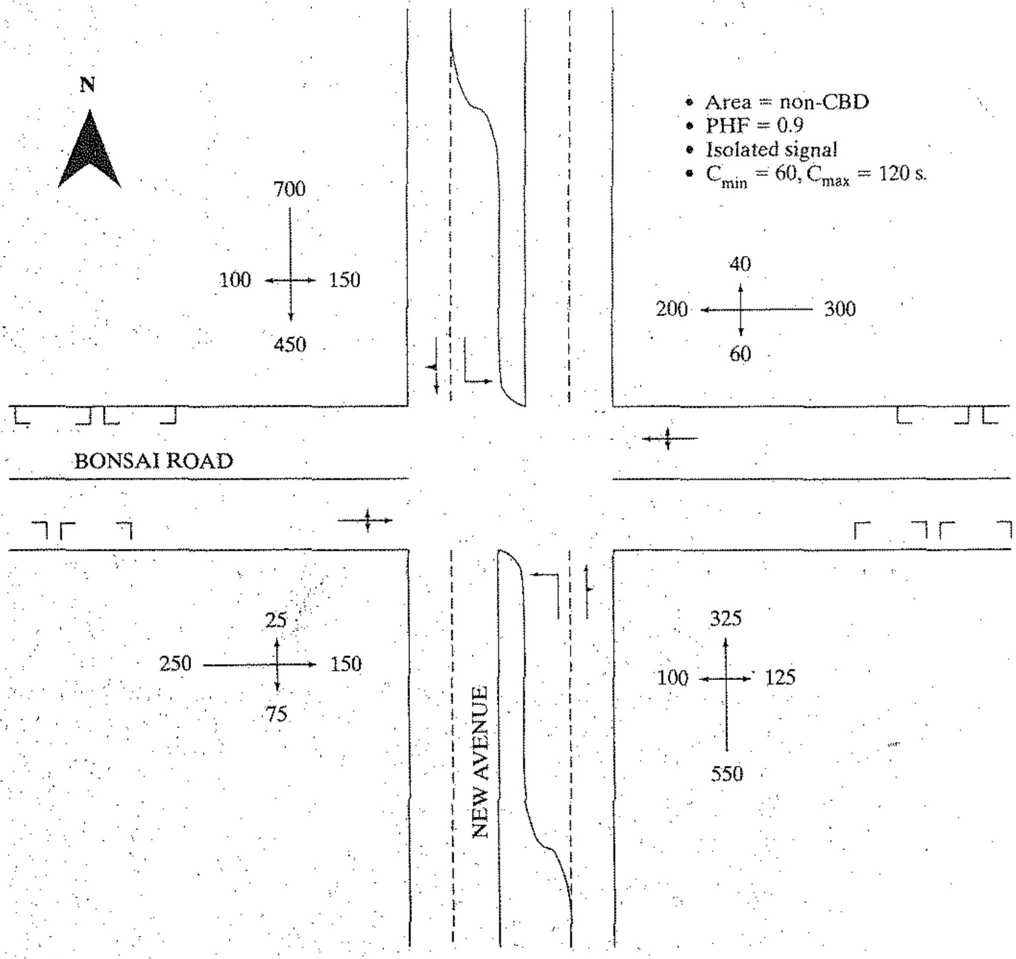


Figure 4.7.4.3 Sample intersection for planning analysis.

TABLE 4.7.4.3(a) Planning Analysis Volume Worksheet

	Step		Procedure	Approach			
				NB	SB	EB	WB
	1	a	LT volume	100	150	25	60
		b	Opposing mainline volume (TH + RT)	550	450	240	225
		c	No. of exclusive LT lanes	1	1	0	0
		d	f_{LT} = 0.95 for single LT, 0.92 for twin LT	0.95	0.95	0.95	0.95
			If LT volume > 90 veh/h, then protected LT is needed if:				
		e	[1a]·[1b] > 55,000 and opposing lanes = 1 [1a]·[1b] > 90,000 and opposing lanes = 2	55,000	67,500	6000	13,500
		f	Specify; protected, permitted, not opposed	PROT.	PROT.	PERM.	PERM.
		g	Adjusted LT lane volume = [1a]/([1c]·[1d]); zero for permitted	105	158	0	0
	2	a	RT volume	125	100	75	40
		b	f_{RT} = 0.85	0.85	0.85	0.85	0.85
		c	Adjusted RT volume = [2a]/[2b]	147	118	88	47
	3	a	TH volume	325	450	150	200
		b	f_p is 0.8, 0.9, and 0.933 for 1, 2, or 3 lanes adjacent to curb parking, respectively, 1.0 otherwise	1	1	0.8	0.8
		c	No. of TH lanes, including shared lanes	1	1	1	1
Exclusive LT	4	a	Approach volume = ([2c] + [3a])/[3b]	472	568	n.a.	n.a.
		b	TH lane volume = [4a]/[3c] = CRITICAL VOLUME	472	568	n.a.	n.a.
Shared LT	4	a′	Approach volume = ([2c] + [3a])/[3b]	n.a.	n.a.	281	300
		b′	f_{LT} (Fig. 9-18, HCM 1997)	n.a.	n.a.	0.717	0.676
		c′	TH lane volume = [4a′]/([3c]·[4b′]) = CRITICAL VOLUME	n.a.	n.a.	392	444
	5	a	Maximum cycle (C_{max})	120			
		b	Permitted LT sneaker capacity = 7,200/[5a]	60			

left-turn signal operation. All rows are self-explanatory and consistent with HCM 1997, except row 4(b′) in which the HCM requires the estimation of a long and detailed worksheet. This may be too burdensome for the planning-level analysis. Considering the overall uncertainty of this type of analysis, a much simpler approximation for the f_{LT} coefficient estimation is needed, such as $f_{LT} = 0.75 - 0.3 \times [1e] \div 55{,}000$, used in this case study.

Once the volume worksheet is completed [Table 4.7.4.3(a)], the analysis proceeds with the signal worksheet. The engineer must select a reasonable phasing scheme that is likely to work well with the planned geometry. For this case study a reasonable plan should take advantage of the planned left-turn bays on the north-south collector. Given the expected volumes, the phasing may be written as follows (using HCM nomenclature):

NSL = NB and SB left turns

STL = SB-through and left turns

NST = NB- and SB-through traffic

EWT = EB- and WB-through traffic (all traffic)

TABLE 4.7.4.3(b) Planning Analysis Signal Operations Worksheet

		North-South			East-West		
Factor		Phase 1 NSL	Phase 2 STL	Phase 3 NST	Phase 1 EWT	Phase 2 —	Phase 3 —
—	Volume 1	105	53	515	392	—	—
—	Volume 2	—	—	472	444	—	—
—	Volume 3	—	—	—	—	—	—
CV	Select critical volume	105	53	515	444	0	0
SCV	Sum of critical volumes	1117					
PL	Lost time per phase	—	—	4	4	—	—
L	Total lost time			8			
CBD	0.9 if in CBD, 1.0 otherwise			1.0			
PHF	Default PHF is 0.9			0.9			
$C_{\max}$	Maximum cycle			120			
Z	$(1\text{-}L/C_{\max})\cdot 1900\cdot\text{CBD}\cdot\text{PHF}$			1596			
X_{cm}	Critical X = SCV/Z			70.0%			
—	Intersection status			under capacity			
RS	1710·CBD·PHF1539						
Ratio	Min(SCV, RS)/RS			0.73			
C'	Cycle length = $L/(1-\text{ratio})$			29.2			
C	Final cycle length			60			
g	(C-L)·(CV/SCV)	4.9	2.5	24.0	20.7	—	—
Check	SUM(g) + L			60.0			
Z_c	$(1\text{-}L/C)\cdot 1900\cdot\text{CBD}\cdot\text{PHF}$			1482			
$X_{c=60}$	Expected critical $X = \text{SCV}/A_C$			75.4%			
—	Expected intersection status			under capacity			

The corresponding volumes are copied from the volume worksheet and the critical volume is selected as the largest for each phase. Lost time is taken equal to 4 s and only for phases without overlapping movements (which tends to produce unrealistically short cycle lengths as shown below). Minimum and maximum cycle lengths are taken as 60 and 120 s, respectively, per HCM 1997. X_{cm} is estimated and a judgment is made whether the design is under, near, at, or over capacity. The desired result is "under capacity" and this design achieves the stated goal.

Preliminary cycle estimation and greens can be estimated as shown in Table 4.7.4.3(b). The cycle estimate of 29.2 s is too small and it is replaced by 60 s. A new Z-factor is estimated and a final X_{cm} is obtained, which also fulfills the stated goal.

A preliminary design of left-turn bay lengths can also be accomplished using a simple graph and table in HCM 1997. Basically, the length of left-turn storage is derived as a function of the passenger-car equivalent volume on the subject left turn movement assuming a fixed cycle length and degree of saturation. A table of adjustment factors is provided for other combinations of X and C. The left-turn requirement may be described as follows:

1.1 ft of storage for each passenger car/hr (pc/h) for the first 150 pc/h, and 0.3 ft of storage for all pc/h in excess of 150 pc/h. These approximate estimates correspond to $X = 0.80$ and $C = 75$ s.

For this case study the derived $X = 0.754$ and $C = 60$ s yield a correction factor (see HCM 1997) of 0.902. Assuming all passenger cars, a 10-year horizon, and a 4% annual growth rate, the left-turn storage requirements are

NB: $100 \times 1.4 = 140$ pc/h and $140 \times 0.902 \times 1.1 = 142$ ft

SB: $150 \times 1.4 = 210$ pc/h and $(150 \times 1.1 + 60 \times 0.3) \times 0.902 \times 1.1 = 182$ ft

4.7.5 Arterial Street LOS and Congestion Quantification

Areawide and corridor analyses require the assessment of the level of service or the quantification of congestion along urban or suburban arterials. Resources for such analyses include the HCM for level of service estimation and NCHRP 398 for the quantification of congestion. This section summarizes the respective procedures after a brief definition of the various classes of arterials according to HCM 2000. These definitions differ somewhat from AASHTO's functional classification of highways and streets illustrated in Table 2.4.1.

Arterials are classified based on their *functional* and *design* category. The functional categorization distinguishes arterials in principal and minor. Unlike AASHTO's definition, principal arterials in the HCM do not include freeways. *Principal* arterials connect important urban centers of activity. *Minor* arterials supplement principal arterials by connecting urban centers to neighborhoods and one neighborhood to another. The design categorization distinguishes arterials to high speed, suburban, intermediate, and urban. A *high-speed* design typically corresponds to a multilane street without parking, exclusive lanes for left turns, absence of curb parking, and fewer than 3 signalized intersections per mile. Shoulders as well as partial separation with medians are often present. The *suburban* design is similar, but it contains more frequent access points and fewer than 5 signalized intersections per mile. An *intermediate* design does not contain medians and shoulders, it may lack left turn bays at some intersections, and there may be sections where curb parking is allowed. Signal density may be as high as 10 signalized intersections per mile. An *urban* design has 6 to 13 signalized intersections per mile, curb parking along most of the length, few exclusive left-turn lanes, and some interference from pedestrians.

Four classes of arterials are recognized based on their functional and design categories [4.3]:

Class I	=	principal and high speed		
Class II	=	principal and suburban	or	minor and suburban
Class III	=	principal and intermediate	or	minor and suburban
Class IV	=	principal and urban	or	minor and intermediate or urban

The level-of-service estimation in the HCM depends on the estimation of the prevailing average travel speed for the time period under analysis. It is recommended that this is established in the field with several runs of vehicles on the through lanes of the subject arterial segment. In the absence of field observations, then, the arterial speed (S_A) is estimated as follows:

$$S_A = \frac{3600 \cdot L}{T_R \cdot L + \Sigma d} \tag{4.7.22}$$

where

S_A = average speed on the subject segment, in mi/h

T_R = total running time per mile of the subject segment, in s; it can be taken from Table 4.7.5.1

L = length of subject segment, in mi

Σd = $\Sigma[d_{1i} \cdot \text{PF}_i + d_{2i} + d_{3i}]$, sum of the control delay for the through movement at all the signalized intersections along the subject segment; $i = 1, \ldots, N$ are the intersections along the subject segment (the formulas for d_1, d_2, and d_3 have been presented earlier)

An important consideration in the analysis of arterials in signal coordination (or arterial progression), which is accounted for by the progression factor PF [see Eq. 4.7.1(e)]. PF for arterial streets is estimated as follows:

$$\text{PF} = \frac{\dfrac{C}{g} - R_p}{\dfrac{C}{g} - 1} \cdot f_{\text{PAG}} \tag{4.7.23}$$

where

g/C = green ratio

f_{PAG} = adjustment for platoon arrival during green; it is taken from Table 4.7.5.2

R_p = platoon ratio taken from Table 4.7.5.2 or estimated as follows:

$$R_w = P \cdot \frac{C}{g} \tag{4.7.24}$$

Once the prevailing speed has been estimated either in the field or with Eq. 4.7.22, Table 4.7.5.3 is used to determine the level of service for arterials in classes, I, II, III, and IV.

TABLE 4.7.5.1 Arterial Street Running Time (T_R) in s/mi

		Arterial Class and Free-Flow Speed					
		I			II		
		56	50	43	43	40	34
Average segment length (mi)	$\frac{1}{4}$	95	101	108	106	109	121
	$\frac{1}{2}$	72	79	92	90	93	105
	$\frac{3}{4}$	69	76	87	87	92	105
	1	64	72	82	82	89	105

Source: Transportation Research Board [4.3]. See HCM for current values.

TABLE 4.7.5.2 Arterial Progression Factors

Arrival type	Progression quality	Platoon ratio (R_P) range	Platoon ratio (R_P) default	f_{PAG}
1	Very poor	≤ 0.50	0.333	1.00
2	Unfavorable	$> 0.50–0.85$	0.667	0.93
3	Random arrivals	$> 0.85–1.15$	1.000	1.00
4	Favorable	$> 1.15–1.50$	1.333	1.15
5	Very favorable	$> 1.50–2.00$	1.667	1.00
6	Exceptional	> 2.00	2.000	1.00

Source: Transportation Research Board [4.3]. See HCM for current values.

TABLE 4.7.5.3 Arterial Street Level of Service

		Class and Typical Free-Flow Speed Range (mi/h)			
		I 43–56	II 34–47	III 31–34	IV 25–34
		Average travel speed (mi/h)			
	A	> 45	> 37	> 31	> 25
L	B	> 35–45	> 29–27	> 24–31	> 20–25
O	C	> 25–35	> 21–29	> 18–24	> 14–20
S	D	> 20–25	> 16–21	> 14–18	> 11–14
	E	> 16–20	> 14–16	> 11–14	> 9–11
	F	≤ 16	≤ 14	≤ 11	≤ 9

Source: Transportation Research Board [4.3]. See HCM for current values.

The HCM states that this methodology is not sensitive to the presence of bottlenecks (e.g., narrow bridge), other lane additions or drops, gridlock conditions, and intersection blockage.

NCHRP 398 [4.10] presents simple procedures based on empirically derived regression equations with which the level of congestion on an arterial can be quantified. The following equation applies to classes I, II, and III of arterial streets and is based on the volume-to-capacity ratio:

$$S_{\text{PH}} = S \cdot (1 + \text{ESD})^{-0.3} \cdot (1 + X^4)^{-0.7} \qquad (4.7.25)$$

where

S_{PH} = peak-hour speed, in mi/h

S = free-flow speed, in mi/h

X = degree of saturation

ESD = effective signal density derived as follows:

$$\text{ESD} = \text{SD} \cdot \left(1 - \frac{\text{B}}{\text{C}}\right) \qquad (4.7.26)$$

where

SD = signal density, in the number of signalized intersections per mile

B = through-band duration, in s

C = cycle length, in s

If X is not available, then the average daily traffic per lane (ADT_L) may be used as a surrogate:

$$S_{\text{PH}} = S \cdot (1 + \text{ESD})^{-0.3} \cdot \left[1 + \left(\frac{\text{ADT}_\text{L}}{F}\right)^4\right]^{-0.7} \tag{4.7.27}$$

where F is 10,000 for class I arterials or 8000 for class II and III arterials.

NCHRP 398 compares its congestion quantification procedures with the level-of-service analyses in the HCM and pinpoints the strengths and applicability of each procedures:

> It should be noted that the design and operation analysis in HCM have different objectives and end products than a congestion procedure. They are better suited to identifying location-specific problems than to assessing route, corridor and areawide congestion levels. Estimating density or delay to estimate a level-of-service, for example, provides information to operations and design personnel, but must be farther manipulated to quantify congestion problems. Just as congestion cannot be used to re-time signals, level-of-service measures cannot support many uses and needs of congestion measures, particularly on a systems basis, nor do they assess the intensity and duration of congestion. Congestion estimates on arterial streets using travel time study data can directly evaluate the effect of coördinated signals and are able to determine the difference between delay due to signal operation and delay due to traffic volume.

4.8 TRAFFIC DATA COLLECTION METHODS

There is a long list of data that need to be collected to apply the *Highway Capacity Manual* procedure for the capacity estimation and performance evaluation of a signalized intersection. The required data may be grouped into two types: static and dynamic (time-dependent). Static data are measured once, whereas dynamic data are collected continuously during the data collection period. Geometric characteristics (i.e., lane widths and grades) as well as pretimed signal timings and the area type are static data. Time-dependent data are traffic volumes, traffic composition (i.e., percent of heavy vehicles in traffic), arrival type (i.e. , percent of vehicles arriving in green), and signal timings of actuated controllers (i.e., cycle length and green time available to each movement). Other time-dependent data are the number of buses stopping at the intersection to serve passengers and the number of parking maneuvers per hour. In addition, actual saturation flows and delays can be measured directly in the field.

There are three common ways for collecting traffic data: (1) image recording with video or film cameras, (2) manual collection with a team of workers, and (3) automated collection with portable or fixed detectors. These range in sophistication from pneumatic tubes (which are still a popular portable counter) to video or radar detectors. (The types of detectors are covered in Section 6.5.4.1). The third approach is preferred only when daily volumes are desired. These volumes are appropriate for evaluating networks as well as for planning applications.

The actual data collected with pneumatic tubes are the number of vehicle axles that cross a specific approach or lane at any time (i.e., hourly, daily, monthly, or annual counts). Dynamic data collected with tubes attached to the pavement cannot supply information on turning movements, arrivals on green, and traffic composition. Data collected with such counters are least appropriate for capacity estimation and performance evaluation.

The videotaping or filming alternative requires image recording equipment as well as some experience for its appropriate location and use. To obtain a sufficiently wide field of view, an elevated point may be necessary. After the recording is complete a substantial number of labor hours is required to translate pictures (frames) into numerical data. Specially manufactured film editors may be necessary because the intervals between frames must be precisely equal to a specified amount of time (e.g., 2 to 20 s).

A major problem with this approach is that the signal indications cannot be viewed at the same time as the queues of vehicles. This is a substantial disadvantage when the traffic signal controller is actuated or when arrivals during green are desired. Multiple cameras may resolve this problem, but then precisely synchronized film editors must be used for the derivation of numerical data from the tapes.

Image recognition technology simplifies these tasks by having a computer connected to the cameras and translate images (i.e., vehicles by type and signal displays) directly into numerical format. Such automated image processing devices entered the market in the early 1990s. Devices such as the AUTOSCOPE, Video Track, Trafficon, and a few others can automatically count volumes from video images and derive a multitude of traffic parameters.

The most common option for data collection is with a team of workers, each of whom takes measurements of a specific traffic element. For a specific approach with variable signal timings (actuated signal controller) the following measurement assignments are necessary for full coverage of all inputs required for signalized intersection analysis:

1. *Service volume* (i.e., vehicles crossing the stop line) *for each movement of traffic* (i.e., through, right, and left turns) *along with the number of heavy vehicles in each movement.* It is preferred to record this information at the end of each cycle (end of green for the approach) because the peak 15 min can be identified accurately. In practice, it is common to take volume measurements every 15 min, following the suggestion in the HCM. Procedures for traffic counts for the analysis of intersections with actuated signal controllers have been devised [4.23, 4.24].
2. *Total number of arrivals as well as the arrivals during green.* This results in the true demand for service for each approach, while it also furnishes accurate inputs for the type of arrival (i.e., if most vehicles arrive at the beginning of green, the arterial progression is good and the delays are reduced). Typically these measurements are not done in small-scale, local applications.
3. *Duration of green for each movement and cycle length.* These measurements are taken and recorded every cycle. If the intersection has a pretimed signal controller, cycle length and green times may be taken only once. Signal timings may also be provided by the city traffic engineer. For actuated controllers, average duration of phases

is derived, which consequently compromises the accuracy of intersection capacity and performance.

The labor needed for the collection of traffic data for intersection capacity analysis varies considerably with the design, operation, size, and load of the intersection. In addition, the experience and reliability of the traffic counts team is an important factor. Usually experienced labor may accurately execute multiple assignments (i.e., volume data collection from more than one approach and for all the movements of traffic).

For intersections with actuated signal controllers the number of people necessary for full coverage (i.e., volume counts from all approaches and signal timing data collection) varies between 3 and 10. Figure 4.8.1 shows a typical intersection configuration with the traffic-count people positioned so as to minimize the personnel needs. The first volume-counting person (VC_1) is responsible for volume counts on the eastbound and southbound

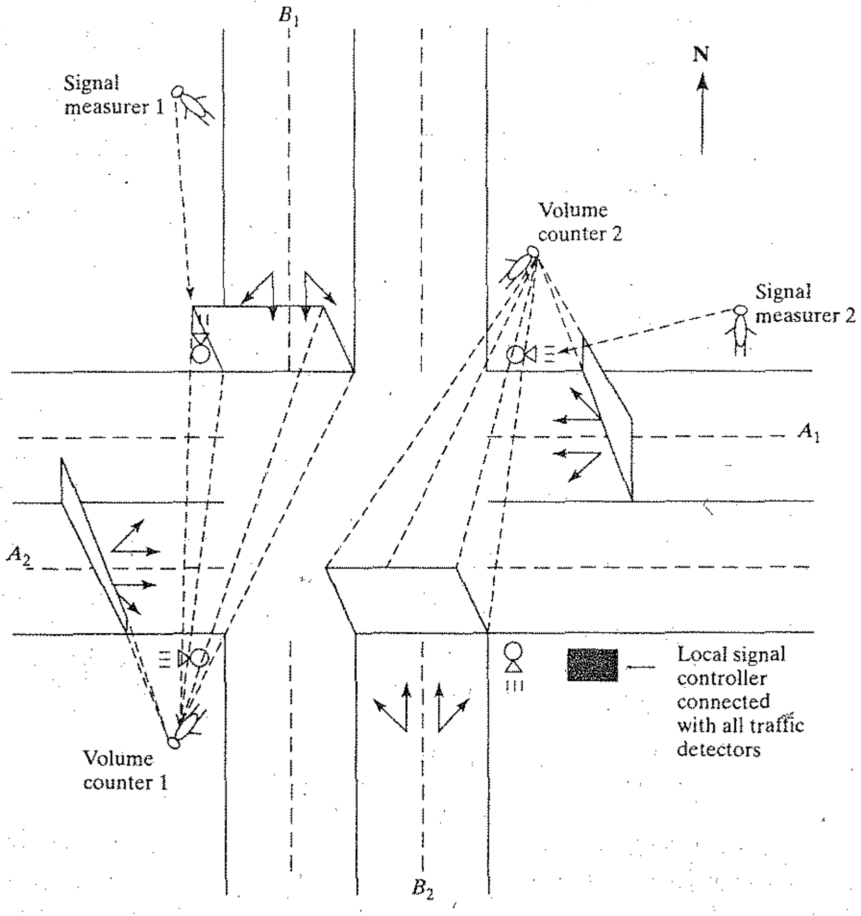


Figure 4.8.1 Personnel assignments for collection of time-dependent data at an intersection with a two-phase actuated signal controller.

approaches, whereas the second volume-counting person (VC_2) takes care of the westbound and northbound volumes. Consider VC_2: When any of the movements on approach A_1 have green, traffic is counted at the stop line of approach A_1 and approach B_2 is ignored because none of the movements on approach B_2 has the right-of-way, except for right-turn-on-red (RTOR), if permitted, and vice versa. Thus one person could take counts from two approaches. The task of traffic counts may be difficult for two persons to accomplish at intersections with multilane approaches and heavy traffic. In such cases one volume-counting person should be allocated for each approach.

Similarly, two to four persons are needed for the collection of signal timing data, if they are variable. In Fig. 4.8.1 the intersection was assumed to operate in two phases. In this case two persons are adequate for measuring signal timings. One person (SM_1) times the duration of green for phase A (east-west traffic, approaches A_1 and A_2) and the cycle length, whereas the other (SM_2) times the duration of green for phase B (north-south traffic, approaches B_1 and B_2). For a complex signal operation with protected left-turn phases and/or other features more persons are required, the maximum being five persons: one for each of the four approaches and the fifth one for the cycle-length measurement [4.24].

Additional labor is required if field-measured saturation flows and stopped delays are desired. Field delays can be estimated with a better than ±10% accuracy by following a simple procedure. Every 10, 15, or 20 s (the length of the sampling interval is defined a priori) the number of stopped vehicles are recorded. (A vehicle is considered stopped when it is within one car length from the vehicle ahead of it in the queue.) Simultaneously volume counts are obtained. These recordings should be taken per lane or lane group. Then field delay (stopped delay), in seconds per vehicle (s/veh), is obtained by using the following formula:

$$\text{Field delay} = \frac{V_s I}{V} \tag{4.8.1}$$

where

V_s = sum of all stopped vehicles counted

I = length of time interval

V = volume count

The field measurement of the saturation flow for a specific lane is simple as well. It requires the accurate measurement of the elapsed time between the fourth and the Nth vehicle as they discharge, with the reference point the stop line. All vehicles from the first to the Nth must be in queue to obtain a valid measurement. Usually the tenth vehicle in queue is utilized (i.e., time between the front bumper of the fourth and the front bumper of the tenth vehicle as they cross the stop line). The field saturation flow can be estimated using the following formula:

$$S_{\text{field}} = \frac{3600}{t_{4\text{ to }N}/(N-4)} \tag{4.8.2}$$

Accuracy of a tenth of a second is essential. An elapsed time between the fourth and tenth vehicle equal to 10.8 s results in a saturation flow of 2000 pcphgpl (passenger cars per hour, green per lane). An error of ±0.5 s results in saturation flows equal to 2100 and

1900 pcphgpl, respectively. Since the formula for estimating field saturation flow is sensitive to the elapsed time, accurate digital chronometers must be used and careful measurements must be taken for representative lanes (one at a time).

4.9 CAPACITY ANALYSIS OF UNSIGNALIZED INTERSECTIONS

4.9.1 Background

The unsignalized intersection is the most common type of traffic intersection. At an unsignalized intersection the service discipline is typically controlled by signs (i.e., stop or yield signs). A primary objective of a traffic engineer studying an unsignalized intersection is to determine its capacity.

There are four types of unsignalized intersections, each one with different flow and traffic control characteristics. The first type consists of one major and one minor street with stop sign traffic control at the minor street. The second type consists of two streets of equal importance where the volume is neither too low (no traffic control required) nor too high (signalization is warranted); in this case traffic is controlled by stop signs on all approaches (four-way stop). The third type consists of two streets (or one street and an off-ramp) where either the flow characteristics or the geometrics (i.e., channelization) warrant yield traffic control for the minor street. The fourth type consists of an intersection where the traffic volume is low. In these traffic facilities the right-of-way is determined by a rule; usually the rule is first-come, first-served; in case of ties the vehicle on the right has the priority. The first two types are commonly subject to capacity analysis.

Estimation of the capacity along the minor street as well as of the left turns from the major to the minor street of an unsignalized intersection is the goal of unsignalized intersection analysis. The most common approach is stochastic (probabilistic) modeling. Implicit in the stochastic modeling are the issues of gap distribution and gap acceptance. The stochastic modeling approach is incorporated in the *Highway Capacity Manual* [4.2] analysis of unsignalized intersections.

Gap distribution represents the distribution of gaps on the major street flow. Long enough gaps give the opportunity of service to minor street traffic. Two alternative types of arrivals are usually assumed for the directional flows on the major street: random (Poisson) and platooned. Platooned arrivals are observed in the case where a signalized intersection exists upstream and/or downstream of the unsignalized intersection; after the stopped vehicles receive the green they arrive at the unsignalized intersection in platoons.

The most common probability distribution representing the headways (gaps) on the major street is the exponential distribution (i.e., random arrivals correspond to exponential interarrival times). In the case of platooned arrivals the lognormal distribution best approximates the distribution of headways [4.25].

Gap acceptance, on the other hand, describes the drivers' behavior, such as the probability of accepting a gap of a certain size given the type of maneuver desired (i.e., cross the major street, or turn right or left on the major street). The driver's physical and mental condition, the perception of risk, and the characteristics of acceleration and handling of the vehicle that he or she drives play a certain role in the decision of gap acceptance or rejection [4.26].

4.9.2 Two-Way Stop-Controlled Intersections

The traffic flow process at an unsignalized intersection is complicated since there are many distinct vehicular movements to be accounted for, all of which operate stochastically. Most of these movements conflict with opposing vehicular volumes. These conflicts result in decreasing capacity, increasing delays, and increasing potential for traffic accidents.

Figure 4.9.1 illustrates conflicts at an unsignalized intersection with stop control on the minor street. t_c^{RT}, t_c^{TH}, and t_c^{LT} denote the critical gaps for the right turn, through, and left turn movements from the minor street, respectively. For a specific movement a gap equal or longer than the critical gap may be accepted by a driver waiting on the minor street. Theory, intuition, and empirical results indicate that usually Eq. (4.9.1) holds. A notable exception is that the gap required for crossing a four-lane major street (6.5 s) is shorter than the gap required for making a right turn onto a four-lane major street (6.9 s).

$$t_c^{\mathrm{RT}} < t_c^{\mathrm{TH}} < t_c^{\mathrm{LT}} \tag{4.9.1}$$

The left turns from the major street to the minor street have the top priority among all the permitted movements. Only after left-turning vehicles from the major street have been served can vehicles from the minor street be served. Sometimes servicing may occur simul-

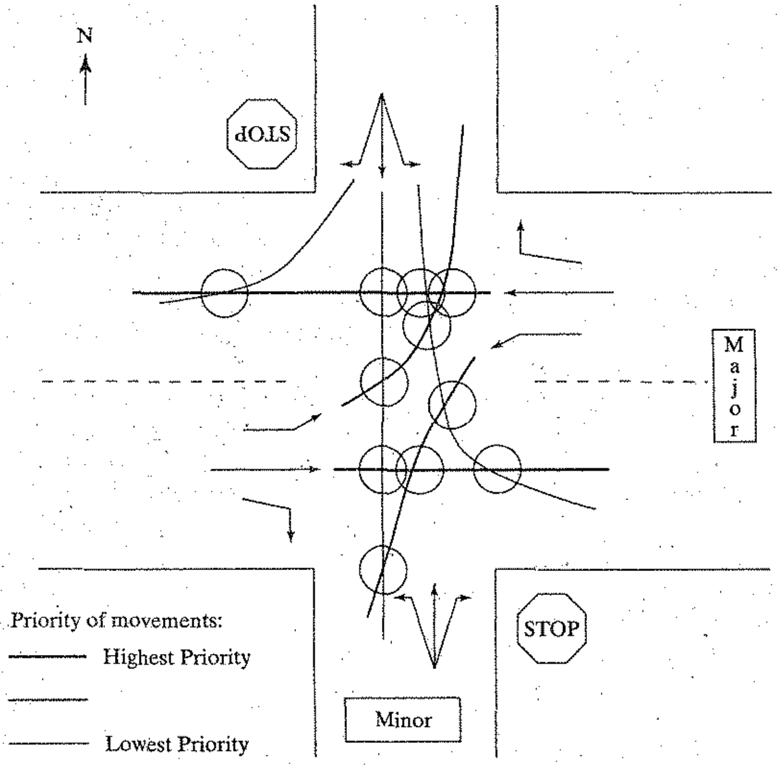


Figure 4.9.1 Identification of conflicts at an unsignalized intersection (northbound movements on the minor street not shown).

taneously, that is, left turn from major and right turn from minor executed at the same time. It is also proper to assume that long gaps may be utilized by more than one vehicle on the minor street (i.e., multiple utilization of gaps [4.27]).

The distribution of headways on the major street along with the gap acceptance behavior enable the derivation of the potential capacity of a stop-controlled minor street. This is the first step in the 1985 HCM procedure for analysis of unsignalized intersections (Fig. 4.9.2). The plots in Fig. 4.9.2 were derived by using the following formula for potential capacity estimation [4.28]:

$$C_p = \frac{e^{-(\alpha-\beta)}}{e^{\beta}-1} V_c \quad \text{with} \quad \alpha = \frac{V_c T_c}{3600}, \quad \beta = \frac{V_c T_s}{3600}, \quad T_s = \frac{1}{2} T_c + 0.5 \qquad (4.9.2)$$

where

V_c = sum of major street traffic volumes that conflict with the subject movement

T_c = critical gap

T_s = follow-up gap; the gap in addition to T_c needed to serve the second, third, and so on, vehicle in multivehicle gap utilization

The potential capacity needs adjustment according to the directional flows and the totals of opposing volumes for each movement on the minor street. The adjustments in the 1985 HCM process have been criticized with regard to their reasonableness and accuracy [4.28, 4.29]. The British have abandoned this approach, whereas the Germans have revised their methodology (the HCM process is a modified version of an early German methodology).

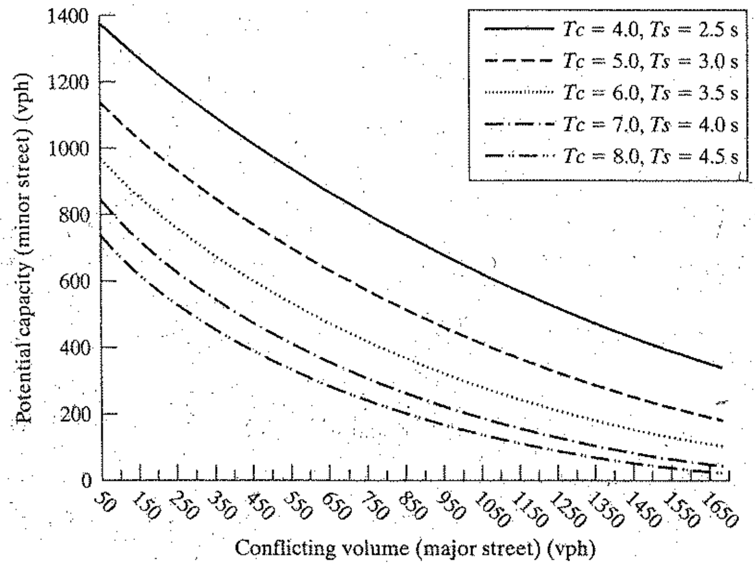


Figure 4.9.2 Potential capacity of a minor street of an unsignalized intersection.

The HCM 2000 procedure includes a number of refinements to the basic methodology presented earlier [4.3]. The potential capacity is given by the following formula:

$$C_p = \frac{e^{-\alpha}}{1 + e^{-\beta}} V_c \quad \text{with} \quad t_f \cong \frac{1}{2} t_c \tag{4.9.3}$$

where

α, β = as defined in Eq. 4.9.2
t_f = replaces t_s (see HCM Chapter 17 for the exact t_f and t_s values)

A table provides base values for t_c and t_f. Subsequently they are adjusted as follows:

$$t_c = t_{c,\text{base}} + t_{\text{HV}} P_{\text{HV}} + t_{\text{G}} G - t_T - t_{3,\text{LT}} \quad \text{and} \quad t_f = t_{f,\text{base}} + t_{\text{HV}} P_{\text{HV}} \tag{4.9.4}$$

where

t_c = adjusted critical gap for the analyzed permitted movement
$t_{c,\text{base}}$ = base value for critical gap
t_f = adjusted follow-up time
$t_{f,\text{base}}$ = base value for follow-up time
t_{HV} = adjustment factor for heavy vehicles
P_{HV} = proportion of heavy vehicles in the analyzed movement
t_G = adjustment factor for grade
G = grade (e.g., 0.04 for 4%)
t_T = adjustment factor for two-stage gap acceptance
$t_{3,\text{LT}}$ = adjustment for minor street left turns at three-leg intersections.

Two-stage gap acceptance is applied to the crossing of arterials with wide medians that permit the temporary safe storage of one or more vehicles.

To aid in the correct determination of conflicting volumes (V_c) for each permitted movement, a ranking order has been established with rank 1 having the highest and rank 4 having the lowest priority. On a typical four-leg two-way, stop-controlled (TWSC) intersection, the ranking works as follows:

- Rank 1: major street TH, major street RT, and pedestrian crossing parallel to the major street.
- Rank 2: major street LT, pedestrian crossing parallel to the minor street, minor street RT.
- Rank 3: minor street TH.
- Rank 4: minor street LT.

HCM analysis for a TWSC intersection proceeds as follows:

1. Summary of inputs
2. Estimation of t_c and t_f separately for each movement
3.a. Adjust for upstream signals (platooned arrivals) and two-stage acceptance, if applicable

3.b. Adjust for two-stage acceptance (no nearby signals), if applicable

4. Impedance and movement capacity

4.a. Shared lane approach, if applicable; this is a rather common condition

4.b. Flared lane approach, if applicable; this permits a more expeditious service to right-turn traffic

5. MOE estimation: delay, queue length, and level of service. The following formulas are used for the estimation of delay and queue length:

$$d = \frac{3600}{c_{m,x}} + 900T\left[\frac{V_x}{c_{m,x}} - 1 + \sqrt{\left(\frac{V_x}{c_{m,x}} - 1\right)^2 + \frac{\left(\frac{3600}{c_{m,x}}\right)\left(\frac{V_x}{c_{m,x}}\right)}{450T}} + 5\right] \tag{4.9.5}$$

where

d = average control delay, in seconds per vehicle (s/veh)

V = volume for the analyzed movement

c = capacity (adjusted, not potential) for the analyzed movement

T = time period of analysis (e.g., $T = 0.25$ for 15 min)

The level of service for both two-way stop-controlled (TWSC) and all-way stop-controlled (AWSC) intersections is to evaluate on a common LOS scale, which is different from the one used for signalized intersections. The HCM criteria for LOS are as follows: $\boxed{A}$ = 0–10, $\boxed{B}$ = 10.1–15, $\boxed{C}$ = 15.1–25, $\boxed{D}$ = 25.1–35, $\boxed{E}$ = 35.1–50, and $\boxed{F}$ = 50.1 or more seconds of control delay per vehicle [4.3]. Once the control delay has been estimated, the average queue length can be derived as follows:

$$Q = dV \tag{4.9.6}$$

where

Q = queue length for analyzed movement

d = delay for analyzed movement (from Eq. 4.9.5) converted, in hours per vehicle (h/veh)

V = volume of the analyzed movement

The HCM makes it clear that this analysis method is based on steady-state conditions; that is, demand and capacity remain constant throughout the period of analysis. Microsimulation is recommended for the proper assessment of time-variant conditions.

Steps 3 and 4 of the analysis procedure are quite complex. They involve the estimation of several probabilities (e.g., probability of blockage by a dominant platoon, probability of blockage by a subordinate platoon, probability that a rank 2 movement will operate in a queue-free state, probability of impedance by pedestrians, etc.) and necessitate the use of 11 worksheets, which makes analysis with pencil and calculator time-consuming, tedious, error-prone, and largely unsuitable for alternative scenario analysis. The use of the HCS or similar software is all but essential.

4.9.3 All-Way Stop-Controlled Intersections

The capacity of all-way stop-controlled (AWSC) intersections is comparatively easy to assess. The analysis of AWSC intersections is easier because all users must stop. Thus the service process becomes more mechanistic and less stochastic, which makes the derivation of representative models easier. Actually the critical entity in AWSC intersection capacity is the average intersection departure headway. Secondary parameters are the number of cross lanes, turning percentages, and the distribution of volume on each approach.

Figure 4.9.3 presents the basics of the 1994 HCM methodology for calculating the capacity of each approach of an AWSC intersection. First, an approach is selected for analysis. This becomes the subject approach. The approach opposite to the subject approach is called *opposing approach,* and the approaches on the sides of the subject approach are called *conflicting approaches.* The superscripts *s, o,* and *c* utilized below stand for *s*ubject, *o*pposing, and *c*onflicting approach, respectively.

The capacity is estimated as follows:

$$c = 1000V^s_{\%} + 700\,V^o_{\%} + 200L^s - 100L^o - 300\text{LT}^o_{\%} + 200\text{RT}^o_{\%} - 300\text{LT}^c_{\%} + 600\text{RT}^c_{\%} \qquad (4.9.7)$$

where

c = capacity of the subject approach, in veh/h

$V^s_{\%}$ = proportion of the intersection volume on the subject approach

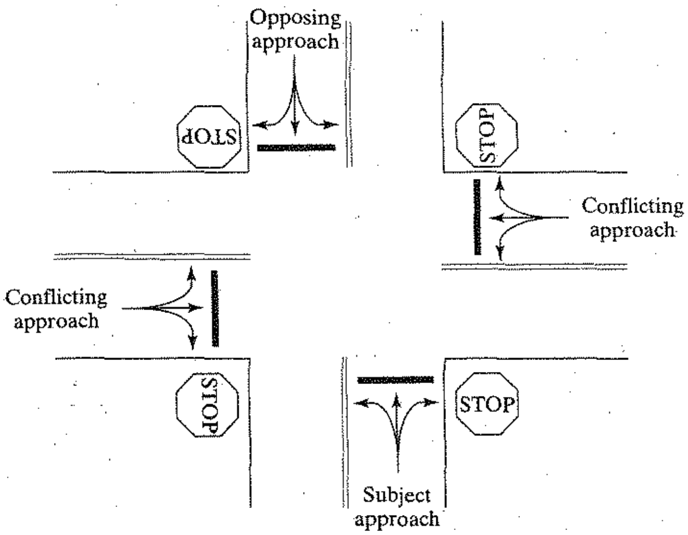


Figure 4.9.3 All-way stop-controlled intersection (AWSC) and example assignment of approaches for capacity and performance analysis.

$V^o_{\%}$ = proportion of the intersection volume on the opposing approach

L^s = number of lanes on the subject approach

L^o = number of lanes on the opposing approach

$\mathrm{LT}^o_{\%}$ = proportion of volume on the opposing approach turning left

$\mathrm{RT}^o_{\%}$ = proportion of volume on the opposing approach turning right

$\mathrm{LT}^c_{\%}$ = proportion of volume on the conflicting approaches turning left

$\mathrm{RT}^c_{\%}$ = proportion of volume on the conflicting approaches turning right

The application of this formula is straightforward, as illustrated in Example 4.8 later in this section. The only required inputs are the number of lanes on each approach and accurate traffic volumes per movement for all approaches. Notably, the process is adaptable to T intersections or intersections with one or two one-way streets (simply, certain factors of Eq. 4.7.3 are not applicable).

The average delay on the subject approach is derived as follows:

$$D = e^{3.8(v/c)} \tag{4.9.8}$$

where

D = delay on subject approach, in s/veh

v = volume on subject approach

c = capacity of subject approach (estimated by Eq. 4.9.7)

After the first approach is done another approach is selected and the roles of subject, opposing, and conflicting approaches are reassigned. The calculations continue until all approaches are analyzed.

This empirically derived methodology for analysis of AWSC intersections should be applied only within a specified range of valid input conditions [4.2].

Example 4.8

An all-way stop-controlled intersection has one lane on each of its four approaches. The following traffic volumes were collected.

	Left turn	Through	Right turn	Total
Eastbound (EB)	75	300	50	425
Westbound (WB)	75	200	50	325
Northbound (NB)	50	250	50	350
Southbound (SB)	50	200	50	300

Estimate the capacity and delay of the NB approach.

Solution

$$V^s_{\%} = \frac{350}{425 + 325 + 350 + 300} = 0.25$$

$$V^o_{\%} = \frac{300}{425 + 325 + 350 + 300} = 0.214 \qquad L^s = 1 \qquad L^o = 1$$

$$\text{LT}^o_{\%} = \frac{50}{300} = 0.167 \qquad \text{RT}^o_{\%} = \frac{50}{300} = 0.167$$

$$\text{LT}^c_{\%} = \frac{75 + 75}{425 + 325} = 0.2 \qquad \text{RT}^c_{\%} = \frac{50 + 50}{425 + 325} = 0.133$$

Substitution in the capacity equation (4.9.7) results in $c = 463$ veh/h. Then $D = e^{3.8(350/463)} = 17.7$ s/veh (LOS C).

HCM 2000 includes a modified version of the previous methodology, and it is applied in four increasingly complex sets of conditions: (1) two one-way street intersections, (2) two two-way street intersections, (3) a general model for intersections with single-lane approaches, and (4) a general model for intersections with 2+ lane approaches. Type 3, which readily applies to types 1 and 2 is described next.

The basic premise of the analysis is the saturation headway or the time elapsed between two successive vehicle departures in the presence of continuous demand. However, as the following notional formula demonstrates, there are a number of interdependencies:

$$h_d = f(\text{lanes per approach, \% HV, } x_O, x_{\text{CL}}, x_{\text{CR}}, \text{\% RT, \% LT}) \tag{4.9.9}$$

where x_O, x_{CL}, x_{CR} are the degrees of utilization for the opposing, conflicting from the left, and conflicting from the right approaches.

The coupling between the capacity of the subject approach and the capacity of the conflicting approaches (through their degree of saturation) is obvious. This necessitates an iterative process based on a system of equations. Given any specific approach as the subject approach, there are five distinct cases of conflicts with the following probabilities of occurrence.

$$\begin{aligned}
P[C_1] &= (1 - x_0)\,(1 - x_{\text{CL}})\,(1 - x_{\text{CR}})\\
P[C_2] &= (x_0)\,(1 - x_{\text{CL}})\,(1 - x_{\text{CR}})\\
P[C_3] &= (1 - x_0)\,(x_{\text{CL}})\,(1 - x_{\text{CR}}) + (1 - x_0)\,(1 - x_{\text{CL}})\,(x_{\text{CR}})\\
P[C_4] &= (x_0)\,(1 - x_{\text{CL}})\,(x_{\text{CR}}) + (x_0)\,(x_{\text{CL}})(1 - x_{\text{CR}}) + (1 - x_0)\,(x_{\text{CL}})\,(x_{\text{CR}})\\
P[C_5] &= (x_0)\,(x_{\text{CL}})\,(x_{\text{CR}})
\end{aligned} \tag{4.9.10}$$

These conflicting cases simply represent the possibilities of vehicle presence on each of the four approaches, as follows:

Conflict case	Subject approach	Opposing approach	Conflicting from left	Conflicting from right
1	Y	N	N	N
2	Y	Y	N	N
3a	Y	N	Y	N
3b	Y	N	N	Y
4a	Y	Y	N	Y
4b	Y	Y	Y	N
4c	Y	N	Y	Y
5	Y	Y	Y	Y

The three probability arguments in $P[C_4]$ represent the three situations 4(a), 4(b), and 4(c). Then the expected value of the saturation headway distribution (h_d) is estimated to be

$$h_d = P[C_1]h_1 + P[C_2]h_2 + P[C_3]h_3 + P[C_4]h_4 + P[C_5]h_5 \qquad (4.9.11)$$

HCM analysis for a AWSC intersection proceeds as follows:

1. Summary of inputs
2. Saturation headway adjustment
3. Probability states
4. Iterative solution for headway so that the difference between successive iterations $\leqslant 0.01$
5. Final h_d and degree of utilization x
6. Capacity, delay, and LOS estimation

The saturation headway is adjusted as follows:

$$h_{\text{adj}} = h_{\text{LT}}\, P_{\text{LT}} + h_{\text{RT}}\, P_{\text{RT}} + h_{\text{HV}}\, P_{\text{HV}} \qquad (4.9.12)$$

where

h_{adj} = adjustment for initial headway

h_{LT} = headway adjustment due to left turns

P_{LT} = proportion of left turn traffic on subject approach

h_{RT} = headway adjustment due to right turns

P_{RT} = proportion of right turn traffic on subject approach

h_{HV} = headway adjustment due to heavy vehicles

P_{HV} = proportion of heavy vehicle traffic on subject approach

In the case of an AWSC intersection with all single lane approaches h_{LT}, h_{RT}, and h_{HV} are equal to 0.2, −0.6, and 1.7, respectively. This concludes step 2 of the methodology. In order to proceed to step 3, it is necessary to subtract the headway adjustment (h_{adj}) from 3.2 s to generate the headway value for the first iteration.

Once the final h_d is determined, the service time is estimated by subtracting the move-up time (m) from the h_d. In the HCM 2000 examples $s = h_d - 2.0$. Then the delay is estimated by applying Eq. 4.9.13.

$$d = s + 900T\left[(x-1) + \sqrt{(x-1)^2 + \left(\frac{h_d x}{450T}\right)}\right] + 5 \tag{4.9.13}$$

As mentioned earlier, the same delay thresholds for the determination of the LOS apply to both TWSC and AWSC intersections. LOS ranges are shown following Eq. 4.9.5.

4.9.4 Roundabouts

Roundabouts in the United States are relatively rare but they are increasing in number. A TRB report, which surveyed 26 municipalities in the United States and Canada, revealed that roundabouts are appealing because of their greater safety, shorter delays, lower costs, and aesthetic attributes. [4.30] A statistically significant before-after reduction of crashes by 51% was observed at the eight small-to-moderate (outside diameter of up to 120 ft) roundabouts analyzed in the report.

Roundabouts vary from tiny circles placed in the middle of an intersection for the purpose of traffic calming (e.g., Seattle style traffic circle) to high-design modern roundabouts such as those in Vail, CO, having three-lane wide approaches [4.30]. Section 2.4.15 covers the basic characteristics of roundabouts.

Because of the scant empirical data from U.S. applications, HCM 2000 suggests that the TWSC formula for the estimation of the potential capacity (C_p) is used for assessing the capacity of a given approach of a roundabout (Eq. 4.9.3). The conflicting volume for each approach (V_c) includes all the conflicting circulating traffic, and in most cases it excludes the right-turn movement from the first (counterclockwise) approach from the subject approach. In dealing with the uncertainty due to the lack of rich field information, the HCM suggests the estimation of both an upper and a lower bound of capacity for each approach based on appropriate values for t_c and t_f. Specifically for upper, $t_c = 4.1$ s and $t_f = 2.6$ s, and for lower, $t_c = 4.6$ and $t_f = 3.1$ s.

If the traffic circle of the roundabout is imposed on a typical four-leg intersection, then the conflicting volumes can be estimated rapidly by following the addition rule shown below. This can be observed in Fig. 4.9.4; the volumes comprising the V_c for the northbound approach are underlined:

V_c for		LT		TH		LT
NB	=	SB	+	EB	+	EB
SB	=	NB	+	WB	+	WB
EB	=	WB	+	SB	+	SB
WB	=	EB	+	NB	+	NB

Brilon and Vandehay [4.31] present various entry approach capacity equations used in Germany for a variety of geometries, such as one or two lanes on the entry approach and

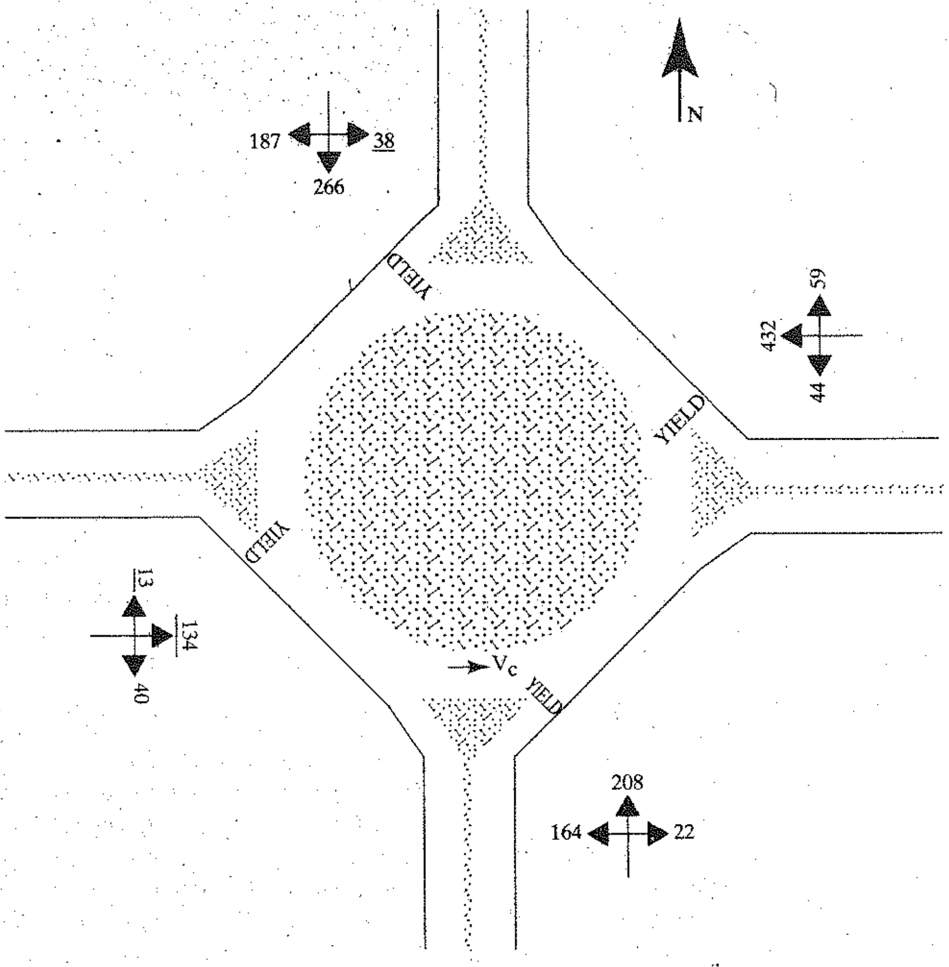


Figure 4.9.4 Simple roundabout for capacity analysis.

one, two, or three lanes around the circle. The formula for the simplest roundabout (one lane approach, one lane around the circle) follows:

$$c = 1218 - 0.74\, V_c \qquad (4.9.14)$$

HCM 2000 does not include a formula for the estimation of delay or a specific procedure for estimating LOS. The method proposed for HCM 2000 terminates in the estimation of the V_c ratio. It is likely, however, that the TWSC delay equation provides an upper-bound estimate of the delay since at roundabouts the typical control is yield instead of stop.

Example 4.9

Given the volumes on the roundabout in Fig. 4.9.4, estimate the upper and lower bounds of capacity per HCM 2000 as well as per Brilon and Vandehay [4.31]. Compare the two.

Solution

				Capacity		
				HCM 2000		
Approach	Movmt.	Volume	V_c	Upper	Lower	Brilon [4.31]
NB	RT	22	185	1198	992	1081
	TH	208				
	LT	164				
SB	RT	187	640	834	667	744
	TH	266				
	LT	38				
EB	RT	40	348	1054	862	960
	TH	134				
	LT	13				
WB	RT	59	385	1023	834	933
	TH	432				
	LT	44				

For this specific example the average of the upper and lower bounds of the HCM 2000 estimates differ only by -0.3 to 1.3% from the estimates of the formula reported by Brilon and Vandehay [4.31].

4.9.5 Signalization Warrants

The type of intersection control is selected from the *Manual on Uniform Traffic Control Devices* (MUTCD) [4.11], which lists the criteria that need to be fulfilled before selecting a type of intersection control. MUTCD lists 11 alternative warrants for judging whether the signalization of an unsignalized intersection is appropriate. If one warrant is met, signal control should be considered. It is recommended that capacity analysis and safety investigation be conducted before a decision to signalize is made: The capacity, performance, and safety under signal control must be assessed and compared to existing conditions.

The warrants for signalization are as follows:

1. *Vehicular volume.* A minimum total volume on the major street and a minimum volume on one of the minor street approaches is required.
2. *Interruption of continuous traffic.* Traffic on the major street is heavy and continuous, which does not allow safe service to the vehicles on the minor street.
3. *Pedestrian volume.* Pedestrians crossing the major street exceed a stated minimum.
4. *School crossings.* The available gaps in traffic are not sufficient in number and length for the safe crossing of schoolchildren.
5. *Progressive movement.* Signalization should be considered if it will enhance the flow between neighboring signalized intersections interrupted by existing unsignalized intersection(s).
6. *Accident experience.* Signalization should be considered if all other measures for accident reduction are not applicable or if they were not effective enough.
7. *Systems.* Signalization may encourage concentration and organization of traffic flow along a signalized intersection network.

8. *Combination of warrants.* Signalization may be justified if 80% of the values stated in warrants 1 and 2 are satisfied.
9. *Four-hour volumes.* This is similar to warrant 1 but only for any 4 h on a typical day.
10. *Peak-hour delay.* Signalization should be considered if undue delay is experienced along the minor street.
11. *Peak-hour volume.* This is similar to warrant 1 but only for the peak hour on a typical day.

MUTCD furnishes values or ranges of values for each of the aforementioned warrants so that existing conditions can be compared with minimum requirements. Compatibility should be established between the unsignalized intersection analysis of the HCM and the warrants of MUTCD. Presently the potential capacity, capacity, and delay estimated with the HCM procedure are not linked to MUTCD warrants.

The MUTCD also specifies three reasons for implementing all-way stop control:

1. A quickly implementable interim measure for an intersection at which a signal is warranted
2. An unacceptable accident experience despite the installation of TWSC and other measures
3. The presence of unacceptable volume and speed levels

4.10 SUMMARY

This chapter presented capacity and performance analysis for:

- Pedestrian facilities
- Bikeway facilities
- Transit facilities and systems
- Highways

operating under both interrupted and uninterrupted flow conditions.

Different measures of capacity and performance are used for each type of facility. Performance is defined by the level of service, which is determined by one or more specific measures of effectiveness. The capacity and the MOE used to define the LOS for each facility are summarized below.

Facility	Capacity Measure	Measure of Effectiveness for LOS
Pedestrian	peds/hr	ft^2/person or delay (s/person)
Bikeway	bikes/hr	meeting and passing events per biker per hour
Transit system	vehicle way capacity in veh/hr or seats/hr	ft^2/pass or pass/seat
Transit station	veh/hr or peds/hr	ft^2/person
Freeway	veh/hr	mean travel speed
Signalized intersection	veh/hr	average control delay (s/veh)
Unsignalized intersection	veh/hr	average control delay (s/veh)

Several of the aforementioned systems are complex or entail complex mathematical formulations in their analysis. In addition, several of them are often analyzed in an integrated fashion, such as a freeway corridor with HOV lanes on the freeway or a parallel arterial with bike lanes and light rail routes and several signalized intersections. For these reasons, a large number of sophisticated software programs are available for individual or integrated facility analysis. Typical capacity analysis software include the Highway Capacity Software (HCS), SIDRA, HCM/Cinema, EZ-Signals, and others. Chapter 15 presents several traffic software applications for capacity and simulation applications.

EXERCISES

1. A two-lane, two-way bike path merges with a two-lane, two-way pedestrian path for about 500 ft. Estimate the LOS for the pedestrian and bikeways at both the merged section and at the adjacent separate sections given the following data; volumes are in units per hour:

Direction	Pedestrian volume	Bicycle volume
EB	102	65
WB	95	88

2. An airport corridor is 30 ft wide. Given a peak demand of 300 pedestrians per minute and an average walk speed of 3 ft/s, estimate the LOS at the corridor.
3. A rapid-transit system employs vehicles that can be connected into transit units. To investigate the effect of vehicular articulation, calculate the capacity (veh/h) and the speed at capacity (ft/s) by varying the number of vehicles per train from $N = 1$ to $N = 5$. Assume a perception-reaction time of 1.5 s, a vehicular length of 40 ft, a normal deceleration of 5 ft/s^2, a clearance length x_o of 4 ft, and a safety regime *a* (Table 3.2.1).
4. Repeat Exercise 3 for safety regime *b*, assuming an emergency deceleration of 15 ft/s^2.
5. Computerize your solution procedure to Exercises 3 and 4.
6. For the system of Example 4.1, calculate the effect of a 5-min decrease in the round-trip time on the fleet size. Also, calculate the before and after headways between vehicles.
7. For the system of Exercise 6, calculate the average headway that would result if the original 34 vehicles were still used.
8. Using the data of Example 4.2, calculate and discuss the station capacity that would result from varying the number of vehicles per train from $N = 2$ to $N = 5$.
9. Computerize the solution procedure for Exercise 8.
10. The peak-hour volumes at two locations were counted and found to be equal. However, the PHFs were 0.85 at the first location and 0.60 at the second. Describe the difference between the two locations if $t = 5$ min.
11. The following 12 consecutive 5-min vehicle counts were taken on a highway:

60 50 40 60 90 80 100 120 140 95 60 30

(a) Plot the histogram of these counts and the histogram of the flow rates computed on the basis of the preceding counts and

(b) calculate the hourly volume and the PHF.

12. Show that for $t = 15$ min the PHF can theoretically range from 0.25 to 1.00.

13. Derive Eq. 4.5.4.

14. A 9-mi segment of a six-lane freeway (three lanes per direction) has a set of characteristics, which are tabulated as follows. Estimate the free-flow speed.

Direction	Lanes	Width (ft)	Right shoulder	% trucks	% RVs	On-ramps	Off-ramps	Terrain
EB	3	12	6	6	1	6	8	Rolling
WB	3	12	6	8	1	6	5	Rolling

15. A number of changes are proposed for the freeway of Exercise 14. Specifically one EB on-ramp will be permanently closed. On both directions the 6-ft left shoulder will be eliminated, the right shoulder will be decreased by 2 ft and each lane will be narrowed by 1 ft. In this way a fourth 11-ft lane will be provided in each direction. Assess the LOS of this plan.

16. For the freeway of Exercises 14 and 15, find the maximum volume on each direction for which a LOS $= C$ is maintained given that PHF is 0.92 and 0.88 on the EB and WB directions, respectively. Does capacity increase when the freeway is fitted with four lanes?

17. A study at an intersection approach found that the approach speed was 30 mi/h. Given that a short-loop detector was located 80 ft upstream of the stop line, calculate (a) the appropriate unit extension and (b) the required minimum green interval.

18. Vehicles are known to approach on a through lane of an intersection at 25 mi/h. For a passage time of 3 s, determine the proper placement of a short-loop detector and the required minimum green interval.

19. Presence detection using a long loop is to be used at an exclusive left-turn lane. For a desired gap of 3 s, design the long loop assuming that its trailing edge will be (a) at the stop line and (b) 10 ft upstream of the stop line. Assume a speed of 10 mi/h.

20. Estimate the optimal cycle length and the green intervals for the intersection shown in Fig. E4.20. Assume that phase A serves the north-south traffic and phase B serves the east-west traffic. Lost time is equal to 3 s per phase and Y + AR is equal to 4 s. The width of each lane is 10 ft and the prevailing saturation flows are as follows:

$$s(\text{TH} + \text{RT}) = 1700$$

$$s(\text{LT}) = 300$$

which reflects a permitted operation.

21. Estimate the signal timings for the intersection of Exercise 20 assuming a four-phase operation. Phase A serves north-south left turns only, phase B serves north-south traffic (no left turns permitted), phase C serves east-west left turns only, and phase D serves east-west traffic (no left turns permitted). Assume 3 s lost per phase and Y + AR is equal to 3 s. The saturation flow for protected left turns is 1700.

22. Evaluate two phasing schemes for the intersection and traffic loads illustrated in Fig. E4.22: (a) a two-phase operation: north-south (phase A), an east-west (phase B) and (b) a three-phase operation north-south (phase A), east-west left turns (phase B), and east-west right turns and through (phase C). Take lost times equal to 3 s per phase and Y + AR equal to 4 s. The following saturation flows prevail: $s(\text{TH, RT, LT}) = 1200$, $s(\text{TH, RT}) = 1700$, $s(\text{TH, LT}) = 500$ (LT permitted), and $s(\text{LT}) = 1700$ (LT protected). Select the best of these two phasing schemes (must furnish quantitative proof other than cycle length, which is not a description of efficiency).

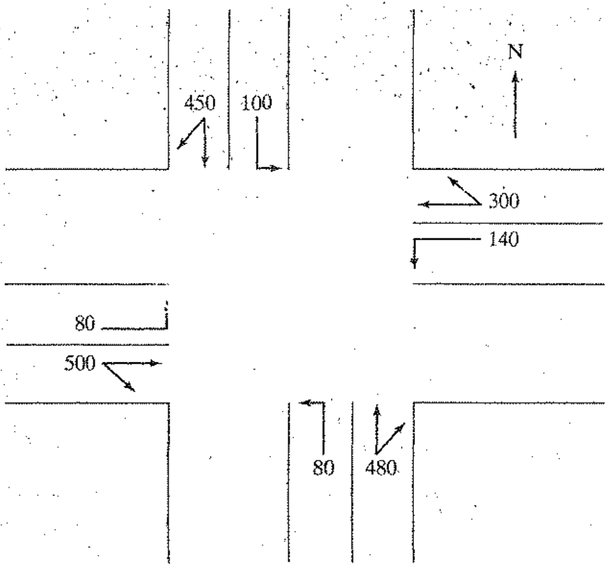


Figure E4.20

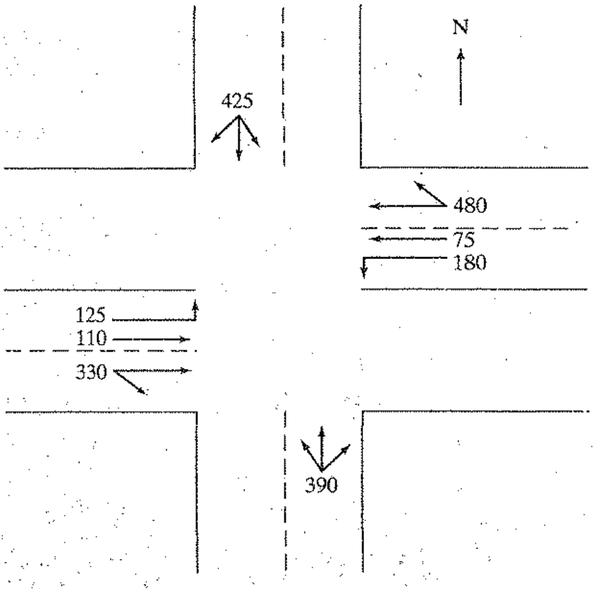


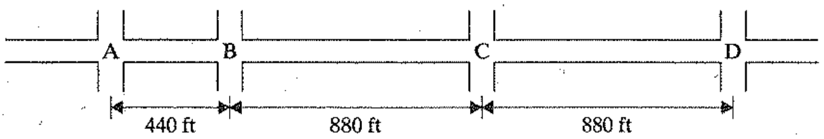
Figure E4.22

23. Derive the prevailing saturation flows for the three case studies tabulated here based on HCM 2000 or on the most current version of the HCM and compare with the saturation flow and the corresponding parameters from the 1985 and 1994 editions of the HCM shown in the table:

	Case 1	HCM parameters			Case 2	HCM parameters			Case 3	HCM parameters		
	Data	1985	1994	2000	Data	1985	1994	2000	Data	~~1985~~	~~1994~~	~~2000~~
No. of lanes and movement(s)	2 LT	2	2	2	3 TH+RT	3	3	3	1 shared	1	1	1
No. of lanes on subject approach		5	5	5		4	4	4		1	1	1
Lane width (ft)	11	0.97	0.967		12	1	1		14	1.07	1.067	
Heavy vehicles	8%	0.94	0.926		5%	0.975	0.952		20%	0.91	0.833	
Grade	−3	1.015	1.015		3	0.985	0.985		−5%	1.025	1.025	
Parking (N_m)	30	1	1		20	0.93	0.933		no	1	1	
Bus blockage (N_B)	20	1	1		30	0.96	0.960		15	0.94	0.940	
Area type	CBD	0.90	0.90		CBD	0.90	0.90		other	1	1	
Right turn	n.a.	1	1		16%	0.957	0.957		23%	0.860	0.860	
Left turn	protected	0.92	0.95		n.a.	1	1		12%	0.881	0.994	
Lane utilization	even	n.a.	n.a.		even	n.a.	n.a.		n.a.	n.a.	n.a.	
Pedestrians	n.a.	n.a.	n.a.		250	n.a.	n.a.		80 RT, 50 LT	n.a.	n.a.	
Base (ideal) saturation flow	s_o	1800	1900		s_o	1800	1900		s_o	1800	1900	
Prevailing saturation flow	s	2759	2952		s	3988	4126		s	1280	1391	

Notes: n.a. = not applicable; for case 3 (1) LT movement is not opposed, (2) turning vehicles turn onto two receiving lanes.

24. The signals at the intersections along the two-way street have been pretimed as shown here [all timings in seconds (s)]. Given that the speed is 30 mi/h, determine the width (if any) of the through bands in each of the two directions and show the through bands on a progression diagram.



	A	B	C	D
green	40	35	35	40
Y+AR	5	5	5	5
red	15	20	20	15
offset	0	5	5	40

25. The pretimed signals at four intersections on a one-way street (from A to D) are tabulated next. Given a speed of 30 mi/h,

(a) Is it possible to coordinate these signals? Why or why not?
(b) Sketch the corresponding time-distance diagram.
(c) Determine the width of the resulting through band (green plus yellow), if any.
(d) Clearly show the trajectories of the first and last vehicles in the through band.
(e) Determine the maximum through-band width that could result by modifying the offsets.

Intersection	Distance from A (ft)	Green (s)	Y + AR (s)	Red (s)	Offset (s)
A	0	25	5	30	15
B	660	30	5	25	20
C	1100	20	5	35	30
D	1760	30	5	25	0

26. The signals at the intersections of a one-way street have been pretimed and coordinated as follows:

Intersection	Green (s)	Y + AR (s)	Red (s)	Offset (s)	Distance from A (ft)
A	40	5	35	5	0
B	50	5	25	60	1800
C	35	5	45	40	5200

Given a design speed of 30 mi/h, determine the width of the resulting through band, if any. Show your calculations.

27. At time $t = 485$ s after the reference time a traffic signal is 22 s into its cycle. Assuming that the cycle length is 90 s, calculate the signal's offset.

28. Show graphically that the answers to parts (a) and (b) of Example 4.6 are correct.

29. For the system of intersections of Example 4.7, calculate the width of the through band that would result from changing the offsets at intersections A, B, and C to 10, 60, and 20 s, respectively.

30. Assuming that the street of Exercise 29 is a two-way street, calculate the width of the through band in the other direction.

31. Given a speed of 45 mi/h and the accompanying data, determine whether a balanced signal coordination exists.

Intersection	Green (s)	Y + AR (s)	Red (s)	Offset (s)	Distance from A (ft)
A	30	5	15	5	—
B	25	5	20	0	660
C		unsignalized			1430
D	20	5	25	25	1980
E	30	5	15	5	3300

32. Construct a computer program that calculates the through-band widths for a series of N intersections given (a) the signal cycles, (b) the signal offsets, (c) the distances between intersections, and (d) the design speed.

33. Given the intersection shown in Fig. E4.33 and the table of adjustment factors, calculate the prevailing saturation flows for each lane, for two levels of ideal saturation flow: 1800 and 2000 vehicles per hour, green per lane (vphgpl). If the cycle length is 60 s and the total lost time is equal to 7 s, what is the effect of the saturation flow estimates on the overall level of intersection utilization (X_c)?

	f_w	f_{HV}	f_g	f_p	f_{bb}	f_a	f_{RT}	f_{LT}
Approach A	0.93	1.00	1.025	0.80	1.00	1.00	1.00	0.85
Approach B	1.00	0.93	1.00	0.70	0.96	1.00	0.85	1.00

34. Repeat Exercise 33 using factors from HCM 2000. Whaι is the new X_c?

35. Conduct capacity and performance analysis for the intersection in Fig. E4.22. Which phasing scheme should be selected on the basis of delay?

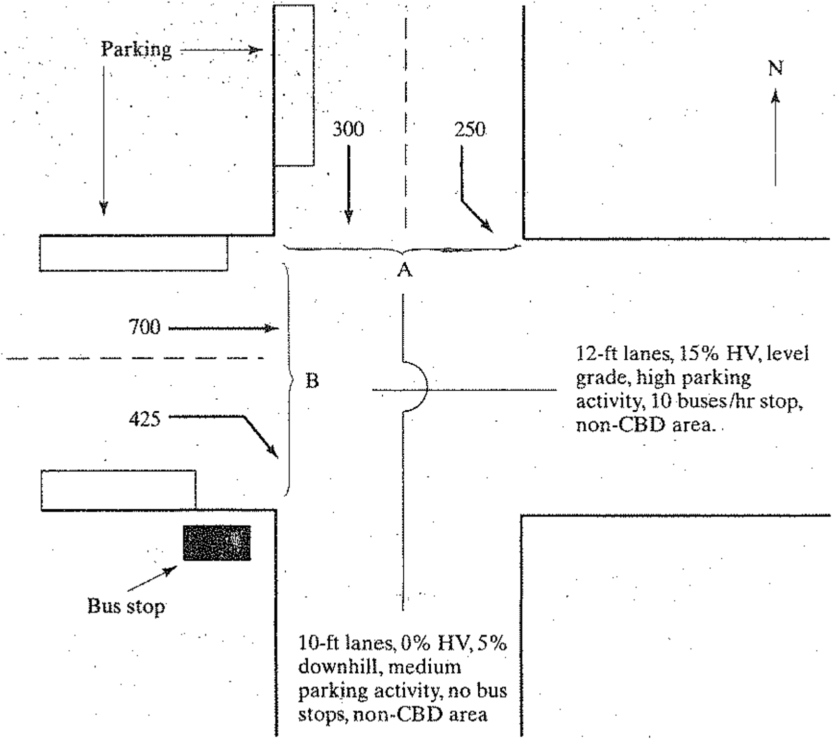

Parking
300
250
N
A
700
B
425
12-ft lanes, 15% HV, level grade, high parking activity, 10 buses/hr stop, non-CBD area.
Bus stop
10-ft lanes, 0% HV, 5% downhill, medium parking activity, no bus stops, non-CBD area

Figure E4.33

36. Conduct a capacity analysis for the intersection approach illustrated in Fig. E4.36. Make a table with the following columns: v, s, v/s, g, C, g/C, c, v/c, d_1, d_2, PF, delay, LOS, whole approach delay, and whole approach LOS. Use a spreadsheet for this analysis.

37. For the intersection shown in Fig. E4.37, conduct analysis for each approach and for the intersection as a whole (delay and LOS). Deliver a table similar to the one in Exercise 36.

38. Derive better signal timings for the intersection in Fig. E4.37. Does intersection performance improve? Maintain the same phasing scheme and assume the total lost time of 10 s. Use of spreadsheet software is strongly recommended.

39. Consider the three factors that play an important role in average delay per vehicle: cycle length (C), green split (g/c), and degree of saturation (X or v/c). Conduct an analysis to identify which of these three factors has the most critical effect on delay. Assume that the progression factor is 1.0 and take $c = 1000$ for the overflow delay only. You need to conduct a sensitivity analysis. Select a range of values for each factor (e.g., 60 to 130 for C, 0.1 to 0.9 for g/c, and 0.2 to 1.2 for X). Hold two factors constant at the midpoint of their range and vary the third factor. Then take the delay estimates and discuss the results.

40. The following data were collected in the field for the two through lanes of an approach. The stopped vehicle counts were

10, 12, 18, 24, 16, 7, 0, 0, 3, 11, 16, 8, 0, 5, 13, 18, 28, 15, 9, 0

every 20 s. If the corresponding vehicle volume is 178, what is the average delay per vehicle?

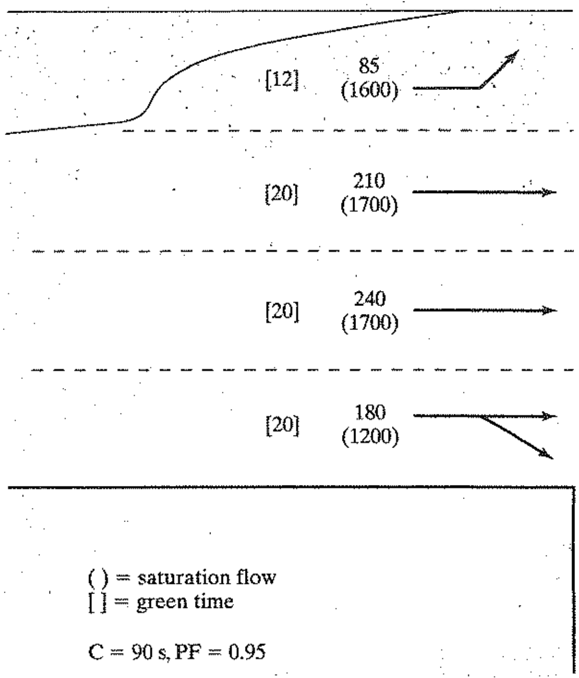


Figure E4.36

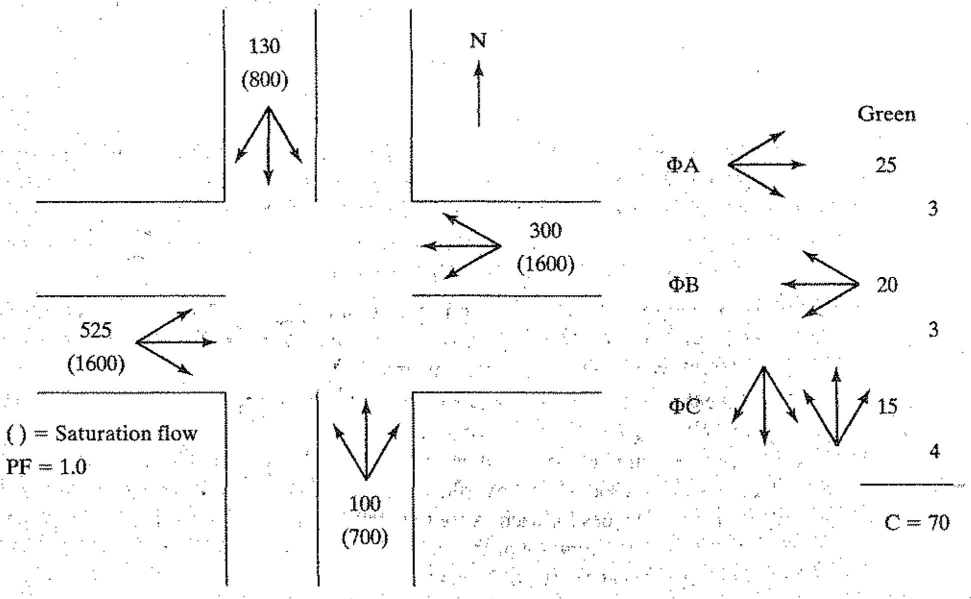


Figure E4.37

Saturation flow measurements: A total of 15 measurements was obtained. They represent the elapsed time between the fourth and the *N*th vehicle in queue. Estimate the mean saturation flow and the range of one standard deviation around the mean.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Time	12.3	11.9	7.6	11.4	11.7	15.0	15.3	16.1	11.7	12.6	12.0	7.8	18.9	11.8	12.2
*N*th	10	10	8	10	10	12	12	12	10	10	10	8	14	10	10

Given these saturation flow measurements, can the hypothesis of $s = 1800$ be statistically rejected at the 95% confidence level? (*Hint:* Conduct a *t* test.)

41. The city agency responsible for the intersection in case study 4.7.4.1 has decided to improve the operations of the intersection by instituting a no-parking policy, removing the bus stops from the intersection, and widening the lanes within the existing width. Assess the improvement in delay and LOS.

42. Review case study 4.7.4.2. Reestimate the signal timings by using volumes that are adjusted with the peak-hour factor. Conduct capacity and performance analyses. Which volumes (V or V_a) yield a smaller overall delay?

43. The field-measured average greens for phases A to E at the intersection of case study 4.7.4.2 are 13, 4, 20, 27, and 25 s, respectively. Estimate the delays with this set of signal timings and state whether the field or the timings estimated in the in-text analysis of case study 4.7.4.2) are better.

44. An important assumption in the estimation of the cycle length with Webster's formula is that of lost time (L) per phase. Use the intersection of case study 4.7.4.2 and vary L from 2.5 to 5 s in steps of 0.5 s and for each L estimate the C_o, X_c, and total delay. What do you observe?

45. Assess the effect of an actuated controllers unit extension (UE) on the d_2 component of delay. Use the intersection of case study 4.7.4.2 and estimate d_2 for all lane groups for the following UE and $k_{\min}$ data pairs:

$$(\text{UE}, k_{\min}) = (2.0, 0.04), (2.5, 0.08), (3.0, 0.11), (3.5, 0.13), (4.0, 0.15), (4.5, 0.19), (5.0, 0.23)$$

46. If $f_{\text{PA}} = f_{\text{PAG}} = f$, show that equations 4.7.21 and 4.7.23 are identical. Show that the C_P Equations 4.9.2 and 4.9.3 are identical (ignore the difference in the definition of t_s and t_f). Are the delay equations 4.9.5 and 4.9.13 identical?

47. Given the unsignalized T intersection illustrated in Fig. E4.47, calculate the potential capacity of each lane on the minor street using the HCM 2000 formula. If the volumes on the minor street are 140 and 190 for the left and right turns, respectively, would the minor street be expected to operate adequately, or are long delays likely to occur? The critical gap sizes are as follows: LT = 7.1 s, RT = 6.2 s, and $t_f = 0.5 \times t_c$.

48. Conduct potential capacity analysis for the intersection illustrated in Fig. E4.48 using the HCM formula and taking critical gaps equal to 6.0 s for turns, 5.0 s for through, and $t_f = 0.5 \times t_c$. If the volumes per movement are as follows: (1) = 200, (2) = 300, (3) = 250, what is the expected performance of each movement? If there is a capacity problem, what can be done about it?

49. Increase the volumes of Example 4.9 by 25% on the NB and SB approaches and by 10% on the EB and WB approaches. Estimate the lower- and upper-bound capacities using both the HCM and Brilon's equation. How do the HCM and Brilon estimates compare? How does the degree of saturation change? Use Eq. 4.9.5 to estimate delay for each approach for a 1-h period.

50. After several months the volumes in Example 4.9 stabilized as in Exercise 49. In addition, local motorists have discovered that roundabouts make U-turns safe and legal. The following U-turn

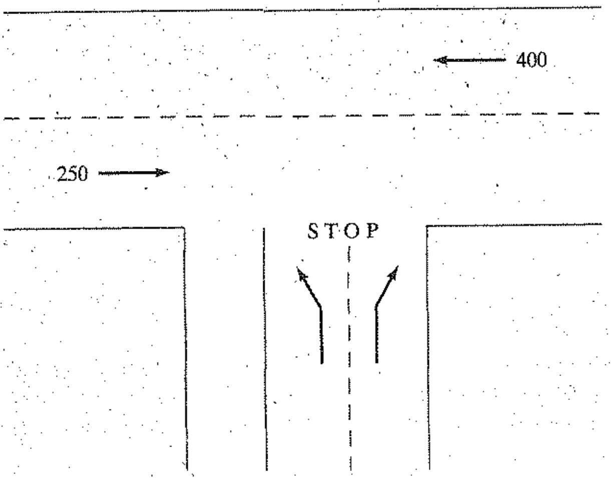


Figure E4.47

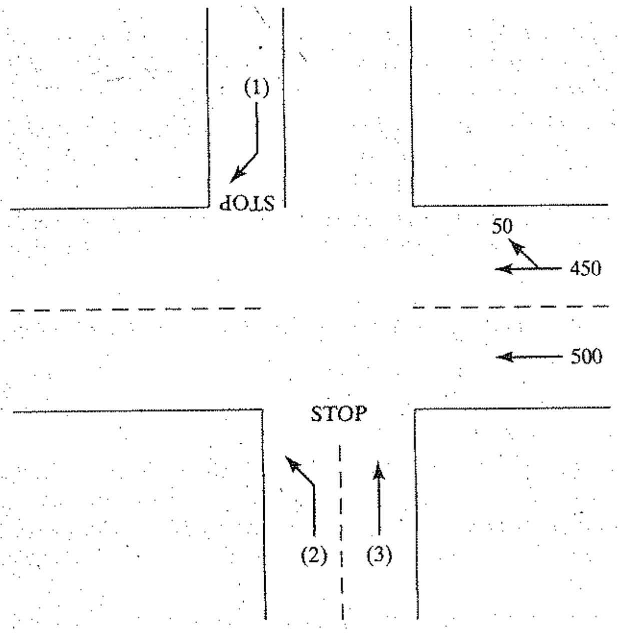


Figure E4.48

volumes have been recorded: NB = 31, SB = 23, EB = 38, and WB = 8 and should be added to the volumes of Exercise 49. Estimate the average capacity based on the HCM method. How much did these 100 additional U–turns affect the degree of saturation?

REFERENCES

4.1 Transportation Research Board, *Highway Capacity Manual,* Special Report 209, 2nd ed., National Research Council, Washington, DC, 1985.

4.2 Transportation Research Board, *Highway Capacity Manual,* Special Report 209, 3rd ed., National Research Council, Washington, DC, 1997.

4.3 Transportation Research Board, *Highway Capacity Manual,* Special Report 209, 4th ed., National Research Council, Washington, DC, 2000.

4.4 Vuchic, Vukan R., *Urban Public Transportation Systems and Technology,* Prentice-Hall, Englewood Cliffs, NJ, 1981.

4.5 Papacostas, C. S., *Energy and Pollution Implications of Bus-Automobile Alternatives,* Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, PA, 1974.

4.6 Dial, R., S. G. Rutherford, and L. Quillian, *Transit Network Analysis: INET,* Report UMTA-UPM-20-79-3, U.S. DOT, Washington, DC, 1979.

4.7 Transportation Research Board, *Highway Capacity Manual,* Special Report 209, 3rd ed. Update, National Research Council, Washington, DC, 1997.

4.8 Highway Research Board, *Highway Capacity Manual,* Special Report 87, National Research Council, Washington, DC, 1965.

4.9 Transportation Research Board, *Interim Materials on Highway Capacity,* Circular 21, National Research Council, Washington, DC, 1980.

4.10 Lomax, T., S. Turner, G. Shunk et al., *Quantifying Congestion,* NCHRP Report 398, Vol. I, TRB, Washington, DC, 1997.

4.11 Federal Highway Administration, *Manual on Uniform Traffic Control Devices,* U.S. DOT, Washington, DC, 1988.

4.12 U.S. DOT, *Traffic Control Systems Handbook,* Implementation Package FHWA-IP-85-11, Washington, DC, 1985.

4.13 Kell, J. H., and I. J. Fullerton, *Manual of Traffic Signal Design,* 2nd ed., Institute of Transportation Engineers, Prentice-Hall, Englewood Cliffs, NJ, 1982.

4.14 Sabra, Z. A., *Traffic Control Equipment and Software,* Demonstration Project 93, Office of Technology Applications, FHWA-SA-93-061, Federal Highway Administration, Washington, DC, 1993.

4.15 National Electrical Manufacturers Association, *NEMA Standards Publication No. TS 1-1983, Traffic Control Systems,* Washington, DC, 1983.

4.16 Yaugh, P. J., *Traffic Signal Control Equipment: State of the Art,* NCHRP Synthesis of Highway Practice 166, TRB, Washington, DC, 1990.

4.17 McShane, W. R., and R. P. Roess, *Traffic Engineering,* Prentice-Hall, Englewood Cliffs, NJ, 1990.

4.18 Berry, D. S., *Notes on Traffic Engineering.* Unpublished, Northwestern University, Evanston, IL, 1978.

4.19 FEDERAL HIGHWAY ADMINISTRATION, *TRANSYT 7-F: User's Manual.* Washington, DC, 1998.

4.20 INSTITUTE OF TRANSPORTATION ENGINEERS, *Traffic Engineering Handbook,* Prentice-Hall, Englewood Cliffs, NJ, 1992.

4.21 JOVANIS, P. P., P. D. PREVEDOUROS, and N. ROUPHAIL, *Design and Operation of Signalized Intersections in Illinois,* Final Report for the Illinois DOT, Northwestern University, Evanston, IL, 1987.

4.22 JOVANIS, P. P., A. J. KHATTAK, and N. ROUPHAIL, *Design and Operation of Signalized Intersections in Small Urban Areas in Illinois,* Final Report for the Illinois DOT, Evanston, IL, Northwestern University, 1988.

4.23 BERRY, D. S., "Volume Counting for Computing Delay at Signalized Intersections," *ITE Journal,* Vol. 57, No. 5 (1987): 21–24.

4.24 PREVEDOUROS, P. D., "Actuated Signalization: Traffic Measurements and Capacity Analysis," Proceedings of the 61st Annual Meeting of the ITE, Milwaukee, WI, September 1991.

4.25 PREVEDOUROS, P. D., "A Model of Unsignalized Intersection Capacity Based on Erlang-3 Gap Distribution," in *Intersections without Signals,* W. Brilon (Ed.), Springer-Verlag, Berlin, 1988, pp. 165–179.

4.26 DAGANZO, C., "Estimation of Gap Acceptance Parameters within and across the Population from a Direct Roadside Observation," *Transportation Research* 15 B (1981): 1–15.

4.27 TROUTBECK, R., "Average Delay at an Unsignalized Intersection with Two Major Streams Each Having a Dichotomized Headway Distribution," *Transportation Science,* Vol. 20, No. 4 (1986): 272–288.

4.28 BAASS, K. G., "The Potential Capacity of Unsignalized Intersections," *ITE Journal* (1987): 43–46.

4.29 KHATTAK, A. J., and P. P. JOVANIS, "Capacity and Delay Estimation Approaches for Unsignalized Intersections: Conceptual and Empirical Issues," *Transportation Research Record 1287,* Washington DC (1990): 129–137.

4.30 GACQUEMART, G., *Modern Roundabout Practice in the United States,* NCHRP Synthesis of Highway Practice 264, TRB, Washington, DC, 1998.

4.31 BRILON, W., and M. VANDEHAY, "Roundabouts—The State of the Art in Germany," *ITE Journal,* November (1998): 48–54.

PART 2

Transportation Systems

5

Transportation Modes

5.1 INTRODUCTION

The preceding chapters were based on the reader's engineering mechanics background and presented topics on the motion of single vehicles and its implications to the design of highways, the flow of multiple vehicles on traffic facilities, and the principles governing the analysis of interrupted and uninterrupted flow. This chapter begins the transition from a vehicle and facility-specific perspective to a systems perspective. It introduces the coexisting, interacting, and competing modes of freight and intercity passenger transportation.

Urban transportation systems including the pervasive problem of traffic congestion are covered in Chapter 6 along with advanced technologies, known as Intelligent Transportation Systems (ITS), which are intended to reduce congestion, to aid the coordination of transportation modes (intermodalism), and to help improve the efficiency of the overall transportation system. After reading these two chapters the reader will be ready to tackle the major topics of planning and demand forecasting in the subsequent Chapters 7 and 8. These chapters also set the stage for the discussion and analysis of the impacts of transportation, presented in Chapters 9, 10, and 11.

Chapters 5 and 6 offer a view of the complexity of transportation. Previous chapters presented the design of highways using *standardized* design vehicles and the assessment of the performance of transportation facilities based on *average* operator/pilot/driver and vessel/aircraft/vehicle characteristics. In reality, however, there is a large number of modes using each facility, serving a multitude of purposes, having a multitude of short- and long-distance destinations, competing with each other for space and time (i.e., this defines car-following and saturation rates in the traffic models) but also for customers (e.g., publicly funded bus versus private vanpool versus taxis, etc.) Each active transportation unit also tries to optimize its travel based on its operator's empirical knowledge, radio-disseminated, other exterior or in-unit guidance. Simultaneously each unit pollutes, makes noise, and risks getting involved in a crash. Transportation units routinely become inactive for loading and unloading people and/or goods, or are stored when not needed. Space as small as a parking stall and as large as an international airport is required for this. Except for walking, all modes of transportation are incapable of functioning in the absence of intermodalism. The simple mailing of a package may involve five or more modes and several stages of processing!

All units in this chapter are in the metric system. All references to tons correspond to 1,000 kg.

5.2 MODES

This chapter begins with an overview of the basic characteristics of major transportation modes (Fig. 5.2.1). The presentation is focused on the transportation industry in the United States. Motor carriers (trucking industry), railroads, pipelines, domestic and international water carriers, air carriers, as well as mail/parcel forwarders are presented. Basic service and cost characteristics are discussed for each carrier or industry, along with their respective advantages and disadvantages. The definitions of fixed and variable costs as well as economies of scale are presented subsequently to facilitate the comparison of major transportation modes.

The total cost of owning and operating a company, including transportation companies, is usually broken into fixed and variable costs. *Fixed costs* do not depend on production levels or the degree of equipment utilization. Aircraft, trucks, trains, computers, and offices cost a fixed amount of money (purchase or lease price) no matter how much they are used (i.e., the fixed cost remains the same irrespectively of whether the equipment is kept idle or is utilized around-the-clock). On the other hand, the more the equipment is utilized, the more labor is necessary to operate them, the more fuel is needed to produce propulsion, and the more maintenance is required due to the increased wear and tear. These costs, which depend on the degree of utilization, are known as *variable costs*. All costs tend to become variable in the long term (5 years or more) as corporations expand or reduce their activities (increase or decrease the acquisition of fixed facilities, equipment, and operations).

The absolute magnitude of the fixed cost as well as the magnitude of the variable cost relative to the fixed cost determine the existence of economies of scale. In simple terms, when economies of scale are present, production increases lower the cost per unit produced and increase the profit per unit. Economies of scale exist first, when the fixed cost is high (thus the more the units over which it is spread, the lower will be the cost per unit) and second, when the variable cost is small compared to the fixed cost. The existence of economies of scale (EOS) is depicted in Fig. 5.2.2.

When EOS are not present, the cost per unit does not change appreciably as the number of units produced (or transported) increases. The opposite also is true. Figure 5.2.2 shows that in the absence of EOS, when production (or the number of units carried) doubles, the cost per unit drops from $0.10 to $0.09, a 10% reduction. By contrast, if EOS are prevailing, then the cost per unit drops from $0.20 to $0.11, a 45% reduction.

Transportation carriers can be private or for-hire. Private carriers are usually the transportation subsidiary of a large parent company (e.g., a manufacturer, a petroleum company, etc.) and carry the cargo of the specific company. For-hire carriers are further categorized in common and contract. Common carriers serve the general public, often on a first-come, first-served basis. Contract carriers provide services to the public on a contract basis only. For-hire carriers can be either regulated or exempt from economic regulation, depending on

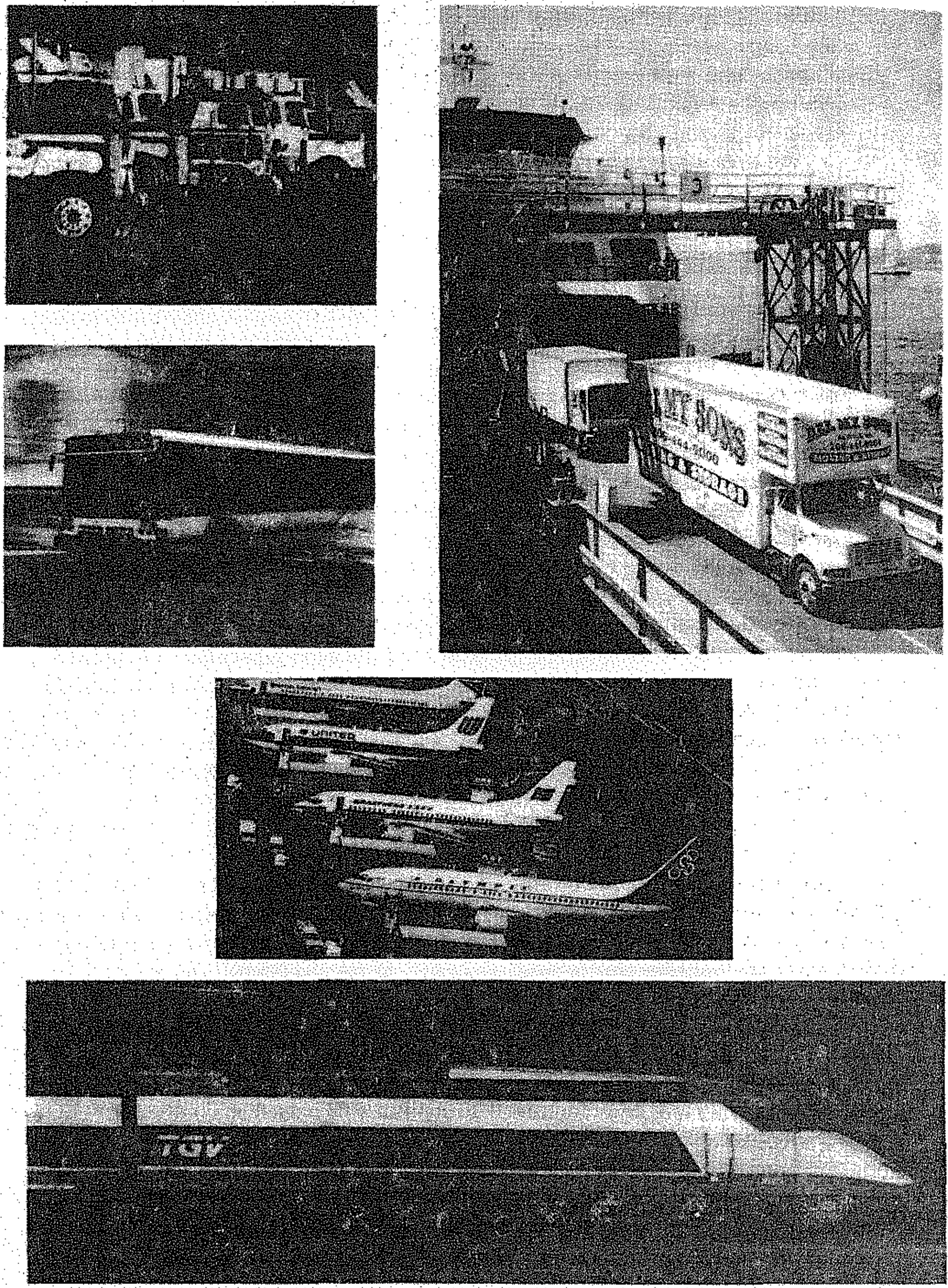


Figure 5.2.1 Major transportation modes.
(From Transportation Research Board, *TR News*, issues 182, 1966; 200, 1999; 169, 1993; and 158, 1992.)

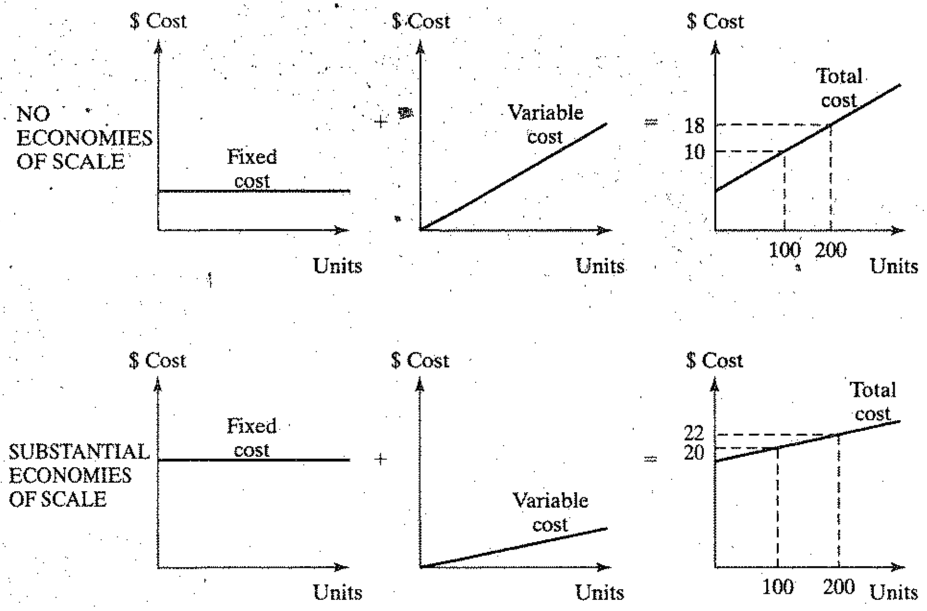


Figure 5.2.2 Example of the principles of economies of scale.

the types of products they carry. The Interstate Commerce Commission (ICC) supervised and regulated the trade among states until the passage of the ICC Elimination Act of 1995. Some regulatory power has been transferred to the federal and state DOTs.

5.2.1 Motor Carriers

Motor carriers (trucking industry) constitute the most ubiquitous mode of freight transportation. The extensive roadway network is the major cause of the popularity and efficiency of this mode (Fig. 5.2.3). Motor carriers have the major advantage of being able to provide door-to-door service to both the shipper and the consignee. The truck is the most common local delivery mode.

The structure of this industry is complex. There are for-hire and private carriers. For-hire carriers may be licensed to operate intrastate or interstate, each of which can be exempt or regulated. Interstate-regulated carriers can be common or contract carriers. The common carriers may serve regular or irregular routes and transport general or special commodities depending on the type of operating certificate they possess (e.g., the State of Indiana issues a special certificate for the intrastate transportation of fertilizers and other agricultural chemicals.) Examples of interstate motor carriers are Yellow Freight, Consolidated Freightways, North American Van Lines, and Roadway.

These regulations and classifications are likely to change dramatically due to the ICC Elimination Act and the Federal Aviation Administration Act of 1994, which preempts states from regulating intrastate transportation with regard to price, route, or service. A new structure is not likely to form soon because both laws were on appeal in 1998. In addition

Figure 5.2.3 Trucks are a small volume, low weight, fast and flexible mode of transportation. (From Transportation Research Board, *TR News*, 154, 1991.)

to economic regulation, there are elements of operational regulation (e.g., affecting the hours of driving, passive or active collision avoidance devices, etc.)

Motor carriers transport a variety of goods, such as agricultural commodities, building materials, forest products, hazardous materials, heavy machinery, household goods, petroleum products, refrigerated goods, retail store items, and vehicles. They also provide other services, such as armored truck service, dump trucking, moving services, and rental services. In the mid-1990s the industry accounted for 25% of the total intercity freight tonnage, but more than 72% of the respective revenues. The average haul is approximately 220 km.

The major advantages of this mode of transportation are high speed and high accessibility. Limitations in volume and weight are the principal disadvantages, and the rates charged are higher compared with the railroads, particularly for heavy hauls over 1500 km. Truck transportation is certainly faster than railroad transportation, and in many cases it is faster than air transportation for hauls up to 1500 km due to the limited flight schedule and the pickup and delivery times incurred by air cargo carriers. Additional advantages are the relatively smooth ride and timely delivery, which makes this type of transportation appropriate for delicate or high value products (i.e., produce, electronic equipment). The high level of integration of motor carriers with all other modes of transportation is another advantage of this industry. Motor carriers link transportation terminals (i.e., ports and docks, airports, railroad yards) with shippers or receivers, and thus provide an essential link for intermodal transportation.

This industry is characterized by low fixed (management, overhead, vehicle fleet) and high variable costs (drivers, fuel, maintenance, insurance, tires, licenses, fleet depreciation). There are no economies of scale in this industry as there are for railroads and pipelines. There are, however, economies of utilization (i.e., large companies tend to utilize terminals and management specialists at a higher rate), as well as economies of equipment acquisition (i.e., volume discounts for the purchase of vehicles, parts, tires, and insurance).

Competition in this industry is strong, as evidenced by the large number of companies in existence (569,000 in 1983 [5.1]). This is largely due to the fact that the cost of entering the industry is low; financed purchase of a truck is sufficient, and the right-of-way, roads and highway infrastructure, is provided by the public sector. Motor carriers are subject to fuel taxes and other user fees. Many argue that motor carriers do not pay their fair share of the bill for the maintenance and repairs of roadway facilities that is commensurate with the damage caused to pavement and bridges. A TRB (Transportation Research Board) report concludes that "reducing the load on an axle by half, for example, 13 tons to 6.5 tons, would reduce the wear it causes by roughly a factor of 16" [5.2]. The same report specifies that in 1978, 89.2% of pavement wear on rural interstates was due to combination trucks with five or more axles; 10 years later this statistic had grown to 92.7%.

Certain characteristics of motor carrier operations, specifically truck loadings, terminals, and types of equipment, are noteworthy. Two common truck-loading schemes are truckload (TL) and less-than-truckload (LTL). In TL one shipment or part of one shipment occupies all cargo capacity, whereas in LTL either smaller shipments are consolidated to the truck's capacity at a terminal or small shipments are picked up on the way, until the truck's capacity is reached. Often shipments do not reach capacity. Also, empty backhauls are not uncommon. Dispatch managers and logistics specialists strive to minimize both of these inefficient states of operation.

Three main types of terminals facilitate the movement of freight by trucks: (1) consolidation terminals, where shipments are sorted and consolidated to form truckloads to specific destinations; (2) break-bulk, where large shipments are partitioned for distribution; and (3) relay terminals, where drivers are relieved by other drivers, given the following federal regulations [60 FR 38748, July 28, 1995], which requires that drivers do not drive for:

1.a). more than 10 hours following 8 consecutive hours off duty; or
1.b). any period after having been on duty 15 hours following 8 consecutive hours off duty;

or,

2.a). any period after having been on duty 60 hours in any 7 consecutive days if the employing motor carrier does not operate commercial motor vehicles every day of the week; or
2.b). any period after having been on duty 70 hours in any period of 8 consecutive days if the employing motor carrier operates commercial motor vehicles every day of the week.

Various cargo space and vehicle configurations are available. The cargo space can be configured as dry van (all sides enclosed), open top, flatbed, tank, refrigerated container, or other cargo-tailored configurations. The general vehicle form varies from a regular two-axle truck (10 to 15 m long), to twin trailers (20 to 33 m long). A handful of states permit the operation of triples comprised of a tractor and three-trailer combination. Economies of scale have caused a noticeable switch to larger rigs. Specifically trucks in the largest class (those exceeding 36 tons) reached 50,000 in 1992, a 180% increase from their 1982 level.

5.2.2 Railroads

In the United States railroads (railroad corporations) are mostly common carriers. There are a handful of private railroads and only one intercity passenger railroad, Amtrak, the operation of which is subsidized by the federal government. Amtrak was formed in 1971 with the purchase of failing passenger services of railroad companies. This section presents freight railroads; Amtrak as well as commuter railroads, and rail rapid transit serving urban areas are presented in Section 5.3.

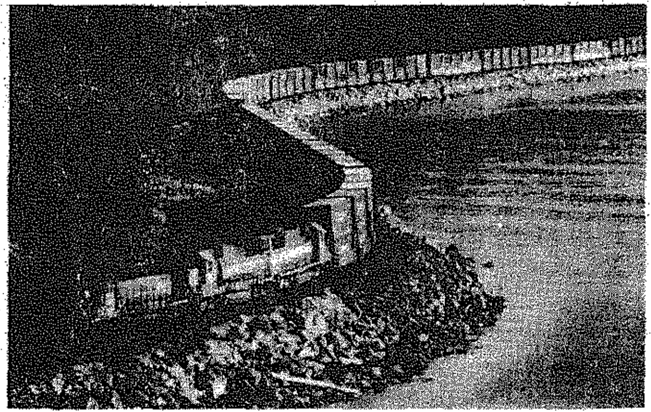


Figure 5.2.4 Railroads are a large volume and high weight mode of transportation. (From Transportation Board, *TR News*, 180, 1995.)

Until 1830 settlements and developments in the United States occurred mostly along coastal and waterway regions. (A comprehensive historical background of railroads is presented in Section 7.2.3.) Railroads are responsible for opening the horizon to western United States. Railroads reached their golden era between 1850 and 1880. Since then they have experienced continuous decline. There were 186 major (class I) railroads in 1920, 31 in 1984, and only 11 in 1995 (Fig. 5.2.4). A similar trend was observed for the total length of line in use, which declined from 378,000 km in 1939 to 266,000 km in 1982 and further down to 200,000 km in 1995. The railroad industry, however, still plays a vital role in the nation's transportation supply. In 1995 it accounted for 26% of the total intercity freight traffic and about 5% of the respective revenues. Based on 1982 annual revenues, the five largest railroads in the United States were Norfolk Southern/Santa Fe, CSX, Burlington Northern, Union Pacific, and Contrail [5.1].

The railroad industry serves all the contiguous states of the United States but individual railroad companies serve specific regions. Inter-regional shipments are switched among railroads at interchange points. This service characteristic tends to create rate discontinuities as well as delays in delivery. In 1997 Union Pacific and Santa Fe merged to form a coast-to-coast railroad.

The large geographic coverage and carrying capacity of railroads and the low rates charged are the major advantages of railroads. Another advantage is that railroads are more energy efficient and friendlier to the environment in terms of energy used and pollution emitted per ton-km carried, compared with motor carriers. Railroads are more suited to transport large volume or weight and low value commodities, such as coal, grain, oil and chemical products, pulp and paper products, forest products, and manufactured products, such as vehicles, machinery, parts, and equipment. The average haul in 1995 was 983 km.

Types of railcars include boxcars for general commodities, tankers for liquids and gases, hoppers for bulk materials, and flatcars. Flatcars are used for the transportation of containers (COFC: container on flatcar) and trailers (TOFC: trailer on flatcar). COFCs, as in Fig. 5.2.5, and TOFCs have increased the integration between railroads and motor carriers to provide intermodal transportation. Railroads transport over the long haul; trucks provide pickup and delivery service between the clients and the railroad terminals. The railroads' ownership of a fixed right-of-way (ROW) poses a service constraint that makes door-

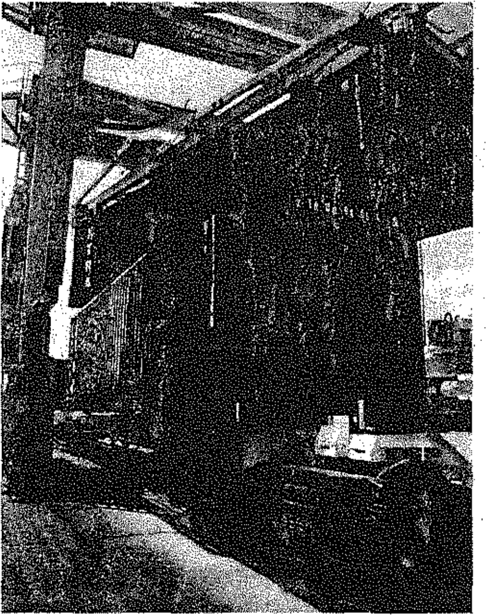


Figure 5.2.5 Containers on flat rail car (COFC). (From Transportation Research Board, *TR News*, 182, 1996.)

to-door service infeasible unless both the shipper and the receiver have side rail lines or rail yards along the railroad mainline.

Railroads are characterized by high fixed costs because they own and maintain their ROW, trackage, bridges, tunnels, switches, terminals (e.g., switching yards, interchanges, maintenance and storage facilities), and rolling stock (e.g., locomotives, cars, repair machinery). Variable costs include labor, fuel, electricity, insurance, taxes, depreciation, and equipment maintenance and upgrading; they are relatively low. Consequently substantial EOS are present.

Worsening financial status and benefits from EOS fueled the tendency for consolidations and mergers (i.e., the 31 class I railroads of 1984 were reduced to 10 by 1998). Also, automation and computerization has helped railroads to overcome major problems of car availability and distribution (i.e., having the right number and type of cars wherever needed), as well as empty backhaul (i.e., cars become available after being transported empty from somewhere else, which is an inefficient operation) [5.3].

5.2.3 Pipelines

Pipelines are mainly an underground form of transportation. They are often referred to as the hidden giants of the freight transportation industry (Fig. 5.2.6). This is because pipelines are both largely unknown to the general public and transport a large share of the intercity freight traffic (i.e., 16.3% of the total ton-km in 1995, but only 3% of the respective revenues).

Pipelines have certain unique characteristics: They transport a very limited variety of commodities that must be in liquid form, have a limited geographic coverage, and provide one-way transportation only. Pipeline corporations are mostly for-hire common carriers, although there are a few private carriers. They operate through a network of trunk (large

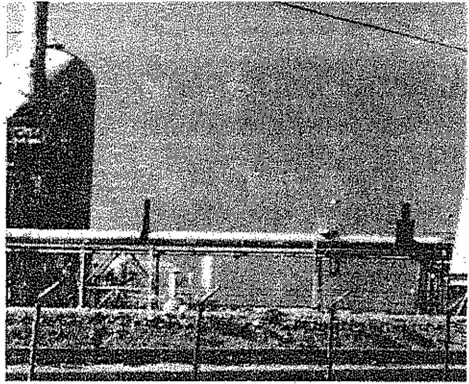


Figure 5.2.6 Above-ground pipelines can usually be seen in industrial complexes. (Photograph by P. D. Prevedouros.)

diameter, long haul) and gathering (smaller diameter, distribution) lines. Trunk lines are laid underground; gathering lines are often laid on the surface.

Typical products carried by pipelines are natural gas, crude oil, petroleum products, liquid chemical products, and coal slurry (crushed coal mixed with water). The average haul is approximately 711 km. Usually a minimum of 500 barrels (1 barrel equals 160 liters) is required for shipment, and rates are on a per barrel basis. Pipeline rates are extremely low; for example, in 1983 one barrel of crude oil could be sent from Texas to New York for $8 or less than 0.5¢ per liter [5.4]. Only ocean supertankers can match the rates charged by pipelines.

The number of pipeline companies is limited largely because of the high capital costs required for establishing a pipeline. These costs include the purchase or lease of land, construction of the pipeline(s) and pumping stations, and control infrastructure and terminal facilities. On the other hand, variable costs which include mostly labor, administration, and insurance are relatively low. For example, the Transalaskan Pipeline, which is among the few federally owned and operated pipelines, was built between 1974 and 1977 at a cost exceeding $9 billion, yet a labor force of 450 is sufficient to operate it [5.5].

The high fixed and low variable costs result in strong EOS. Parties interested in pipelines tend to consolidate and start with a large initial investment that tends to yield higher payoffs, partly because of EOS and partly because of inherent performance characteristics (i.e., a 30-cm pipe operating at capacity transports three times the liquid transported by a 20-cm pipe [5.1]). A typical trunk-line diameter is 75 cm (30 in.)

Sophisticated monitoring of facilities with computers as well as significant protection from the elements result in minimal loss and damage (e.g., quick detection of leaks) and in highly reliable delivery schedules. A negative characteristic is the slow service. However, the high accuracy and reliability of forecasted delivery times diminish the need for safety stock at the receiving end, whereas in essence pipelines offer free storage for as long as the order is on the way to delivery.

5.2.4 Water Transportation

Water transportation is the oldest form of mass freight transportation over seas or long distances (Fig. 5.2.7). Traditionally vibrant economic and industrial centers as well as population settlements were developed around sea ports and harbors (e.g., Alexandria in Egypt, Los Angeles, Mumbai, New York City, Singapore, Yokohama) and lakes and navigable

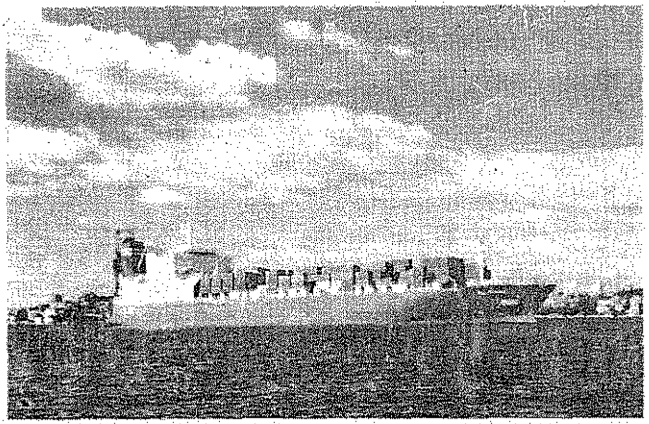


Figure 5.2.7 Water transportation is a very large volume and tonnage mode of transportation. (Photograph by P. D. Prevedouros.)

rivers (e.g., Chicago, Detroit, London, Paris, Moscow.) At present water transportation is an important mode for shipping raw materials, crude oil, and manufactured products among domestic and international points of trade.

Water transportation accounted for 24.4% of the total intercity ton-km of freight in 1995, and 4.1% of the revenues. Barges are the primary vessels of inland water transportation. Deep-sea water transportation includes shipments across the seas and between coastal areas. Common vessels used are liners (containers and break-bulk shipments), nonliners (bulk bottom) and tankers. Liners follow fixed routes and schedules and charge according to published tariffs. A special type of liner is the RORO ship (roll on, roll off), which carries vehicles and rolling equipment (i.e., construction equipment) much like a ferry boat. Tramp ships are those that can be hired, rented, or leased on a short-term basis, much like a taxi or a rental car.

The structure of the domestic water carrier industry is similar to that of the motor carrier industry. Domestic water carriers are either for-hire or private. The former can be either regulated or exempt, carrying bulk commodities. For-hire, regulated water carriers are either common or contract. Domestic water carriers operate in three distinct areas: (1) inland navigable waterways mostly rivers and canals, (2) the Great Lakes, and (3) coastal ports. Waterway and lake service is occasionally affected by ice formation and drought.

In general, water transportation offers low cost but slow service. Domestically operating carriers transport at a speed of approximately 8 km/h upstream and 16 km/h downstream along the Mississippi River and its tributaries. Both the shipper and the receiver need to have access to the waterway or port, otherwise connections with railroads or motor carriers are necessary. Since the capacity of vessels far exceeds the capacity of railcars and trucks, warehousing is needed for storage. Specifically the capacity of one 1350-ton barge is equivalent to 15 jumbo hopper railcars or 60 semitrailer trucks; the equivalent of the capacity of a 20,000-ton liner is 225 railcars or 900 semitrailer trucks [5.6].

Large harbors are primary intermodal facilities that in addition to warehousing provide the physical infrastructure necessary for freight transfer from sea vessels to railroads and trucks, and vice versa (Fig. 5.2.8.) Ports and docks are usually owned and operated by port authorities, the largest of which is PANYNJ (Port Authority of New York and New Jersey). These authorities provide comprehensive planning and development through their

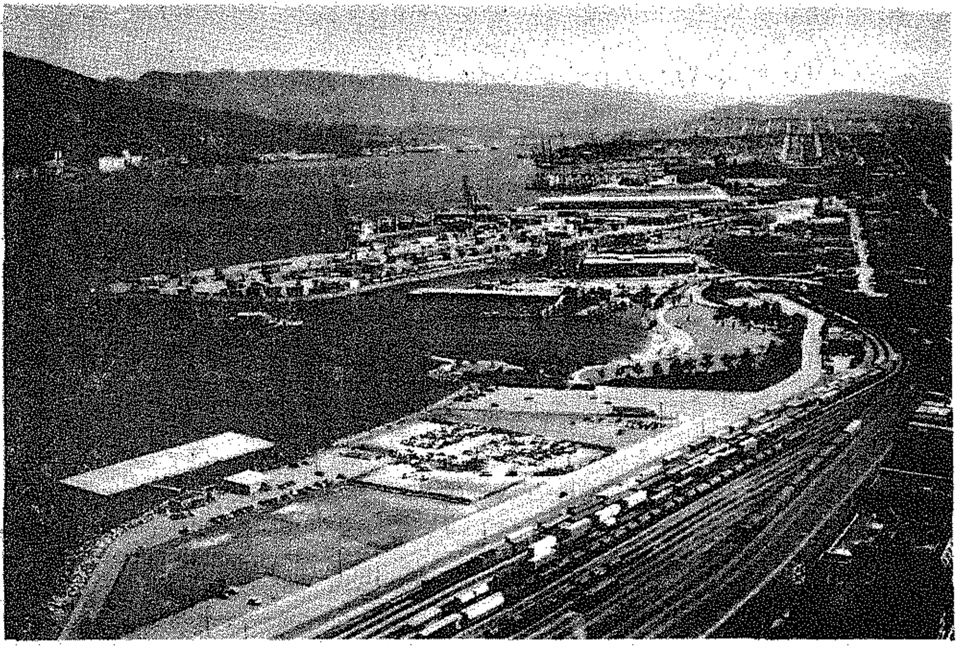


Figure 5.2.8 Intermodal operations among sea vessels, railroads, and motor carriers in Vancouver, B.C. harbor.
(Photo by P. D. Prevedouros.)

ability for substantial investments as well as promotion of trade and integration of industrial and shipping activities.

Typical products carried by domestic water carriers are coal, coke, iron, steel, grains, lumber, sand, gravel, stone, chemicals, petroleum products, paper, waste, and scrap material. Ocean vessels transport sugar, coffee, grains and foods, oil, petroleum products and chemicals, machinery, automobiles, and consumer products. Usually freight is subjected to multiple handlings and to rough waters; therefore expensive protective packaging is necessary for certain types of shipments.

Water transportation is the second least labor-intensive compared with other modes. The 1989 million ton-km per employee were 0.7 for motor carriers, 5.1 for railroads, 7.8 for water carriers, and 24.8 for pipelines [5.7]. The domestic waterway transportation industry is characterized by low fixed and high variable costs. Casualty and insurance make up a substantial part of the variable costs; they are necessary to cover loss and damage from the elements of nature. Part of the reason for the low fixed cost is that water carriers operate in free (deep-sea operations) or publicly financed (waterways, ports) media. Often private firms handling large amounts of commodities or special shipments invest in dock and terminal construction. By contrast, the fixed costs of deep-sea operations are substantially higher compared with inland water operations, and strong economies of ship utilization are possible.

Domestic water carriers compete with railroads for the shipment of dry, bulk commodities and with pipelines for the shipment of liquid commodities. The rates of international water carriers, primarily ocean liners, are set by cartel-like bodies called *steamship conferences*. This arrangement hinders competition but offers stability with respect to fluctuating currencies, fuel, and labor rates. Tariffs (rates) are usually made on a weight or measure (W/M) basis; that is, the shipment cost is based on the largest between *weight ton* (1000 kg) and *cubic ton* (12 m^3, or 40 ft^3.)

Due to benefits in taxes, labor, and safety requirements, ships are registered in countries that provide such shelters and economic benefits to ship owners. Cyprus, Liberia, and Panama are examples of countries that provide the so-called *flags of convenience.*

Given that a ship generates revenue only when it travels with load, empty backhauls and long docking times become costly. Containerization and mechanization of port operations reduced dock times for loading and unloading a ship's cargo from five days to less than a day.

5.2.5 Air Carriers

Historically, air travel is the newest mode of transportation (Fig. 5.2.9) and has been growing steadily since the first commercial flight. The previous edition of this book mentioned that "The worldwide expected growth in air travel from 1989 to 1998 will be as high as 5.6% per annum" [5.8]. This estimate was conservative; the actual 1996 and 1997 growth rates were 6.6 and 6.7%, respectively. For the first decade in the twenty-first century the Federal Aviation Administration (FAA) predicts an annual growth of about 4.2% [5.9], whereas airframe manufacturers Airbus Industrie and Boeing Commercial Aircraft predict annual growth of no less than 5.5%. The major explanation for this growth is the speed and convenience provided by air travel and the expansion of global business and tourism. The FAA estimates that in the mid-1990s about 60% of the population in the United States resided within 50 km of one of the 28 major hub airports.

The structure of this industry is simple. Air carriers are either private or for-hire. The commercial U.S. fleet approached 6000 aircraft by the turn of the century and the general aviation fleet exceeded 171000 aircraft. For-hire carriers are classified according to both their size or the type of service they provide. Size is determined by the annual revenues; three types are recognized: majors, nationals, and regionals. The 1997 majors were (listed alphabetically) Alaska, America West, American, Continental, Delta, Northwest, Southwest, TWA, United, U.S. Airways, FedEx, and UPS, the latter two being all-cargo carriers. Types of service include cargo only; air taxi, which offers passenger service on demand; commuter, which offers passenger service based on published timetables; charter for which the route and schedule are negotiated in a contract; and international. The establishment of

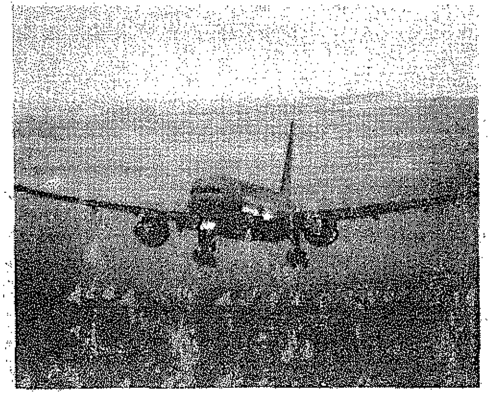


Figure 5.2.9 Air transportation provides the fastest mode of transport for people and high-value goods.
(From Transportation Research Board, *TR News*, 182, 1996.)

international routes requires treaties among countries and involves difficult negotiations involving both governments and airlines.

The advantages of this mode are fast terminal-to-terminal transportation, reliable service (except under extremely poor weather conditions), and attention to the customer (in-flight services and entertainment). Limited frequency of flights, capacity restrictions, and poor service to small cities are disadvantages. Long travel times to and from the airports, which are traditionally located at the outskirts of urban areas, as well as often long wait times (e.g., check-in, boarding, taxiing, baggage claim) increase the overall travel time.

Air cargo is growing fast worldwide, that is, 610 million ton-km in 1970 and 1520 million in 1988 [5.8]. In 1995 air cargo in the United States accounted for only 0.1% of the total domestic ton-km of freight, but for 2.3% of the respective revenues. The average cargo haul was approximately 2000 km. In 1996 the top five cargo carriers (listed by freight ton-kilometers, FTK), FedEx, Lufthansa, UPS, Air France, and Korean Air carried 25% of the worldwide FTK.

Advantages of air cargo include the smooth ride along with the automated and efficient handling facilities, whereas the high cost and the limited capacity are disadvantages. In general, high-value, emergency, and low weight items are shipped via air carriers. Such items include mail and documents, photographic equipment, parts and electric components or devices, perishables such as flowers and newspapers, medical components, and human organs. Airlines contract motor carriers to provide door-to-door service.

The airline industry is characterized by low fixed and high variable costs. Fixed costs include the aircraft fleet and maintenance facilities, computer reservation systems (CRS), management, logistics, airport counters, gates and baggage handling facilities, as well as offices in cities. Several of these, including aircraft, may be leased for short periods, and this makes them semivariable in nature. Variable costs include landing fees, which cover the use of local, state, or federal facilities (e.g., airport facilities, roadway access networks, aircraft traffic controls), labor and fuel (which combined account for 65% of the total variable costs [5.10]), maintenance, and commissions to travel agents.

There are EOS in the form of aircraft size utilization; this is usually evaluated on the basis of cost per seat-km, also known as ASK or available seat kilometer. Typically the use of larger aircraft, which have a lower cost per ASK, results in higher profit margins, provided that there is enough demand to fill the seats. Table 5.2.1 presents selected characteristics for four widely used commercial jet aircraft as of 1998. Aircraft seating configurations (the number of total seats as well as the number of seats per class) vary widely and airlines select different payload/range configurations (e.g., larger tanks provide longer range but reduce the

TABLE 5.2.1 Selected Characteristics of Commercial Aircraft

Aircraft	Seats	MTOW (kg)	Payload (kg)	Range (km)	Type
Boeing 747-400	416	397000	60000	13200	4 engine/2 aisle/wide body
Airbus A340-300	295	257000	51000	10800	4 engine/2 aisle/wide body
Boeing 757-200	235	116000	26000	7000	2 engine/1 aisle/narrow body
Airbus A319-100	145	75000	18000	4500	2 engine/1 aisle/narrow body

Source: Boeing Commercial Aircraft and Airbus Industrie Internet sites.

payload.) Thus the values are shown for illustration purposes. MTOW is the maximum take-off weight.

Large aircraft contribute to the solution of the increasingly pressing problem of airport congestion. Airport congestion is observed when the number of arriving and departing aircraft reaches or exceeds the capacity of a field. The capacity of an airfield is defined by a maximum number of landing and take-off slots in a given time period. At congested airports arriving aircraft are placed on a holding pattern (usually spirals in the airspace near the airport) and departing aircraft are queued on taxiways. Larger aircraft require longer but fewer landing and take-off slots for serving a fixed number of passengers. For example, 400 passengers served by one B747, which requires one landing slot and one gate, can be served by three A319 flights with respective requirements for landing slots and gates. Therefore consolidation of flights and use of larger aircraft less frequently may offer some relief to congested airports by decreasing the required number of operations per passenger served.

The flipside of this is that *frequency of departures* (which, along with *on-time performance*, constitute the two most important attributes of air travel for business travelers) is reduced. Given the conflict between airport congestion and departure frequency, airports (which have no power over the type of aircraft airlines choose to operate) institute congestion pricing (i.e., inflated landing fees) during peak hours in an attempt to shift demand to less congested hours. A good portion of these fees are usually passed through to the travelers in the form of increased fares.

There is strong competition among airlines for the acquisition of rights over high volume routes (airport and trade constraints determine the maximum number of flights allowed), as well as for passengers, through pricing. A consequence of the former is that low-density routes tend to be abandoned, therefore the service offered to small cities deteriorates. This is one of a few major drawbacks caused by the Airline Deregulation Act of 1978 [5.11]. On the other hand, the General Accounting Office (GAO, which is the investigative branch of the U.S. Congress) estimated that in 1994 inflation-adjusted air fares compared with 1974 air fares were 8 to 11% lower.

The SuperSaver fare system was first implemented by American Airlines in 1977 as a part of its yield management system (e.g., maximization of the yield per seat by using a time-variant pricing for reserving the seats of aircraft; as the day of departure nears, more discounted and complementary frequent flier seats are made available). These fares are accompanied by several restrictions which tend to make them unattractive to customers who can afford to pay the full fare price; primarily business travelers. In addition, frequent flier programs offering free trips or upgrades to a higher class of service have been developed to stimulate customer loyalty to a particular airline. Travelers have the freedom to select the lowest fare airline that is serving their travel plan, but by doing so they forego the opportunity of being awarded free trips after a sufficient number of points has been accumulated in their account.

Deregulation also fostered the development of hub and spoke networks (Fig. 5.2.10), where in essence travelers are consolidated at hub airports such as Atlanta, Chicago, and Denver and then flown to their destinations. This is a significant departure of the traditional linear network. The system was pioneered by Delta Airlines and refined by American Airlines [5.12].

The analysis of airline operations is complex. First, flights, departure times, and connections are developed. Then aircraft and crews need to be assigned to each flight. The problem becomes complex because the demand for each origin and destination pair needs

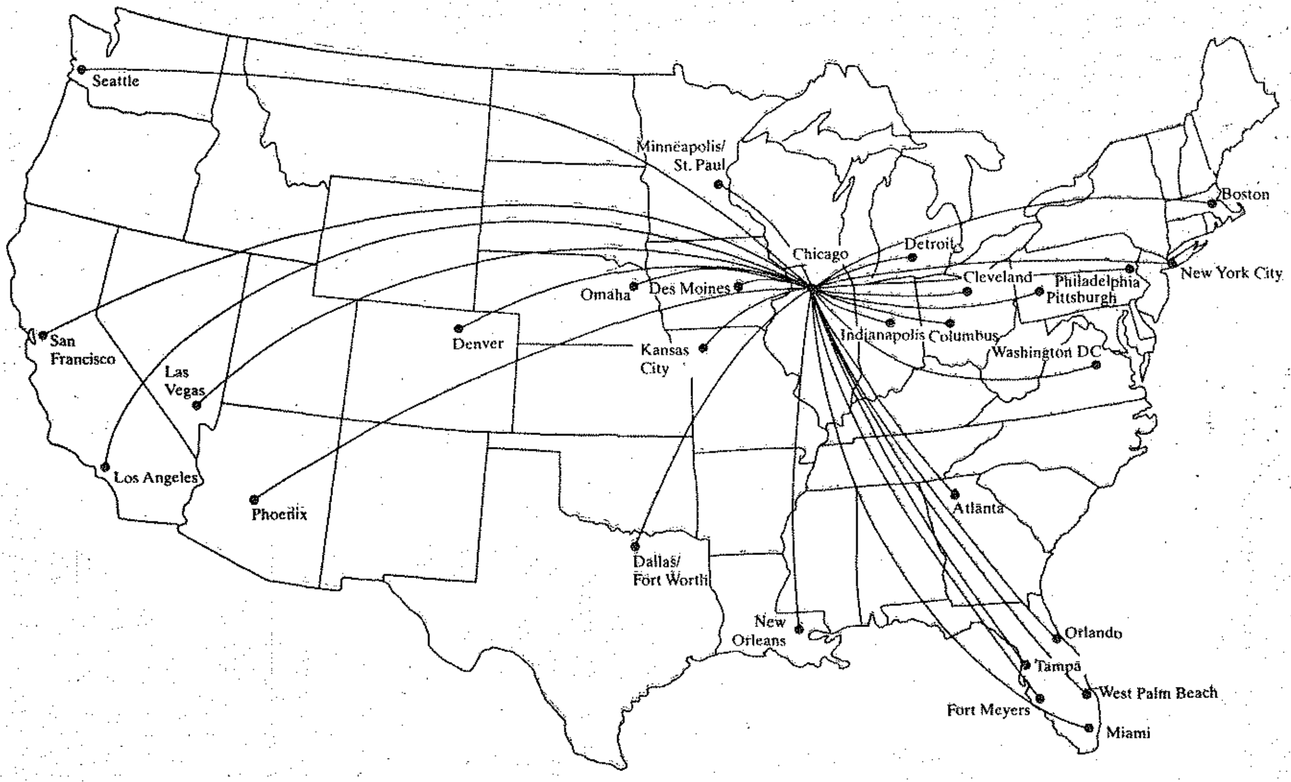


Figure 5.2.10 1988 routes of a regional U.S. commercial air carrier; the hub and spoke structure is clear.

to be satisfied, subject to constraints of variable aircraft sizes and capabilities, crews qualifications to fly only specific types of aircraft, and the preferences and seniority of crews.* For example, in 1997 United Airlines had 2200 daily flights spanning the entire globe conducted with 565 aircraft of 15 different types operated by 27800 flight crew (cockpit crew and flight attendants) and maintained by 24000 engineering staff. Scheduled aircraft maintenance and maximum crew hours of service are additional constraints.

A 1990s trend has been the formation of global alliances among airlines permitting code-sharing, gate-sharing, and coordinated scheduling. In 1998 *The Economist* reports that the world's 221 international air carriers were forming major alliances. The largest such alliance, the Star Alliance consisting of United-Lufthansa-Thai-SAS-Air Canada-Varig-SAA airlines flew 35.2% of the total revenue passenger kilometers (RPK) carried by alliances in 1996 [5.13]. This trend produced some economies and added flexibility in equipment use for the airlines and some added convenience for the traveler. Critics cautioned that alliances may stifle competition in specific areas. For example, in 1998 the American-British-Canadian-JAL-US Airways-Qantas airline alliance controlled 64% of the seats available between London and the entire American continent.

Through the FAA the federal government provides control of runway, taxiway, and flight operations with a dense network of air traffic control (ATC) facilities. It is important to realize that all twentieth-century commercial aircraft are not equipped so that the control crew knows about the traffic in the area where they fly. Visual identification of neighboring aircraft is hardly feasible given the speeds realized. Only ATC operators have the ability to channel and separate air traffic, both vertically and horizontally, so that operations commence safely. Given the large volumes of passengers carried daily and the disastrous outcomes of an accident, the role of the ATC system is of the utmost importance to air traffic operations. An alternative to traditional ATC control is the GPS-based *Free Flight* concept, which is discussed in Section 5.3.

As large harbors are typical intermodal facilities for freight, large airports are intermodal facilities for passengers. They provide connections between air and land modes. Within the grounds of airports, a large variety of conventional and custom-made vehicles (Fig. 5.2.11) perform a host of activities such as passenger, luggage and cargo transfers, aircraft refueling, inspection and maintenance, cabin cleaning and supply replenishment, ground guidance (e.g., the "Follow Me" car), security, and so on.

5.2.6 Express Package Carriers

Express package carriers are essentially a form of privately owned and operated mail service, which serve the general public in a way similar to the U.S. Post Office. Well-known U.S. express package carriers are FedEx (previously known as Federal Express), United Parcel Service (UPS), and DHL. The latter is actually the older air forwarder; it was established in 1969 in Honolulu by Dalsye, Hillblom, and Lynn for air freight transport to California. These couriers as well as smaller competitors expanded tremendously since the early 1980s, in both domestic and international markets. This growth is largely due to the speed, efficiency, and reliability of service provided (Fig. 5.2.12). Some express package carriers began service as *air forwarders* (e.g., FedEx and DHL), others began as *couriers* offering

*This is a typical union-negotiated item that applies to other modes as well.

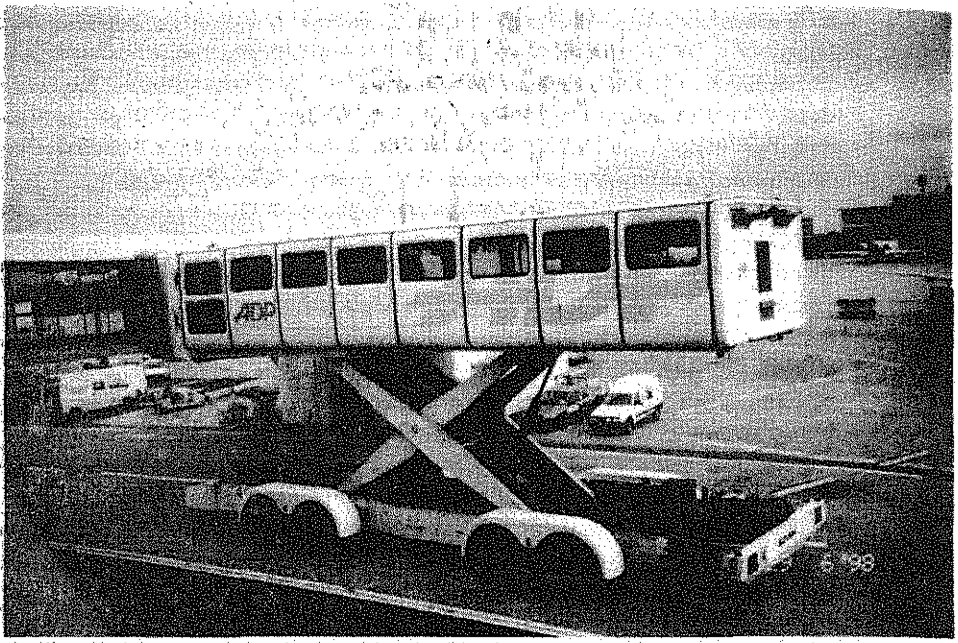


Figure 5.2.11 Custom-made buses at Paris airports (Aeroports de Paris) facilitate the increase of flight operations without terminal expansion. (Photograph by P. D. Prevedouros.)

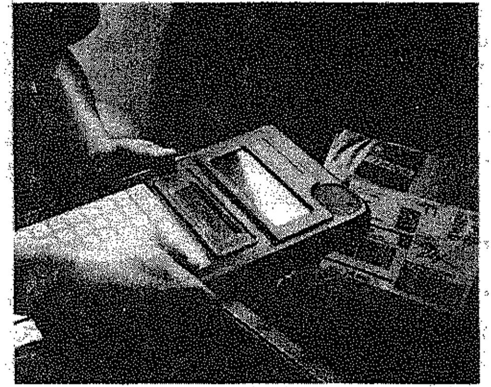


Figure 5.2.12 Delivery person uses portable tracking device to scan the bar-coded parcel prior to delivery. (Photograph by P. D. Prevedouros.)

express land service (e.g., UPS). Through expansions and acquisitions, these three major express package carriers have created a network of operations that offers worldwide door-to-door transportation service for packages up to 25 kg.

The equipment utilized by couriers includes a large number of sorting terminals, trucks, vans, and all-cargo aircraft. Vans are used for pickup and delivery, then packages and documents are sorted at the terminals, and then they are shipped via trucks or aircraft for the long haul. The incorporation of advanced electronics and package coding facilitates the real-time tracking of packages by the company, but also by the sender and the receiver via phone or computer.

In 1998 UPS shipped more than twice the number of packages shipped via regular mail. As of 1998 (for comparison, 1989 figures are given in parentheses), UPS served 200 (80) countries by utilizing a fleet of 147000 (116000) trucks and vans, and 197 (100) aircraft. Furthermore, 339000 (238000) employees manage an annual volume of 3.1 (2.8) billion packages [5.14, UPS on the Internet].

Express package carriers typically charge higher rates compared with the U.S. Post Office mail service (which until the mid-1990s was federally subsidized.) Their dependability, however, has made them the fastest and most reliable mode for shipping business documents and parcels.

5.3 INTERCITY PASSENGER TRAVEL

5.3.1 Major Modes

The basic purposes generating intercity passenger transportation are business, vacation, and personal reasons (e.g., visit family or friends, medical emergency). Travel modes that are available to serve intercity travelers include air travel via scheduled airlines, chartered flights (tourist groups), or private aircraft. Travel agencies, taxi and limousine service companies, mass transit authorities, airport and terminal authorities, car rental companies, local sightseeing services, hotels and restaurants, as well as the entertainment industry facilitate and complement intercity transportation modes.

Travel by bus is provided by two privately owned and operated national carriers, Greyhound and Trailways and several regional operators. Bus transportation has the most extensive geographic coverage; most cities with a population of 1000 or more are served.

Rail service is provided by Amtrak along a series of corridors connecting large urban areas. The most heavily utilized corridor is the one between Boston and Washington, DC, which also includes the cities of New York and Philadelphia. Between Boston-Washington, DC, and Chicago-New York City trains operate on upgraded lines at speeds between 140 and 200 km/h.

Cruising on passenger ocean liners has regained its popularity for vacation travel in the 1980s. Cruise ships offer all types of entertainment. Favored cruise regions are the Caribbean Islands, the coasts and islands of the Mediterranean Sea, the Hawaiian Islands, and the coasts of Alaska and British Columbia.

The automobile (i.e., private, rented, or company car) is the most readily available mode. This mode is among the slowest for long-distance trips. Part of the reason for the large volume of intercity travel by autos is the convenience of its use as well as the people's perception of costs. People tend to recognize out-of-pocket costs such as gasoline, tolls, and parking, and ignore other important costs, such as insurance, maintenance, and depreciation [5.15]. On the other hand, high utilization produces lower cost per km.*

*This simple principle has lead to an interesting practice in Singapore, as reported in *The Economist* [5.16]: "On an island that measures barely 35 km east to west and 20 km north to south, the average car, using some of the world's most expensive petrol, clocks up 20,000 km a year—much the same as in America. The reason is not far to see. A Mercedes E200, valued at about $35,000 before fees and taxes, would cost a Singaporean buyer a whopping $180,000."

5.3.2 Choice of Mode

The choice of mode for long-distance travel is heavily dependent on the sensitivity of the traveler with respect to time and cost. By and large, business travel is time-sensitive, vacation travel is price-sensitive, whereas travel for personal reasons may be either time- or price-sensitive, or both. The basic attributes of each mode are schedule, speed, cost, service offered, and perceptions regarding the service offered.

Schedule and speed prescribe the ability of the mode to serve passengers at the times they want and at the speed (or travel time) they require; for example, a same day round-trip from Chicago to New York can be accomplished by air travel only. Also, the location of the origin and destination points may restrict the mode choice set, or it may require the use of more than one long-distance transportation mode (i.e., air and bus).

Cost is a major consideration for most passengers. For a given distance rail and bus are the least expensive, with private or rented car following, and air travel coming last as the most expensive means of travel. Advance purchase of discounted fares may reduce the air transportation cost substantially. For example, in January 2000 the SuperSaver Honolulu-Chicago round-trip fare was between $610 and $850 depending on the itinerary, the coach class fare was around $1330 and the first class fare was around $4100.

Service is another important factor. Travel by private or rented car offers the convenience of having a car available at all times, which may be essential for some travelers (i.e., representatives and salespeople). Bus or rail offer few amenities on board. In contrast, airlines offer a wide variety of services on board (e.g., drinks, meals, minimart, multichannel music, and screen entertainment). Perceptions of passengers regarding the overall service offered by a mode compared with other modes (comparison between modes, e.g., auto versus bus or rail), or among providers of the same mode (i.e., American versus Delta Airlines, Avis versus Hertz Car Rental) affect the choice of modes and carriers.

Setting costs aside, the competitiveness of modes can be judged by their ability to provide fast service from origin to destination on a door-to-door basis (i.e., from the office in town A to the meeting place in town B, or from the house in town X to the hotel room in town Y). All modes except private auto and rented or company car provide terminal-to-terminal service. There are several time-consuming components before and after the main haul as well as in the terminals.

Typical travel-time components for rail and air transportation are listed here, along with the assumed time durations. These approximations are based on experiences in large urban areas, such as Cincinnati, Honolulu, Milwaukee, and Portland. Several of the following travel-time components are expected to be longer in large metropolitan areas with very busy airports (e.g., Atlanta, Athens, Chicago, Los Angeles, London, New York, San Francisco, Sydney, Toronto, Tokyo, etc.)

Rail	
Access origin terminal	20 min
Wait for train and board	15 min
. . . trip (terminal-to-terminal main haul) . . .	
Leave train and walk to exit from terminal	5 min
Access destination point	20 min

Air	
Access origin terminal	30 min
Check-in, walk to gate, and wait	30 min
Board and time until plane leaves gate	15 min
Taxiing and stops until takeoff	10 min
. . . trip (terminal-to-terminal main haul) . . .	
Landing and taxiing to gate	5 min
Permit to open doors, exit, and walk to baggage claim	10 min
Wait for luggage	15 min
Walk to exit terminal	5 min
Access destination point	30 min

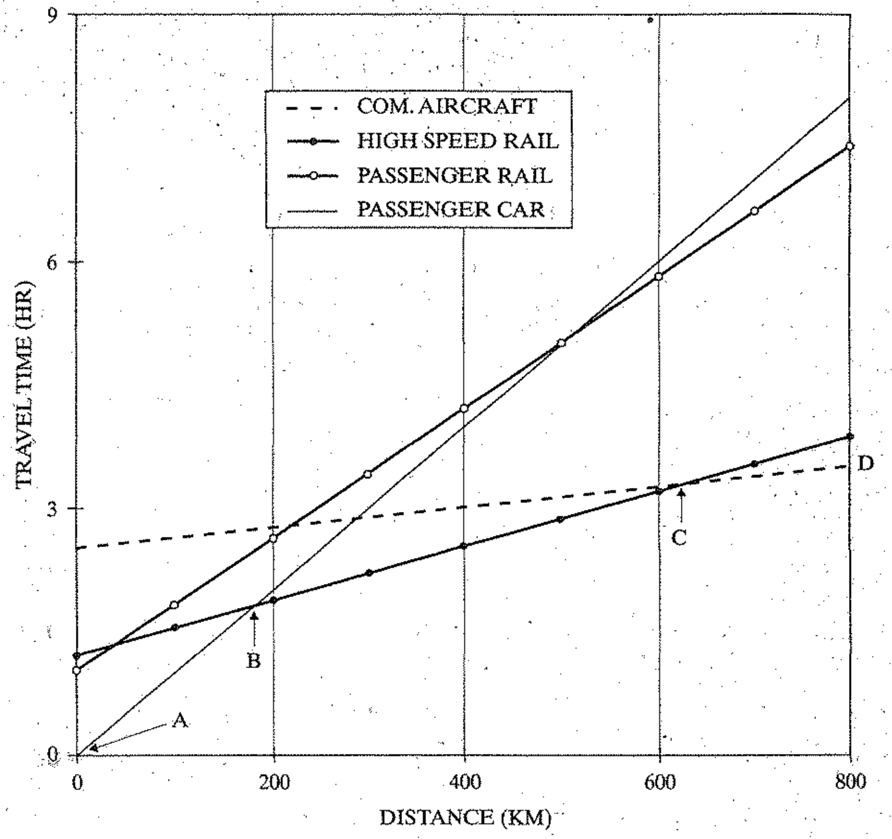
Hence the total nontrip time by rail is about 1 h long (which is similar for bus), whereas the total nontrip time by air is about 2 1/2 h long. These time estimates vary; they depend on origin and destination locations, transportation systems congestion levels, weather and equipment condition, terminal size and efficiency, and so forth. It is normal to expect, however, that a 2-h trip (terminal-to-terminal) by airplane should take at least 4 h from origin to destination (door-to-door). Considering the door-to-door time frame, it is likely that several modes may offer competitive service.

Fig. 5.3.1 compares four alternative modes of intercity transportation. The following average main haul speeds were assumed: 100 km/h for passenger car, 125 km/h for regular rail, 300 km/h for high-speed rail (a description is given later in this section), and 800 km/h for subsonic jet aircraft. Zero access and wait times were assumed for private and rented auto because this mode is, in most cases, readily available. Access and wait times equal to 1.0, 1.2, and 2.5 h were assumed for rail, high-speed rail, and air travel, respectively. (Note the value of the corresponding y-axis intercept.)

Auto results as the fastest mode for trips up to about 200 km, high-speed rail is most competitive for distances between 200 and 600 km, and air travel is the fastest for all trips exceeding 600 km. Regular rail and bus (not shown) are not competitive for any trip distance with respect to minimum travel time. These approximations are supported by the findings of the 1995 American Travel Survey conducted by the Bureau of Transportation Statistics (BTS) as shown in Table 5.3.1. In this table mode shares add up horizontally. They may not add up to 100% because the modes of chartered/tour bus and ship/ferry have not been included in the table.

TABLE 5.3.1 Intercity-Trip Distribution by Length and Mode

Approximate one-way trip length (km)	Distribution of trips based on length (%)	Car or similar vehicle (%)	Commercial aircraft (%)	Intercity bus (%)	Passenger train (%)
<500	29.6	95.5	0.7	0.3	0.5
500–800	26.6	91.6	4.1	0.4	2.6
800–1500	21.3	76.3	19.1	0.6	0.7
>1500	22.5	35.9	60.6	0.3	0.5



AB = For a distance of up to about 200 km, the passenger car is the fastest mode.
BC = For a distance of about 200 to 600 km, high speed rail is the fastest mode.
CD = For any distance exceeding about 600 km, air travel is the fastest mode.

Figure 5.3.1 Door-to-door travel-time comparison of four passenger transportation modes.

The same survey of 80000 U.S. households revealed that:

- The average intercity one-way trip length was 450 km for car, 640 km for bus, 660 km for rail, and 1750 km for air travel.
- Airport access was by:
  - private or rented car: 87% at the point of origin and 75.8% at the point of destination
  - taxi: 5.6% at the point of origin and 11.6% at the point of destination
  - limousine or shuttle: 5.9% at the point of origin and 10.6% at the point of destination
  - transit: 1.3% at the point of origin and 1.7% at the point of destination
- People aged 25 to 64 made two-thirds of the trips.

- People with a college degree made 42% of the trips, whereas people without a high school diploma made only 6% of the trips.
- Almost one-half of the trips were made by people in households with incomes of $50,000 or higher in 1995.
- About one-third of the trips occurred in the 3 months of July to September.

5.3.3 Emerging Intercity Modes

Advances in intercity transportation are expected to come in the form of high-speed trains capable of reaching 500 km/h, and second-generation supersonic (the Anglo-French Concord aircraft consists of the first generation) and/or suborbital aircraft capable of traveling from New York to Tokyo in less than 4 h. Thus the major objective of advanced intercity transportation technologies is the substantial reduction of travel times through high cruising speeds. Additional objectives are the reduction of fuel consumption, pollution, and noise.

High-speed rail is defined as a passenger rail transportation service with operating speeds of at least 200 km/h. High-speed rail debuted in 1964 in Japan (Shinkansen or Bullet train) and was followed in 1983 by France's TGV (or Train à Grand Vitesse). Amtrak's Metroliner passed the 200-km/h threshold in 1986. In 1994 Le Shuttle for vehicles and Eurostar for passengers were inaugurated in the Chunnel (the tunnel under the channel between the United Kingdom and France). These services reduced travel time between London and Paris from 7 to 3 h and caused a 40% reduction to air travel between these two cities.

U.S. Congress' 1991 Intermodal Surface Transportation Efficiency Act (ISTEA) required the commercial feasibility study of a high-speed ground transport (HSGT). In response to this, three levels of rail technology were assessed in a number of studies tailored to the needs of eight specific corridors (in California, Texas, Florida, the Northeast, etc.). They included *Accelerail*, which are technologies for the substantial upgrading of existing services, *New HSR*, which includes the latest advancements in traditional steel-wheel-on-steel-rail technology (such as the latest version of Shinkansen and the TGV), and *Maglev* [5.17].

Maglev (abbreviation for "magnetic levitation") trains essentially float on a magnetic cushion. Superconducting magnets interact with aluminum coils fixed on the guideway. Magnetic repulsion on the vertical plane lifts the train 3 to 13 cm from the guideway. Lateral magnetic repulsion on the horizontal plane enables the train to snuggle the guideway, thereby averting derailment. Longitudinal magnetic attraction and repulsion generate forward and backward propulsion [5.18]. Major advantages of these systems are low energy consumption, no emissions, practically noiseless operation, and minimal wear and tear due to the frictionless operation. German and Japanese industrial consortia experiment with real-world, full-scale magnetic levitation trains; at least one consortium from each of these two countries markets an implementation-ready system (Fig. 5.3.2). Urban transit as well as intercity versions cruising at speeds in excess of 500 km/h have entered revenue service.

Similar to high-speed rail, in air transportation there are two dominant new technologies: one fully applied, the other at the pilot-testing stage. Both are heavily dependent on electronics, such as computerized controls, satellite geolocation, and so on. The former is the fly-by-wire technology commercially introduced by the European aircraft manufacturing consortium Airbus Industrie (model A320 and derivatives). The technology utilizes electronic signals to command mechanisms that adjust control surfaces (i.e., flaps, ailerons,

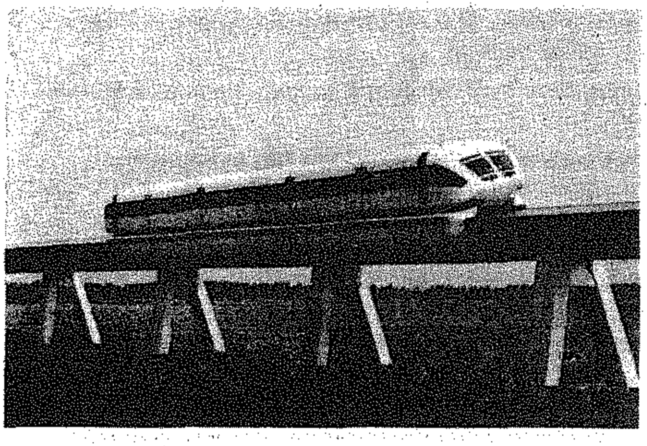


Figure 5.3.2 Magnetic levitation train prototype by Transrapid of Germany. (From Pennsylvania High Speed Intercity Rail Passenger, *Final Report*, 1990.)

rudder, and air brakes), whereas in conventional aircraft the pilot moves the control surfaces directly by operating levers or other mechanical devices [5.8]. Multiple on-board computers inspect flying conditions and pilot commands and suggest optimal actions or warn about potentially erroneous judgments. Fewer improved technology engines are necessary to propel the aircraft, whereas emissions, noise, and consumption continue to decline. Extensive use of composite materials further reduces aircraft weight. All these developments on conventional commercial jet aircraft make air travel more efficient, safer, and less harmful to the environment.

The other aviation advancement is Free Flight. Free Flight is preceded by a small-scale application called Flight 2000. Flight 2000 and Free Flight are technological evolutions of the National Airspace System (NAS) managed by the FAA. Free Flight permits users (captains of commercial and private aircraft) flexibility to plan and to fly their preferred route with limited or no interaction with the ATC. This is a major departure from current conditions according to which most if not all movements of aircraft are guided by the ATC. The Flight 2000 project is scheduled to begin in 2001; it integrates information via digital communications, automatic surveillance and broadcasts, weather processors, navigation satellites, advanced cockpit displays, and modified ATC and flight planning procedures. Free Flight is seen as the means for enhancing the capacity, efficiency, and safety of the air space. Free Flight architecture is expected to harmonize the global aviation system.

Although Free Flight can augment the capacity of air space, airport congestion will not be significantly improved. In response to this several large airports (e.g., Chicago's O'Hare Field) have banned all general aviation traffic, which is diverted to *reliever* airports. Some relief also could come from the tilt-rotor aircraft technology as applied to short-haul intercity air travel. The advantage of tilt-rotor aircraft is that they can take off and land vertically (VTOL), thus saving runway landing slots and delays for taxiing, and cruise as regular propelled aircraft (i.e., the engine and oversized propeller group gradually tilts from the vertical position of a helicopter to the horizontal position of a fixed-wing propeller aircraft). The military version (V-22 Osprey) of Boeing-Bell tilt-rotor aircraft can carry 8 to 40 passengers in various commercial configurations and travel at speeds of up to 600 km/h [5.19].

5.4 SUMMARY AND COMPARISONS AMONG MODES AND COUNTRIES

This chapter presented the major modes of transportation. Motor carriers are ubiquitous and provide door-to-door service. Railroads are best suited for transporting bulky products in large quantities. For liquid commodities pipelines offer fast, reliable, and inexpensive, transportation. Intercontinental transportation of freight is almost exclusively made by ocean liners and tankers. Air carriers provide fast transportation of people over long distances and high value, low volume goods. Express package carriers offer fast and guaranteed delivery mail and package shipping service.

Table 5.4.1 provides a quantitative summary based on Bureau of Transportation Statistics reports [5.20, 5.21] of various freight transportation modes, including truck, rail, combination of truck and rail, pipeline, water, air, and courier. Rail and truck account for 50% of the ton-km transported and more than 75% of the value of the shipments.

TABLE 5.4.1 Summary of Selected Characteristics of Major Freight Transport Modes

Freight transport mode	Metric ton-km (million) 1995	Metric ton-km (%) 1995	Selected size characteristics (United States) 1995	Value of shipment (U.S. $/kg) 1993	Average haul distance (km) 1995	Value of shipments 1993 (%)

Truck (for hire & private)	1260000	24.0	58 million light trucks;
			6.9 million freight trucks;
			250000 km of National Highway System roads
				1.54	220	71.9

Rail	1365770	26.0	11 class I companies;
			18812 locomotives,
			1.2 million freight cars;
			200000 km of track
				0.45	983	4.0

Truck + rail	54590	1.0	—	4.63	1493	1.4
Water	1283940	24.4	40000 vessels under U.S. flag (combined Great Lakes, inland, and ocean fleets)	0.45	670	4.1
Pipeline	859110	16.3	Liquid: 160 companies and 320000 km of pipe Gas: 150 companies and 2 million km of pipe	0.45	711	2.9
Parcel, postal, courier	19060	0.4	—	67.70	1121	9.2
Air (includes truck + air)	5810	0.1	681 airports serving large certif. carriers; 5567 certificated air carrier aircraft; 86 large carriers	100.50	2056	2.3
Other	408620	7.8	—	—	655	4.2
Total	5256900	100.0	—	—	480	100.0

Source: Ref. [5.20]

The safety statistics among modes also vary widely, as shown in Table 5.4.2. About one half of all accidental deaths in the United States are attributable to transportation [5.21]. This table shows that almost all transportation crashes (i.e., 95.8% in 1995) involved motor vehicles. Crash rates for all modes have been decreasing, but a notable slowing has occurred in the 1990s. Only general aviation displays a consistently decreasing crash trend.

TABLE 5.4.2 1970–1995 Trend in U.S. Transportation Fatalities

Year	Air carrier(1)	General aviation	Motor vehicles(2)	Rail, transit	Water transp.(3)	Pipeline
1970	146	1310	52627	785	178	30
1975	221	1252	45442	575	243	25
1980	143	1239	51924	584	206	19
1985	639	955	44407	454	131	33
1990	96	766	45297	938	85	9
1995	229	732	42377	841	46	21

Notes: (1) Includes commuter and taxi service, (2) includes accidents at rail crossings, (3) excludes recreational water accidents.

Source: Ref. [5.21].

Selected general and transportation characteristics of 12 countries and the United States are compared in Table 5.4.3. The United States is among the least densely populated countries (which has necessitated a vast network of highways and railways and has made the airplane the primary long-distance mode.) The United States also is among the most urbanized nations and has both the highest auto ownership per capita and among the cheapest prices of fuel (which encourage the use of automobiles, often by single occupants).

TABLE 5.4.3 Selected Characteristics of 12 Countries and the United States

	Country	Area (000 km^2)	Population (000; 1994)	% popul. growth (1985–1994)	People/km^2 (1994)	% urban popul. (1994)	GDP growth (1970–1994)	Cars/km^2 (1994)	Fuel price in mid-1996 ($/liter)
1	Brazil	8512	159100	1.8	19	77	4.8	1	n.a.
2	Canada	9976	29251	1.3	3	77	3.6	1	0.42
3	China	9561	1190918	1.4	125	29	8.7	<1	n.a.
4	France	549	57960	0.5	106	73	2.5	45	1.20
5	Germany	357	81407	0.5	228	86	n.a.	112	1.04
6	Hungary	93	10161	−0.4	109	64	2.2	22	0.81
7	India	3288	913600	2.0	278	27	4.5	1	n.a.
8	Italy	301	57190	0.1	190	67	2.7	99	1.21
9	Japan	378	124960	0.4	331	78	3.7	113	0.96
10	Mexico	1973	88402	2.2	45	75	3.4	4	0.32
11	Russia	17705	148366	0.5	8	73	n.a.	1	n.a.
12	UK	245	58375	0.3	238	89	2.3	97	0.92
13	USA	9373	260651	1.0	28	76	2.8	14/20[a]	0.33

TABLE 5.4.3 Selected Characteristics of 12 Countries and the United States—*Continued*

			Transport characteristics per 1,000,000 population				
Country	Cars per capita (1994)	% paved	Railroad tracks (km)	Roads, all (km)	Inland waterways (km)	Pipelines, liquid (km)	Airports, all
Brazil	0.08	10	192	10164	314	36	22
Canada	0.49	30	2672	29038	103	806	47
China	0.00	17	55	864	92	9	0
France	0.43	54	588	26073	258	130	8
Germany	0.49	79	534	7816	64	93	8
Hungary	0.20	44	766	15620	160	118	8
India	0.00	49	68	2156	18	6	0
Italy	0.51	91	341	5340	42	67	2
Japan	0.34	68	219	8899	14	3	1
Mexico	0.09	35	277	2741	33	450	23
Russia	0.05	78	1038	6295	681	425	17
UK	0.41	100	289	6168	39	67	9
USA	0.51/0.73[a]	61	817	24155	157	1235	70

[a] For USA, the second number also includes light trucks.

Caution: The Cars column numbers are approximate because the definition of vehicle types differ among countries. Some of the paved roads include only graveled roads (e.g., Russia). Most statistics are from 1995; some are from the first half of the 1990s.

Source: Ref. [5.21]

Canada has a roadway kilometrage commensurate to its area (which is larger than the United States and it boasts the highest kilometrage of railway track. The UK seems to be the only nation which has paved all its public roads. Russia has the longest kilometrage of navigable inland waterways. The United States has the longest kilometrage of pipelines and the most airports. The reader is cautioned that several of the statistics are per one million population. For example, in total the United States has more than 18000 airports, air fields, and public landing strips.

EXERCISES

1. The Concrete Products Corporation (CPC) is considering the purchase of a transportation company to facilitate the distribution of its products. The options are (1) a trucking company with fixed assets of $3 million, variable costs of 5¢/ton-km, and annual fixed costs of $250,000 and (2) a small railroad company with fixed assets of $15 million, variable costs of 3¢/ton-km and annual fixed costs of $600,000.

   Given that CPC will be shipping 30,000 tons of products over a 1000-km corridor each year for the next 10 years, which transportation company should CPC purchase and what other factors of motor carrier and railroad transportation should it consider before the purchase? Costs are expected to increase by 3% and shipments by 8% per annum. Assume that assets remain constant over the time period considered.

   (*Note:* To estimate the present size of quantities increasing for the next *n* years, consult Chapter 12.)

2. The cost function of a large railroad corporation is $Y = 10^7 + 0.5 \cdot T$, where Y is the total cost of shipping in U.S. dollars and T is the tons shipped. Last year the company charged on average 88¢ for each ton of freight. Their annual shipments totaled 48 million tons. This year they are considering geographical expansion through the purchase of a smaller railroad corporation that last year shipped a total of 21 million tons. Economists estimated that the total cost function (for the merged corporations) will be $Y = 10.5^7 + 0.3 \cdot T$, while 10% more freight should be expected due to the better geographic coverage, at a price discounted by 8¢.

   Show that the large and the merged railroad realize substantial economies of scale (EOS). Which railroad realizes greater EOS? Use a numerical example or a graphic for proof. Show numerically that the large railroad should merge with the smaller one.

3. Coal Distributors Corporation (CDCorp) is considering adding one coal slurry pipeline between their main facility and a location M. The length of the line is 200 km. The cost of placing a pipeline varies by size:

Gathering line:	\$250,000 per km (throughput: 100 l/s)
Trunk line:	\$400,000 per km (throughput: 400 l/s)

   A third option of CDCorp is to lease trucks at a cost of \$1/km-s at a full truckload. One truckload is equivalent to a throughput of 50 l/s (liter per second). Only full truckload shipments will be made.

   The demand for the first 5 years is estimated at 150 l/s; it is expected to drop to 80 l/s after that point and to diminish after another 5 years. If the operating cost is 2¢/l (per liter) and 1.5¢/l for the trunk and the gathering pipeline, respectively; and if CDCorp charges a flat rate of 3¢/l, which option should CDCorp choose? (Round all monetary estimations to the closest million.)

4. Safeway motor carrier has a contract with Byte Computer Company, which corresponds to 0.24 million tons of freight per annum. Safeway operates a fleet of trucks with 12-m trailers, each providing a capacity of 80 m^3. The total cost per kilometer for each tractor-trailer unit is \$2. Safeway charges Byte 25¢/ton-km. The typical shipment from Byte is a full truckload transported over 500 km. Recently Byte requested a rate decrease from 25 to 22¢. Safeway has agreed to lower the rate. Given that the freight density of Byte's shipments is 120 kg/m^3, should Safeway keep its current fleet or upgrade to a fleet of 15-m trailers with a capacity of 95 m^3 each and a total cost equal to \$2.10/km, which includes the cost of upgrade?

5. The risk of an accident during a commercial airliner flight may be assumed as follows: 36% during takeoff and climb, 5% during cruise, 56% during decent, approach, and landing. The remaining 3% risk is during loading, unloading, and taxiing; this component should be ignored in this exercise.

   Considering the commercial jet aircraft data supplied here, and assuming that on average 80% of the seating capacity is utilized and that the average trip is two-thirds of the maximum range, calculate the risk factor for each aircraft for every one billion passenger-km. Then set the highest risk estimate equal to 100 and scale the other three estimates. Interpret the results based on the scaled estimates.

Aircraft	Passenger capacity	Maximum range (km)	Takeoff, climb, distance (km)	Decent, approach, land distance (km)
B-747	380	8100	100	70
DC-10	250	7500	100	70
A-310	200	1500	65	55
B-737	105	1800	65	55

Hint: Estimate the flights required to serve one billion passenger-km and estimate the risk for each trip segment according to its length.

6. An airline serving the Hawaiian Islands is planning to introduce service between Honolulu (Oahu) and Kahului (Maui). The airline wants to choose the right size of aircraft for a particular set of new flights. Two types of aircraft are available for scheduling:

McDonnell Douglas MD-80 with 95 seats and $45 cost per seat
McDonnell Douglas MD-11 with 220 seats and $32 cost per seat

The cost per seat is specific to the one-way trip between the two cities that are considered.

Given the following set of four flights and the expected passenger demand, which aircraft or combination of aircrafts should the airline select in order to minimize the actual cost per seat?

	Expected number of passengers				
Flights	1	2	3	4	5
Honolulu-Kahului	120	150	175	130	100
Kahului-Honolulu	130	140	185	120	215

Flights cannot be combined. Passengers who cannot get in a flight are lost to competitors.

7. An international air carrier plans to purchase a jumbo-class aircraft to serve the Chicago-London city pair. Restrictions in landing slots at Heathrow Airport allowed the purchase of five landing/take-off slots, one per weekday. The carrier does not intend to offer weekend service. The following table presents passenger demand estimates and fares charged by competitors:

Trip	Five-weekday demand	One-way fare ($)
Chicago-London-Chicago	975	400
Chicago-London	230	680
London-Chicago-London	880	475
London-Chicago	400	775

The round-trip Chicago-London-Chicago is 13,000 km, and the aircraft choices and costs are:

Aircraft	Seating capacity	Oper. Cost seat-km	Fixed cost/year
B-747	385	2.57¢	$4.5 million
DC-10	235	2.97¢	$3.7 million

Should the airline enact service and which aircraft should be chosen? Solve based on profit maximization.

8. The OTR bus company must replace or rebuild their fleet of 100 buses because they have reached their useful life. The three options available are to purchase a new standard bus, to purchase a superbus, which is more costly but lasts longer, and to rebuild available buses. The costs of these options are given below: (O&M is operating and maintenance.)

Option	Capital cost	Life years	Annual O&M cost	Overhaul cost	Overhaul year
Standard	$215,000	12	$40,000	$50,000	5th
Superbus	$275,000	20	$47,500	$25,000	5th, 10th, 15th
Rebuild	$111,000	6	$45,000	none	none

Which bus is preferred at a 6% discount rate? (*Hint:* Annualize costs and sum up; see Chapter 12 for the appropriate formulas.)

9. A businessman residing in Chicago considers his options for a trip to Detroit. His options are private car, rental car, bus, or airplane. Given the following data, suggest the best mode for his travel. Distance between cities (one way) = 425 km. Estimated access travel at origin and destination = 38 km at each city; the access distance and access trips are the same for all modes.

Costs: (1) private auto: 20¢/km (all costs combined); no access mode required; (2) rental car (2 days): $50/day plus $12/day for insurance and tax; 11 l/100 km fuel efficiency and gas price is 40¢/l; no access mode required; (3) bus: round-trip fare $55; access mode required; (4) air: round-trip fare $100; access mode required.

Access modes and costs: in Chicago taxis charge $2 plus 20¢/km, buses charge $1 per ride (assume two rides); in Detroit taxis charge $2.40 plus 15¢/km, buses charge $1.5 per ride (assume four rides). The door-to-door travel times by mode are as follows:

	Private auto	Rental car	Bus	Air
Best	4.5	5.0	6.0	1.5
Worst	6.0^{a}	6.5^{a}	8.0^{b}	3.0^{b}

[a] Accounts for potentially congested conditions.
[b] Use of bus for access.

In order to make his selection, the businessman assumed a disutility function (a measure of "discomfort" due to the cost and travel time encountered):

$$\text{Disutility} = (\text{total trip cost})/5 + 8 \cdot (\text{one-way travel time})$$

Which mode did the businessman select? (Round out all cost estimates to the nearest integer.)

10. Select three countries of your choice excluding the United States from Table 6.2.4 and make a narrative and quantitative comparison among them and with the United States in about 500 words.

11. A study in California concluded that high-speed rail is an inferior alternative to air and car travel between Los Angeles and San Francisco, even when the social costs of accidents, noise, and air pollution are taken into account. These costs are much smaller for high-speed rail than for air and car travel. The researchers actually accounted for zero accident costs based on the no-accident operation of both Shinkansen and TGV. The per passenger-km cost estimated by this study is shown in the following table [5.22]. It is suggested that even if the passenger (possibly optimistic) forecasts for high-speed rail are doubled, the amount of required subsidy would still be very high: $19 per high-speed rail passenger versus $2.5 and $0.75 for air and car travel, respectively, assuming a 600-km distance between these two cities. What other reasons could

necessitate the planning and implementation of high-speed rail despite the inferior economic statistics shown here?

Cost per Passenger-Km of Three Intercity Modes

Mode	Total cost (¢)	Revenue (¢)	Subsidy (¢)
Air	2.43	2.02	0.41
Car	2.05	1.93	0.12
HSGT	15.60	6.00	9.16

REFERENCES

5.1 COYLE, J. J., E. J. BARDI, and J. L. CAVINATO, *Transportation*, West Publishing Company, 2nd ed., 1986.

5.2 TRANSPORTATION RESEARCH BOARD, *New Trucks for Greater Productivity and Less Road Wear,* Special Report 227, National Research Council, 1990.

5.3 HAGHANI, A., and M. DASKIN, "A Combined Model of Train Routing, Makeup and Empty Car Distribution," *The Logistics and Transportation Review,* Vol. 23, No. 2 (1987): 173–188.

5.4 AMOCO EDUCATIONAL SERVICES, *Oil on the Move,* Chicago, IL., 1983.

5.5 WOOD, D., and J. JOHNSON, *Contemporary Transportation,* Petroleum Publishing, Tulsa, OK, 1975.

5.6 CHATTERJEE, A., G. P. FISHER, and R. A. STALEY (Eds.), *Goods Transportation in Urban Areas,* American Society of Civil Engineers, 1989.

5.7 ENO FOUNDATION FOR TRANSPORTATION, *Transportation in America: A Statistical Analysis of Transportation in the United States,* 8th ed., Washington, DC, 1990.

5.8 AIRBUS INDUSTRIE, *Market Perspectives for Civil Jet Aircraft,* Toulouse, France, 1990.

5.9 FEDERAL AVIATION ADMINISTRATION, *FAA Aviation Forecast: Fiscal Years 1995–2006,* U.S. DOT, 1994.

5.10 AIR TRANSPORT ASSOCIATION OF AMERICA, *Air Transport 1983,* Washington, DC, 1983.

5.11 TRANSPORTATION CENTER, NORTHWESTERN UNIVERSITY, *Transportation Deregulation and Safety,* Conference proceedings, Evanston, IL, 1987.

5.12 AVIATION WEEK AND SPACE TECHNOLOGY, *American's Carty Walks in Crandall's Footsteps,* 1998, p. 37.

5.13 THE ECONOMIST, Business: *Come Fly with Me,* June 20, 1998, pp. 69–70.

5.14 UNITED PARCEL SERVICE OF AMERICA INC., *1989 Annual Report to Shareholders,* Greenwich, CT, 1989.

5.15 METCALF, A., "The 'Misperception' of Car Running Costs and Its Impact on the Demand for Energy in the Transport Sector," *Proceedings of the World Conference on Transport Research (1980):* 1583–1603, London.

5.16 THE ECONOMIST, "A Survey of Commuting: To Travel Hopefully," p. 15, September 5, 1998.

5.17 FEDERAL RAILROAD ADMINISTRATION, *High Speed Ground Transport: On Track for the Future,* U.S. DOT, 1998.

5.18 HEIRICH, K., and R. KRETZSCMAR, *Transrapid MagLev System,* Hestra-Verlag, Darmstadt, Germany, 1989.

5.19 The Boeing Company, *Civil Tilt Rotor (CTR) 2000*, October 1994.

5.20 Bureau of Transportation Statistics, *Transportation in the United States: A Review*, U.S. DOT, 1997.

5.21 Bureau of Transportation Statistics, *Transportation Statistics Annual Report 1997*, BTS97-S-01, U.S. DOT, 1997.

5.22 Kanafani, A., "Balancing Act: Traveling in the California Corridor," *Access*, No. 11, University of California Transportation Center, Fall 1997.

6

Urban and Intelligent Transportation Systems

6.1 INTRODUCTION

This chapter begins with a historical sketch of urban development and urban transportation modes in the United States (Fig. 6.1.1). This is followed by a presentation of contemporary urban transportation modes. Urban transportation issues are presented with particular emphasis on traffic congestion and congestion alleviation strategies.

The chapter includes an in-depth presentation of intelligent transportation systems (ITS) with descriptions of user services, architecture, and mature ITS applications such as detectors, traffic signal systems, freeway management (automatic incident detection, incident management, and ramp metering), electronic road pricing, and automatic vehicle classification.

This chapter concludes the transition from a vehicle and facility-specific perspective to a systems perspective, which was initiated in Chapter 5. After reading these two chapters the reader will be ready to tackle the subsequent topics of planning and demand forecasting as well as the analysis of transportation impacts. All units in this chapter are in the metric system.

6.2 DEVELOPMENT OF CITIES AND TRANSPORTATION MODES

The movement of people and goods within cities is a special area of transportation that has several unique characteristics. Transportation is one of the most important components of urban infrastructure that is necessary for ensuring the vitality of an urban area. An efficient network of transportation services is required to support the complex activity patterns within cities. Furthermore, there is a strong connection between transportation and city growth. Transportation can promote or hinder development and vice versa; that is, vibrant,

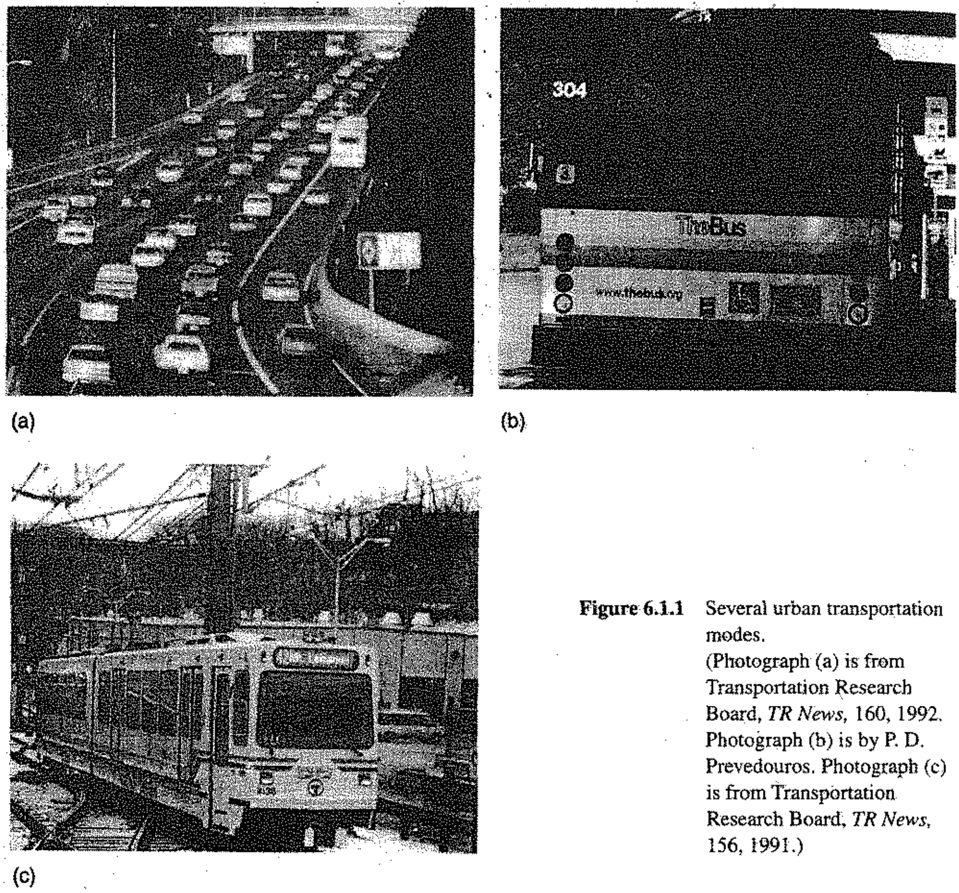


Figure 6.1.1 Several urban transportation modes. (Photograph (a) is from Transportation Research Board, *TR News*, 160, 1992. Photograph (b) is by P. D. Prevedouros. Photograph (c) is from Transportation Research Board, *TR News*, 156, 1991.)

growing urban areas invite expansion or implementation of new transportation facilities and services.

The historical evolution of urban areas suggests that population settlements first occurred beside accessible harbors, lakes, canals, and rivers. These settlements evolved into cities. Later on cities developed at crossroads of major railroad lines and highway routes.

The intimate interaction between transportation and urban development can be best put in perspective by observing the historical growth of U.S. cities and their transportation networks. Figures 6.2.1(a) and 6.2.1(b) illustrate the *parallel* chronological evolution of cities and urban networks. In the beginning towns consisted of a main street where most businesses and services were located. Residences were scattered along secondary roads. These main streets were usually the "urban" parts of trails connecting neighboring settlements.

Expansion over time, particularly after industrialization, created a central city core where most businesses were concentrated. In many postindustrial cities the core was served by a grid-shaped road network on which horse-drawn carriages and trams were rolling. The city core was surrounded by residential neighborhoods. Public transportation (carriages and trams) connected the neighborhoods with the city core. Industrial sites were typically found

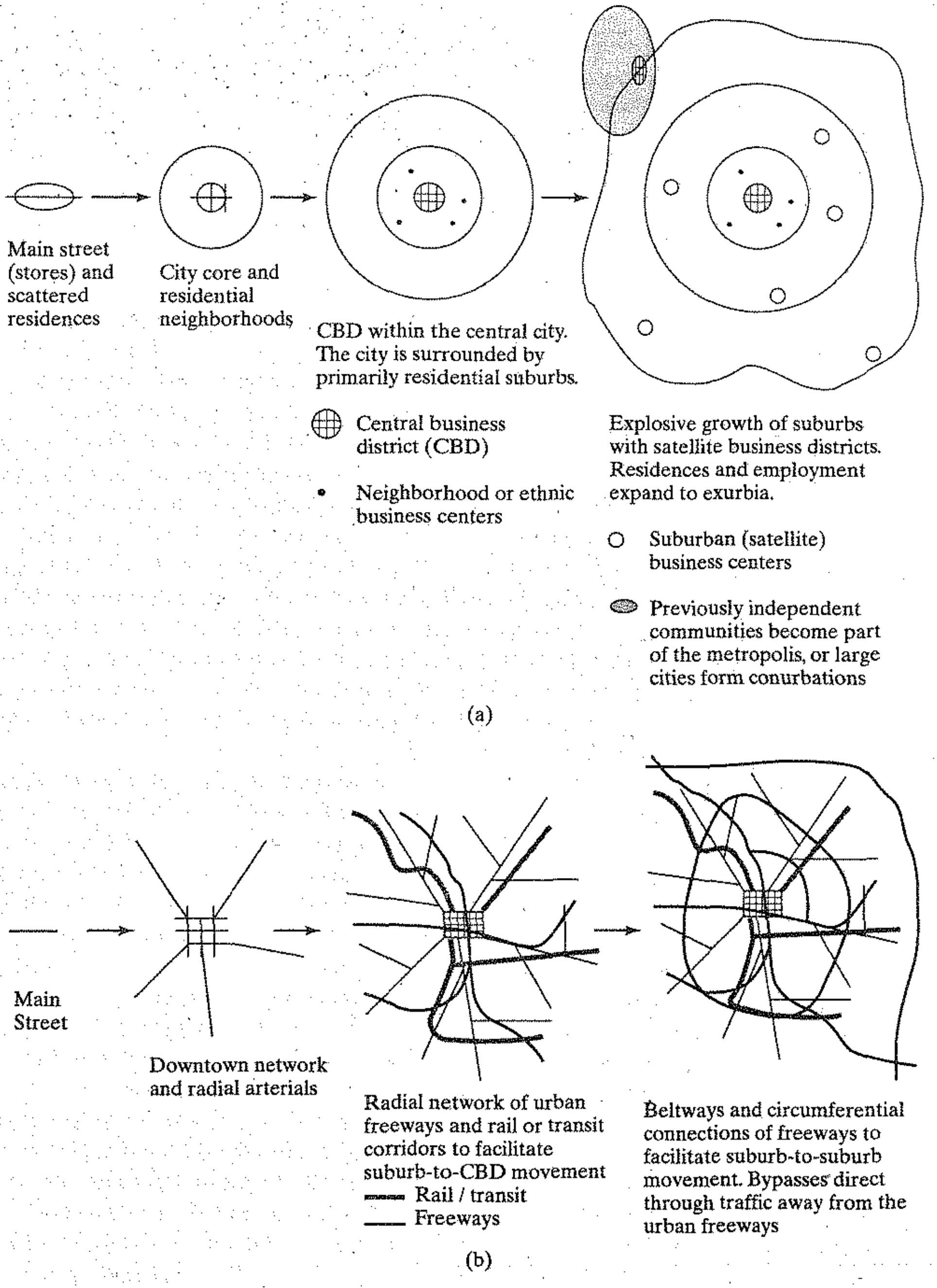

Main street (stores) and scattered residences
City core and residential neighborhoods
CBD within the central city. The city is surrounded by primarily residential suburbs.
⊕ Central business district (CBD)
• Neighborhood or ethnic business centers
Explosive growth of suburbs with satellite business districts. Residences and employment expand to exurbia.
○ Suburban (satellite) business centers
⬬ Previously independent communities become part of the metropolis, or large cities form conurbations
(a)
Main Street
Downtown network and radial arterials
Radial network of urban freeways and rail or transit corridors to facilitate suburb-to-CBD movement
▬ Rail / transit
— Freeways
Beltways and circumferential connections of freeways to facilitate suburb-to-suburb movement. Bypasses direct through traffic away from the urban freeways
(b)

Figure 6.2.1 Historic evolution of (a) cities and (b) parallel evolution of urban transportation networks.

adjacent to the city core within easy access to the work force and transportation connections. As cities expanded, nearby suburbs began legal wars against encroachment by the central city, thereby setting limits to the expansion of the central city. This is the prevalent way in which city limits were established.

The next stage of evolution brought cities close to the shape recognized today. The city core became an exclusive business center (central business district: CBD), often including high-rise office buildings. Improved transportation accessibility, the availability of relatively inexpensive land, and concerns about air quality pushed industrial zones to the outskirts of the urban area. Ethnic and neighborhood business centers took their place within the city limits.

The major trend at this stage was an exodus of affluent residents to the suburbs, which were viewed as quiet bedroom communities, providing a socially desirable setting for raising a family. The transportation network fostered as well as followed this trend. Radial corridors of public transit and highways were developed to bring the suburban workers to their workplaces in the central city where most employers were located. Downtown areas began experiencing traffic congestion problems, pollution problems, and lack of adequate parking space. The flight of affluent families to the suburbs deprived the central cities of their tax base. Without the necessary resources, cities experienced a deteriorating infrastructure and the worsening of conditions within slum areas where the less affluent became trapped. In response to this nationwide trend, during the mid-1960s government embarked on a major urban renewal program, including the areas of housing and urban transportation. Part of this effort addressed subsidies for declining urban public transportation systems and an ambitious research program toward the development of advanced urban transportation systems [6.1, 6.2]. The oil crises of the 1970s gave a major boost to the development and expansion of public transit systems.

The next stage represents present times in which cities still reshape and adjust to demographic, social, and economic trends. The worsened levels of pollution, density, services, and safety of central cities further encouraged people to move to the suburbs. An outer ring of suburbs growing at high rates started developing in the 1980s. At this stage businesses and employers followed the residents in the exodus to the suburbs. As a result, satellite business districts were developed in the outer suburban ring, for example, Tysons Corner in Washington, DC, Naperville and Schaumburg in Chicago, and Orange County in Los Angeles. Furthermore, the expansion of metropolitan limits swallowed small independent towns of the past and minimized the distance between neighboring metropolitan areas, for example, Washington, DC and Baltimore; Chicago, Gary, and Milwaukee.

The parallel evolution of urban modes can be summarized as follows:

Initially modern-era cities were pedestrian-oriented. Even after the development of mechanized long-distance transportation systems such as railroads, the size of most cities was sufficiently small for people to walk to most places. Private transportation in the form of horseback and animal-drawn carriages was sufficient for longer distances. Public transportation in the form of *sedan chairs* in European cities and *jinrikisha* (rickshaw) in Japan was the exception rather than the rule. The first public transportation service per se has been attributed to the French mathematician Pascal, who in 1662 began to offer a horse-drawn service in Paris. However, horse- and mule-drawn *omnibuses* (derived from "omnis," meaning "all," i.e., offering services to the general public) did not come into their own until the mid-nineteenth century. These services spread widely in Europe and the United States and remained a major mode of urban public transportation until early in the twentieth century. In 1832 the first horse-drawn rail streetcar began service in Harlem, New York, and portended the eventual replacement of the omnibus (which was driven on cobblestone pavements) by rail-supported *horse-drawn streetcars* that offered a much more comfortable ride. They operated along designated routes in mixed traffic at relatively low speeds and made frequent stops to take on and discharge passengers. Since propulsive power was the greatest limitation of horse-drawn streetcars, alternate power sources were sought. An early contender was the *cable car,* in which the vehicle is propelled by attaching it to a continuously moving cable. The cable is kept in motion by a stationary source of power. One of the most famous cable car systems opened in San Francisco in 1873, and a few cable car systems are extant to this day, mostly at special locations such as steep inclines and ore mines. Experimentation with the steam engine also occurred, but the major power supply breakthrough came late in the nineteenth century in the form of rail-supported *electric streetcars,* which received their power from overhead wires. At about the same time intercity railroads began to extend their lines into a few major central cities like London and Boston. These urban extensions are known as *commuter railroads* since their urban service is confined to moving commuters between suburban areas and the city during the morning and evening peak hours. Unlike the typical streetcar lines, these heavy rail systems offered limited express service without many intermediate stops and operated on their own rights-of-way. The superior service of these exclusive pathway lines encouraged the development of heavy rail *rapid-transit systems* that were capable of moving large numbers of passengers quickly within the elsewhere congested city. The first underground steam engine rapid transit line opened in London in 1863, and the first elevated urban railroad line, also using steam, was inaugurated in New York City 5 years later. Both cities subsequently converted their systems to electricity. Many large, high-density cities followed suit. Most other cities relied exclusively on electric streetcar lines.

The next chapter in the evolution of transportation in general and public transportation in particular belongs to the adaptation of the internal combustion engine to motorized transportation. In the area of urban public transportation the *motor bus* began to make inroads into the electric streetcar market around 1920. Coupled with an increasing willingness of government to support the construction of streets and highways and with comparatively low fuel costs, the city bus emerged victorious over the electric streetcar, in some instances after a transition to the hybrid *trolley bus,* which operated on rubber tires but gathered its power from overhead wires. The conversion to city buses occurred despite a courageous attempt by the Electric Railways Presidents Conference Committee (PCC) in the 1930s to systematically develop a superior streetcar, the marvelous PCC car. The same technology that replaced the electric streetcar also marked the beginning of the demise of its successor, the city bus, and public transportation in general. The source of this demise was the private automobile, which attracted patronage from public transportation systems. To add insult to injury, a few entrepreneurs even began to use their automobiles to offer competing for-hire services by seeking customers at transit stops. These *jitney* services (which still operate in several places, notably in New Jersey) are undoubtedly the precursors of the modern *taxi,* which now operates in a differently regulated environment.

As explained earlier, during the 1960s several societal changes encouraged a reevaluation of the automobile-based urban system and led to a revision of the existing highway-oriented federal transportation policy to include support for the improvement, research, and development of public transportation systems.

Despite these efforts, the ridership of central city buses and subways has been declining due to the increasing number of people working in the suburbs, whereas the ridership of commuter railroads and suburban bus operations has been slowly increasing. Largely due to urban sprawl, the shares of both transit and walk have decreased between 1960 and 1990. Specifically transit share decreased from 13 to 5%, and walk share decreased from 10 to 4% [6.3.]

6.3 URBAN TRANSPORTATION MODES

Urban transportation needs are served with a multitude of modes. The *intracity* or *urban* distribution of freight is predominantly accomplished by the highway subsystem using vans and trucks of various sizes. The major movements within urban areas are related to the travel undertaken by people. Waterbased urban transportation is found in only a few cities, and air transportation is largely unsuited for urban travel. Thus the means of travel available for urban passenger transportation are mainly land-based and include private transportation (walking and private motor vehicles) and various public transportation services, of which some are highway-based (i.e., regular city buses), others are not (e.g., urban rail transit systems). The latter operate on an *exclusive right-of-way* unrestricted from the interference caused by highway vehicles; however, systems that are commonly thought to use *shared right-of-ways,* such as buses, can also operate on exclusive facilities, thus improving their service quality to levels that rival those of certain other exclusive pathway systems. The primary access mode of transportation in urban areas is walking, but its share is negligible for line-haul trips (i.e., home to work trips). Line-haul urban trips are served by the following modes or combinations of these modes, which may be classed as public or private modes, and also by the type of right-of-way (e.g., roadway versus fixed guideway) and technology. The latter two are used in subsequent descriptions.

6.3.1 Roadway Modes

Private automobile, which can be in the forms of drive alone, drive with passenger(s), or passenger in a private car. Carpooling is the organized commute to work by car with a minimum of two people aboard. The most common form of carpooling is intrahousehold where two or more members of the household utilize the same car. Carpooling between coworkers is less popular despite its potential to reduce the net amount of traffic substantially.

Vanpooling is the voluntary or company-organized commute to work. Either a group of individuals agree to hire a commercial vanpool provider (thereby cutting down their commuting costs by foregoing the purchase and/or use of a car) or companies provide transportation to and from work. A notable company-organized vanpool program was initiated in 1973 by the 3M company in Minneapolis. The program was motivated by the decision to expand the work force without the construction of costly parking [6.4]. In several areas that experience severe congestion problems, ordinances mandate that employers reduce the number of trips generated by their employees through the institution of car and vanpools.

Taxis are owner-operators or private companies that provide transportation to the general public. Major markets for taxis are tourists, visitors, and businesspeople.

Buses usually belong to a public-sector transit system or to private companies. Several cities hire private firms to operate and manage their transit system. Most bus service in London and other large cities has been privatized; that is, it is owned and operated by private firms. Urban bus services include scheduled public bus service, tour services, schoolchildren transportation, and labor transportation. Buses provide high accessibility because they can run through neighborhoods, but lower the level of service compared with fixed guideway systems because they often use congested routes. They also can respond to shifting demands by modifying routes, adding routes, and redistributing the fleet of buses on routes. An *articulated* bus consists of two sections connected by a flexible joint similar to those connecting rail passenger cars; it is also known as a *bandy bus* from the German word for tapeworm (*bandwurm*). Buses can provide rapid transit service by operating on exclusive right-of-way busways (e.g., Pittsburgh.) Lanes on arterial streets or entire streets (e.g., State Street in Chicago) can be converted to exclusive busways.

6.3.2 Fixed Guideway Modes

Fixed guideway transit systems consist of vehicles affixed to a guideway and include *dual-rail, monorail,* or *rubber-tired* systems. Fixed guideway systems are either operated by on-board operators or without the intervention or supervision of on-board operators, such as the entire rapid transit system in Lyon, France (automated guideway transit).

Fixed guideway buses are a technological innovation that increases the capacity and level of service of both regular and articulated buses (Fig. 6.3.1.) This technology provides an exclusive right-of-way for buses, primarily along congested corridors. Slightly modified buses can operate on exclusive guideways. These systems require much lighter infrastructure and provide more flexibility than rail systems because the same buses can operate off the guideway on regular streets. Guided bus systems have been implemented in Essen, Germany, and Adelaide, Australia [6.5]. In early 2000, the U.S. Federal Transit Administration (FTA) initiated a major program to demonstrate advanced Bus Rapid Transit (BRT) in several cities.

Light rail is the modern name for fixed guideway trolleys or electric streetcars. Several North American cities have modern light rail systems in operation today (i.e., Baltimore, Boston, Calgary, Cleveland, Dallas, Edmonton, Los Angeles, New Orleans, Philadelphia, Pittsburgh, Sacramento, San Diego, San Francisco, San Jose, and Toronto had multiple corridor systems as of mid-1998). All these systems opened after 1981 and resemble more rapid transit systems (i.e., they have exclusive, often elevated, right-of-way) rather than streetcars. More traditional light rail is common in older European cities, where in some locations it carries the majority of trips to CBD, enjoys priority treatment through signal preemption, and so on. Light rail operations can be inefficient in congested urban sections because they operate on arterials along with vehicular traffic. Their bulky and slow nature may worsen traffic conditions [6.6], but they do offer pedestrian accessibility because they have relatively short station spacing.

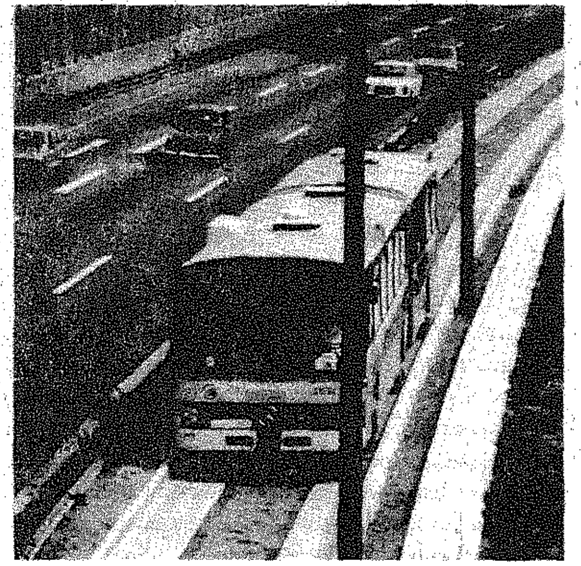


Figure 6.3.1 Fixed guideway bus system in Essen, Germany. (From Essener Verkehrs AG [6.5].)

Rail rapid transit systems are common in large U.S. cities. Atlanta, Boston, Chicago (Fig. 6.3.2), Los Angeles (service was inaugurated in 1990), New York City, San Francisco, and Washington, DC have such systems. Operations include trains of four to ten railcars with stations every 0.5 to 5 km depending on densities. At the present time all U.S. systems involve the classic technology of electric-powered cars with steel wheels running on steel rails. The major advantage of these systems in terms of operating efficiency is that they operate on an exclusive right-of-way, which isolates them from other traffic and gives them the ability to offer relatively fast and on-time service. Major disadvantages of these systems are the high implementation cost, the high levels of subsidy required for their implementation and operation (i.e., typically only a fraction of operating costs is recovered from the fare boxes), and their inflexibility in following shifts in demand. The long station spacing often requires special collection and distribution support, frequently in the form of buses or park-and-ride facilities.

Commuter (or regional) rail systems connect primarily distant suburbs with financially affluent population to suburban centers and the CBD of a major metropolitan area. This is reflected in the average one-way-trip length of commuter rail passengers, which was about 35 km in 1988 [6.3]. Such systems are unique to a handful of cities (i.e., Chicago, Los Angeles, Miami, and New York City). Commuter railroads use large passenger cars and often run on freight railroad lines. Commuter systems offer much shorter travel times compared with private automobiles and rapid transit systems due to the fewer stops and the high average speed between stations.

Personal rapid transit refers to systems that operate on exclusive pathways employing small vehicles to allow for frequent service and scheduling flexibility; they may be described as horizontal elevators. They have found applicability in major activity centers and airports.

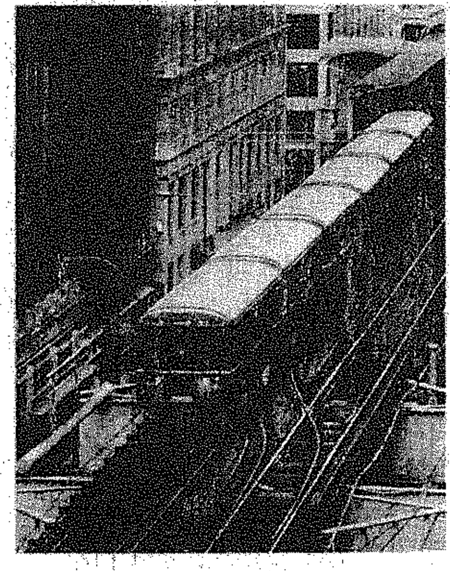


Figure 6.3.2 Elevated rapid transit rail system (Chicago). (Photograph by P. D. Prevedouros.)

6.3.3 Demand-Responsive, Dual-Mode, and Other Modes

Demand-responsive systems have the flexibility in route or time scheduling or both that permits them to respond to the actual demand placed on them. These systems represent an attempt to rival the flexibility of the private automobile, in contrast to the traditional *fixed route-fixed schedule* transit systems. Taxis are naturally demand-responsive, but other systems have been developed, such as dial-a-ride and prescheduled systems that allow for the dispatching and rerouting of common carriers to serve temporarily changing demands. These types of *paratransit* systems have found applicability as specialized services for elderly and handicapped persons.

Dual-mode systems can be of two types, which are not mutually exclusive. The first type of vehicle can operate on both a guideway and on a regular street, such as the fixed guideway bus presented earlier. The second type of vehicle can operate under different power sources, such as internal combustion engine (ICE) or electric drive powered internally from batteries or fuel cells, or externally by attaching to an overhead power cable like a trolley bus. ICE propulsion is typically used on outer city or suburban parts of a route and electric drive propulsion is used in pollution and noise-sensitive areas such as dense neighborhoods, archaeological districts, and downtown areas.

Usual combinations of modes are *park-and-ride,* according to which individuals drive their autos to transit terminals, park and use public transportation for the line-haul trip to work, and *kiss-and-ride,* according to which a car passenger is dropped off at a terminal. This mode is common among family members, particularly in families with fewer cars than workers.

Air and water services are provided in a few metropolitan areas. Transportation of corporate executives and government officials with helicopters is quite common, whereas passenger ferry services are offered in Hong Kong, New York City, Seattle, Sydney (Fig. 6.3.3,) Vancouver, and so on.

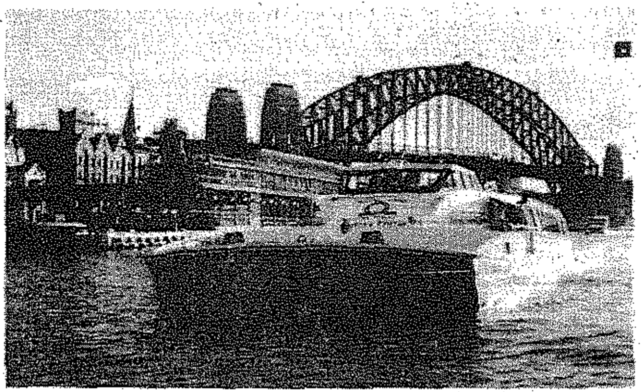


Figure 6.3.3 Frequent, fast, and high capacity passenger ferries are a commuting alternative in Sydney, Australia. (Photograph by P. D. Prevedouros.)

The use of various modes is heavily dependent on the pattern of origins and destinations, which is itself constantly changing. The changes have a strong effect on the number and types of trips needed as well as on the modes chosen.

6.4. URBAN TRANSPORTATION ISSUES

6.4.1. General

About 80% of the U.S. population resides in urban areas and about 50% resides in metropolitan areas with a population of one million or more. The 1990 National Personal Transportation Survey (NPTS) found that about 70% of the total annual travel in terms of person-km (about 13,300 km per annum) took place on local networks and the balance took place on long-distance travel, that is, on trips exceeding 120 km; 88% of the local travel was made using cars [6.7]. In addition, heavy trucks carry freight into and out of cities, and smaller trucks and vans are used for local delivery. Many cities have ports and airports that are the primary entry or exit points for passengers and cargo originating from or destined to a broad area. All these functions generate substantial traffic loads on the urban networks. Transportation problems affect most of a nation's population and get considerable political attention.

Several transportation problems plague contemporary urban areas, the most prevalent of which is traffic congestion. Other urban transportation problems, some of which are of primary concern in certain locations, include the following:

- The inefficient utilization of public infrastructure systems and transportation services caused by the normal weekday peaks and valleys in travel demand that necessitate wide roads, large bus fleets, more drivers, and so on for about 2 h during the morning and afternoon peaks, and only a fraction of these capacities for the rest of the time and during weekends and holidays
- Infrastructure financing with difficult choices, such as capacity expansion versus rehabilitation, highway versus transit investment, and the share of financing among local, state, federal, and private sources
- Special transportation provisions for the elderly, disabled, and low-income people

- Environment concerns of emissions and noise pollution, as well as balancing the conflicting demands for environment quality and efficient and affordable transportation
- Safety and security for all residents on all public spaces and transportation modes
- Institutional and operational changes for efficiency improvement
- Legislated requirements without the commensurate financing for implementation

Countermeasures for some of these are readily available, but they also involve implementation hurdles. For example, in most cities the morning and afternoon peaks cannot be accommodated by one work shift for drivers and train conductors. The requirement of two shifts and/or overtime is very costly. In 1978, the Seattle Metro successfully negotiated with the transit operators union to allow for part-time drivers to be assigned to special peak-period "trippers," thus obviating the need to pay them for 8 h [6.8]. By contrast, the public bus agency in Athens, Greece, proposed a split-shift arrangement that would necessitate a mandatory long break (off-duty) period for a portion of the drivers, on a rotating basis.This proposal faced strong union opposition, and was not enacted.

Services for the handicapped, the disabled, and the elderly are available, but they require a major financial commitment. The costs of services requiring transportation* vary with respect to densities (of people and activities), vertical buildup, and city size. There is a continuous debate regarding the size of the city at which the total cost of services per inhabitant is minimized. A unique answer to this issue is elusive, partly because of definition problems (i.e., decision makers tend to minimize public costs only, but the minimization of both public and private costs results in better public welfare), as well as complex relationships among costs of services, densities, types of infrastructure, and so on [6.9].

The difficult choice for transportation investment is indirectly addressed in Chapter 11 on "Evaluation." It defines the goals, objectives, criteria, and measures as well as the base (effectiveness or efficiency) for assisting the choice among transportation investment options. The remaining urban issues previously listed are largely outside the scope of this text.

6.4.2 Traffic Congestion

> [*Traffic jams*] *demonstrate that people by the hundreds of millions are getting two things they badly want: a chance of a prosperous urban life rather than a poorer rural one, and a private car. For it is the combination of these two desires that has made congestion so universal at the end of the 20th century.* [*The Economist,* "A Survey of Commuting: To Travel Hopefully," p. 3, September 5, 1998.]

Traffic congestion occurs on fixed capacity† road networks when traffic grows beyond about 90% of the capacity. As a result, the level of service, which is usually a measure of speed or delay, deteriorates to unacceptable levels (Fig. 6.4.1). The capacity of

*Essential services that require transportation include emergency medical services, street and highway maintenance, police, mail collection and delivery, fire protection, utility connections and repairs, snowplowing, street cleaning, refuse collection and disposal, school bus service, and so on.

†Fixed capacity actually varies primarily with respect to specific cross-sections, but also time. The latter affects lighting conditions as well as traffic composition (e.g., % heavy vehicles). Although a 3.5-m wide freeway lane has a "fixed" capacity of 2200 veh/h, short-period capacity may exceed 2500 veh/h. As discussed in earlier chapters, capacity is affected by factors, such as alignment (e.g., grade and curvature), obstacles (e.g., shoulders and objects causing lateral displacement), and weather and pavement conditions.

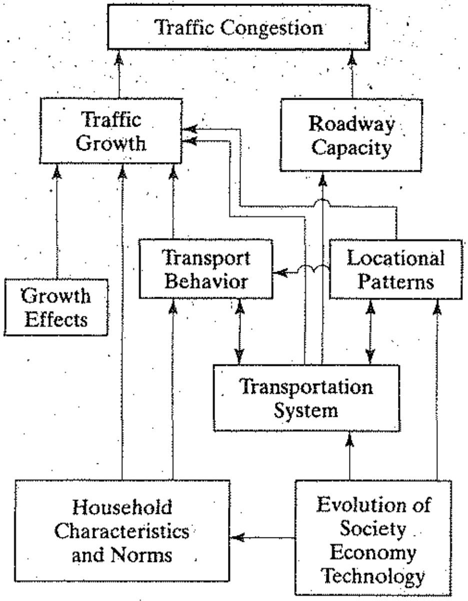


Figure 6.4.1 Generation of traffic congestion.

the roadway system to a large extent is determined by the transportation system characteristics and policies, such as geometry, signalization characteristics, and traffic management restrictions (i.e., high occupancy vehicle lanes, bus lanes, reversible lanes). To a lesser extent, roadway system capacity is determined by the driving habits of the population, the size and average acceleration rate of vehicles, the weather conditions, and so forth.

The growth of traffic is affected by four major factors: (1) the natural growth of the population, (2) locational patterns (i.e., spatial distribution of residence, work, shopping, and entertainment places), (3) transportation characteristics and policies, and (4) transport behavior of individuals and households manifested in their mode, departure time, and route choices. Furthermore, most of the factors affecting traffic growth are time-dependent.

The U.S. DOT measured several changes affecting travel that have occurred between 1970 and 1995 [6.10]:

- The national population grew by 29%.
- The metropolitan population grew by 49%.
- The personal disposable income grew by 56%.
- The number of households grew by 56%.
- The number of workers grew by 59%.

As a consequence of these developments, during the same time span the number of automobiles and light trucks grew by 86%, and the amount of passenger-km traveled grew by 49%, both of which accelerated congestion levels in metropolitan areas. Congestion also is compounded by the improper operation of traffic devices. For example, a 1994 General Accounting Office report states that 90% of the signals in metropolitan areas were not functioning *at a minimum standard* due to poor maintenance [6.11].

The effects of traffic congestion are multiple; they include:

1. Loss of productive time
2. Loss of fuel
3. Increases in pollutants (because of both the additional fuel burned and more toxic gases produced while internal combustion engines are in idle or in stop-and-go traffic)
4. Increases in the wear and tear of automobile engines
5. High potential for (usually low impact) traffic accidents
6. Slow and inefficient emergency response (Fig. 6.4.2) and delivery services
7. Negative impact on people's psychological state, which may affect productivity at work and personal relationships

The summation of all these effects yields a considerable loss for the society and the economy of an urban area.

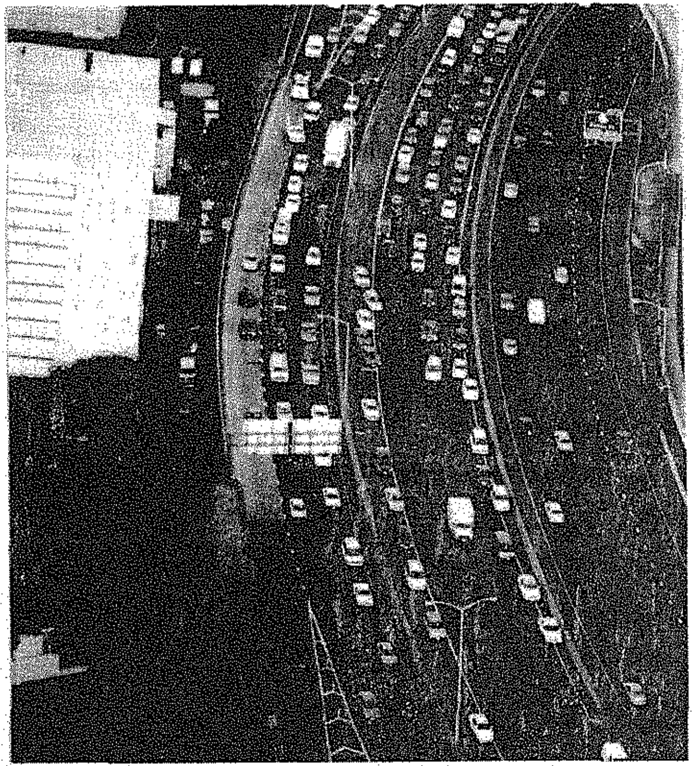


Figure 6.4.2 Medical and other emergency services are delayed by traffic congestion. Pictured is an ambulance making way on the Moanalua Freeway in Honolulu.
(Photograph by P. D. Prevedouros.)

Several researchers have tried various approximations, typically with the use of composite indices, to assess the severity of congestion or the difficulty of connecting productions and attractions (or trip origins and destinations) in urban areas. Samples from two noted studies are presented in Table 6.4.1. The Wharton study measure [6.12] is based on the access time in minutes, whereas the TTI index [6.13] is an index that computes systemwide congestion level on the basis of traffic volumes and the proportion of daily volume occurring during the peak periods. The TTI index appears to be more appropriate for comparative analysis. Since the metrics are different, the results of the two studies do not coincide. For example, Houston seems to be having poor accessibility (rank 2), but its level of congestion is relatively moderate (rank 13). A more complete assessment of the methods and measures for assessing traffic congestion is available in NCHRP report 398 [6.14]. The report includes models for the macrolevel assessment of speeds and travel times on various types of highways, which are shown in Sections 4.5.5 and 4.7.5.

Public transit service and ridership are a perennial issue, particularly in the United States. Between 1985 and 1995 ridership on mass transit systems fell by 11%, mostly due to heavy ridership losses by bus systems [6.10]. The same report shows that in the same time period metropolitan areas with passenger rail systems increased from 14 to 22, and route kilometrage increased by 18%. Despite these additions, the average age of railcars increased from 12.3 to 19.8 years, which created more service, comfort, and reliability problems.

Congestion countermeasures are basically classified into supply and demand measures. *Supply measures* add capacity to the system or make the system operate more efficiently. Their focus is the transportation system. *Demand measures*, on the other

TABLE 6.4.1 Measures of Roadway Traffic Congestion and Accessibility for 15 U.S. Cities

	City[a]	Wharton accessibility measure, 1993		TTI congestion index, 1994	
		Measure	Rank 1	Index	Rank 2
1	Atlanta	56.2	17	1.18	10
2	Boston	50.2	28	1.08	18
3	Chicago	52.7	24	1.28	5
4	Denver	57.4	13	1.07	19
5	Honolulu	46.4	39	1.13	12
6	Houston	69.5	2	1.12	13
7	Los Angeles	62.5	6	1.52	1
8	Miami	56.6	16	1.32	4
9	New York City	55.2	18	1.15	11
10	Philadelphia	40.2	50	1.05	24
11	San Francisco	42.5	46	1.33	3
12	San Jose	35.9	54	1.06	21
13	Seattle	44.5	45	1.26	6
14	St. Louis	59.4	10	0.98	30
15	Washington, DC	48.7	30	1.43	2

[a] Sorted alphabetically; higher rank means more congestion and more difficult access.

Note: Rank 1 = among the 60 largest MSAs; rank 2 = among the 50 urban areas.

Source: Refs. [6.2, 6.13].

hand, focus on motorists and travelers and attempt to modify their trip-making behavior. In either case actions that affect either supply or demand tend to influence both. This is because the interaction of supply and demand results in a new equilibrium between them. An additional, longer-term tool for action against traffic problems is land-use planning and policy. It has the potential (1) to control the number and growth of major traffic generators along congested corridors, (2) to establish sensible allocations of land for future development given present constraints and expansion plans for the transportation network, and (3) to enforce balanced employment and residential development, thus reducing long home-to-work trips [6.15]. Known as *growth management,* these land-use decisions are often difficult to implement as they affect control of private ownership rights. Several rapidly growing states, including California and Florida, have enacted legislation to empower such actions.

No single measure can "solve" traffic congestion problems. A combination of measures typically helps to stabilize delays for several years. Large improvements have been realized but they are confined to small network sections or narrow corridors rather than to entire regions. Before presenting supply and demand measures, two unconventional views of congestion should be mentioned. One suggests that traffic congestion is a positive measure of urban vitality. Prosperous regions have traffic congestion, whereas decaying urban regions do not. The other view is that congestion is a self-limiting problem. In other words, if congestion is left uncontrolled, at worst, the roads will be congested during the better part of the day. Because of this, people "naturally" will change modes, cancel trips, schedule activities differently, or relocate to neighborhoods closer to work, schools, and shops, or to other cities altogether, to cope with the situation.

6.4.2.1 Supply Strategies

Supply strategies for resolving congestion include the development of new or expanded infrastructure, small-scale infrastructure efficiency improvements. All actions in this category *supply capacity* so that demand is better satisfied and delays and queuing are lessened.

Major infrastructure improvements include civil projects, such as new freeways (Fig. 6.4.3), transit lines, ferry boat docks, and so on. They also include large-scale modifications, such as lengthy road widening, bridge replacements, permanent freeway lane conversions, technology conversions (e.g., a new rail technology, a modernized bus fleet, and intelligent transportation systems.)

Small-scale capacity and efficiency improvements fall under the classification of transportation system management (TSM), which was a 1976 Urban Mass Transportation Administration (now Federal Transit Administration, FTA) requirement for metropolitan planning organizations (MPO). They include all types of small infrastructural and operational improvements. Examples include bottleneck elimination through channelization and spot-widening, signal system upgrades and coordination, freeway ramp metering and high occupancy vehicle (HOV) ramp-metering bypasses, contraflow with coning and/or overhead signals (Fig. 6.4.4) and other lane management schemes such as the concrete/movable barrier systems in Boston, Dallas, Honolulu, and the Golden Gate Bridge in San Francisco. TSM also applies to transit systems and involves equipment upgrades, scheduling and dispatching improvements, route evaluation, and improvement, including the relocation of bus stops. Demand-responsive public transportation is a nonscheduled passenger service aimed at helping people with mobility problems or serving low-density locations where regular

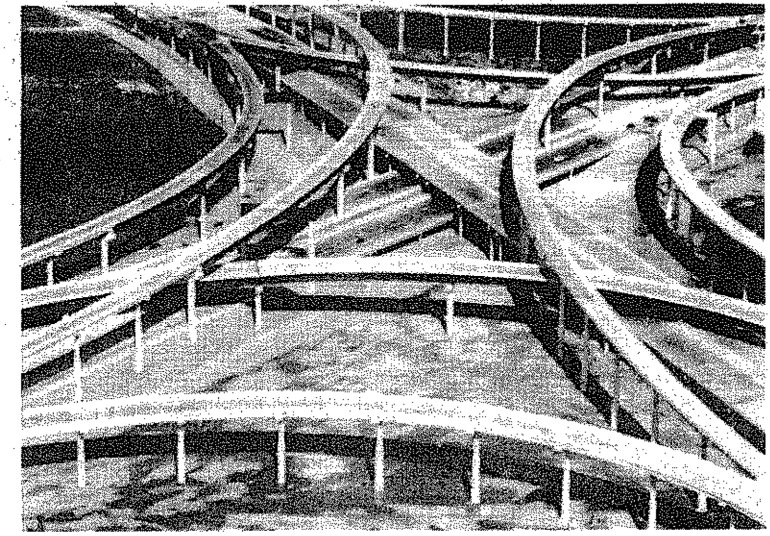


Figure 6.4.3 New roadways and freeway interchanges such as this one in Houston, Texas are a traditional strategy for congestion relief. (FHWA, *Our Nation's Highways.*)

Figure 6.4.4 Contraflow operations are common on many busy bridges worldwide. Sydney's Harbor Bridge is pictured. (Photograph by P. D. Prevedouros.)

public transportation would be wasteful and less attractive to the public (i.e., fixed route stops may be too far from residences).

6.4.2.2 Demand Strategies

Demand strategies for resolving congestion include congestion pricing, parking pricing, restrictions on vehicle ownership and use, and other incentive and disincentive policies. All actions in this category *aim to modify travel habits* so that travel demand is lessened or switched to other modes, other times, or other locations that have more capacity to accommodate it.

Congestion pricing is the imposition of a direct charge on motorists for the true cost of their trip (as a function of both infrastructure cost and, importantly, congestion and environmental consequences). It is based on the peak-period pricing principle that has seen widespread use in the airline, vacation, restaurant, and utility (telephone and electricity) industries. It has been estimated that in the United States "optimal" congestion pricing for congested freeways was 10 to 20¢ per vehicle-km, in mid-1990s prices, and twice as high for congested arterials [6.16]. Congestion pricing can:

1. Divert travelers to other modes of travel (transit, car pools, taxis)
2. Cause the cancellation of nonessential trips during peak periods and change the departure time or route of vehicular trips
3. Collect sufficient funds for major upgrades of highways
4. Cross-subsidize public transportation modes

Parking pricing and availability restrictions also discourage the use of private vehicles to specific areas. Capacity restrictions that are grossly disproportional to demand (as in many old European cities), however, may cause congestion due to excessive circulation in search of an empty parking stall. Another option, *employee parking cash-out,* was mandated by California's legislature in 1993 in an effort to combat congestion and pervasive air quality problems by providing a cash amount in lieu of a parking space. Statute 43845 reads as follows:

- In any designated nonattainment area, each employer of 50 persons or more who provides a parking subsidy to employees, shall offer a parking cash-out program.
- Parking cash-out means an employer-funded program under which an employer offers to provide a cash allowance to an employee equivalent to the parking subsidy that the employer would otherwise pay to provide the employee with a parking space.
- A parking cash-out program may include a requirement that participants will comply with guidelines designed to avoid neighborhood parking problems.

Vehicle ownership and use policies include ownership restrictions in the form of heavy import duties (e.g., China, Israel) or a separate licensing requirement (e.g., Singapore, specific areas in Japan). Heavy annual fees, strict periodic inspections, and expensive fuel prices (see Table 5.4.3 for a comparison among countries) also act as restraints to private vehicle acquisition and use. Car-producing countries usually impose fewer and less stringent restrictions to automobile acquisition and use.

Telecommuting [6.17] is an attempt to reduce congestion by providing satellite offices for neighboring employees who transmit their work to the central offices. In this way long commutes to the central location are replaced by telecommuting connections. In other words telecommuting provides office space instead of road space. Telecommuting also includes work at home for professions whose product of work is transmittable through electronic devices (i.e., writers, reporters, engineers, designers, data or orders processing personnel, help/technical assistance, etc.).

Other incentive and disincentive policies usually fall under the classification of transportation demand management (TDM), which became a requirement for MPOs in 1982. They include incentives in the form of TSM actions (e.g., implementation of bus and HOV lanes), as well as transit and pedestrian malls with restricted access to all or to single-occupant private vehicles. Other TDM measures include free or reduced tolls for car pools and vanpools, preferential and/or cheaper parking for car pools and vanpools, guaranteed-ride-home provisions, flextime or staggered work hours, employee parking cash-out, restricted area access through cordoning and permitting by licensing (e.g., Singapore) or even/odd vehicle license number scheme (e.g., cities in Brazil, Greece, Italy). HOT lanes (HOV/Toll lanes) are free for vehicles with a minimum of three occupants and available to vehicles with fewer than three occupants at a cost (toll). HOT lanes are intended to bridge the gap between over- and underutilized HOV lanes [6.18].

The 1991 Intermodal Surface Transportation Efficiency Act mandated the enactment of congestion management systems (CMS) by all state DOTs. Work plans were finalized by 1994 and most state DOTs are working on their CMS. The surface transportation management process includes nine major steps [6.11], which should be followed in a "round-robin" fashion continuously over time:

1. Monitoring
2. Performance evaluation
3. Identification of improvement strategies
4. Evaluation of strategies
5. Prioritization
6. Programming and funding
7. Implementation
8. Operation
9. Maintenance

In 1998 the Transportation Efficiency Act for the twenty-first century (TEA-21) extended the CMS requirement for another 6 years.

Intelligent transportation systems (ITS) provide tools for implementation of both supply and demand congestion countermeasures. Supply type ITS tools include early incident detection and resolution, optimized signal operation based on real-time demand, freeway management with ramp metering, accident avoidance with variable message signs warning of upcoming conditions (e.g., congestion, fog, etc.), and bus system coordination. Demand-type ITS tools include the provision of real-time traffic congestion information at various places (e.g., home, work, at the shopping center, etc.) for informed travel decisions.

Car travelers with some degree of flexibility may postpone a trip, delay it, or make it in a different mode, if the roadways are congested. Also, in-vehicle devices may switch demand from a congested route to an alternate route, thereby improving the performance of the entire corridor system. ITS applications are covered in the next section.

6.5 INTELLIGENT TRANSPORTATION SYSTEMS

In 150 years information technology progressed from the transmission of a few bytes per second to the transmission of billions of bytes per second. Much of this progress has occurred in the last 20 years, as shown in Table 6.5.1 [6.19].

Some hoped that the progress in telecommunications would lessen the need for transportation, particularly for business travel and business document shipping. This did not occur. Data from the United States and France (which was the pioneer in card-phones and advanced communications via regular phones, e.g., teleshopping) indicate that both transportation and communications grew at similar rates throughout the 1990s. This suggests the presence of latent demand for human interaction in the form of combined communication and transportation rather than mere substitution of communications for transportation [6.10]. This phenomenon also was true earlier in the twentieth century when the telephone was first introduced.

The dramatic increase in performance and the almost as dramatic reduction in real costs of both computing and communications technologies have enabled engineers to address, among other things, traffic congestion problems and to make transportation operations more efficient. Traditional objectives have not changed, but information technologies (IT) offer new ways for achieving them. For example, cities like Austin, TX optimized emergency vehicle location and routing in the early 1980s [6.20]. Technology now permits this to be done in real time and with full consideration of existing traffic conditions [6.21].

Data transmission for the transfer of information and the communication links among pieces of equipment are key to ITS. At the end of the twentieth century a lot could be accomplished through telephone lines. For example, the integrated services digital network, ISDN, is a telephone service that is able to transmit voice, video, data, and text as well as other supplementary services. The basic ISDN transmission rate of 160 kilobytes per

TABLE 6.5.1 Timeline of the Evolution in Communications

Year	Development
1847	Telegraph
1877	Telephone
1920	Sound
1930	Telex, fax, TV
1960	Hi-fi stereo, color TV, mobile telephone
1975	Medium speed data transmission, paging
1984	High-speed data transmission, telemetry, computer networks, video-conferencing
2000	Wide-band data transmission, high definition TV, voice-activated controls, . . .

TABLE 6.5.2 Sample Properties of Selected Telecommunications Technologies

Technology	Medium transfer	Rate/channel	Information types
Twisted wire pair	Copper wire	1.2 to 3.1 Kbps	Data, voice, slow scan TV
Fiber optics	Glass or plastic fibers	Up to 2.4 Gbps	Data, voice, slow scan TV
CATV	Coaxial cable	Up to 7.5 Mbps	Data, voice, analog TV
Radio networks	Atmosphere	9.6 Kbps	Data
Terrestrial microwave	Atmosphere	Up to 7.5 Mbps	Data, voice, analog TV

Source: Ref. [6.22]

second (Kb/s) is 50 to 100 times faster than a regular telephone transmission and makes video-conferencing and telecommuting possible. Table 6.5.2 shows a few key characteristics of the most common telecommunications technologies.

Elements of what came to be known as ITS had been incrementally deployed in many localities over the years, particularly during the 1980s and 1990s. Among these were actuated signal controls and supporting components (see Sections 4.6 and 6.5.4). Despite the established functionality of such elements, some contended that ITS and its precursor, intelligent vehicle-highway systems or IVHS, came about as "a solution looking for a problem." Such objections notwithstanding, beginning in 1993 the U.S. DOT invested heavily in an elaborate effort to establish a national ITS architecture and to adopt a set of ITS communication standards in order to provide a systematic framework for planning, defining, and integrating ITS implementations at the regional level and to ensure that travelers are presented with consistent user interfaces across the nation [6.23].

This effort began with the identification of the required *user services* that were to be supported by ITS serving transportation users ranging from pedestrians and bicyclists to large trucking companies and airport operators. User services were then grouped into a set of subsystems constituting the *physical architecture* that defined the interfaces between physical entities in terms of three layers: The *communications* layer addressed the methods of transferring information between subsystems; the *transportation* layer that defined the type of information transferred by each subsystem; and the *institutional* layer that defined the necessary supporting institutional structure and policy. Each subsystem of the physical architecture was decomposed into the set of functions it needed to perform. Some of these functions were found to be required by more than one subsystem, and this knowledge could aid in the avoidance of unnecessary duplication when designing a regional subsystem. Thus at progressively lower levels of detail, lied the multilayered *logical architecture,* the purpose of which was to identify the specific functions to be performed and the data flows between these functions.

With ITS came a proliferation of new acronyms to describe components and processes. Some of the most common ITS and other relevant transportation abbreviations are shown in Table 6.5.3 for quick reference. The problem with these abbreviations is that several are not unique. For example, CMS may mean changeable message sign or congestion management system. In such cases the applicable term is spelled out in this text to avoid misunderstandings.

TABLE 6.5.3 Common Abbreviations

AHS	Automated highway system
APTS	Advanced public transportation services
ATIS	Advanced traveler information services
ATMS	Advanced traffic management services
AVL	Automatic vehicle location
CCTV	Closed circuit television
CVO	Commercial vehicle operations
DAB	Digital audio broadcasting
DSRC	Dedicated short-range communications
EMS	Emergency medical service
ERP	Electronic road pricing
ETC	Electronic toll control
GIS	Geographic information system
GPS	Global positioning system
GPWS	Ground-proximity warning system (aviation)
GSM	Global system for mobile communications
HAR	Highway advisory radio
ICC	Intelligent cruise control (also, ACC = adaptive)
ITS	Intelligent transportation systems
IVRG	In-vehicle route guidance
RDS	Radio data system (an FM subcarrier)
SCATS	Sydney coordinated adaptive traffic system
SCOOT	Split-cycle-offset optimization tool
TCC	Traffic control center
TMC	Traffic management center, also traffic message channel (RDS-TMC)
UTC	Urban traffic control (traffic signal system)
VMS	Variable message sign (also CMS = changeable and DMS = dynamic)
WIM	Weigh-in motion
WWW	Worldwide web (part of the Internet)

The presentation in this section includes the following main components: user services, ITS architecture components and standards, ITS in Europe and Japan, and mature applications. The latter include detectors, traffic signal systems, freeway management, electronic road pricing, and automatic vehicle classification. ITS safety, environment, and liability issues also are discussed in brief.

6.5.1 User Services

In order to systematically design advanced technologies in the field of transportation and to reap benefits for travelers and goods, a set of 30 user services have been defined.* Each user service is composed of a set of hierarchically arranged user service requirements. Depending on their basic objectives, user services have been grouped into six bundles. Tables 6.5.4(a) and (b) show the six groupings of the 30 user services and sample applications for each user

*More user services are likely to be developed. Specifically a 31st user service, the Archived Data User Service (ADUS) is designed to fulfill the need for an historical data archive. ADUS requires ITS-related systems to have the capability to receive, collect, and archive ITS-generated operational data for historical, secondary, and non-real-time uses. For example, traffic control data can be applied in transportation administration, policy, safety, planning, operations, and research. Other user services under discussion before the turn of the century were multijurisdictional emergency management, weather data sharing, and intermodal freight logistics.

TABLE 6.5.4(a) User Services (1 to 15)

Bundle	User services	Explanation	Sample elements
1.0 Travel and transportation management	1.1 Pretrip travel information	Provides pretrip information on traffic and transit LOS so travelers can decide on their travel (route, mode, time-of-day, or trip cancellation).	At home/work information outlets: Internet, kiosks, telephone, etc.
	1.2 En-route driver information	Provides drivers information about traffic conditions, incidents, construction, weather conditions, hazardous road conditions, and safe speeds while enroute. This information allows midtravel changes.	VMS, radio, HAR, pagers, mobile telephone, in-vehicle navigational systems
	1.3 Route guidance	Provides travelers with instructions on how to reach their destinations. Identifies a preferred route to a destination. Public transit guidance could be determined from bus schedules or real-time information through AVL.	IVRG, AVL, GIS maps, real-time traffic reports
	1.4 Ride matching and reservation	Provides convenient ride-matching information and reservations to potential users.	At home/work information outlets: Internet, kiosks, telephone, etc.
	1.5 Traveler services information	Provides information on the location, operating hours, and availability of food, parking, auto repair shops, hospitals, police facilities, and other modes of transport from home, work, shopping centers, airports, etc.	Internet, kiosks, interactive telephone, television, IVRG
	1.6 Traffic control	Provides for the integration and control of freeway and surface street systems to improve the flow of traffic, to improve safety for vehicular and nonvehicular travelers, to give preference to transit/HOV.	UTC, ramp-metering, signal preemption
	1.7 Incident management	Provides technologies integrated into traffic surveillance systems to reduce incident-induced congestion by improving authorities' ability to detect and clear incidents.	Surveillance, automatic incident detection, highway patrol
	1.8 Demand management and operations	Provides strategies promoting increased use of HOV and public transit. Calls for the development of supportive regulations and policies.	Variable work hours, compressed workweeks, telecommuting, congestion pricing, parking fees
	1.9 Emissions testing and mitigation	Provides information for monitoring air quality and develops air quality improvement strategies. Advanced vehicle emissions testing systems determine when the quality of air approaches critical levels.	Remote sensing of vehicle emissions; integration with UTC
	1.10 Highway-rail intersection	Provides advanced warning to drivers and by implementing improved crossing control and warning devices for at-grade crossing sites.	IVRG, RDS-TMC, ICC
2.0 Public transportation operations	2.1 Public transportation management	Provides real-time (schedule adherence) and facility (passenger loading, running times, mileage) information to automate operations and to assist in the planning and management of public transit services.	AVL, computer-aided dispatching
	2.2 En-route transit information	Provides real-time information to public transit users relating to schedule changes, delays, and others at key transfer locations.	Automatic on-board and at-station VMS or audio announcements
	2.3 Personalized public transit	Provides on-demand routing to pick up passengers and to deliver them to their destinations by implementing advanced technologies for dispatching and routing the vehicles. Fleet may include courtesy vans, taxicabs, etc.	AVL, computer-aided dispatching
	2.4 Public travel security	Provides a more secure environment for public transportation patrons and operators by monitoring transit stations, bus stops, parking lots, and on-board vehicles with security cameras.	Automated alarms, CCTV, police
3.0 Electronic payment	3.1 Electronic payment services	Provides automated means for paying for transportation services. Smart-cards, etc. can be used to increase the efficiency of toll payments, public transit fares, and parking services (prepayment or postbilling).	ETC, smart-card

TABLE 6.5.4(b) User Services (16 to 30)

Bundle	User services	Explanation	Sample elements
4.0 Commercial vehicle operations	4.1 Commercial vehicle electronic clearance	Provides automated inspection/weigh facilities for commercial vehicles at check points and border crossing without delay, after inspecting their safety status, credentials, and weight to be within acceptable limits.	AVL, transponders, GPS
	4.2 Automated roadside safety inspection	Provides real-time access to commercial vehicle safety performance records (including previous problems) and minimizes the time required for roadside inspections.	AVL, transponders, vehicle condition inspectors
	4.3 On-board safety monitoring	Provides monitoring systems that sense the safety status of a commercial vehicle and responds/reports them at mainline speeds (e.g., warning systems for the driver, the carrier, and/or enforcement officials.)	AVL, transponders, vehicle and driver condition inspectors
	4.4 Commercial vehicle administration processes	Provide for the automatic collection and recording of travel distance, fuel purchased, and trip and vehicle data. Reduces preparation effort for fuel tax and registration reports for affected jurisdictions.	AVL, enhanced trip computers
	4.5 Hazardous materials incident response	Provides an immediate description of the hazardous material to the emergency responders in the event of an incident involving a vehicle transporting hazardous materials.	Automatic emergency notification
	4.6 Commercial fleet management	Provides commercial drivers and dispatchers with real-time routing information in response to congestion or incidents.	CB radio, pagers, mobile telephones, in-vehicle displays of dispatch center information
5.0 Emergency management	5.1 Emergency notification and personal security	Provides immediate notification of an incident and a request for assistance. Notice may be given manually or automatically. Automatic notification sends information on crash location, nature, and severity.	AVL, "mayday" function (on mobile phones or dynamic IVRG),crash/disablement sensors
	5.2 Emergency vehicle management	Provides fleet management, route guidance, and signal priority to reduce the time it takes emergency vehicles to respond to an incident after its detection.	AVL, IVRG, computer-assisted dispatch
6.0 Vehicle control and safety systems	6.1 Longitudinal collision avoidance	Provides technology for preventing rear-end vehicle collisions.	Radar, ABS
	6.2 Lateral collision avoidance	Provides crash warning and controls for potential lateral collisions between two vehicles in adjacent lanes or between a vehicle and obstacle(s).	Radar, warning systems
	6.3 Intersection collision avoidance	Provides a warning for impending collisions when approaching a signalized intersection. Also provides a warning when the right-of-way at the intersection is ambiguous.	Vehicle-to-roadside communications, enhanced traffic control with trajectory forecasts
	6.4 Vision enhancement for crash avoidance	Provides a warning for potential upstream collisions with other vehicles or obstacles in the roadway which are not yet visible to the driver.	Infrared (or other) in-vehicle forward-looking sensor, "night-vision" head-up display
	6.5 Safety readiness	Provides unobtrusive monitors which warn if the driver is becoming drowsy or inattentive. Monitors the vehicle; alerts the driver to impending malfunctions (some cars already have sophisticated systems checks).	Same as user service 4.3; this one tailored to noncommercial vehicles
	6.6 Precrash restraint deployment	Provides advance responses to an impending collision, such as tightening safety belts, deploying air bags optimally, etc., based on the velocity, mass, and trajectories and occupant characteristics.	Computerized in-vehicle safety devices (several are in the marketplace)
	6.7 Automated vehicle operations	Provides a system for automated vehicle operations where vehicles are guided along a roadway without driver assistance.	Visual lane keeping, high-speed off-ramps, radar, ICC

TABLE 6.5.4(c) User Service Requirements

Service bundle	1.0	Travel and traffic management
User service	1.1	Pretrip travel information
	⋮	
	1.6	Traffic control
	⋮	
	1.10	Highway-rail intersection
1st-level requirements		
	⋮	
	1.6.2	Traffic control shall include a traffic surveillance function.
	⋮	
2nd-level requirements		
	⋮	
	1.6.2.2	Traffic surveillance shall include a **data collect** function to provide the capability to collect data that are needed for determining traffic flow and prediction.
	⋮	

service [6.23]. Table 6.5.4(c) illustrates the decomposition of user service 1.0 into a hierarchy of requirements.

Not all user services are mutually exclusive. Several share common subobjectives and many share the same information infrastructure. Some may be inexpensive to implement (e.g., a phone and computer-based ride-matching service) and others may be very complex and expensive (e.g., intersection collision avoidance).

Currently one of the most accessible outcomes of ITS user services are Internet traffic reports. Visual information (traffic camera image capture, color-coded traffic conditions or live video) is typically received pretrip at home or at work enabling a wiser choice of mode and time of departure, even trip cancellation. However, conditions at the site of the surveillance camera may be much different when the traveler arrives there, which may undermine the validity of such services [6.24]. This issue may be partly resolved with color-coded congestion maps depicting near-future traffic conditions from traffic forecasting algorithms and origin-destination information received from vehicles with dynamic route guidance equipment.

Real-time route guidance [6.25] involves the keying-in of origin and destination points by the motorists on the in-vehicle guidance equipment (e.g., miniature display terminal as in Fig. 6.5.1, processor, and keypad). Essential components of route guidance are

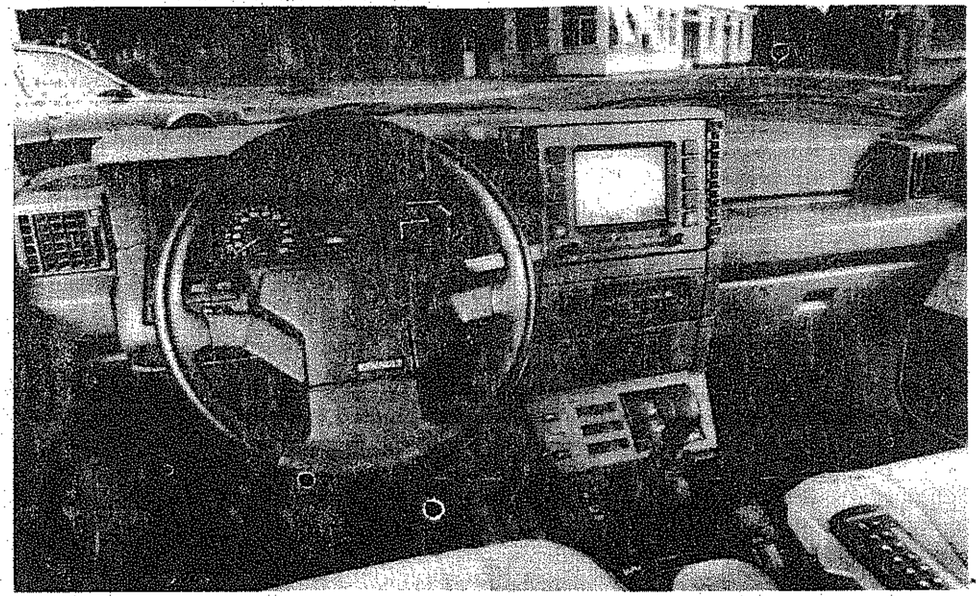


Figure 6.5.1 Example of dashboard layout including route guidance devices. (From *TR News*, No. 152, Transportation Research Board, National Research Council, Washington, D.C., 1991.)

the automatic vehicle identification and the communication of the in-vehicle guidance unit with a central computer that oversees the road network operations and advises individual motorists (through the in-vehicle equipment) or groups of motorists (with variable message signs). Navigation capabilities are based on GPS, geolocation or dead-reckoning applied to digital road maps stored on CD-ROMs for use by the in-vehicle processor. Navigation suggests the shortest path between two points with respect to distance, whereas dynamic route guidance uses information from the central computer to suggest the shortest path in terms of travel time.

Many user services depend on specific standards for successful countrywide deployment. The FHWA and ITS America support the approval of 75 MHz of spectrum in the 5.8 GHz band for dedicated short-range communications (DSRC), which are essential for many ITS services. Europe adopted this band as its standard in the mid-1990s. Eight channels of 6 Hz each have been defined in the ITS-America/FHWA specifications, as follows:

Channel	Use
1, 2	In-vehicle advisories and intersection applications (e.g., preemption, collision avoidance)
3, 4	Commercial vehicle operations on public facilities, including electronic clearance (border crossings, remote sensing for vehicle inspection, WIM)
5	Roadside-vehicle interrogation for tolling and access purposes
6	Automated highway system
7, 8	Commercial vehicle operations on private facilities

TABLE 6.5.5 Present and Future Levels of Driver Assistance

Level of assistance	Type of service
Situation analysis	Traffic information (e.g., VMS, RDS)
Situation assessment	Warnings (similar to aviation's GPWS)
Action selection	Navigation (route guidance)
Action support	Partial vehicle heading control (e.g., ICC)
Responsibility	Autcmated driving (e.g., AHS)

Source: Ref. [6.26]

TABLE 6.5.6 VMS Character Legibility: Duration (s.) of VMS in View

Size and distance conditions	Speed = 60 km/h	Speed = 100 km/h
Character height = 10 cm and Legibility distance = 60 m	3.6	2.1
Character height = 25 cm and Legibility distance = 150 m	9.0	5.4

Source: Ref. [6.27]

Certainly the most futuristic user service is the AHS, or user service No. 6.7 in Table 6.5.4(b). Based on Table 6.5.5, all five levels of driver assistance are currently available in terms of technology. Several components of AHS are already in the marketplace and large-scale experiments have been conducted in California and in GM's labs. Technological and liability issues may push implementation of AHS well into the twenty-first century.

On the other hand, ITS issues can be simple and similar in nature with standard highway design applications, such as the example on VMS character legibility shown in Table 6.5.6.

6.5.2 Architecture Components and Standards

The user services presented earlier describe detailed ways in which ITS can serve and improve transportation systems. To accomplish this in an efficient way, an architecture is necessary to define the role and interoperability of components. Electronic circuitry, data compatibility (e.g., digital versus analog signals), and data transfer protocol issues are well outside the scope of this text. Simply put, ITS compatibility should be much like the compatibility of expansion cards and peripherals for personal computers, or common video and sound systems in the marketplace that enable the purchase of components with different functions and from different manufacturers to create a well-integrated computer or a home entertainment system, respectively. In addition to compatibility and interoperability issues, architecture addresses the functional elements of ITS, which are discussed next.

A simple architecture is presented in Fig. 6.5.2. It shows that ITS serves specific transportation objectives (which have been used to form user services) and through data

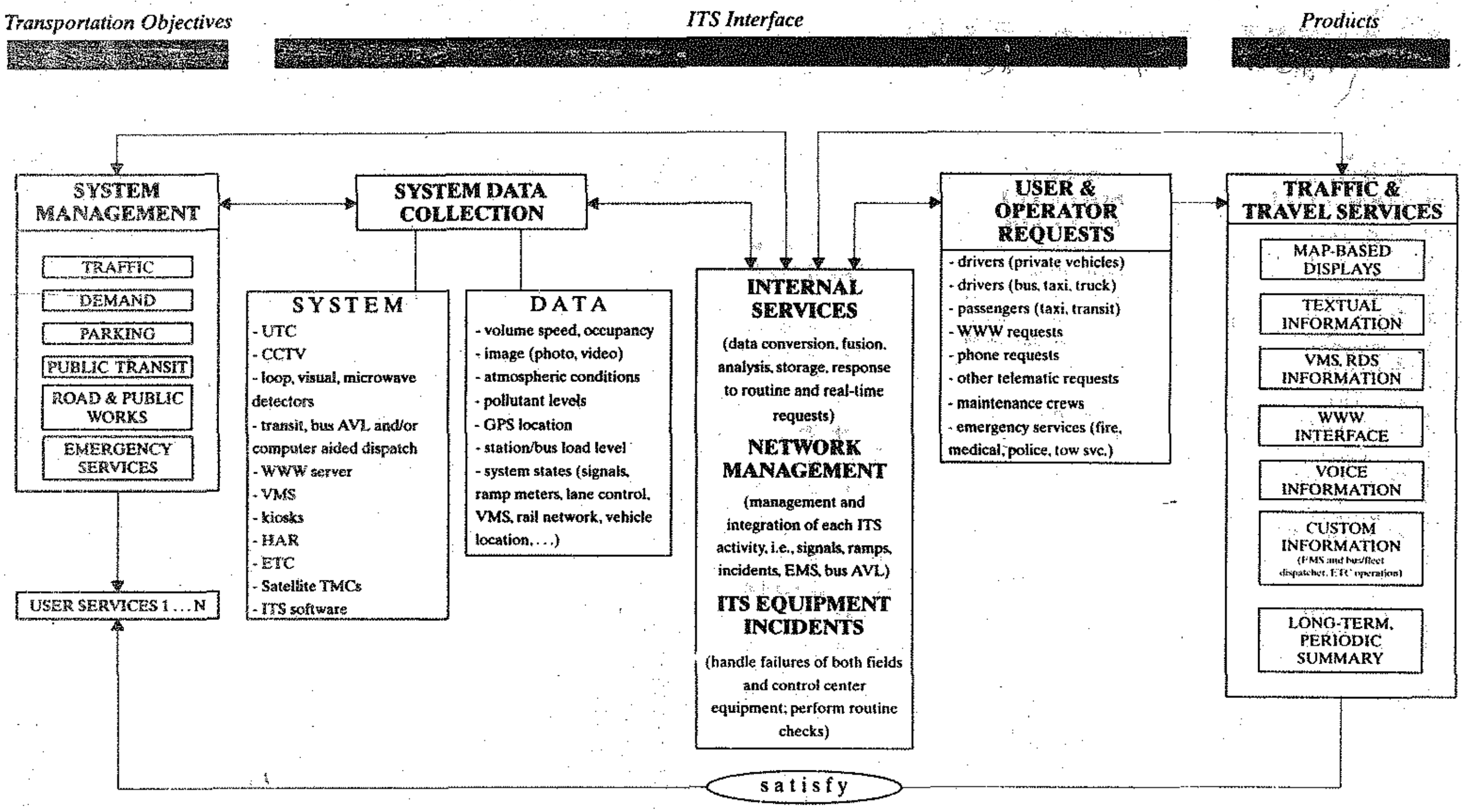

Transportation Objectives
ITS Interface
Products
SYSTEM MANAGEMENT
TRAFFIC
DEMAND
PARKING
PUBLIC TRANSIT
ROAD & PUBLIC WORKS
EMERGENCY SERVICES
USER SERVICES 1 . . . N
SYSTEM DATA COLLECTION
SYSTEM
- UTC
- CCTV
- loop, visual, microwave detectors
- transit, bus AVL and/or computer aided dispatch
- WWW server
- VMS
- kiosks
- HAR
- ETC
- Satellite TMCs
- ITS software
DATA
- volume speed, occupancy
- image (photo, video)
- atmospheric conditions
- pollutant levels
- GPS location
- station/bus load level
- system states (signals, ramp meters, lane control, VMS, rail network, vehicle location, . . .)
INTERNAL SERVICES
(data conversion, fusion, analysis, storage, response to routine and real-time requests)
NETWORK MANAGEMENT
(management and integration of each ITS activity, i.e., signals, ramps, incidents, EMS, bus AVL)
ITS EQUIPMENT INCIDENTS
(handle failures of both fields and control center equipment; perform routine checks)
USER & OPERATOR REQUESTS
- drivers (private vehicles)
- drivers (bus, taxi, truck)
- passengers (taxi, transit)
- WWW requests
- phone requests
- other telematic requests
- maintenance crews
- emergency services (fire, medical, police, tow svc.)
TRAFFIC & TRAVEL SERVICES
MAP-BASED DISPLAYS
TEXTUAL INFORMATION
VMS, RDS INFORMATION
WWW INTERFACE
VOICE INFORMATION
CUSTOM INFORMATION
(EMS and bus/fleet dispatches, ETC operation)
LONG-TERM, PERIODIC SUMMARY
satisfy

Figure 6.5.2 Abstract ITS architecture.

collection and internal data manipulation produce a number of products, the aim of which is the improvement of existing transportation services. An ITS is a system of many components, such as the traffic signal system (UTC), freeway ramp control, surveillance, computer-aided public transit dispatch, and so on. These systems produce visual and (mostly) numerical information. All these data form the backbone for the provision of user services. Most of these data need to be processed to become useful information for travelers and other users. Other data are direct replies to user requests (e.g., graphic display to the request "plot all No. 23 buses en route that are 5+ min behind schedule.")

Figure 6.5.2 lists a subset of products that ITS can produce. All of them are designed to satisfy specific user service requirements. Less obvious but essential ITS activities are the integration of ITS components, such as the combination of freeway and arterial traffic data that also are likely to be under different authorities and/or jurisdictions and the repetitive system checks, which detect the condition of system components and safeguard the integrity of the system, its data, and its products.

The major components of the U.S. national ITS physical architecture are depicted in Fig. 6.5.3. It recognizes four fundamental subsystems for travelers, vehicles, management centers, and roadside equipment. Each subsystem includes major components, several of which provide numerous functions (logical architecture). The figure shows four important functions provided by the traffic management component that (1) manages incidents, (2) provides traffic surveillance, (3) provides device control, and (4) is responsible for the overall management of travel demand. At lower levels these functions are further decomposed and the types of data that flow between them are explicitly described.

As mentioned in the introduction, a key property of ITS is communications. Figure 6.5.3 shows that four major types of communications are recognized: vehicle-to-vehicle, wide area wireless communications, wire-line communications, and short-range wireless communications. All subsystems require interfaces for communication and interoperability. Both interface devices and communications need to operate under specific standards that will make them widely compatible.

In parallel to the development of a national ITS architecture and in cooperation with standards development organizations a major effort has been launched to establish open standards (i.e., freely available) to support the interfaces between architectural subsystems. Each set consists of three types of standards: data element (D) standards in the form of data dictionaries (e.g., the traffic management data dictionary or TMDD), message sets (M), and communications profiles (C). For example, the National Transportation Communications for ITS Protocol (NTCIP) falls under the auspices of a joint AASHTO, ITE, and NEMA committee and represents a family of standards of all three types (D, M, C.) The Institute of Electrical and Electronic Engineers (IEEE) plays the lead role in the development of message sets for DSRC, AVI, and incident management, whereas the Society of Automotive Engineers (SAE) is concerned with the standard traveler information dictionary and message set and the vehicle location referencing specification among other standards. Also involved in the setting of ITS standards are the Society for Testing and Materials (ASTM) and the American National Standards Institute (ANSI.)

ITS are typically deployed at a local or regional level. As such, the decision as to what center subsystems, center subsystem functions, roadside equipment, and types of vehicles served require institutional coordination. One municipality may focus on freeway management, another on public transit management, and so forth. Large cities and metropolitan

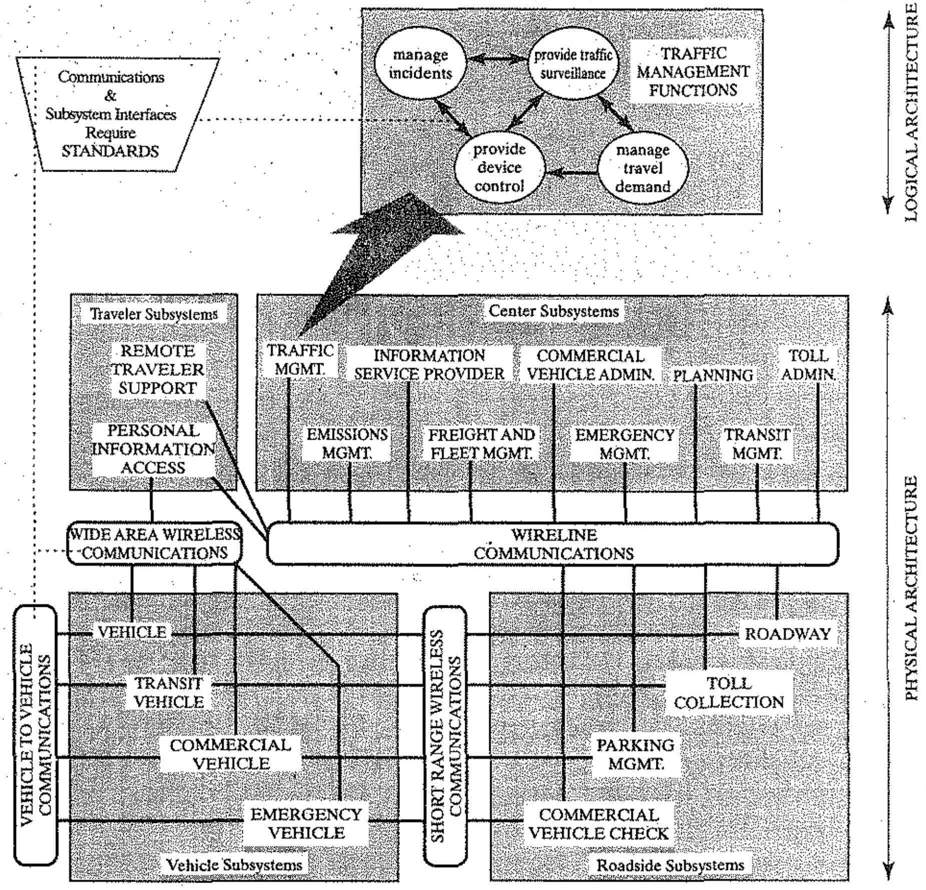


Figure 6.5.3 Major components of the U.S. national ITS architecture.

regions tend to proceed with a multicomponent approach. Based on this reality, U.S. deployment of ITS can be classed into four broad categories [6.10]: ATMS, ATIS, APTS, and CVO.

ATMS or advanced traffic management systems, such as freeway management, urban traffic control (UTP)/signal systems, incident management, and the like. A major large-scale deployment of ATMS occurred in Atlanta, GA during the 1996 summer Olympics. Major foci were incident management and signal control. Incidents were detected in less than a minute, whereas most of them would have taken 5 to 10 min to detect without ATMS.

ATIS or advanced traveler information systems, such as traffic conditions on roadside variable message signs and traffic reports at information kiosks, on the Internet, through interactive TV, and so on.

APTS or advanced public transit systems, such as the GPS-based automated vehicle location (AVL) systems, which along with computer-aided dispatching (CAD) improve

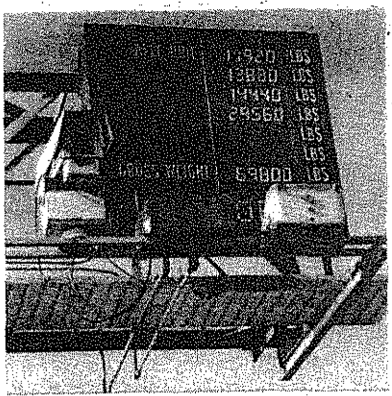


Figure 6.5.4 Weights per axle derived in Minnesota's Guidestar ITS program.

schedule adherence, lessen the "bus bunching" problem, and feed useful information to public transit ATIS systems. Kansas City reported real savings of $1.5 million in rolling stock acquisition as well as an annual $0.4 million in operating expenses due to the purchase and operation of a $2.3 million AVL/CAD system [6.28]. Furthermore, the Minnesota Travlink project, which includes both CAD and AVL as well as information signs, kiosks, and on-line information for home or office access via a modem link, reported a 6% ridership increase among Travlink users compared with the control group members who had no access to Travlink [6.28].

CVO or commercial vehicle operations focus on automated fleet management. In many respects the systems are similar to AVL/CAD for buses, but include additional features, such as speed monitoring, driver duty cycles monitoring, and electronic clearance at weight stations (Fig. 6.5.4) and borders.

ITS implementations usually begin with a modular system, which becomes progressively more integrated to take advantage of efficiencies gained by centralized processing. Systems expand in terms of geographic coverage, types of ITS equipment, and user service coverage. The majority of large ITS systems are centralized but allow for distributed functions. Multiple sources credit good agency coordination as the most important aspect of successful ITS deployments. Good coordination in management and data sharing improves efficiency and the public outlook of the agencies involved.

6.5.3 ITS in Europe and Japan

European ITS applications since the late 1980s have quantified a number of benefits, some of which are listed below. (These statistics should be viewed cautiously due to the small scale and site-specific scope of some of the projects.)

- 30% accident reduction with VMS showing traffic and weather information (85% on foggy days), and 10% CO_2 emission reduction through delay reduction
- 25% travel-time reduction for all urban travelers when traffic management, public transit priority, and real-time traveler information are offered in combination
- 41% fewer crashes due to driver monitoring systems in commercial vehicles

- 21% increase to mean motorway speeds due to monitoring strategies (automatic speed control)
- 12% improvement in crash survival rates due to GPS and mobile phone EMS notification

Based on these promising early assessments, the European Union's Council of Ministers made ITS an integral part of the overall European transportation infrastructure. Europe's ITS architecture aims to integrate urban ITS with rail transportation, such as the European rail traffic management system (ERTMS), commercial aviation, such as the global navigation satellite system (GNSS), as well as water transportation, such as the integrated vessel traffic management and information systems (VTMIS) [6.29]. Five priority areas have been identified:

- Traveler information services
- Automated fee collection
- Transport data exchange and management
- Human machine interface
- Systems architecture

Institutionally, the European approach is more top-down compared with the U.S. approach. This is necessitated by the multitude of peoples, cultures, technologies, and level of development in each country. For example, in traveler services major emphasis is given to symbolic icon-based instructions that are understandable to the citizens of any country. This requires considerable interaction among traveler services, route guidance, and human machine interfacing. Also, route guidance CD-ROMs typically include libraries that permit them to display the stored information in several languages.

Results from several large-scale projects in Europe are available, particularly from traveler information and public transportation applications. Traveler information services are focused chiefly on RDS-TMC, IVRG, and parking guidance. Public transit ITS also have multiple foci; two prominent ones are passenger information and automatic fare payment using smart-cards. Highlights of ITS deployments in Europe on these subjects are summarized next.

RDS-TMC findings of six projects in Europe [6.30] are:

- 70% were satisfied with the service.
- 70% used the service for pretrip information; 85% for on-trip information.
- 50% were willing to pay up to $175 (1995) for an RDS-compatible radio.
- Most preferred spoken rather than displayed messages.
- 20 to 24% diverted in response to a congestion warning.
- RDS radio offered warnings for 64% of relevant queues, whereas radio did 39%.

IVRG (in-vehicle route guidance) findings of five projects in Europe [6.31] are:

- 90% of users required travel-time savings; 60% required assurance for not getting lost.
- Depending on the city, 40 to 90% of the drivers reported improved comfort and reduced stress.

- IVRG had no influence on departure time but 42% of the drivers found the suggested routes better than their choices.
- Lateness of information, poor correspondence with the actual network, and unclear recommendations were the main reasons cited for noncompliance.
- About 40% of the drivers would pay $1750 (1995) for the equipment and 60% would pay about $200 for the annual service fee.

Navigation results reported by BMW from several trials using the CARIN navigation device of Phillips compared unfamiliar drivers using regular road maps with ideal navigators who were drivers very familiar with the area [6.26]. The results, displayed as "% worse than ideal navigator," show that CARIN does at least twice as good a job as an unfamiliar driver with a map. Such a navigator would be particularly useful to visitors and as a training tool to taxi and delivery companies.

	Route length (km)	Travel time (min)	Errors
Driver + road map	26%	129%	105%
CARIN navigation	15%	37%	38%

Parking guidance systems have taken a hold in large European cities where it has been found that one-third of the travel time is often spent in searching for parking. Few U.S. cities face similar problems. Several cities in Germany and Ireland as well as in Singapore, Toyota City, and St. Paul, MN [6.32] have implemented parking guidance systems. This technology is also useful for large airports and has been applied in Amsterdam and Dallas/Ft. Worth, TX.

Public transport passenger information findings of 11 projects in Europe [6.33] follow:

- Radio beacon systems performed better than GPS for AVL.
- Minimum forecast reliability for the arrival of the next bus was 75%, which is unacceptable for practical applications.
- 57 to 90% of the users were supportive of the information systems, but support was lost quickly when the systems became unreliable.
- 82% perceived the information accurate but only 28% would rely on it for travel decisions.
- When information was accurate (i.e., London), passengers perceived an improvement in travel times, even when in actuality the travel times had worsened.
- When the system indicated delays of 15 min, 10 to 26% of the passengers left the bus stop.

Smart-card payment for transportation service findings of four projects in Europe [6.34] are:

- The requirement of transaction times of less than 100 ms was difficult to achieve in the first half of the 1990s.
- The rate of error during the trials was 12 times above the desired one in one million.

- A high rate of card failure due to sticking, bending, melting, and so on was observed.
- Lack of international (or regional) standards make operators reluctant to approve electronic payment.
- 85% of the users found that the system was better than cash.

The European research project CALYPSO [6.35], which is an acronym for "Contact and contact Less telematics platform Yielding a citizen Pass integrating urban Services and financial Operations," with partners such as IBM, RATP (the Paris regional transit authority), and SNCF (France's intercity rail provider including the high-speed rail TGV) is an effort to merge transit fare payment (contactless pass) with banking and payment for services (electronic purse). The CALYPSO products that have been upgraded to correct most of the smart-card limitations listed previously are used in Constance, Lisbon, Paris, and Venice.

Similar to FHWA's user services for ITS deployment in the United States, the ITS deployment in Japan is based on 20 user services, with a heavier emphasis on commercial vehicle operations, IVRG, and explicit objectives for the guidance and safety of pedestrians. Advanced traffic management and traveler information systems, including giant color-coded overhead highway signs depicting congestion levels on the freeway network ahead have been available in Japan since the late 1980s (see picture on rear cover). Without doubt, Japan is well ahead of both the United States and Europe in several types of ITS applications. A case in point is route guidance. Cumulative car navigation system installations in Japan exceeded 2.1 million devices by mid-1997 [6.36]; 90% of these are static (map-based) guidance. The balance consists of dynamic guidance devices such as the vehicle information and communication system or VICS.

6.5.4 Mature ITS Applications

Basically these are traditional transportation processes, equipment, or services in which ITS elements were adapted reliably and produced considerable and consistent improvements. Manual and in-pavement traffic count devices have progressed to nonintrusive, portable, adjustable sensors, some of which can yield a wealth of data. Pretimed traffic signals can be replaced by demand-actuated signals with sophisticated platoon-detection capabilities. Infrequently patrolled freeways can be fully managed with ramp control, surveillance, automatic incident detection, and management, as well as real-time traffic density and speed depictions. Person and coin/barrier toll operations that severely restrict flow can be replaced by electronic toll collection systems, some of which operate at freeway speeds and do not require heavy infrastructure. These applications are presented next.

6.5.4.1 Detectors

Traffic detection is the cornerstone of many ITS services. It is accomplished with a number of sensors, the most common of which are the inductive loop detectors that were presented in Chapter 4, Section 4.6.3. To this day they serve as the benchmark for the evaluation of other types of sensors because when properly installed, their count accuracy exceeds 99%. Well-known disadvantages of loops include the expensive installation that disrupts traffic flow, work crew exposure, failures due to weather and repeated traffic loads, and destruction during construction (including pavement resurfacing). As a result, other more flexible sensors gained market acceptance and most have found use in traffic applications throughout the world.

A 2-year, large-scale study was conducted in Minnesota for the FHWA [6.37]. It focused on the evaluation of nonintrusive traffic detection devices. Nonintrusive devices are those installed overhead, at a side pole (sidefire), or pushed under the pavement from the shoulder. Eight technologies and several makes within each technology were evaluated in extreme summer and winter conditions. The fundamental principle of data collection for each technology follows:

Passive infrared sensors compare the infrared energy naturally emanating from the pavement with the energy caused by the presence of a vehicle. The actual change in heat triggers the detection of a vehicle.

Active infrared sensors emit one or more low energy laser beams at the pavement and measure the elapsed time between emission and return. When this elapsed time is shorter than usual, a vehicle is detected.

Passive magnetic sensors detect the change in the earth's magnetic field caused by the presence of a vehicle.

Active magnetic sensors are similar to inductive loops in the sense that they pass electric current through a small coil of wires and detect the inductance drop when a vehicle is present.

Doppler, radar, and millimeter microwave sensors detect either the frequency shift or the time delay of the returned signal due to the presence of a vehicle. These sensors, radar in particular, can also assess the speed of objects (including stationarity) in the detection field.

Passive acoustic sensors are basically microphones which detect the sound energy from vehicles.

Ultrasonic sensors can be a pulse or Doppler type. Pulse detection relies on elapsed time and Doppler detection relies on frequency shift.

Video sensors use video from CCTV cameras and machine vision to survey traffic. Two image processing analysis techniques, trip line and tracking, are used to "see" the traffic. The foımer technique detects the presence of a vehicle within a user-defined section of the video image, whereas the latter utilizes algorithms to identify and track vehicles within a user-defined section of the image. Commercially available devices use either or both techniques.

Video-based devices are the most intuitive as they are the closest to the "what you see is what you get" (WYSIWYG) principle. Advantages of image detection and processing include the detailed and immediate analysis of an incident by operators, the electronic (and manual) detection of queues and accidents, the emulation of loop detectors, and the derivation of traffic parameters such as occupancy and speed [6.38]. More specifically, Autoscope developers Michalopoulos and Anderson [6.39] estimated that for three-lane freeway mainlines, the cost per detector is about $3300 for conventional loops and $1000 for Autoscope-based detectors (which typically measure more flow parameters than loops).

The results on detector performance vary. Duckworth et al. [6.40] tested a large number of traffic monitoring devices including video camera, Doppler radar, Doppler ultrasound, pulsed ultrasound, passive acoustic, and passive infrared. Three performance measures were used: volume count, vehicle classification, and speed accuracy. They concluded that "with the exception of the video camera, no one sensor tested provides good performance in all three performance categories." An FHWA study [6.37] was less positive: "video required extensive installation and set-up time and is not as accurate as other technologies," but it extolled the video technology's flexibility, wealth of data, and surveillance capability (Fig. 6.5.5.)

Detailed information on the accuracy of each detector device can be found in Ref. 6.37. Selected important findings of this study are reproduced verbatim below:

- Pulse ultrasonic, Doppler microwave, radar, passive magnetic, passive infrared, and active infrared have been found to count within 3% of baseline volume data. The count results are more varied at the intersection test site. The pulse ultrasonic, passive acoustic, and video devices were generally within 10% of baseline volume data.
- Speed data were collected from active infrared, passive magnetic, radar, Doppler microwave, passive acoustic, and video devices at the freeway test site. In general, all of the devices were within 8% of the baseline data. Radar, Doppler microwave, and video were the most accurate technologies at measuring vehicle speeds.
- Video and radar devices have the advantage of multiple-lane detection from a single unit.
- Weather variables were found to have minimal direct effect on device performance, but snow on the roadway caused some vehicles to track outside of their normal driving patterns, affecting devices with narrow detection zones.
- Lighting conditions were observed to affect some of the video devices, particularly in the transition from day to night.
- Extremely cold weather made access to devices difficult, especially for the magnetic probes installed under the pavement.
- Traffic conditions, including heavy congestion, were found to have little effect on device performance.
- In general, the differences in performance from one device to another within the same technology were found to be more significant than the differences from one technology to another. It is more important to select a well-designed and highly reliable product than to narrow a selection to a particular technology.

Besides accuracy, a number of attributes and characteristics should be considered when selecting traffic detection devices. They are listed in no particular order because,

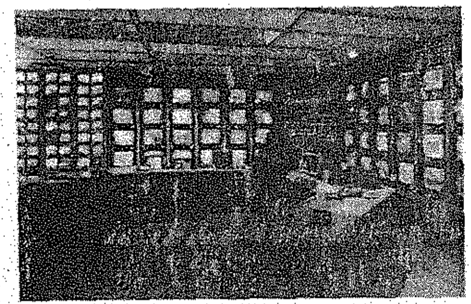


Figure 6.5.5 Traffic surveillance center. (From Transportation Research Board, *TR News*, 160, 1992)

depending on the application and technology choice-set, some may be critical and others may be irrelevant. These attributes and characteristics include:

- Expertise requirement and set up calibration time
- Reliability, typically represented by the mean time between failures (MTBF)
- Number of lanes detected by each detector
- Mounting (i.e., overhead, side fire, and height requirements)
- Installation difficulty
- Transportability
- Solar/battery power capability
- Traffic data types (i.e., counts, classification, speed, occupancy)
- Effects of light, weather, and traffic conditions on performance
- Purpose of the detection (e.g., data collection versus controller actuation; the latter obviously requires higher precision and reliability levels)

6.5.4.2 Traffic Signal Systems

Traffic signal systems are a prevalent feature of the transportation system in both large and small urban areas. For example, the results of the 1996 survey on traffic signal systems in the United States and Canada established by ITE's District 6 reflect about 150 responding municipal traffic agencies controlling more than 33,000 signalized intersections [6.41].

Characteristic	Type	Magnitude
Signal controller	NEMA	66.4%
	Type-170	19.9%
	Electromechanical	7.2%
	Other	6.4%
Detection	Loops	95.4%
	Other	4.6%
Phasing	Permissive	46.3%
	Permissive/protected	26.5%
	Protected	21.5%
	Other	5.7%
Maintenance	Annual visits/signal	5.5
	Annual emergency calls/signal	4.6
	Loop life (years)	~7
Signalization (construction) cost	Minimum	$18,000
	Maximum	$200,000
	Average	$75,000
Annual O&M cost	Average	$2700

Section 4.6.3 presented signal controllers including those that can be responsive to traffic demand, and subsequent sections in Chapter 4 showed the basic characteristics of signal timings and arterial progression. FHWA's urban traffic control system (UTCS) in the early 1970s was the first large-scale effort for the adaptive control of traffic signals. In the 1990s FHWA began pursuing adaptive control systems (ACS), which initially were referred to as RT-TRACS (real-time traffic control systems). ACS attempts to develop a suite of

adaptive control strategies that are able to respond to recurrent congestion, isolated demand spikes, extensive oversaturation, and incident and accident conditions on both arterial and grid networks. Traffic signal system evolution is described in terms of four generations: 1, 1.5, 2, and 3:

First-generation signals are the well-known pretimed (fixed timing) systems that still exist in large numbers in several metropolitan areas.

Generation 1.5 signals estimate signal timings in a quasi on-line mode. Data are uploaded from detectors and extrapolated to a standard period (15 or 60 min). Then optimization based on TRANSYT-7F is conducted and operator-audited results are downloaded to controllers. Manual supervision is often necessary to provide missing inputs and to correct for errors because these systems are usually applied wi.h limited detector coverage.

Second-generation signals offer traffic adaptive control based on averages of data collected by traffic detectors. Typical response time (for cycle, split, and offset adjustment) is about 5 min. The data gathering and signal computation process is "on-line."

Third-generation signals provide a much quicker response to demand fluctuations and as such, they are particularly capable of handling demand spikes. Their sensitivity, however, makes them unstable.

The principal components of advanced traffic adaptive systems include actuated signal controllers, traffic detection, controller interconnection, and centralized system supervision. Two types of adaptive systems are:

- Cyclic systems are based on cycle length and green splits. These systems are designed to work on a subnetwork basis (e.g., 5 to 100 signals). Examples include SCOOT and SCATS, which are discussed below.
- Acyclic systems do not depend on a cycle length. Instead they use a 30 to 300 s rolling horizon and an optimization function (e.g., a delay minimization function) to optimize traffic conditions. They work best independently. Examples include OPAC, PRODYN, UTOPIA, and SPPORT [6.42].

FHWA [6.28] reported benefits in the ranges of 8 to 15% decrease in travel time and 14 to 22% increase in speed due to advanced signal systems. Large reductions in vehicle stops (which also reduce emissions) are possible. These outcomes are corroborated by an in-house evaluation by the city of Los Angeles' Department of Transportation [6.43]. The Los Angeles automatic traffic surveillance and control system (ATSAC) was found to reduce travel time by 18% and to increase speed by 16% over the old (pretimed) system. They estimated an annual benefit per intersection of more than $230,000 and an annualized per intersection cost for upgrade to advanced standards (including both equipment and manpower) of about $7500. ATSAC has been a successful application customized for Los Angeles. Two well-known traffic responsive systems, SCOOT and SCATS, are presented next.

SCOOT (split-cycle-offset-optimization technique) was conceived in the early 1970s. As of the late 1990s, more than 170 implementations of SCOOT have taken place worldwide [6.44]. SCOOT is a real-time signal-timing optimization tool that is based on TRANSYT's optimization logic. (The TRANSYT software is presented in Chapter 15.) Besides traffic improvements, SCOOT also reduces the costs associated with the periodic revision of signal timing plans.

SCOOT's operation [6.45] is based on cyclic flow profiles (CFP), which are measured by loops or other sensors. Specifically data are collected four times a second from presence detectors placed midblock on every significant link in the network. Using the CFPs, the offset optimizer calculates the queues at the stop line. Then two other optimizers calculate the most suitable split and cycle. Cycle optimization usually occurs in small steps every 5 min, and the progression band is elastic; it stretches and shrinks depending on competing flows and queue conditions.

Applications and evaluations of SCOOT can be found dating back to the early 1980s [6.46]. More recent ones include those in Minneapolis and in Toronto. Minneapolis has implemented adaptive control based on SCOOT on a 65-signal portion of the CBD. This application uses a newer miniaturized version of Autoscope, Solo, which performs image collection, processing, and traffic detection within the camera housing. The project deployed about 140 Solo/traffic detection cameras. Early results [6.47] showed that drivers had a hard time adjusting to a nonfixed type of signal operation (large phase lengths seemed to lead them to believe that the signal was malfunctioning) and that SCOOT performed better when signals were operated in a fixed sequence (e.g., no phase extensions).

Much better results were achieved over time in Toronto. In 1990 Toronto installed the SCOOT traffic signal control system as a demonstration project at 75 intersections. Toronto's evaluation of SCOOT on two corridors and the CBD network found that, compared to a "best effort" signal-timing plan, travel time decreased by 8%, vehicle stops decreased by 22%, vehicle delay decreased by 17%, fuel consumption decreased by 6%, and CO and HC emissions decreased by 4 to 5%. It was estimated that the costs and benefits of SCOOT indicate that its installation has a payback period of less than 2 years. It was also found that the following characteristics yield higher SCOOT benefits:

- Linear networks rather than grid networks
- Roads with higher volumes and roads that are congested
- Roads with highly variable traffic flows and roads with special event or heavy diverted traffic
- Roads interacting with freeways
- Roads with exclusive turn lanes
- Roadways without on-street parking, bus stops, and pedestrian crossings

Partly because of SCOOT's unusual requirement for midblock sensors, SCATS (Sydney coordinated adaptive traffic system) was conceived by the Roads and Traffic Authority of New South Wales in the late 1970s to take advantage of many signal systems with stop-line traffic sensors [6.48]. SCATS' objective is to equalize saturation flows among conflicting approaches. As a result, it usually does not minimize delay, and major arterials may exhibit deterioration of traffic conditions during peak loads, as some results from the FAST-TRAC ITS deployment in Michigan indicates [6.49]. The FAST-TRAC project in Oakland County, MI consisted of 1000 image sensors (typically CCTV cameras tilted about 45° downward) feeding 275 Autoscopes which perform vehicle detection [6.50]. The vehicle data are fed to the SCATS system, which controls the signals in the area.

Both SCOOT and SCATS have been enhanced to accommodate bus signal preemption requirements, for example, priority treatment for mass transit vehicles. Signal preemption for buses, trams, trolleys, and LRTs is popular in Europe where they carry more than half of the commuting trips. The applicability of this concept in areas where only a small number of commuting trips are by bus; transit has been questioned, at least during peak periods [6.51]. This was the experience in Ann Arbor where "[I]t was found that in all cases signal preemption disrupts traffic progression and thus increases overall vehicle delay." [6.52] A comprehensive review of selective vehicle priority systems in the urban environment (SPRUCE project) was conducted at the University of Leeds in the UK [6.53]. The report acknowledges that some systems make an attempt to compensate for the delays to nonpriority vehicles, but it concludes that "Many of the interventions [for priority service] are made directly by the local controllers on the street and no attempt is made to compensate nonpriority vehicles for the extra delay incurred by the passage of the priority vehicle."

6.5.4.3 Freeway Management

The principal components of freeway management systems (FMS) are ramp metering and incident detection, including driver advisories for route diversion. According to the FHWA [6.28], benefits include travel time decreases by 20 to 58% and speed increases by 16 to 62%. In addition, smoother freeway flow improves throughput, and capacity can increase by 17 to 25%. Variable speed limits have produced positive outcomes in the Netherlands (Fig. 6.5.6.) and the UK where they are used as an indirect warning of downstream congestion or other hazards such as an incident or fog. Three established components of freeway management are discussed next: automatic incident detection (AID), incident management, and ramp metering.

6.5.4.3.1 Automatic Incident Detection

Incident detection is the foundation of incident management. Early detection sets the mechanism of incident management in motion to provide both relief to the distressed and to curtail congestion buildup due to capacity reduction. Nonautomated incident detection includes information from passing motorists, patrol officers, airborne surveillance, and so on. Automatic incident detection is the systematic monitoring of flow at specific cross sections (or

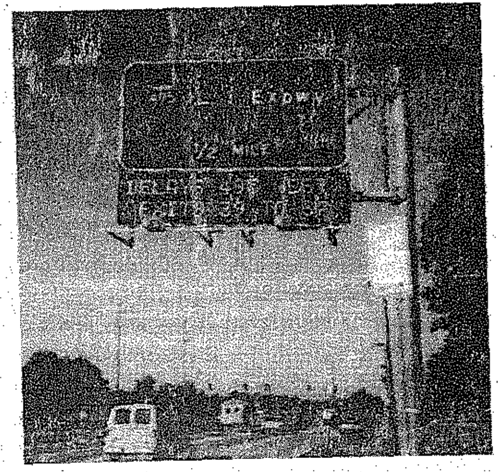


Figure 6.5.6 Electronic variable message sign capablc of providing real-time information. (From Transportation Research Board, *TR News*, 165, 1993.)

areas, if done by a video-and-image-processing technology) for the detection of any sudden changes in the characteristics of the flow, and the issuance of an alarm if specific threshold and repetition (persistence) limits are exceeded. For example, a sudden increase in speed (or decrease in occupancy) likely reflects a capacity restriction (probably an incident) upstream of the detection zone. A sudden decrease in speed (or increase in occupancy) likely reflects an incident downstream of the detection zone. Typically volume (flow rate) is not used for automated incident detection because it cannot discriminate between congested and uncongested conditions (e.g., in Fig. 3.6.1, a flow of 1200 veh/h can be sustained in both the congested (right) and uncongested (left) sections of the curve).

Techniques vary from simple occupancy observation to elaborate vehicle-matching techniques which identify vehicles along a series of detection stations and calculate section speeds. The data needed for automated incident detection are collected by detectors and fed into algorithms, which are classified in two categories: pattern recognition algorithms and forecasting algorithms. Pattern recognition algorithms examine the measured data and find values or combination of values (which are usually time-of-day dependent and they may also be day- and date-dependent) that are historically typical. When large deviations that exceed a specified threshold are observed between the historical data and the real-time data, an alarm is issued. To reduce the false alarm rate, an alarm may not be raised unless the triggering condition persists for a specified number of time periods (e.g., three or four periods for field data reception every 30 s.) Forecasting algorithms issue alarms by comparing real-time data to forecasts from recent measurements. For example, a time series of data is collected at $t_1, t_2, t_3, \ldots, t_N$. From it a forecast is made for t_{N+1}. If the actual data received at t_{N+1} differ from the forecast by more than a specified amount, an alarm is issued. Again, a signal persistence over several periods may be included to reduce the false alarm rate.

Unfortunately, the desired properties of high *incident detection rate* and short mean time to detect, and the undesired property of high *false alarm rate* go together. *Detection rate* is the ratio of incident detected over the total number of incidents. Mean time to detect is the average elapsed time between incident occurrence and incident alarm. The false alarm rate is the ratio of incident-free intervals for which an alarm was raised over the total intervals. Actually improvements in detection are followed by logarithmic increases in the false alarm rate. Evaluations of the TSC-2 algorithm (see below) in California in 1979 showed that a 56% detection rate came with a 0.005% false alarm rate, whereas a 66% detection rate came with a 2% false alarm rate [6.54]. As a result, successful algorithms settle for a 60 to 70% detection rate and mean detection times between 1 and 3 min because higher rates are accompanied by unacceptable levels of false alarm rates [6.55]. During the mid-1990s Dia and Rose created a large database containing more than 100 freeway incidents on Melbourne's freeway and developed incident detection models based on artificial neural networks with detection rates exceeding 80% and false alarm rates under 0.1% [6.56]. False alarm rates must be very low, otherwise the alarms are impractical. For example, consider a system that collects data every 20 s from ten freeway stations and has a 0.1% false alarm rate. The operators of the system will receive about two false alarms per hour ($10 \cdot 3 \cdot 0.001 \cdot 60 = 1.8$), on the average; if the false alarm rate is 1%, then they will receive one false alarm every 3 min, on the average! Video surveillance is a desirable tool of an incident detection system because it provides both confirmation of incidents and relief from false alarms.

Among the better known algorithms are the California algorithm (or TSC-2) and the modified California algorithm (or TSC-7). TSC-2 is commonly used as a benchmark for judging the performance of newer algorithms. It uses various occupancy indices as well as a decision tree logic with five possible states to arrive at the conclusion of whether or not an incident alarm should be triggered. The five indices are as follows [6.54]:

Index	Description	Definition
OCC(i, t)	Occupancy at station i for interval t	
DOCC(i, t)	Downstream occupancy	OCC($i + 1$, t)
OCCDF(i, t)	Spatial occupancy difference	OCC(i, t) − DOCC(i, t)
OCCRDF(i, t)	Relative spatial occupancy difference	OCCDF(i, t) / OCC(i, t)
DOCCTD(i, t)	Relative temporal difference in downstream occupancy	$\dfrac{\text{DOCC}(i, t-2) - \text{DOCC}(i, t)}{\text{DOCC}(i, t-2)}$

An incident typically causes a large difference between up- and downstream occupancy readings. OCCDF and OCCRDF capture these effects. In addition, a quick drop in the downstream readings may be observed. This is captured by DOCCTD. TSC-7 introduced a persistence element (i.e., repetitive detection for at least 2 min) and less reliance on temporal differences in occupancy, which may be due to congestion-and-relief shock waves generated naturally at busy periods.

There are several dozen automated incident detection algorithms, some already embedded in ITS equipment, such as Autoscope, Trafficon, and other detection devices. Incident detection has been much easier to accomplish on limited-access facilities such as freeways as opposed to open access facilities such as city arterial streets. Late versions of SCOOT inclue ASTRID, the automated SCOOT traffic information database that collects detailed data for each link. A number of analyses of historical data can be conducted with ASTRID. An additional utility is real-time incident detection on arterial streets. ASTRID identifies links with current travel times that are much higher than historical travel times. This information is relayed to operators who may use surveillance equipment or other means to confirm the event [6.57].

6.5.4.3.2 Incident Management

Although recurring congestion is a recognized problem, motorists usually compensate for it by planning their trip based on past travel-time knowledge. However, the effect of nonrecurring congestion on travel times cannot be anticipated. Nonrecurring congestion is due, for the most part, to incidents. It has been estimated that 60% of all congestion-induced delay is caused by incidents [6.57, 6.58]. The costs of incidents on the traveling public is staggering. The California Department of Transportation (CALTRANS) estimated that .2 million vehicle-hours of delay are caused by incidents each day [6.59].

An incident is defined as "any nonrecurrent event which causes reduction of roadway capacity or abnormal increase in demand" [6.60]. The majority of incidents involve stalled vehicles or accidents. Incidents also include flow interruptions caused by debris or spills on travel lanes or malfunctioning traffic signals. Minor incidents account for 65% of incident delay, whereas major incidents account for the rest [6.60].

Incident management is the process of minimizing delays caused by nonrecurrent congestion through quick detection and clearance, efficient on-site management, and prevention of incidents on major roadways. Although strategies for dealing with incidents vary from location to location, successful programs share four major components [6.61]: (1) quick detection and verification of an incident, (2) dispatch of the proper response to the incident site, (3) efficient clearance of the obstruction, and (4) recovery and management of roadways affected by the incident, which includes on-site traffic control, management of parallel arterials, and public notification of the incident.

Incident management programs may be classified as areawide, corridor, or spot locations. Areawide systems manage a number of freeways in and around a central city. Corridor systems manage selected freeways, ramps, and frontage roads, as well as parallel freeways in the area. Spot location programs are usually located on critical network components, such as bridges, tunnels, or locations where many incidents have occurred in the past. Selected features of major incident management systems in the United States are summarized in Table 6.5.7. Experience has shown that intra- and interagency coordination is the most important attribute of successful incident management programs.

6.5.4.3.3 Ramp Metering

Ramp metering was installed on the Eisenhower Expressway in Chicago in 1963 following successful metering applications in New York City tunnels and lane closures in Detroit [6.62]. By 1995 ramp-metering controls had been installed in the freeway systems of 23 metropolitan areas. Improvements of 5 to 6% in volumes carried on the freeway over premetering conditions have been observed in several areas [6.62]. The primary objective of ramp metering is to preclude freeway flow from entering into the congested regime with the subsequent sharp reductions of capacity and speed (refer to fundamental flow characteristics, Fig. 3.4.5).

A major benefit of ramp metering is the transitioning function of splitting up platoons for merging with the freeway mainline so that freeway flow disruption is lessened [6.63]. A positive concomitant outcome of ramp metering is the diversion of some short trips from the freeway. Several applications of ramp metering have reported good results:

- Seattle has experienced many benefits from their freeway management system, which includes ramp metering: Despite a 10 to 100% growth of traffic on various segments on the I-5 freeway, speeds have remained steady or increased up to 48% and accident rates have fallen to a level of 62% based on the prefreeway management period [6.64].
- The Minnesota DOT observed a capacity increase to the level of 2200 vehicles per hour per lane (veh/h/l) compared to 1800 veh/h/l prior to metering; average speeds rose from 55 to 74 km/h and accidents on a segment of Freeway I-35W dropped from 421 to 308 per year [6.62].
- Denver realized speed increases of up to 58% and a decrease in accidents of 5% during the periods when metering was on as opposed to an increase of 16% during the nonmetered periods. Furthermore, much larger than anticipated capacity gains with freeway flows of 2450 veh/h/l and a less than expected diversion to local streets were observed. Actually the prevalent form of change observed was in arrival time (at the ramp locations), thus a concomitant benefit of ramp metering was peak spreading [6.65].
- The implementation in France and the Netherlands of ALINEA, a local ramp-metering strategy, improved the base conditions without ramp metering and in several occasions was superior to coordinated ramp metering. ALINEA was found also to improve traffic conditions on the arterial network adjacent to the freeways [6.66].

TABLE 6.5.7 Sample Characteristics of IM Systems in 12 Areas

	Control and Information^{ab}						
Area	VMS	HAR	Radio station	Ramp meter	Cable TV	Kiosk	Intenet RTTR^{c}
Atlanta	41	12		5		140	yes
Baltimore/Wash., DC	173	51	ok	26	ok		yes
Chicago	20	11	ok	113	ok		yes
Detroit	57	4		57			yes
Houston	ok		ok	106		ok	yes
Los Angeles	87	12	ok	896	ok		yes
Minneapolis	60	ok	ok	420	ok		yes
New York	ok	ok	ok	ok			yes
Philadelphia	4		ok	31	ok		no
San Antonio	89				ok		yes
Seattle	51	7	ok	83	ok	ok	yes
Northern Virginia	156	19	ok	ok	ok		yes

	Incident Detection Technology^{ab}							
Area	Induct. loops	Radar	Patrol	CCTV	Call box	Mobile phone	Video detect.	Air surveil.
Atlanta	55	50		452			ok	1
Baltimore/Wash., DC	577	153	47	94		ok		ok
Chicago	2238		35	3		ok		
Detroit	2000		4	157			10	
Houston	ok	ok		ok		ok	ok	
Los Angeles	4402	89	156	218	4378		2	
Minneapolis	3300		ok	180	ok	ok	38	3
New York	ok		ok	ok	ok	ok	ok	
Philadelphia	ok	ok	ok	29	ok		1	ok
San Antonio	876			290			2	
Seattle	2600		4	205	208		3	
Northern Virginia	200		56	200		ok	10	4

Note: ^{a}ok = available but no specific information.

^{b}As of April 1996.

^{c}RTTR = real-time traffic report.

Source: Intelligent transportation infrastructure deployment at itsdeployment.ed.ornl.gov.

- A study by JHK Associates [6.67] estimated that ramp metering can reduce passenger-hours by 6.3% if there exist "good parallel arterials" (plenty unused capacity) along the metered corridor. If there are "average parallel arterials" (some unused capacity), the reduction dropped to 1.4%. In addition, ramp metering was estimated to yield undesirable results if there are "no parallel arterials" (includes arterials that operate at capacity): A 3.1% increase in passenger-hours was estimated.

Yagar [6.68] expertly summarized the benefits and disbenefits of ramp metering. Benefits include a good potential for minimizing total travel time, improvement in capacity utilization, avoidance of routes that increase system or societal costs, application of some order and control over merging maneuvers, improvement of corridor travel-time consistency, and (in several cases) public acceptance. Disbenefits include the lengthening of an average trip, the reduction of land values, the preferential treatment for through traffic (e.g., favor suburbanites), the alteration of the historical status quo, and the metering system installation and operation costs.

Several studies also present specific drawbacks to ramp metering:

- Simulation using INTRAS [6.69] on the Garden Grove Freeway in Orange County, CA concluded that a significant amount of diversion from the metered ramps must occur in order to improve the overall network performance with ramp metering, and that requires a supply of alternative routes with sufficient capacity. Best results were obtained when all overflow queues at metered ramps were diverted to arterials. Even in their best case scenario, however, the improvements were characterized as modest and nowhere near the 40 to 50% improvements shown in other studies. The authors stress that studies tend to ignore the details of the alternate routes and caution that no improvements may be realized if the alternate route network is poor.
- Researchers at the University of California at Irvine investigated the optimal ramp control problem [6.70]. They concluded that ramp metering does not improve freeway conditions when the demand-to-capacity ratio exceeds 0.8, and that under these conditions it "can have a deleterious impact on the surface street network."
- Banks [6.71] conducted ramp-metering research on San Diego's freeway system and concluded that ramp metering can eliminate mainline queuing and delay only if metering rates are set low enough to keep flows below the mainline capacity. He also asserted that "there is substantial risk that metering will be counter-productive unless it is precise" [6.72].
- Hellinga and Van Aerde [6.73] used the INTEGRATION traffic simulation software to investigate ramp-metering strategies. They added support to Banks' contention that ramp metering needs to be precise in order to be effective. In one case study they found that "initiating ramp metering just 2 minutes earlier than optimal can negate any metering benefits."

In addition to these concerns, major design elements make the successful implementation of ramp metering problematic. The maximum discharge flow of a metered, single-lane on-ramp is 900 veh/h; the metering of ramps with higher volumes is problematic and requires extensive analysis [6.62] to assess the impact of traffic diverted onto the surface network. The three primary elements of successful metering in addition to moderate demand are storage space, adequate acceleration distance, and sight distance [6.62].

Despite this "mixed bag" of results, ramp metering is often seen as a successful and seasoned ITS component. Ramp-metering systems integrated with manual or automated incident detection and response as well as metered ramp bypasses for high occupancy vehicles are growing. Negative neighborhood reaction, jurisdictional disputes, and substantial implementation costs are the main reasons for the relatively few applications worldwide.

6.5.4.4 Electronic Road Pricing and Automatic Vehicle Classification

Tolls have been in existence since the creation of the first improved roadways in the United States. These roads were private and their development and maintenance costs were paid through the collection of tolls. The collection was automatic with the assistance of a mechanical device (called a turnpike). Many modern freeways in the eastern United States are still called turnpikes. After World War II many roadway projects were financed through tolls. Much later, in the 1970s, the concept of concession financing* arose in Europe. This scheme of fixed-duration, for-profit, private roadway development requires the collection of tolls.

Three basic characteristics of toll systems are:

- Charges are distance based and are differentiated by type-of-vehicle (vehicle classification).
- Closed toll systems have tolls at all entrances and exits; the toll is determined at the exit point based on the point of entry.
- Open toll systems have tolls that are located along segments of the facility; a toll is assessed for each segment traversed.

Tolls have been implemented using various traditional technologies such as manned toll collection, automated (electromechanical) coin or token collection and area licensing schemes (e.g., special stickers on vehicles). A major drawback of these systems is their inefficiency (i.e., they tend to develop congestion and pollution of their own) and expense (i.e., high cost of manned operations, machine and/or violator inspectors). A recent development, electronic toll collection (ETC), has become the most successful real-world application of ITS due to its ease of installation and very large and immediate cost savings [6.16, 6.74].

ETC consists of both in-vehicle and roadside equipment. The in-vehicle device for the automatic vehicle identification (AVI) is a transponder (transmitter and responder). Three systems of AVI transponder technology are distinguished; they are listed here in a chronological order of development:

1. *Read-only.* In-vehicle transponder transmits information to roadside unit; all records remain with the tolling authority only.
2. *Read-write.* In-vehicle transponder transmits information to roadside unit and can store information from the roadside unit; both the user and the tolling authority retain a record of the transaction.
3. *Smart-card.* Integrates the above with other user services (i.e., public transit and parking payment, credit and banking card) and offers advanced encryption option, "mayday" function, and so on.

*Concession financing for motorways is quite common in France and Spain. Several small sections of motorway concessions exist in the UK and other countries are considering concessions. In the United States, California's SR-91 State Highway is a major success. SR-91 is open only to AVI-tagged vehicles and the tolls are congestion-based, that is, the toll charge increases as demand approaches SR-91's capacity so that congestion is controlled and a good level of service is offered.

The roadside equipment consists of antennas installed on an existing overpass or a light metal structure called a *gantry.* Redundant antennas as well as video enforcement devices and network lines to the supervising computers complete the basic infrastructure. Some systems also include detectors for the arrival and departure of vehicles in the detection area. Communication between the roadside antennas and the in-vehicle transponder takes place several times per second in the detection area. After identification is made a charge is made and confirmed. Actual payment occurs through either periodic bills sent to subscribers, subtraction from the prepaid card, or debiting to a smart-card linked to a bank account. Most modern ETC systems are entirely paperless and able to operate almost flawlessly (e.g., one error in one million transactions) at freeway speeds.

ETC in urban areas combines naturally with other ITS services and forms what is known as electronic tolling and traffic management (ETTM). For example, AVI transponders used for tolling can be polled by strategically located receivers (probe vehicle sampling).

This type of time and space data can be used to assess the level of congestion along specific roadway segments, which in turn can be disseminated to motorists (i.e., user service 1).

In 1998 there were 29 U.S. metropolitan areas with ETC systems. Based on 1995 U.S. DOT statistics from 12 tolling authorities, the benefits of ETC systems include a 90% decrease in operating expenses; a 250% increase in capacity; 6 to 12% decrease in fuel consumption; and a 40 to 80% decrease in CO, HC, and NOx emission (per affected kilometer). Congestion pricing is the force behind the expanding application of toll systems in urban areas. An application of congestion pricing during 1994 to 1995 in Stuttgart yielded the following results [6.29]:

- 12.5% of peak-period trips shifted to the off-peak period.
- Up to 5% of weekday trips and 15% of Saturday trips shifted from auto to transit.
- HOV trips increased by 7%.

There are, however, several barriers to the implementation of ETC. *Privacy* typically refers to the public's fear of being recorded by the government. This does not appear to be an issue with the (several) existing systems even in countries where personal freedom is a primary constitutional guarantee (e.g., the United States). *Payment enforcement* for the capture of violators usually relies on automated video capture and license number recognition.* *Coordination among different toll agencies/AVI compatibility* is a problem among countries in Europe and states in the United States. No short-term solution is available, other than the adequate provision of manned or coin toll lanes for motorists with no or different AVI equipment. *Deployment in existing plazas* causes the loss of a large part of the advantages of ERP (e.g., tolling at near free-flow speeds). Speeds drop well below free-flow levels, safety risk increases as motorists speed through toll booths, and capacity drops by more than 30%. *Lack of technology standards* is continuously debated because both sides of the issue have advantages. Specifically, standardization may accelerate deployment but the standardization of a (possibly inferior) dominant technology may hinder technological progress and longer-term efficiency. Lastly, *equity* applies mostly to congestion pricing. At issue here is the distribution of benefits and costs within the society, which is a major con-

*Similar techniques are employed for capturing motorists who violate speed limits or enter an intersection while the signal displays red (Fig. 6.5.7).

sideration, particularly for decision makers. A road pricing charge is viewed as inequitable to the poor because it constitutes a larger portion of their disposable income (this is akin to regressive taxation, which is any fixed tax). Others argue that congestion pricing imposes a direct charge on motorists for the true cost of their trip (as a function of both infrastructure cost and, importantly, congestion and environmental consequences).

Automatic vehicle classification (AVC) is an integrated system of detectors and processing units that permits the identification of the class* of vehicle so that (1) the proper toll is charged with an automated tolling system, (2) AVC double checks the toll collection at person-operated plazas (fraud deterrent), and (3) accurate data for planning and design can be collected (i.e., for pavement rehabilitation, commerce, and tax studies, etc.) Automated toll collection and toll fraud/error reduction are the primary uses for AVC.

There is no common way for vehicle classification. Some toll authorities have many vehicle classes (e.g., CALTRANS has 17), whereas others have few (e.g., Toronto's 407 Highway has 3). As a result of the variable requirements illustrated earlier, AVC systems need to detect a series of characteristics including length, height (highest point, or specific height over each axle), vehicle profile, axles, distance between axles (wheelbase), number of tires and tires per axle, and weight. Speed is used to correct the frame capture information collected with optical curtains for vehicle profiling.

AVC can occur before or after the toll booth. The first is called *pre-classification* and helps the operator in double-checking the vehicle's class; it can also be used to display the proper toll to the motorist. Because of this, the AVC equipment must be placed in advance of the booth. For example, an advance length of 30 m would suffice if light curtains are used, but a distance of 100 m may be needed in the case of high approach speeds. The second is called *post-classification* and is used as a verification of the assessment either by the human toll operator or for the determination of the correct charge in an ERP (electronic road pricing) system [6.74].

Various types of equipment, usually in combination when detailed classification is required, are used for AVC (Fig. 6.5.8); they include:

- *Inductive loop detectors,* some forms of which can detect axles.
- *Treadles,* can be electromechanical, optical, resistive rubber, or piezoelectric. They typically detect axles. Diagonal placement helps to detect single- or double-wheel axles. Treadles can also be used for weight-in-motion detection (WIM).
- *WIM devices,* such as bending plates, capacity strips, or piezoelectric sensors, can weigh vehicles.
- *Light beams* can detect vehicle presence and height based on infrared light interruptions.
- *Light curtains* are an extension of light beams. They can produce a longitudinal two-dimensional profile of a vehicle.
- *Scanning devices* such as ultrasonic, infrared, and laser scanners can produce vehicle profiles.
- *Video image processing* usually can classify vehicles on the basis of length.

*Vehicle class is typically defined on the basis of vehicle size, number of axles, and actual or maximum allowable weight.

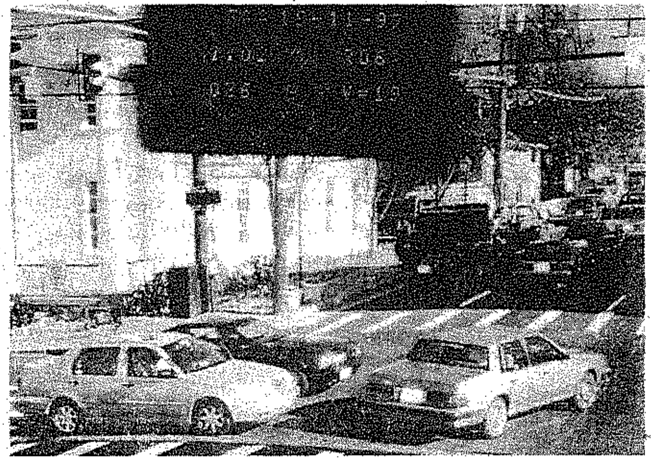


Figure 6.5.7 Automated red-light running and speeding enforcement. (From Transportation Research Board, *TR News*, 201, 1999.)

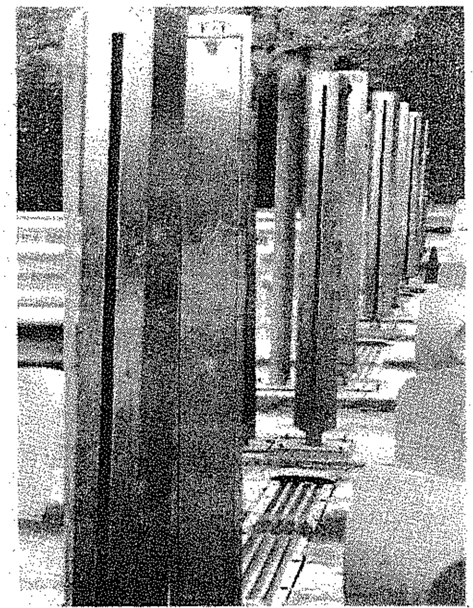


Figure 6.5.8 Pre-classification consisting of optical sensors and treadles at an intercity highway toll plaza in Greece. (Photograph by P. D. Prevedouros.)

Most AVC systems that combine at least two detection subsystems are reported to be able to offer a 98% or better vehicle class detection accuracy. This has important implications for tolling authorities, which can realize reductions in revenue loss and practically eliminate mistakes and fraud.

6.5.5 Safety and Liability

According to the National Highway Traffic Safety Administration (NHTSA), roadway crash costs exceeded $150 billion in 1994; the estimate represents the lifetime loss of those injured and killed in traffic crashes. The Bureau of Transportation Statistics (BTS) reports that traffic crashes in 1995 injured 3.4 million people, 428,000 of whom were incapacitated [6.10].

Collision warning, collision avoidance, night vision, and drowsiness detection are primary ITS technology components aimed at improving roadway safety. Collision detection systems utilize a telesensing device (typically a radar) that gathers information about the conditions around the vehicle (usually within 10 m sideways and up to 300 m ahead and behind). Audio or visual collision warnings are given if the driver fails to adjust the speed or position of the vehicle to maintain a safe distance or to avoid a collision (e.g., abort lane changing maneuver or commence deceleration). Collision avoidance systems are more sophisticated. An in-vehicle computer controls engine and brake functions, which enable it to intervene and adjust speed, acceleration, or deceleration characteristics so that a collision is avoided, or its impact is less severe. Intelligent or adaptive cruise control is a current application of this genre of ITS technologies.

Night vision and artificial vision are systems that are able to detect hazardous conditions ahead, which the driver ignores or has no ability to realize, for example, pedestrians, approaching too fast to a curve, hidden or approaching too fast to a stop sign, yield sign, or red signal, particularly under reduced visibility conditions such as at night, rain, or fog. Drowsiness warning consists of a sensor attached to the interior rear-view mirror, which monitors the rate of eye blinking. Blinking intervals tend to last longer and to occur at longer intervals at the onset of drowsiness. A warning is issued when a hazardous blinking pattern is detected.

The basic architecture of ITS deployment shows that information is generated as a means to achieve the objectives of user services. It is clear that several of these objectives go well beyond the realm of information and well into the realm of control, for example, the adaptive cruise control discussed earlier. An important issue, therefore, is the liability involved with ITS services [6.75]. Let us first give a very brief description of liability.

U.S. liability law is a part of tort law that governs the resolution of disputes for wrongful acts. The basic premise of liability is negligence (i.e., failure to exercise due care) on the part of the involved parties, which typically are a subset of the following parties: the driver(s), the operator(s) of the subject roadway(s) and equipment (e.g., traffic signals), the contractors and subcontractors of the roadway and its parts (e.g., lane striping and traffic signing), and the manufacturer(s) and maintainer(s) of the vehicle(s) and their individual components.

Most of ITS user services are not different from other widely available electronic products and services. However, due to the breadth of applications, a large variance in the liability of ITS services exists. For example, electronic tolling and in-home traveler advisories are practically liability-free. In-car traveler services as well as mobile telephony are likely to have a somewhat higher level of liability. Intelligent cruise control and automated highway systems that essentially transfer all or part of vehicle control to computers and machinery are likely to translate into substantial liability for the manufacturers of the vehicle, the intelligent equipment, and the operator of the highway facility. Further analysis of this important subject is well beyond the scope of this book.

6.6 SUMMARY

Urban transportation is an important area of transportation study, given that the vast majority of populations in developed countries reside in such areas. A brief overview of the historical, parallel evolution of cities and transport was given. Readers may want to relate this to the section on urban transportation planning in the next chapter. In this way they will have a rough historical background of the evolution of urban areas, urban modes, and urban transportation planning and engineering.

A universally vexing urban problem is traffic congestion. Thus the presentation focused on its consequences and on well-known congestion reduction strategies. Intelligent transportation systems offer novel solutions and assist in making transportation systems safer and more efficient. All ITS are local and require extensive cooperation among professionals (planners, civil and electrical engineers for starters) and agencies (at a minimum, the city department of public works, county traffic office, state DOT, and all emergency and police departments serving the subject area). In response to the complexity of ITS and in an attempt to aid local jurisdictions in the selection of ITS components based on local priorities and probable success scenarios, the FHWA sponsored the development of the ITS deployment analysis system (IDAS) software, which integrates planning inputs with traffic modules adjusted to respond to locally selected ITS components [6.76].

EXERCISES

1. Review the user services as well as the mature ITS applications. Which of each exist in your area (or in the metropolitan area closest to your location)? Select a user service or an ITS application that exists in your area and describe it in detail (e.g., generic and specific features, public/motorist response, years in operation, scale of application, etc.).
2. Make a detailed list of all the modes available in your urban area separated by major uses or classes such as air-international, truck-container, and so on. A thorough list could be surprisingly long. Honolulu, which does not have light, rapid, or commuter rail, has about 40 modes of transportation!
3. ACT 290/July 1, 1997 of the Hawaii Revised Statutes is reproduced here. Can you locate the flagrant error in the wording of this law in its attempt to provide incentives for electric vehicles?

ACT 290 (S.B. NO. 1160) A Bill for an Act Relating to Electric Vehicles

Be It Enacted by the Legislature of the State of Hawaii:

SECTION 1. The legislature finds that the State relies primarily on the consumption of imported oil to satisfy its energy needs. The legislature further finds that because oil is a limited resource, the State must develop and implement mechanisms to reduce the consumption of oil and other petroleum-based products in Hawaii.

The legislature further finds that the residents of the State consume a large quantity of gasoline for motor vehicle use. According to recent statistics, there are over 900,000 registered motor vehicles on Hawaii's roads and highways. Because of this, Hawaii's drivers consumed over 375 million gallons of gasoline in 1990.

One possible mechanism of reducing the consumption of petroleum products is to promote the use of newer technologies in everyday life. The legislature recognizes that many advances have already been made in the field of transportation. The emergence of alternatives to fossil-fueled vehicles has the potential to significantly reduce our dependency on petroleum-based products.

The purpose of this Act is to:

1. Improve the transportation of people and goods through the expanded use of electric vehicles by undertaking a program of financial and regulatory incentives designed to promote the purchase or lease of such vehicles;

2. Obtain the benefits to the state economy of lessened dependence on imported petroleum products through greater reliance on vehicles that utilize domestically-produced electricity as a source of energy; and
3. Preserve and enhance air quality by encouraging the widespread use of vehicles that are emissions-free in operation.

SECTION 2. It is the policy of the State to support the development and widespread consumer acceptance of electric vehicles within the State. This policy is intended to accelerate the use of a substantial number of electric vehicles in the State to attain significant reductions in air pollution, improve energy efficiency in transportation, and reduce the State's dependence on imported oil or petroleum products. Exempting electric vehicles from various requirements applicable to conventional, internal combustion engine-powered vehicles may encourage operators to choose electric vehicles.

SECTION 3. The department of transportation shall:

1. Establish and adopt rules pursuant to chapter 91, Hawaii Revised Statutes, for the registration of electric vehicles in this State; and
2. Establish and issue a special license plate to designate that the vehicle to which the license plate is affixed is an electric vehicle.

SECTION 4. An electric vehicle on which a license plate described in section 3 is affixed shall be exempt from:

1. The payment of parking fees, including those collected through parking meters, charged by any governmental authority, other than a branch of the federal government, when being operated in this State; and
2. High occupancy vehicle restrictions or other traffic control measures.

SECTION 5. For a period of 5 years from the effective date of this Act, the motor vehicle registration fee and other fees, if any, assessed upon or associated with the registration of an electric vehicle in this State, including any fees associate with the issuance of a license plate described in section 3, shall be waived; provided that the department of transportation shall review the incentive program every two years to determine the proper level of incentives for continuation of the program.

SECTION 6. This Act shall take effect on July 1, 1997.

4. Selected characteristics from a handful of large ETC systems are presented below (all reflect 1997–1998 conditions). Search literature sources as well as the Internet to update the information shown here:

Singapore. The Phillips Singapore-Mitsubishi consortium was awarded a contract after extensive trials of three competing consortia. It received an order for 1.06 million AVI tags and the implementation of a ETC system with 60 gantries. Motorcycles, vibration, and radio-frequency interference have been reported as challenges to achieving accuracy goals.

Toronto. The modern boothless ERP system on Highway 407 is open for vehicles with and without AVI tags. They charge 10¢/km for passenger cars during peak, 7¢ off-peak, and 4¢ at night time. Vehicles without transponders pay the same tolls plus a fixed charge of $1 (all figures in Canadian currency); they are identified and billed with a camera-based license plate recognition system.

New York. EZ-pass implementation of ERP through conventional plazas. Traffic speed improved from a crawling 12 to 18 km/h to a flowing 40 km/h. The people-operated lanes on the Tappan Zee Bridge toll plaza serve 350 to 400 veh/h; AVI toll lanes have a top service rate of 950 to 1000 veh/h.

Orlando-Orange County Expressway. The Florida DOT has contracted Amtech to expand the SunPass ETC system with AVI through existing toll plazas at a cost of $39 million for 455 toll lanes. The measured throughput of the dedicated AVI lane increased by 154%.

Oklahoma Turnpike. ETC in operation since 1991 has helped to achieve savings by attrition (retirements), not layoffs. The annual cost to operate a automatic lane is estimated at $15,800 versus $176,000 for a manually supervised lane.

5. Use the Internet site referenced in Table 6.5.7 and the ETTM-on-the-web and California PATH program's Learning from the Evaluation and Analysis of Performance (LEAP) Internet database to identify other locales with ETC and to develop summaries similar to those in Exercise 4.

REFERENCES

6.1 MABEE, N. B., and B. A. ZUMWALT, *Review of Downtown People Mover Project Proposals,* The MITRE Corporation, Urban Mass Transportation Administration Report No. UMTA-IT-06-0176-77-1, 1997.

6.2 HOEL, L. A. (Ed.), *Advanced Urban Transportation Systems,* Transportation Research Institute, Carnegie-Mellon University, Pittsburgh, PA, 1970.

6.3 ENO FOUNDATION FOR TRANSPORTATION, *Transportation in America: A Statistical Analysis of Transportation in the United States,* 8th ed., Washington, DC, 1990.

6.4 RIBNER, R. H., *Ridership Operations,* Transportation Research Board, Special Report 193, 1981.

6.5 ESSENER VERKEHRS, A. G., *Spurbus Essen: Information on the Research and Development Project Guided Bus Essen,* 1986.

6.6 EUROPEAN CONFERENCE ON MINISTERS OF TRANSPORT, *Changing Patterns of Urban Travel,* Paris, 1985.

6.7 U.S. DEPARTMENT OF TRANSPORTATION, *1990 Nationwide Personal Transportation Survey,* Washington, DC, 1991.

6.8 PAVLOU, S., *Negotiating for Part-Time Operators in Seattle, Washington,* Report CETP-TS-78-2, University of Hawaii, Honolulu, HI, 1978.

6.9 THOMSON, W., *A Preface to Suburban Economics,* in the Urbanization of the Suburbs, Sage Publications, pp. 409–430, 1973.

6.10 BUREAU OF TRANSPORTATION STATISTICS, *Transportation Statistics Annual Report 1997,* BTS97-S-01, U.S. DOT, 1997.

6.11 URBANIC, T., *Management of Surface Transportation Systems,* National Cooperative Highway Research Program, Synthesis 259, Vol. 1, TRB, National Research Council, 1998.

6.12 ALLEN, W., D. LIU, and S. SINGER, "Accessibility Measures of U.S. Metropolitan Areas," *Transportation Research,* 27B (1993): 439–449, Pergamon Press, Oxford, UK.

6.13 TRANSPORTATION RESEARCH BOARD, *Curbing Gridlock,* Special Report 242, Vols. I and II, National Research Council, 1994.

6.14 LOMAX, T. et al., *Quantifying Congestion,* National Cooperative Highway Research Program, Report 398, Vol. 1, TRB, National Research Council, 1997.

6.15 CERVERO, R., *Suburban Gridlock,* Center for Urban Policy Research, Rutgers University, New Brunswick, NJ, 1986.

6.16 Learning from the Evaluation and Analysis of Performance (LEAP) Internet Database at www.path.berkeley.edu/~leap/EP/Electronic_Payment/, Partners for Advanced Transit and Highways (PATH), 1996.

6.17 STATE OF HAWAII, DEPARTMENT OF TRANSPORTATION, *Hawaii Telework Center: Demonstrating Innovative Ways to Reduce Traffic Congestion,* 1989.

6.18 FIELDING, G., and D. KLEIN, "Hot Lanes: Introducing Congestion-Pricing, One Lane at a Time," *Access,* No. 11, University of California Transportation Center, Davis, CA, Fall 1997.

6.19 JAGODA, A., and M. DEVILLEPIN, *Mobile Communications,* John Willey, Chicester, England, 1993.

6.20 EATON, D., M. DASKIN, D. SIMMONS, B. BULLOCH, and G. JANSMA, "Determining Emergency Medical Service Vehicle Deployment in Austin, Texas," *Interfaces,* 15(1): 96–108, 1985.

6.21 SCRASE, R., "Smarter Fire Fighting: In-Cab Computers Improve Effectiveness and Safety," *ITS Int'l,* May/June 1997.

6.22 REISS, R., and R. GORDON, "Telecommunication Design: A Logical Approach," *COMTrans,* August /September 1996.

6.23 U.S. DOT, *The National Architecture: A Framework for Integrated Transportation into the 21st Century,* CD-ROM, Washington, DC, 1998.

6.24 LYONS, G., and M. MCDONALD, "Traveller Information and the Internet," *Traffic Engineering and Control,* January 1998, pp. 24–31.

6.25 ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Route Guidance and In-Car Communication Systems,* Road Transport Research Series, 1988.

6.26 REICHART, G., "Driver Assistance: Concepts of the Future Individual Mobility," *Traffic Technology Int'l,* 1997.

6.27 STAINFORTH, R., "LEDs Light the Way," *ITS Int'l,* September/October 1997.

6.28 The MITRE Corporation, *Intelligent Transportation Infrastructure Benefits: Expected and Experienced,* Report to the FHWA, 1996.

6.29 HARRIS, R., *Transport Telematics,* Report for European Community Directorate Generale XIII, The Lingfield Press, Ltd., 1997.

6.30 KATTELER, H., *Acceptance and Impacts of RDS/TMC Traffic Information: Results of the ATT Cross-Project Collaborative Study,* CORD AC16, Vol. 3, Brussels, 1995.

6.31 BECCARIA, G., and M. HOOPS, *In-Vehicle Route Guidance: Results of the ATT Cross-Project Collaborative Study,* CORD AC16, Vol. 4, Brussels, 1996.

6.32 PALMER, D., "Managing Urban Parking Space," *ITS Int'l,* March 1996.

6.33 CASSIDY, S., and G. TANNER, *Public Transport Information: Results of the ATT Cross-Project Collaborative Study,* CORD AC16, Vol. 6, Brussels, 1996.

6.34 BLYTHE, P. T., and N. THORPE, *Smart Cards in Transport: Results of the ATT Cross-Project Collaborative Study,* CORD AC16, Vol. 7, Brussels, 1996.

6.35 EUROPEAN COMMISSION, DIRECTORATE GENERALE XIII, *CALYPSO,* Telematics Applications Programme, 1998.

6.36 HIGHWAY INDUSTRY DEVELOPMENT ORGANIZATION, *Handbook: Intelligent Transport Systems in Japan,* 1997.

6.37 MINNESOTA DOT and SRF CONSULTING GROUP, INC., *Field Test of Monitoring of Urban Vehicle Operations Using Non-Intrusive Technologies,* Report FHWA-PL-97-018, 1997.

6.38 BOGAERT, M., L. CYPERS, and F. LEMAIRE, "Safety and Efficiency by Video," *Traffic Technology Int'l,* 1997.

6.39 Michalopoulos, P., and C. Anderson, "Costs and Benefits of Vision-Based, Wide-Area Detection in Freeway Applications," *Transportation Research Record* 1494 (1995): 40–47.

6.40 Duckworth, G. et al., "A Comparative Study of Traffic Monitoring Sensors," *Proceedings of 1994 ITS America Meeting,* Atlanta, GA, April 1994.

6.41 McCourt, R. S., *Traffic Signal Maintenance and Design Survey,* ITE District 6, 1998 (Visit the ITE District 6 website for updated survey summaries.)

6.42 Conrad, M., F. Dion, and S. Yagar, "Real-Time Traffic Signal Optimization with Transit Priority: Recent Advances in the SPPORT Model," Presented at 1998 Annual Meeting of the TRB.

6.43 City of Los Angeles Department of Transportation, *ATSAC Evaluation Study,* June 1994.

6.44 Bretherton, D., K. Wood, and N. Raha, "Traffic Monitoring and Congestion Management in the SCOOT UTC System," Presented at 1998 Annual Meeting of the TRB.

6.45 Harris, R., and M. Juha, "Making Sense of Traffic Surges," *ITS Int'l,* No. 6: 82–85, September 1996.

6.46 Hunt, P. B., D. I. Robertson, R. D. Bretherton, and R. L. Winton, *SCOOT: A Traffic Responsive Method of Coordinating Signals,* Report 1014, TRRL, Crowthorne, UK, 1981.

6.47 Ragsdale, P., "AUSCI-SCOOT: A Transatlantic Partnership for Minneapolis," *Traffic Technology Int'l,* December/January 1998.

6.48 Lowrie, P., SCATS, *Sydney Co-Ordinated Adaptive Traffic System: A Traffic Responsive Method of Controlling Urban Traffic,* Roads and Traffic Authority of New South Wales, Traffic Control Section, 28 pp., 1990.

6.49 Wolshon, P., and W. Taylor, "Analysis of Intersection Delay under Real-Time Adaptive Signal Control," Presented at 1998 Annual Meeting of the TRB.

6.50 Vinger, S., "AUTOSCOPE and SCATS Together in FAST-TRAC," *Traffic Technology Int'l,* August/September 1997.

6.51 Garrow, M., and R. Machemehl, "Development and Evaluation of Transit Signal Priority Strategies," Presented at 1998 Annual Meeting of the TRB.

6.52 Al-Sahili, K., and W. Taylor, "Evaluation of Bus Priority Signal Strategies in Ann Arbor, MI," *Transportation Research Record* 1554 (1996): 74–79.

6.53 Institute for Transport Studies, *Selected Vehicle Priority in the UTMC Environment,* University of Leeds, 1998 (www.its.leeds.ac.uk/projects/spruce).

6.54 Solomon, M., *A Review of Automatic Incident Detection Techniques,* ADVANCE Project, Northwestern University, Evanston, IL, August 1991.

6.55 Stephanedes, Y., and J. Hourdakis, "Transferability of Freeway Incident Detection Algorithms," *Transportation Research Record* 1554 (1996): 184–195.

6.56 Dia, H., and G. Rose, "Development and Evaluation of Neural Network Freeway Incident Detection Models Using Filed Data," *Transportation Research,* Pergamon Press, Part C, Vol. 5, No 5, (1997): 313–331.

6.57 Bretheron, D., "Current developments in SCOOT: Version 3," *Transportation Research Record* 1554 (1996): 48–52.

6.58 Transportation Research Board, *Freeway Corridor Management,* National Cooperative Highway Research Program, Report 177, National Research Council, 1992.

6.59 Transportation Research Board. *Freeway Incident Management,* National Cooperative Highway Research Program, Report 156, National Research Council, 1990.

6.60 Reiss, R. A., and W. M. Dunn, *Freeway Incident Management Handbook,* FHWA, 1991.

6.61 Kasamoto, K., and P. Prevedouros, *Incident Management in Honolulu,* Research Report UHM/CE-98-05, Department of Civil Engineering, University of Hawaii, Honolulu, HI, 1998.

6.62 Piotrowitz, G., and J. Robinson. *Ramp Metering Status in North America: 1995 Update.* FHWA, U.S. Department of Transportation, 1995.

6.63 Elefteriadou, L., R. Roess, and W. McShane. "Probabilistic Nature of Breakdown at Freeway Merge Junctions," *Transportation Research Record* 1484 (1995): 80–89.

6.64 Henry, K., and O. Meyhan, *6 Year FLOW Evaluation,* Washington State DOT, District 1, 1989.

6.65 Concoran, L., and G. Hickman, "Freeway Ramp Metering Effects in Denver," *Compendium of Technical Papers,* ITE Annual Meeting, 1989.

6.66 Papageorgiou, M., H-. Haj-Salem, and F. Middelham, "ALINEA Local Ramp Metering: Summary of Field Results," Paper presented at 1997 Annual Meeting of the TRB, Washington, DC.

6.67 Alexiadis, V., and J. Schmidt, "Ramp Metering: A System Concept Design Methodology," *Proceedings of ITS America Annual Meeting,* Vol. 2 (1994): 861–866.

6.68 Yagar, S., "Predicting the Impacts of Freeway Ramp Metering on Local Street Flows and Queues," *Compendium of Technical Papers,* ITE Annual Meeting, 1989.

6.69 Nsour, S., S. Cohen, J. Clark, and A. Santiago, "Investigation of the Impacts of Ramp Metering on Traffic Flow with and without Diversion," *Transportation Research Record* 1365 (1992): 116–124.

6.70 Zhang, H., S. Ritchie, and W. Recker, "On the Optimal Ramp Control Problem: When Does Ramp Metering Work?" Paper presented at 1995 Annual Meeting of the TRB, Washington, DC.

6.71 Banks, J., "Performance Measurement for a Metered Freeway System," Paper presented at 1988 Annual Meeting of the TRB, Washington, DC.

6.72 Banks, J., "Two-Capacity Phenomenon at Freeway Bottlenecks: A Basis for Ramp Metering?" *Transportation Research Record* 1320 (1991): 83–90.

6.73 Hellinga, B., and M. Van Aerde, "Examining the Potential of Using Ramp Metering as a Component of an ATMS," *Transportation Research Record* 1494 (1995): 75–83.

6.74 *ETTM on the Web* at www.ettm.com.

6.75 Roberts, S., " Liability and ITS," *Traffic Technology Int'l,* October/November 1997.

6.76 Lee, R., "IDAS: Planning ITS into the Mainstream," *Traffic Technology Int'l,* August/September 1998.

7

Transportation Planning

7.1 INTRODUCTION

Much has been written about the subject of planning and the role of the professional planner in various societal functions. One hears of urban, economic, financial, corporate, industrial, water resource, environmental, and many other kinds of planning. In the field of transportation, professional designations, such as highway planner, airport planner, and urban transportation planner, are common. Clearly planning is considered an important function in modern society, and whatever this function is, it has a specific focus, that is, it concentrates on particular areas, subjects, or systems.

For the purposes of this book, planning may be defined as the activity or process that examines the potential of future actions to guide a situation or a system toward a desired direction, for example, toward the attainment of positive goals, the avoidance of problems, or both. As the conceptual, premeditative process that precedes a decision to act in a certain way, planning is a fundamental characteristic of all human beings. However, as a focused professional discipline, planning is viewed in a wider, yet bounded, context.

The most important aspect of planning is the fact that it is oriented toward the *future:* A planning activity occurs during one time period but is concerned with actions to be taken at various times in the future. However, although planning may increase the likelihood that a recommended action will actually take place, it does not guarantee that the planned action will inevitably be implemented exactly as conceived and on schedule. Another time element of importance to planning's forward-looking perspective is the lag between the time when the action is to be taken and the time when its effects are felt. This time lag depends on many factors, including the scope and magnitude of the contemplated action.

It is often said that everything is related to everything else. Therefore any event or human action affects everything else, ultimately in ways that are beyond the limits of human comprehension. As a matter of practicality, planning is not a search for ultimate answers but only a means to specific ends based on the proposition that better conditions would result from premeditative as opposed to impulsive actions. How much premeditation is necessary (i.e., how much planning is good planning) in a particular situation is always an open question: Too little planning is almost like no planning, and too much planning is self-defeating, as it leads to inaction.

By necessity, any particular planning effort has a limited scope and is oriented toward bringing about specific desirable ends. Whereas desirability cannot be divorced from the value system of human beings, planning is necessarily directed toward the satisfaction of the goals and objectives of particular groups of people. Within its social context, however, planning cannot afford to ignore the reactions of other groups; it must in fact, anticipate these responses as well. In addition, when the group on behalf of which planning is undertaken is heterogeneous, the planning effort must deal with the presence of internal conflicts relating to specific objectives and aspirations. This is especially critical when the government participates in or regulates the planning effort.

The fundamental purpose of transportation is to provide efficient access to various activities that satisfy human needs. Therefore the general goal of transportation planning is to accommodate this need for mobility. Within specific contexts, however, such questions as whose mobility, for what purpose, by what means, at what cost and to whom, and who should do the planning and how, are not amenable to easy answers. Contemporary responses to these questions are largely rooted in history and have been influenced by a confluence of many factors, including *technological innovations, private interests,* and *governmental policies.*

The purpose of this chapter is to illustrate the dynamic nature of transportation planning by briefly tracing this evolutionary process for the case of land transportation. The main objective of the chapter is to bring about an appreciation of the complexity of the transportation planning methodology, which certainly has not reached a historical finality, and to highlight, in broad strokes, its most fundamental elements. More detailed coverage of the colorful history of the U.S. transportation system and of evolving transportation planning issues may be found in the literature (e.g., Refs. [7.1–7.8]).

Section 7.2 presents the development of the major intercity and urban transportation systems in the United States. The reason that this material is included here goes beyond a mere interest in history for its own sake. The intent is to help the reader follow how the elements of the contemporary transportation planning process and related planning methods evolved.

For each historical period the reader is encouraged to contemplate several important questions, particularly those relating to the effect of technology and the ever-evolving dynamic relationship between the roles of government and the private sector. Among these are the following:

1. What was the extent and technology of the transportation system?
2. What were the pressing transportation concerns of the time?
3. Who had the responsibility for the planning, design, and operation of transportation facilities and services?
4. What direct and indirect government actions at the local, state, and federal levels influenced the development of the transportation system and the establishment of transportation planning requirements?
5. What was the rationale for these governmental actions?

A review of Section 1.3, which discusses the role of government in general terms, including its motives and instruments, is recommended at this point as it provides a general framework for understanding the specific governmental actions discussed in this chapter.

Section 7.3 addresses the development of a formal urban transportation planning process. It shows that modern urban transportation planning has, to a considerable degree, evolved in response to pressing social, economic, and environmental concerns.

Section 7.4 provides the background to the development of contemporary transportation planning methods and techniques to place the material covered in Chapter 8 in its proper perspective. Section 7.5 presents some wider issues related to transportation planning applications, including land-use modeling.

7.2 HISTORICAL DEVELOPMENT IN THE UNITED STATES

7.2.1 Colonial Era

The migration of European settlers to North America, which led to the establishment of cities, occurred by sea. Thereafter water transportation using the coastline and natural inland channels was the major form of long-distance transportation. Chartered privately owned ferries offered for-hire service on the rivers. Movement over land utilizing human and animal power was cumbersome and was inhibited by topographic obstacles. Following English practice, the responsibility for planning, building, and maintaining roads rested with local jurisdictions. The use of these primitive roads by travelers was free of charge and their construction and maintenance were accomplished primarily by statute labor, an English practice that required all men over 16 years of age to work on the roads on appointed days.

7.2.2 Turnpikes and Canals

Following the War of Independence, the new nation was underdeveloped, and its immediate transportation needs were primarily related to simple *accessibility* between the cities lying primarily on the eastern seaboard and toward the unexplored lands to the west. Available technology consisted of wagons and coaches for land transportation and boats and barges for movement on rivers and canals. Two major problems that had to be overcome were topography and finance. Because of debts incurred during the war, the states were not in a financial position to accommodate increasing demands to provide the needed facilities. Some states adopted the practice, introduced in England a century earlier, of allowing private companies to plan and construct transportation linkages and to charge tolls for their use. This practice ushered in the turnpike and canal eras, during which hundreds of companies were chartered by the states to operate as regulated monopolies. Turnpikes were named for a pike, or pole, that was turned to allow access to the roadway after the payment of the toll. Minimum design standards such as roadway width and maximum gradients were usually included in the charter requirements. The states participated in varying degrees in the construction of these facilities by conferring the right to eminent domain for the taking of land and building materials, by subscribing to company stock, and in some cases by providing direct subsidies. The most notable canal, the Erie Canal, opened in 1825.

At the beginning the role of the federal government was confined to the building of military roads. Otherwise a strict interpretation of the U.S. Constitution commanded respect for the sovereignty of the states in matters of local concern including transportation. Indirectly, however, the federal government aided transportation development through land surveys of its territorial holdings, the sale of which was its major source of income. The first national transportation facility inventory was undertaken in 1807 by the secretary of the treasury, Albert Gallatin, who in his report a year later clearly recognized the importance of a good national transportation system to the growth and unity of the nation. Gradually the federal government took carefully measured steps toward land grants to the states for roads and canals. It also moved toward the allocation of proceeds from land sales to the states to be used for transportation development, among other purposes. A notable exception to the hands-off federal policy was the construction, after bitter debate, of the first national road, the Cumberland Road, through Maryland, Virginia, and Pennsylvania. This road was later extended to Ohio, Indiana, and Illinois. Inadequate congressional appropriation for maintenance and a determination that the federal government was not empowered to charge user tolls led to the transfer of the road to the aforementioned states.

On the technological side, turnpike pavement construction of the heavily traveled routes initially employed the French method of building a heavy stone structural foundation. By 1820 the macadam method (named after the Scottish engineer John McAdam) was preferred. This method relied on the strength of the native soil, over which thin layers of small stones were packed for protection. In the area of water transportation Fulton's successful, although not original, demonstration of steam power on the Hudson in 1807 enhanced the efficiency of this mode and eliminated the need to pull canal boats by horses, which walked along the shore. The network of individually planned turnpikes and water lines formed the basic long-distance transportation system of the nation until the arrival of the railroads.

7.2.3 Railroads

In their infancy during the early part of the nineteenth century railroads were expected merely to provide a better roadbed for animal-drawn vehicles. The introduction of the railroad steam engine altered this notion and gave rise to railroad companies essentially as they are now known. Railroad planning and development followed the paradigm of turnpikes and canals but on a grander scale. The states extended charters to railroad companies on a line-by-line basis. When a standard gauge evolved, short lines were consolidated into fewer but larger entities. The superior performance and efficiency of the new technology caused the demise of most private turnpikes and canals.

The federal government contributed to the development of the railroads in the West via land surveys conducted by the U.S. Army Corps of Engineers between 1824 and 1838, by the imposition of tariffs on imported iron in 1832 to support the United States iron industry, and, beginning in 1850, by land grants to the railroads. In return, the recipients of land grants agreed to carry the U.S. mail and troops. The first transcontinental railroad was ordered by the Pacific Railroad Act, signed into law by President Lincoln in 1862, and was completed in 1869 with the driving of the golden spike at Promontory, UT. The oligopolistic advantage enjoyed by the railroad companies began to be moderated by a series of government regulatory actions, which led to the passage of the 1877 Interstate Commerce Act. With regard to the railroads, this law provided for the regulation of rates and included several anticollusion clauses. The Interstate Commerce Commission (ICC) was established by this act to carry out its provisions.

As mentioned earlier, the railroads were favored by the government through the granting of eminent domain, that is, taking privately owned land for "public" purposes at "just" compensation. This, along with the fact that each railroad operates on its own right-of-way, generated excess capacity because of a multitude of parallel lines developed between destinations. Until 1975 the railroad industry was heavily regulated by the federal government through the ICC whose power was strengthened by the Hepburn Act of 1906 and again by the Transportation Act of 1920. By the 1960s the railroads were encountering serious competition from the trucking industry that were supported by the large federal outlays toward the construction of the interstate highway system (see the next section). Hampered by heavy economic regulation, the railroad industry appeared to be on its deathbed. Due to their decline, abandonment of trackage was a necessary step for a large number of railroads. However, abandonment had the potential to cause major economic impacts to the regions served by the lines proposed to be abandoned; thus there was substantial resistance for the approval of such proposals.

The federal government came to the rescue with the signing of the Rail Passenger Service Act in the fall of 1971. This act established the National Railroad Passenger Corporation. Initially called "Railpax" but eventually named *Amtrak* ("American, Travel and Track"), this quasi-public corporation was chartered by Congress to operate almost all intercity rail passenger service under contract with the railroads and with heavy capital and operating subsidies from the federal government.

On the freight transportation side, the Regional Rail Reorganization Act of 1974 created the Consolidated Rail Corporation, *Conrail*. The intent of this act was to allow Conrail to rebuild and operate the rail freight network in the Northeast and Midwest regions of the country and, upon reaching profitability, to return control to the private sector. Conrail commenced operations in 1976. In the rest of the country, particularly west of the Mississippi, a series of mergers and acquisitions facilitated by loosened regulatory structures, resulted in a small number of consolidated railroad companies, such as Union Pacific (UP) and Burlington Northern and Santa Fe (BNSF).

The Railroad Revitalization and Regulatory Reform (4R) Act of 1976 made abandonment of trackage easier for railroads, while it was shown that hardship in the affected areas was hardly experienced because the motor carrier industry quickly filled the void. Furthermore, the 1980 Staggers Act gave the railroad industry the ability to compete under free-market conditions by making major changes in the regulation of rates. The next year marked Conrail's first year of profitability, and the system was sold by the government group of private investors in 1986.

Two major federal actions were taken during the mid-1990s. In December 1994, faced with severe budget deficits, the U.S. Congress debated whether to continue its support of Amtrak and concluded by directing the company to take actions toward self-sufficiency by the year 2002. In response Amtrak developed a strategic plan that included the strengthening of partnerships with state departments of transportation, expanding nonpassenger revenues from retail operations at major stations, leasing its right-of-way for the placement of fiber optic cables, and expanding its mail and express service operations. It also announced the selection of a vendor to construct, by late 1999, the first high-speed rail service capable of reaching speeds of 150 mi/h between Boston and Washington, DC. In August 1998 Amtrak unveiled the Midwest Regional Rail Initiative, a visionary plan conceived in cooperation with several states and the Federal Railway Administration. This plan included a 3000-mi high-speed (110 mi/h) network centered in Chicago to be fully implemented by the year 2006.

The second major federal action that affected freight and passenger rail services was the passage of the 1995 Interstate Commerce Commission Termination Act. This act abolished the ICC, eliminated certain onerous ICC functions, and transferred the remaining economic regulations to the Surface Transportation Board, an independent adjudicatory body administratively housed within the U.S. DOT. It was under this new environment that the major initiatives and consolidations described earlier occurred.

7.2.4 Rural Highways

Under local control, statute labor laws, and other sources of local support, urban streets were kept in a reasonably adequate condition. Outside the cities, however, the same could be said only for short roadway spurs connecting farms and towns. The importance of good roads was appreciated but the means necessary to plan, finance, construct, and maintain them was lacking. Toward the end of the nineteenth century a good-roads movement swept the country. Newly formed associations of recreational bicyclists who had begun to brave the countryside played a pivotal role in this movement, which was consistent with the later strengthened U.S. tradition of organized citizen participation in public planning and decision making.

The states assumed an active role in 1880 by extending aid to the counties and municipalities for the construction of public highways and by establishing highway or public roads commissions empowered with varying degrees of advisory, supervisory, and planning responsibilities. The Commonwealth Highway plan enacted by the Massachusetts legislature in 1894 was perhaps the first attempt at planning a connected statewide network of public roads.

Two experimental programs, both approved in 1893, marked the formal reentry of the federal government in the planning of rural highways. First, the U.S. Congress approved a mail-delivery experiment on specially designated rural routes beginning in 1896. Later program expansions provided a strong incentive to the states to improve certain roadways in order to qualify for mail-route designation. Second, a temporary Office of Road Inquiry (ORI) was established within the U.S. Department of Agriculture to undertake research in road-building methods and to disseminate its findings to the states. The ORI's 1899 successor, the Office of Public Road Inquiry (OPRI), was merged in 1905 with the Division of Tests of the Bureau of Chemistry that had played an instrumental role in road-building materials research to form the Office of Public Roads (OPR), the precursor of the Bureau of Public Roads (BPR). Much later, in 1966, the BPR was absorbed into the Federal Highway Administration within the then-formed U.S. Department of Transportation.

The early planning-related pioneering work of the ORI included the preparation of a national inventory of macadamized roads, the Good Roads National Map, and the compilation of statistics relating to road usage, including quantified data on trip lengths and user costs for the transportation of farm products. To help meet its charge to disseminate its findings relating to road-building technology, the ORI initiated the construction of short segments of demonstration roads. In the meantime technological advances were rapidly bringing the motor vehicle (electric, steam, and gasoline-powered) to the forefront of U.S. transportation. In 1912 a congressional appropriation was approved, which authorized an experimental determination of the ability of good roads to effect savings to the U.S. Post Office's rural mail delivery. The first full-fledged federal highway aid was extended to the states via the Federal Road Act of 1916, which established a 50–50 construction-cost sharing between the federal government and the states, with the federal contribution to each state determined by a formula. During World War I the federal government seized the railroads and established the Federal Railway Administration to operate them. In addition, it designated a number of military highway routes for use by an increasing number of heavy trucks. This development had three major effects on highway planning after the war: (1) It supported the growth of long-distance trucking; (2) it renewed the need for improved roadbuilding; and (3) it emphasized the need for a physical continuity of the intrastate and interstate highway systems.

The Federal-Aid Highway Act of 1921 required each state to designate up to 7% of their existing highways as part of a national system that would be eligible for federal aid. The BPR established cooperative agreements with some states to aid in the planning and location of this system. In the process innovative survey methods and studies were developed that had a profound effect on the shaping of transportation planning. These studies included:

> the ownership of motor vehicles; the seasonal, monthly and daily variations in traffic; the origin and destination of cargoes; the size and weight of trucks. . . . In later studies, they examined driver behavior—the average speeds of drivers traveling freely on the highway and their observance of traffic laws, such as those prohibiting passing on hills and curves. In Maine, the researchers discovered a historical relationship between vehicle ownership, population and traffic. By projecting historical trends ahead, they were able to make fair estimates of traffic 5 years in the future [7.1, p. 122].

Highway construction benefited during the Great Depression from federal work-relief programs. In 1934 the Hayden–Cartwright Act permitted the expenditure of up to 1.5% of federal highway funds to be used for "surveys, plans, and engineering investigations of projects for future construction." This led the states, through their highway departments, to undertake massive needs studies by projecting population, traffic volumes, and vehicle ownership trends into the future. The projections were used to identify highway capacity deficiencies, which in turn guided the overall highway planning effort.

The Federal-Aid Highway Act of 1944 set the nation's sights toward a national system of interstate highways and provided for urban extensions of this system. In 1945 the American Association of State Highway Officials (AASHO), a cooperative organization established in 1914 during the good-roads movement, adopted a set of geometric design policies, which became the precursors of the contemporary standards discussed in Chapter 2.

The worldwide political instabilities that followed World War II led to a requirement in the 1948 Federal-Aid Highway Act for a cooperative federal–state study to assess the nation's highway system from the perspective of national defense. This study revealed significant deficiencies in many aspects, including a lack of adequate and uniform geometric designs, and encouraged a stronger federal role in the planning of national highways. About the same time and because of insufficient highway revenues collected from user charges, the states embarked on the construction of high-standard toll highways that mostly followed or paralleled the national interstate system. Most of these publicly owned turnpikes opened between 1948 and 1954 when this type of state financing appeared to be the direction of the future. This trend, however, was reversed after the extension of federal funding to the secondary system of highways and the passage of the Federal-Aid Highway Act of 1956. This act and its companion Highway Revenue Act redefined the roles of the federal government and the states in the area of highway planning and had a profound effect on the evolution of a formalized planning process. Among its major provisions were the following:

1. It mandated the construction of the *national system of interstate and defense highways,* the largest single public works undertaking of its kind, in accordance with high and uniform design standards.
2. It established a 90–10% federal–state funding basis for this interstate highway system and provided that the larger federal portion be paid from revenues collected in the form of user taxes and charges that were to be placed in a special *highway-trust fund* for this purpose.
3. It required the conduct of supportive planning studies and extended the requirement for the conduct of related *public hearings* in relation to project location, first set forth in 1950, to all federal-aid projects.

7.2.5 Urban and Regional Transportation Planning

The extension of federal highway aid to urban areas that began with the 1944 Federal-Aid Act brought about a division of interests between state concerns for interstate highway system continuity on one hand and urban concerns related to local circulation and the growing urban traffic problem on the other. This fusion of perspectives contributed to the refinement of formalized planning processes and methodologies.

During the preindustrial era many European cities were laid out according to the conceptual designs of notable city planners. These physical plans reflected a primary concern with the aesthetic qualities of city form and the dominant location of symbolic structures such as cathedrals and palaces. In this connection several schools of thought evolved. Early American cities can be traced to these city planning traditions. Williamsburg, VA, Savannah, GA, Philadelphia, PA, and L'Enfant's plan of the nation's capital are notable examples of this trend.

During the late eighteenth and most of the nineteenth centuries the growth patterns of U.S. cities were driven by a speculative fever, and city planning, as practiced earlier, disappeared from the scene. Most new cities and towns were built on land hastily subdivided into gridiron street patterns that followed the lines of government surveys. The same spirit permeated the growth of older cities as well; the original city plans were all but ignored, and random development of every available parcel of land ensued. Aided by the technological breakthroughs of the industrial revolution, cities entered a period of rapid expansion by absorbing a massive influx of people from rural areas and from abroad. Industrial and economic forces and the centralizing influence of railroad transportation necessitated the concentration of the population in cities. Among the major technological innovations were the steam engine, which facilitated the growth of the railroads, and the elevator and frame construction method, which made the skyscraper possible.

Urban planning reemerged as a professional discipline during the later part of the nineteenth century. This reemergence was aided by four developments that redefined the government's *power of eminent domain* (i.e., the power of taking private rights and property for public purposes with just compensation) and the *police power* (i.e., the power to regulate the use of private property). Not necessarily in chronological order, the first development affecting modern city planning was related to an increasing involvement of city governments in the alleviation of slum conditions and the urban ills associated with them. New York enacted its first tenement law in 1867 to regulate building structures in an attempt to minimize fire hazards and to enhance the living conditions of the population. The origin of modern *building codes* can be traced here. The second trend was the practice of *districting*, or *zoning*, where the government assumes the power to regulate the use of land, for example, for commercial, industrial, or residential purposes. Zoning was first applied in Germany in 1884 and spread through Europe before reaching the United States. The term *zone* originally referred to a concentric ring, or belt, with the central city at its center but was later taken to mean a land area of any shape and location. The first comprehensive zoning code in the United States was enacted by the city of New York in 1916. The third precursor of modern city planning is evident in the *parks movement* advocated by an emerging group of planners and landscape architects, who, praising the virtues of pastoral life, saw public parks as having a beneficial influence on the otherwise drab existence of urban populations. Land for the first major urban park in the United States, New York City's Central Park, was acquired in the 1850s. The fourth major antecedent of modern urban planning was an advocacy for the beautification of *public buildings* within dominating civic squares, the descendents of the cathedral and palace of the earlier planning philosophy. Taking its lessons from European city forms, this movement developed during the closing years of the nineteenth century. The *city-beautiful* planning movement of the early 1900s was a planning philosophy distilled from these trends, which emphasized the construction of monumental civic centers and urban park systems connected by wide boulevards. It helped to legitimize the need for planning and brought into the planning process the civil engineering profession and its road-building and structural-engineering techniques. During this time the cities began to establish planning commissions composed of influential civic leaders to guide the formulation and implementation of plans, relying on a cadre of consultants to carry out the work. Both practices continue to this day.

The introduction of streetcar services (see Chapter 6) exerted a decentralizing influence on the urban form by facilitating an outward expansion that reached beyond the city limits in a spokelike pattern. This radial growth left an indelible impression on the way planners thought of cities for decades and directed their attention to the urban region rather than merely to the city proper. Public transportation was provided by private companies that were franchised as public utilities and operated on city-owned streets. Proponents of city planning argued that the award of franchises should be made in an orderly fashion and in accordance with regional city plans.

Consistent with the age of mechanization, a new planning perspective gradually emerged that saw the city as a machine. Although more complex than other mechanical devices, the efficient functioning of its interrelated parts was considered to be amenable to scientific treatment. According to the developing *city practical* (also known as *city efficient* and *city scientific*) planning movement of the 1920s and 1930s, transportation was seen not only as the skeleton of a static city plan but as a force that affects the future shape of the city. As such, it was reasoned, transportation planning should become an integral part of urban planning.

The decentralizing influence of public transit was not confined to residential suburbanization but extended to industrial decentralization as well. Many new towns were built along the principles set forth in Graham R. Taylor's book *Satellite Cities* [7.9], which was influenced by Ebenezer Howard's earlier work on garden cities [7.10]. These principles are reflected in the design of modern suburban communities.

In step with the scientific approach to planning was an increasing reliance on measurement and prediction, for example, the conduct of physical, economic, and demographic inventories; the collection of transportation usage data; and the attempt to discern the relationships among these factors. This kind of planning activity was contemporaneously occurring in relation to intercity highways.

By accelerating residential and industrial decentralization not only along fixed radial routes but also in the spaces between them, the motor vehicle caused another revision of planning practice and also the evolution of a specialized profession, *traffic engineering*, the scope of which was initially restricted to the orderly expansion of street capacity, parking facilities, and traffic control strategies to accommodate the quality and safety of ever-increasing automobile flows. Among the major contributions of traffic engineering were advanced traffic and driver-behavior studies and the modeling of land-use, population, and traffic-demand relationships. A rudimentary model of land use and traffic was applied in San Juan, Puerto Rico, to plan a freeway system serving a new airport. The invention of the high-speed digital computer made possible the analysis of large quantities of data and the development of more sophisticated planning methods.

By the 1940s urban planning became an established function of city governments. The *master planning* era of the next two decades emphasized the production of comprehensive regional zoning maps and specified the planned location of infrastructural systems, including transportation, water supply, and sewage facilities. Figure 7.2.1 shows a portion of the 1950 zoning map of the city of Honolulu, HI [7.11].

The extension of federal aid to the urban portions of the national highway system in 1944 marked the definite entry of the states into the urban planning scene. This development brought together three overlapping professional perspectives, the interaction of which was fundamental in the evolution of transportation planning: through their highway departments, the states that were predominantly concerned with the connectivity of the intrastate and interstate highway network; the city traffic engineering departments that were primarily concerned with accommodating the efficient and safe operation of the urban street network; and the city (or in some instances, regional) planning departments, which were concerned with regional land-use planning, housing, and urban public transportation. Out of this interaction emerged a generally shared, quantitatively based land use-transportation planning methodology.

7.3 DEVELOPMENT OF A FORMAL PLANNING PROCESS

7.3.1 Housing Policies

Direct federal involvement in the area of housing is evident in the Home Loan Bank Act of 1932 and in the National Housing Act of 1934, which established the Federal Housing Administration (FHA). The Housing Act of 1949 set up the Housing and Home Finance Agency (HHFA) and appropriated funds for the elimination of urban blight through the clearance of slums and the redevelopment of the areas occupied by them. The Housing Act of 1954 shifted the emphasis of this program from leveling and rebuilding to urban renewal,

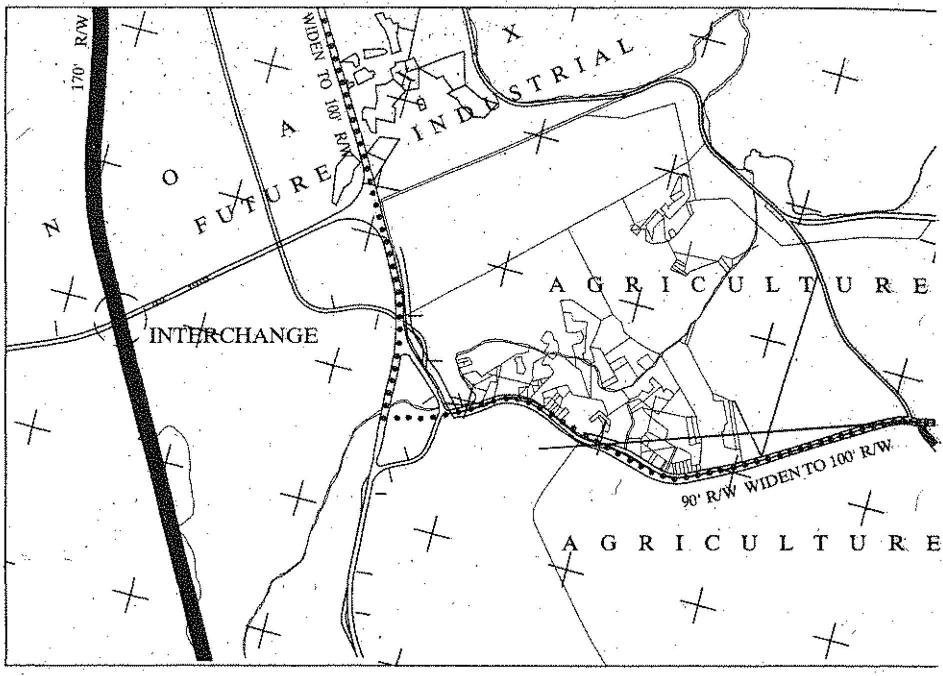


Figure 7.2.1 Portion of a land-use map.
(From City and County of Honolulu [7.11].)

which included the rehabilitation and preservation of existing structures in accordance with a general plan for each locality. Grants for transportation planning, including comprehensive traffic surveys and studies, were provided by the Housing Act of 1961 and administered by the HHFA, which in 1965 became part of the newly formed Department of Housing and Urban Development (HUD). The Demonstration Cities and Metropolitan Development Act of 1966 included additional assistance to urban transit and required the designation by each urbanized area of a regional organization to oversee the orderly development of the program.

7.3.2 The 3C Process

The federal-aid highway program was also gaining a planning perspective that was being adapted to the conditions found in urban areas. The Office of Planning was established within the BPR in 1961, and the Federal-Aid Highway Act of 1962 mandated that after 1965 state eligibility for federal highway aid in major cities would be conditioned on the existence of long-range plans that would:

> . . . be based on a *continuing, comprehensive* transportation planning process carried out *cooperatively* by states and local communities . . . [emphasis added]

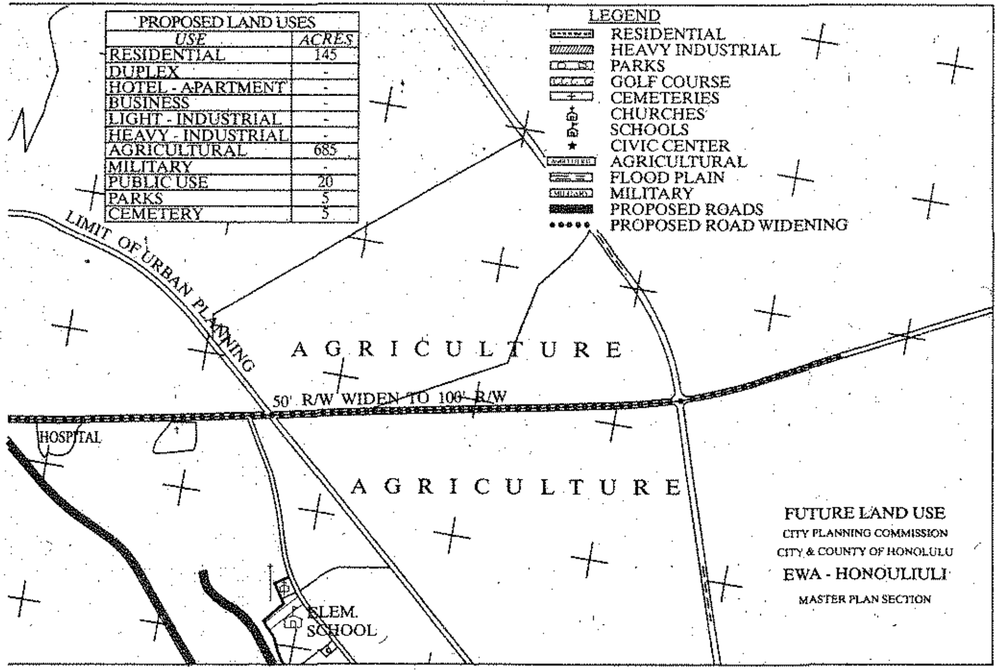


Figure 7.2.1 *continued*

The act also declared that it is:

> . . . in the national interest to encourage and promote the development of transportation systems, embracing various modes of transport, in a manner that will serve the states and local communities effectively and efficiently.

This act officially established the requirement for the *3C planning process.*

7.3.3 Social Concerns

Beginning with the movement for racial equality, which led to the passage of the Civil Rights Acts of 1964 and 1968 and other housing and equal opportunity laws, the decade of the 1960s was a decade of social concern. Social activists held the view, sometimes justified and sometimes exaggerated, that the slum clearance and urban highway programs were signs of a "bulldozer mentality" that was insensitive to social needs. These groups learned to organize and lobby at the local, state, and federal levels and to seek judicial relief of their grievances. The provision of public transportation was viewed as one means of addressing the needs of disadvantaged groups.

The Urban Mass Transportation Act of 1964 provided capital assistance for urban public transportation to be administered by HHFA. Federal responsibilities for urban mass transportation remained with HUD until 1968, when the Urban Mass Transportation Administration (UMTA) was established within the Department of Transportation (DOT).

The Federal-Aid Highway Act of 1968 amended the requirements of earlier highway acts relating to the economic effects of highway location to:

> . . . economic and social effects for such a location, its impact on the environment, and its consistency with the goals and objectives of . . . urban planning as has been promulgated by the community.

This law allowed the use of federal highway aid for the construction of fringe parking facilities and earmarked funding for traffic engineering measures known as the Traffic Operation Programs to Increase Capacity and Safety (TOPICS). To be eligible for matching funds, fringe parking and TOPICS projects had to be based on the 3C process.

The Urban Mass Transportation Assistance Act of 1970 provided a multiyear commitment for capital (but not operating assistance) for transit projects and encouraged the provision of mass transportation services to the elderly and handicapped, having recognized that:

> . . . elderly and handicapped persons have the same right as other persons to utilize mass transportation facilities and services.

Two decades later the 1990 Americans with Disabilities Act (ADA) found that more than 46 million Americans had one or more physical or mental disabilities and declared:

> . . . to provide a clear and comprehensive national mandate for the elimination of discrimination against individuals with disabilities.

Among its many provisions, this act required that access to public transportation vehicles and roadway designs be retrofited to accommodate the needs of this particular group.

7.3.4 National Environmental Legislation

The passage of the landmark National Environmental Policy Act (NEPA) of 1969 consolidated several trends relating to social, economic, environmental impact, and citizen participation in public decisions that were developing in various areas of federal concern including transportation. The act required that proposals for "Federal actions significantly affecting the quality of the human environment" be accompanied by:

> a detailed statement by the responsible official on—
>
> (i) the environmental impact of the proposed action,
> (ii) any adverse environmental effects which cannot be avoided should the proposal be implemented,
> (iii) alternatives to the proposed action,
> (iv) the relationship between local short-term uses of man's environment and the maintenance of long-term productivity, and
> (v) any irreversible and irretrievable commitments of resources which would be involved in the proposed action should it be implemented.

Major transportation proposals required the preparation of such an *environmental impact statement* (EIS), the expressed purpose of which was a full and objective disclosure of positive and negative environmental effects in order to aid the decision-making process. Indirect transportation consequences (e.g., traffic congestion and needed capacity) were to be included among the impacts covered by the EIS for nontransportation actions. An important provision of NEPA was the requirement to analyze alternatives to a preferred action including a baseline (or do-nothing) alternative.

The Clean Air Act of 1970 established national ambient air quality standards and required the states to develop plans to meet these standards. Motor vehicle emissions were identified as major contributors to the problem. The Environmental Protection Agency (EPA), which played a central role in subsequent transportation laws, rules, and regulations, was created by this act, which was followed by a series of environmental laws relating to noise, management of coastal and other environmentally sensitive areas, protection of endangered species, water quality, and so on. The 1982 version of the guidelines issued by the Federal Highway Administration (FHWA) relating to the conduct of environmental assessments is included in Appendix A to illustrate the breadth of impacts and transportation alternatives that have been incorporated in the requirements of the 3C planning process. Of special impact to transportation planning is the so-called Section 4(f) regulation (see Appendix A). This regulation requires that certain projects can be approved by the secretary of transportation only in the absence of "feasible and prudent" alternatives and, even then, only if all possible planning to minimize harm is undertaken and documented in the EIS. The 4(f) projects are those using publicly owned land that is a refuge for wildlife; is being used as a public park/recreational area; or has local, state, or national historic significance.

The 1990 Clean Air Act Amendments tightened the pollution standards for motor vehicles and established a severity classification system for geographical areas not meeting ambient air quality standards. Nonattainment areas (as they are called) could be placed into the categories of marginal, moderate, serious, severe, and extreme. Depending on the area's category, increasingly severe requirements are set along with a timetable for compliance with the standards.

The Federal Water Pollution Act of 1972 along with its amendments, such as the Clean Water Act of 1977 and 1987, which incorporate wetland protection and the national pollutant discharge elimination system, have also affected the planning, construction, and maintenance of transportation facilities [7.12].

7.3.5 Toward Planning Coordination

The Intergovernmental Cooperation Act of 1968 recognized a need for a mechanism by which projects seeking federal aid could be reviewed by the various interested and affected agencies. A year later the Bureau of the Budget issued Circular A-95, which set forth a requirement to designate specific state and metropolitan agencies as clearinghouses to facilitate the project-review process, which was thereafter referred to as the "A-95 review."

A series of highway- and mass-transportation-related laws enacted between 1970 and 1974 extended federal support to mass transit in urban and rural areas and increasingly placed federal aid for both highway and transit projects on essentially identical planning requirements. Both had to be produced by the 3C process; address the same social, economic, and environmental impacts; ensure community participation; and undergo similar agency and public reviews. This and functional overlaps between highways and transit systems using the highways motivated a closer degree of coordination between the FHWA and UMTA program requirements. In 1975 the two agencies issued joint regulations, which required each urban area to designate a single metropolitan planning organization (MPO) with a widely based membership to coordinate the planning activities of the local communities and modal planning agencies within their respective regions. The MPOs were to be certified annually by both FHWA and UMTA to ensure the presence of a satisfactory 3C process. One of the duties of the MPOs was the preparation of an annually updated multimodal transportation plan for the entire metropolitan area consisting of:

1. A long-range element addressing a time horizon of the order of 20 years
2. A transportation systems management element (TSME) containing the region's plan for low-cost operational improvements
3. A transportation improvement plan (TIP) specifying the region's 5-year priorities drawn from the other two elements and including an annual element (AE), which listed the programs and projects scheduled for the following year

Like its precursor (i.e., the 1969 TOPICS program), the TSM element addressed the need for short-term, low-cost operational improvements aiming for a better use of existing facilities, but in addition to traffic engineering measures, it included additional options such as carpooling, the use of taxis and other demand-responsive services, automobile restraints, changes in work schedules, and the like. The term *transportation-demand management* (TDM) is often used to describe such options. TSM planning was first required by a 1976 UMTA policy statement as a possible alternative to major transit projects for which the UMTA purse was becoming insufficient. Eventually the requirement for a separate TSM element in the annual plan was dropped, but by that time a regional TSM option became the de facto baseline alternative against which other proposals were compared. The attempt to coordinate the FHWA and UMTA programs is also seen in the title of the Surface Transportation Assistance Act of 1978. This act shifted the emphasis of the highway program from construction of new facilities to the renovation of existing highways and authorized additional funding for transit development. Partly because of a financial inability to fund all heavy rail rapid-transit proposals, UMTA issued a policy toward rail transit in 1978, which promulgated its intent to fund such systems on an incremental basis rather than in toto and only when an alternatives analysis has shown them to be superior to all-TSM, light rail, busway, and other options.

Following a 1973 oil embargo by the Organization of Petroleum Producing Countries (OPEC), the Highway Trust Fund also experienced difficulties because of decreasing revenues from gasoline taxes and because of price inflation. Responding to this problem, the U.S. Congress enacted the National Transportation Assistance Act of 1982, which imposed an additional federal tax of 5 ¢ on each gallon of fuel, of which 4 ¢ was earmarked for highway purposes and 1 ¢ for transit assistance.

Toward the latter part of the 1970s the nation showed signs of a change in mood away from federal intervention in the private sector. The federal government embarked on a trend to deregulate many sectors of the economy and to transfer the responsibility for many programs back to the states.

7.3.6 Intermodal Surface Transportation Efficiency Act of 1991

The Federal Intermodal Surface Transportation Efficiency Act (ISTEA), which was signed into law in December 1991, consisted of eight titles as follows:

I: Surface Transportation (mainly highways)
II: Highway Safety
III: Federal Transit Act Amendments of 1991
IV: Motor Carrier Act of 1991
V: Intermodal Transportation
VI: Research
VII: Air Transportation
VIII: Extension of Highway-Related Taxes and Highway Trust Fund

This comprehensive transportation law gave states and urban areas unprecedented flexibility with respect to the transfer of funds between highways, transit, and other projects. The major provisions of the act included the following:

1. It mandated the designation, by 1995, of a national highway system (NHS) consisting of approximately 155000 mi of roadways. The NHS would include the entire interstate highway system, a large portion of urban and rural principal arterials, and other major roads that, collectively, are considered to be critical to interstate and international travel, national defense, and intermodal connectivity.
2. It provided increased funding for research and development in the areas of new technology including intelligent vehicle-highway systems (IVHS), high-speed ground transportation systems, magnetic levitation technologies, and electric vehicles. The IVHS was subsequently renamed intelligent transportation system (ITS) to more accurately reflect its intermodal nature.
3. It stipulated that each state must establish a statewide planning process and six management systems in the areas of highway pavement maintenance, bridge management, highway safety, traffic congestion, public transportation, and intermodal transportation facilities.
4. It required metropolitan planning organizations (MPOs) to incorporate in their transportation improvement programs (TIPs) and their long-range plans considerations of land-use policies, intermodal connectivity, enhanced transit service, and management systems.
5. It permitted the use of federal transportation funds for projects aimed toward enhancing the environment such as wetland and wildlife habitat protection, air quality improvement measures, and highway beautification.
6. It strengthened the level of federal support for toll roads and allowed for private entities to own such facilities.
7. It renamed the Urban Mass Transportation Administration (UMTA) to the Federal Transit Administration (FTA) to reflect more accurately the broadened transit initiatives of the act, including an added emphasis on rural and intercity transit services.

8. It extended the Mass Transit and Highway Accounts of the Highway Trust Fund (HTF) by 4 years to Fiscal Year 1999, and established a National Recreational Trails Trust Fund. The federal share was set at 90% for interstate construction and maintenance projects, at 80% for most other highway-related projects and for transit capital improvements, and at 50% for transit operating assistance.

7.3.7 Transportation Equity Act for the Twenty-First Century

The ISTEA expired in October 1997 but was extended to May 1998 while a debate for a reauthorization bill was taking place in the U.S. Congress. At issue in this debate was the question of "equity" in the distribution of funds from the Highway Trust Fund (HTF) in proportion to the amounts that each state contributed to the fund.

The *Transportation Equity Act for the Twenty-First Century,* known as TEA-21, was signed into law on June 9, 1998. Along with the *TEA-21 Restoration Act,* which was enacted in July 1998 to effectuate technical corrections, TEA-21 represented a record-breaking $217 billion authorization for highways, transit, highway safety, and motor vehicle carrier programs over the 6-year period 1998–2003.

A major provision of this act was an allocation process "to ensure that no State's return from such Trust Fund is less than 90.5 percent." Moreover, the HTF was placed in a special budgetary category that prevented it from being used, as had been the case with earlier legislation, for purposes other than transportation (e.g., budget deficit reduction). Also, TEA-21 guaranteed that, at a minimum, 94% of the total authorization is distributed to the states.

TEA-21 maintained essentially the same program structure and flexibility that had been established by ISTEA. The federal-aid highways title of the act retained its major emphasis on six programs. These were the interstate maintenance program, the national highway system, the highway bridge program, the surface transportation program (for non-NHS elements) and the congestion mitigation and air quality improvement (CMAQ) program. New initiatives were added to the highway title: the national corridor and development and coordinated border infrastructure programs aimed at corridors of significance with respect to national and international trade.

Highway Safety (Title II of the act) received added support. This included incentives to states enacting stringent blood alcohol concentration (BAC) laws and passenger restraint (i.e., seat belt) requirements. Safety funds were also made available for special safety programs, alcohol-impaired driving countermeasures, and improvements in the collection and use of safety-related data.

Title III of TEA-21 addressed programs of the Federal Transit Administration. This part emphasized the rehabilitation of existing facilities and rural transportation accessibility programs, and addressed the needs of elderly individuals and individuals with disabilities. A notable new provision in this title was a special program to enhance the accessibility of low-income residents of the urban core to suburban jobs.

Transportation research and technology (Title V) were also given a prominent place. Among the major provisions of TEA-21 were training and education programs, state planning and research, advanced vehicle technologies, the deployment of intelligent transportation systems (ITS) consistent with a national ITS architecture and standards, commercial remote sensing products, and spatial information technologies.

Other provisions of TEA-21 included a continued emphasis of transportation enhancement projects (including bicycle and pedestrian facilities, environmental protection, and highway beautification), encouragement of joint public-private initiatives, conversion of segments of interstates to toll roads, and streamlined motor carrier regulations.

7.4 PLANNING STUDIES AND METHODS

7.4.1 Background

The preceding section traced the evolution of a formalized transportation planning process, by which transportation plans are produced, revised, and selected for implementation. It also identified the groups of participants in the process and the factors that were deemed relevant to the proper execution of the planning function.

The participating groups include bodies of elected officials, public agencies that have leading and supportive roles in the process, officially appointed citizen advisory commissions and committees, private and public transportation system operators, voluntary citizen and professional associations, and interested individuals. These groups and individuals bring into the process differing, often conflicting, and also changing goals and objectives. To complicate matters, not all these groups are particularly interested in a continuous and intensive participation in all aspects of the process; on the contrary, they often feel free to enter or exit the process at will. Moreover, the membership of these groups exhibits a considerable amount of fluidity. It is not unusual, for example, for a member of a voluntary organization to be elected or appointed to public office or for an agency representative to belong to a professional organization and to reside also in the path of a proposed facility.

Institutionalized procedures, such as the requirements for planning documents, interagency reviews, and public hearings and other means of citizen participation have evolved within the larger sociopolitical system specifically to ensure that the factors considered to be relevant to a particular situation are adequately addressed and in order to facilitate the formation of local consensus in an orderly manner. Transportation engineers and planners participate in various aspects of this complex process. One aspect of involvement that merits further treatment here is the conduct of supportive planning studies that attempt to model and estimate *some* of the many travel, economic, social, and environmental factors that have been deemed important to transportation planning.

7.4.2 Antecedents to Planning Studies

The first step toward the development of the contemporary transportation planning methodology may be traced to the conduct of land surveys that supported the layouts of cities and towns and the locations of turnpikes, canals, and, later, railroads. The second step was the need to conduct facility inventories, such as the first national inventory of 1807. The third step commenced when the Office of Road Inquiry, toward the end of the nineteenth century, extended data collection efforts to include information relating to facility use, that is, traffic levels, trip lengths, and user costs. The expanded usage studies that followed the prescription of the Federal-Aid Highway Act of 1921 to plan a connected national network and the transition to studies emphasizing highway planning to meet future needs made possible by the Hayden–Cartwright Act of 1934 established the fundamental elements of transportation planning.

7.4.3 Planning for Future Needs

A major breakthrough of the needs studies of the 1930s and 1940s was the recognition that planning highway network extensions should not be based merely on the static criterion of connectivity but also on continuous efforts to anticipate future demands for travel. Initially this was accomplished by projecting current traffic measurements into the future using traffic growth factors based on discerned relationships between population and economic growth on one hand and traffic levels on the other. For example, based on annual rates of growth in the gross national product (GNP), traffic growth factors in the range of 3 to 4% were considered to be reasonable. The projected traffic levels could then be checked against the capacity of existing highways to anticipate future capacity deficiencies and, within financial constraints, to plan and schedule capacity improvements accordingly.

7.4.4 Large-Scale Urban Travel Surveys

Significant differences in the patterns of urban travel necessitated the development of more refined techniques. An important difference was (and still is) the fact that in urban areas, street capacities between various parts of the city involved multiple rather than single routes. If needed, capacity enhancements should consider this combined supply of roadways. A *desire line diagram,* which shows the region divided into smaller sectors, or *traffic (analysis) zones,* and the flows between these zones irrespective of individual roadway links are preferable. To obtain this type of information, new travel survey and data reduction methods were developed during the 1940s, including the *origin-and-destination (O-D) surveys* consisting of home interviews, truck interviews, taxi interviews, and parking surveys. The data on travel habits obtained from interviewing a sample consisting of 4 to 5% of the total households in the region and about 20% of the truck and taxi companies were expanded to the overall population by computer-based statistical techniques, and the actual traffic counts crossing selected *screen lines* were used to check the accuracy of the statistical expansion of the sample data. The first large-scale travel survey of this type was conducted in Detroit.

At the present time travel surveys have become an indispensable tool for planning. In 1996 the *Travel Survey Manual* authored by Cambridge Systematics, Inc. was released. The manual is a product of the Travel Model Improvement Program (TMIP) and was sponsored by the FHWA and the EPA.

7.4.5 Travel-Demand Forecasts

Initially the projection of the interzonal *trip distribution* toward the target year was accomplished by applying simple growth factors to the base-year travel desire volumes in a manner that was similar to rural highway practice. Gradually, however, it became evident that the need for added capacity and parking facilities in urban areas was not uniform throughout the region but was dependent on the specific types (e.g., residential, commercial, or industrial) and intensities (residential density, workers per acre, shopping floor space, etc.) and the *land uses* found in each zone. Moreover, the expected regional growth of the population and the economic system was unevenly distributed among the zones owing to differences in the availability and suitability of developable land for various purposes, urban planning policies (such as zoning), and accessibility. The first computer-based quantitative *land use* and *socioeconomic projection models* were developed by transportation planners in this connection and were later adopted eagerly by other urban planners. Mathematical *trip-generation models* relating the trip-producing capability of residential areas and the trip-attracting potential of various types of nonresidential land-use classes were postulated, calibrated, and validated.

Because the emphasis of these studies was placed on the urban highway system, transit trips had to be subtracted from the projected total interzonal traffic volumes to arrive at an estimate of future highway demands. *Modal split models,* such as the one illustrated in Fig. 7.4.1, were developed to help divide the total flows between the two modes, highway and transit [7.13]. The planning for mass transit services was generally left to the operators and considered to be of secondary importance since transit patronage was experiencing a steady decline.

Of relevance to urban highway design was a prior knowledge of the degree to which arterial street traffic would be attracted to new freeways. Having this knowledge before designing a new facility was important in determining the capacity (e.g., the number of lanes) that it should provide. Models of traffic diversion from arterials to freeways similar in shape to the model choice curve of Fig. 7.4.1 resulted. These route choice models were later extended to cover large networks and became known as *traffic assignment models.* Thus trip-generation, trip-distribution, mode choice, and traffic assignment models evolved, each intended to describe and forecast a different component of travel behavior.

The Chicago Area Transportation Study (CATS) was the first to combine land-use and socioeconomic projection models with these travel-demand models to analyze regional long-range transportation alternatives. This urban transportation planning methodology was then applied to other U.S. metropolitan areas and was also taken to major cities throughout the world by U.S. consulting firms. In the process the methodology was further refined and applied to various planning contexts.

Figure 7.4.2 is a simplified flowchart of steps involved in applying the original methodology after the conduct of planning inventories and surveys (e.g., land-use data, economic investigations, and travel surveys) and the postulation and calibration of models forecasting land use and travel demand to fit local conditions.

Step 1: Forecasts for the target year of the regional population and economic growth for the subject metropolitan area.

Step 2: Allocation of land uses and socioeconomic projections to individual analysis zones according to land availability, local zoning, and related public policies.

Step 3: Specification of alternative transportation plans partly based on the results of steps 1 and 2.

Step 4: Calculation of the capital and maintenance costs of each alternative plan.

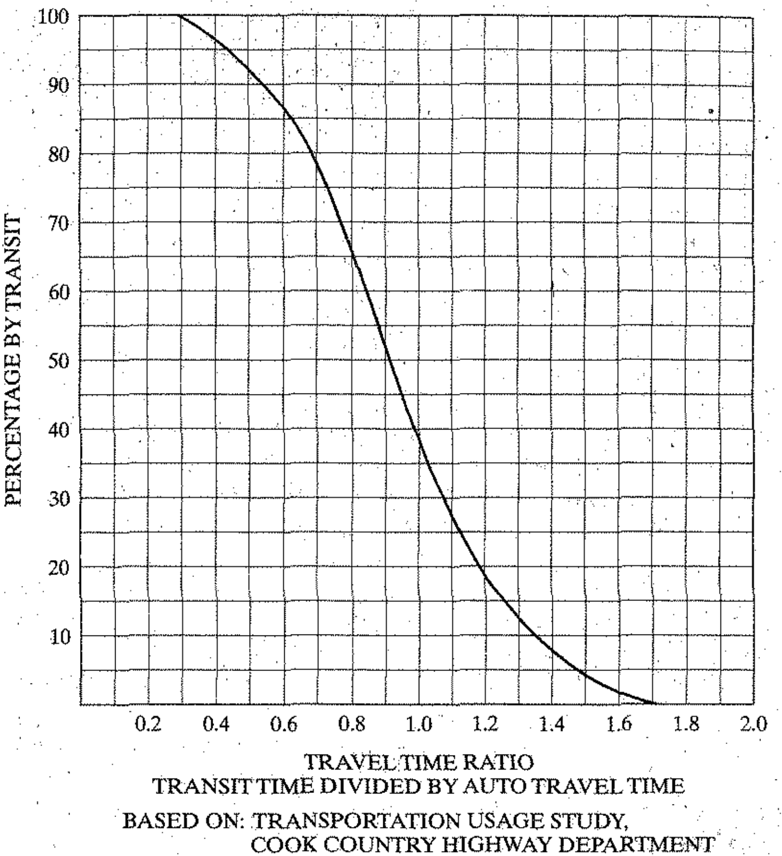


Figure 7.4.1 Early modal split curve.
(From Voorhees and Morris [7.13].)

Step 5: Application of calibrated demand-forecasting models to predict the target-year equilibrium flows expected to use each alternative, given the land-use and socioeconomic projections of step 2 and the characteristics of the transportation alternative (step 3).

Step 6: Conversion of equilibrium flows to *direct user benefits,* such as savings in travel time and travel cost attributable to the proposed plan.

Step 7: Comparative evaluation and selection of the "best" of the alternatives analyzed based on estimated costs (step 3) and benefits (step 6).

This methodology was refined and expanded to cover additional social, economic, and environmental benefits and costs; to admit a wider range of multimodal transportation alternatives; to be more sensitive to the relationship between land-use and transportation planning; and to admit multiagency and public participation.

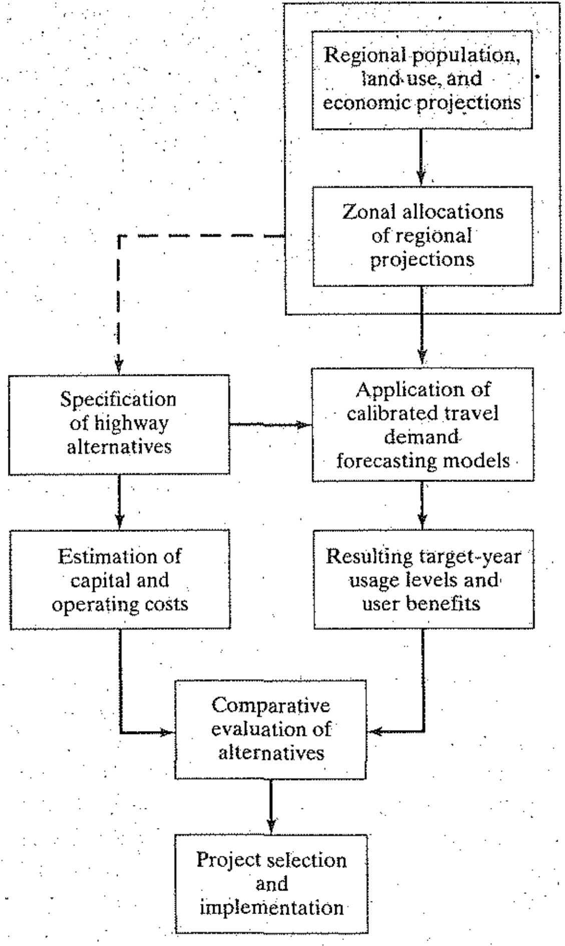

Regional population, land use, and economic projections
Zonal allocations of regional projections
Specification of highway alternatives
Application of calibrated travel demand forecasting models
Estimation of capital and operating costs
Resulting target-year usage levels and user benefits
Comparative evaluation of alternatives
Project selection and implementation

Figure 7.4.2 Simplified version of the original urban transportation planning process.

7.5 OTHER PLANNING ISSUES

7.5.1 Background

A significant problem of the long-term transportation procedure as typically applied during the 1960s and early 1970s was the problem of joint transportation and land-use planning, the treatment of uncertainty about future policy changes, and the problem of coördinating long-range and ongoing planning. This section addresses these issues in general terms.

7.5.2 Transportation and Land Use

The fact that an intimate relationship exists between transportation and land development has been understood for centuries [7.14, 7.15]. An understanding of this interaction is also evident in the principles of the *city practical* approach. In that case the city form preferred by certain planners was closely related to the transportation technology of the day: radially expanding urban development along heavily traveled fixed-route streetcar and heavy rail lines. The subsequent proliferation of the ubiquitous automobile all but eliminated the search for ideal city forms and facilitated a wider range of urban structure possibilities. Gradually the prescriptive nature of early master planning gave way to an adaptive approach that aimed instead at providing guidance for an orderly development of the region. To this day zoning maps accompanied by zoning ordinances and other land-use policies remain the major tools of this approach. The zoning map specifies the type (e.g., residential, manufacturing, mixed, open space) and intensity (e.g., high-rise, low-density) of the land uses permitted at designated locations within the region. However, if developed to the limits permitted by the zoning map, the population and land-based activities in the region would far exceed the growth that is typically anticipated within normal planning horizons. The land-use policy expresses the community's principles and procedures that are intended to guide land development in desirable directions. Among these rules are provisions for the issuance of *zoning variances* (special permission to private- or public-sector developers to deviate from the specifications of the current zoning map). In addition to zoning map revisions, major changes in land development policies take place in response to changing economic, social, and political conditions.

Because the basis of long-term regional transportation prediction models is the future land-use pattern, and because the zoning map provides room for a variety of future patterns, there exists a need to forecast the most likely pattern subject to zoning constraints and current land-use policy. This is precisely the purpose of the land-use prediction phase shown in Fig. 7.4.2. However, the accumulated effect of zoning variances, the sequence in which major projects are implemented, and possible revisions of the underlying land-use policies introduce a significant degree of uncertainty to a single land-use forecast. Moreover, if implemented, major recommended modifications in the regional transportation system that are partly based on these forecasts would have an effect on the future land use by making certain areas within the region more accessible than others. This gives rise to the problem of model consistency on the grand scale: The predicted transportation demand may not be consistent with the assumptions of the land-use projection. Conceptually, the modeling of this dynamic effect may be rectified by introducing a feedback link between the output of the travel-forecasting procedure and the land-use prediction phase to facilitate an iterative equilibrium solution.

An alternative to the iterative equilibrium-solution method just described is an approach that examines several alternative land-use scenarios in connection with the alternative transportation options. In this manner, the analysis of combined transportation and land-use alternatives may provide some guidance for joint planning. Although not attempting to predict a single future pattern, this approach can address a larger set of alternatives, including land-use and transportation combinations that appear to be consistent with each other. For example, a fixed-guideway rapid-transit system may be examined in relation to a potential land-use policy that guides land development along the transit corridor and especially in the vicinity of transit stations [7.16].

A relatively recent development in connection with land-use and transportation planning is a practice first begun in California: assessing land developers special *traffic impact fees* as a precondition for approving land development. Traditionally this practice has been used in connection with providing other public services, such as sewers, where those residents that would gain from the new service are asked to pay part of the associated costs. In connection with traffic impact fees the courts have held that a municipality can impose them only if it can show that the proceeds from the assessment will be used to ameliorate traffic impacts that are attributable to the proposed land development. This relationship is legally known as *rational nexus* between the development and the traffic impact fees. Such requirements, coupled with the rules and regulations adopted by EPA to implement the 1990 Clean Air Act Amendments [7.17] and the provisions of federal transportation law (i.e., ISTEA and TEA-21), induced urban areas and states to redouble the attention placed on the interaction of land-use transportation development.

7.5.3 Operational Land-Use Models

Land-use allocation models were initially developed to provide inputs to the travel demand forecasting models. The most notable of the early models was the 1964 Lowry model [7.18] that had a profound influence on land-use modeling for decades. The Lowry model views a "metropolis" as consisting of three sectors: the basic (or export) sector, the nonbasic sector, and the population sector. The basic sector consists of those industrial activities producing goods and services intended mainly for export and thus contributing to the weatlth of the region. Nonbasic activities (such as retail) are those activities that essentially serve the needs of the region's population, and consequently tend to follow residential markets. Basic industries, on the other hand, are assumed to choose their location irrespectively of where the population is located. Given this view of urban development, the Lowry model begins with an empty region subdivided into analysis zones. It then selects the location of basic employment exogenously, perhaps guided by land-use policy, land availability, and similar considerations. The next step involves the distribution of the population that supplies the labor to basic industries via a gravity model formulation.* Following the initial allocation of population to the zones that make up the region, the model proceeds to locate nonbasic activities at the rates needed to support that population distribution. Through an iterative procedure, the model balances the available land between residences and nonbasic activity and ensures that zonal population meets the constraints imposed by the land-use policy in effect. The latter may include population density or the maximum number of households permitted in each zone. The end result is the equilibrium target-year spatial distribution of population, housing, and employment within the region.

In the decade or so that followed significant land-use modeling activity occurred that attempted to extend the simplistic Lowry model to capture more complex phenomena associated with residential and industrial location. These included supply and demand for land through market clearing mechanisms, developer behavior, tax policies, and so forth. Considering the computer technology of the time, these models soon became very large and computationally demanding to the point of earning the disfavor of "practical" transportation planners. In a milestone paper published in 1973 Lee [7.19] declared a "requiem" for them.

*The gravity model is covered in Chapter 8. In this application it essentially allocates population with respect to accessibility to employment: The more accessible a zone is, the more likely the zone will attract residential activity.

As Wagener [7.20] explains in a seminal paper, land-use modeling had entered a period of semidormancy, with only a few individuals and research centers spread around the globe contributing to its development. He attributes a recent resurgence of activity in the subject of integrated land-use models to the "urgency of environmental debate" as described earlier. Wagener reviewed 20 operational models (in the sense that they had been applied to at least one policy decision in at least one real urban area) in terms of their attributes such as:

1. *Comprehensiveness.* The degree to which the model addressed various subsystems including population, employment, travel, and so on
2. *Model structure.* The degree to which the subsystems were integrated within the model
3. *Model theory.* The theoretical underpinnings of the model and the method used for its calibration
4. *Modeling technique.* The way by which the model treated the interaction between transportation system characteristics and location decisions by individuals and firms

Of the 20 models examined only four incorporated a multimodal transportation network. Nevertheless, great strides had been made toward enhancing the state of the art.

In 1999 the national cooperative highway research program (NCHRP) issued a guidebook on the subject motivated by two objectives:

1. To improve the practice of land use forecasts
2. To identify tools and procedures for realistically evaluating the land-use impact of transportation investments and policies [7.21].

The report reviewed several of the most widely used land-use models at the time, including the following:

1. *DRAM/EMPAL* consisting of three components: the disaggregated residential allocation model (DRAM), the employment allocation model (EMPAL), and a model of travel demand. The first two were essentially of the Lowry type. This model happened to be available at the time of dire need for integrating transportation and land-use analyses. Consequently it immediately found wide application. However, subsequent experience has found it lacking.
2. *MEPLAN* and *TRANUS* share the same ancestry, are based on classical macroeconomic theory, and are *evolutionary* or *dynamic* in nature. They produce incremental results over simulated time in contrast to the single-target-year equilibrium solution produced by Lowry-type models.
3. *METROSIM* is said to be a *unified* model, meaning that it solves its component submodels (basic and nonbasic employment, traffic assignment, housing, and commercial real estate) simultaneously. It is based on random utility* and microeconomic theory.
4. *HLFM II+* is a simplified version of DRAM/EMPAL, which is for use by small urban areas.
5. *UrbanSim* incorporates interactions among transportation, land-use actions, and various development policies in a dynamic, rather than equilibrium, approach. It is a dis-

*Random utility models are discussed in Chapter 8.

aggregate model (i.e., it operates at the individual household and individual firm level) and provides linkages to external travel demand models, thus allowing for the consideration of major transportation system changes on land use over time.

Each of these as well as other land-use models have strengths and weaknesses. Moreover, the most widely used among them are under continuing enhancement in response to the specific needs of their users. The decision to adopt a land-use model by a regional transportation planning organization should be reached very carefully and should consider the modeling requirements of the area and the resources (both human and material) that are available for data collection, calibration, validation, continued maintenance, adjustment, and update.

7.5.4 Project, System, and Operational Planning

Long-term regional strategic planning may disclose a need to enhance the capacity of a certain travel corridor either in general terms of specifically in terms of highways or transit. In either case it would normally not be able to supply sufficiently detailed information for *project-level* planning. Depending on the way in which the problem is expressed and its specificity regarding the admissible range of alternative options, a series of increasingly detailed planning studies leading to the stage of final design would normally be undertaken. The following quotation from a request for consultation services issued by the Honolulu Department of Transportation Services (DTS) in July 1985 illustrates this point:

> CONCEPTUAL ENGINEERING FOR HONOLULU RAPID TRANSIT. To reactivate specific planning and engineering elements of the Honolulu Area Rapid Transit Study. Develop architectural, engineering, and operating system design criteria. Develop elements of definitive engineering, architectural and operating system plans, and geotechnical investigations as necessary. Develop elements of supplemental EIS as necessary to handle major alignment and/or station shifts from current approved HART EIS. Develop materials for the conduct of public information and citizen involvement programs to keep people apprised on program development and to obtain inputs to aid system design. Perform new ridership estimates based on land use changes based on the approved Development Plans and the OMPO Long-Range Program ([7.22, p. D-10]).

OMPO is the metropolitan planning organization that has been designated as the 3C process-coordinating agency for the island of Oahu, where the city of Honolulu is located. Close study of this terse statement can help the reader to appreciate the issues discussed in this and earlier chapters of the book.

In addition to multimodal regional sketch planning and planning for new capital-intensive projects, a variety of ongoing planning studies exists at various geographical scales and time horizons for the modal components of the regional system. To illustrate this point, consider the following requests for consulting services issued by the Honolulu DTS at the same time as the one quoted earlier:

> COMPREHENSIVE ISLAND-WIDE TRANSIT SYSTEM STUDY. Develop short-, intermediate-, and long-range bus transit plans including improvements for bus fleet, maintenance facilities, and transit operations. Other tasks include on-board bus survey, evaluation of existing system, transit financing feasibility of use of small shuttle bus concept and contract supplemental bus service.

With regard to traffic operations:

> OAHU VEHICLE COUNTS AND TRAVEL DATA. Collect and process current traffic data in order to permit evaluation of both the effectiveness of short-term improvements and the accuracy of travel forecasts on both State and County facilities.

The three planning studies just described would be undertaken more or less at the same time by different consultants under the coordination of the DTS, which is one of the many agencies participating in the 3C planning process. Moreover, other such land-use and transportation-related issues that were being addressed at the time in Honolulu included a major freeway project that was in litigation in the federal court system, a major deep-draft harbor, the expansion of the Honolulu International Airport, the selection of a site for a general aviation airport, and several proposals for major industrial centers and residential developments, some of which involved applications for zoning variances. The sequence in which some of these projects would be implemented (if at all) might require revisions of the regional long-range plan to incorporate committed, programmed, and implemented projects. The last sequence of the quotation relating to planning for the Honolulu rapid-transit system illustrates that currently uncommitted projects may subsequently have to be restudied.

7.5.5 Planning at the Statewide Level

The examples given in the preceding subsection were drawn from the urban context. Planning for single-mode and multimodal transportation systems in larger areas (i.e., statewide) are also undertaken by various planning entities. The ISTEA of 1991 imposed the additional requirement that:

> States are required to carry out statewide planning in coordination with metropolitan planning and to meet its [sic] responsibilities for the development of the transportation portion of the SIP as required by the CAA. States are required to develop a plan and a program that addresses all modes of transportation [7.12].

As a consequence the states began adapting the travel-forecasting tools discussed in Chapter 8 to meet this requirement [7.23]. In some cases it became necessary to develop special context models as well.

The need for strategic, project-level, and operational planning is expressed in this context as well.

7.6 SUMMARY

In this chapter we defined planning as the forward-looking, organized, and premeditative process that precedes the undertaking of actions intended to guide a particular situation or system in desirable directions but not as a search for the ultimate. The fundamental objective of transportation is to provide the efficient and safe levels of mobility required to support a wide spectrum of other human needs for a heterogeneous variety of societal groups. Because these needs, goals, and objectives are continuously changing, transportation planning is also an ever-evolving process.

The evolution of contemporary transportation planning in the United States was traced along the historical path of land transportation, with particular attention to the confluence of three important factors: technological progress, private interests, and changing governmental policy.

The merging and interaction of three disparate planning perspectives (the facility orientation of intercity highway planning, the traffic operations-oriented traffic engineering approach, and the social consciousness of urban planning) produced the basic elements of the contemporary urban transportation planning process, incorporating technical analyses, widely based citizen participation, and a concern for a large variety of social, economic, and environmental impacts in addition to connectivity and accessibility.

EXERCISES

1. Obtain an EIS for a major transportation project and write a short report summarizing its major contents.
2. Prepare a synopsis of the transportation impacts covered in an EIS for a nontransportation project, such as a residential development, an industrial plant, or a commercial center.
3. Present your own arguments for or against governmental actions to improve the mobility of the elderly and handicapped.
4. What is the major difference between the *needs* studies described in this chapter and the urban transportation planning methodology illustrated by Fig. 7.4.2?
5. How can the methodology of Fig. 7.4.2 be applied to aid in the planning of a statewide system of airports?
6. Prepare a list of the major ways by which the federal government has been involved in planning land transportation.
7. Review the material included in Appendix A, and in your own words discuss the potential impacts of building a freeway in a large city.
8. In your own words describe the major consequences of implementing a high capacity rapid-transit system in a major urban area.
9. Review an article from the technical literature that addresses the topic of TSM.
10. Discuss the major advantages and disadvantages of privately owned highways.
11. Discuss the major advantages and disadvantages of privately owned urban bus systems.
12. Compile a dossier containing clippings from your local newspaper of transportation-related stories over a 2-week period. Arrange this material in an organized way of your choosing.
13. If a major transportation-related decision is pending in your city or state, identify the major groups involved and briefly describe the thrust of their arguments. Distinguish between qualitatively and quantitatively supported claims.
14. What kinds of transportation impacts do you think that construction of a multistory residential building in a densely populated area would possibly have?
15. Prepare a short report describing the zoning ordinances of your city or town.
16. List the various types of planning-related surveys discussed in this chapter.
17. Make a list of alternative strategies that have the potential of alleviating urban traffic congestion.
18. Briefly compare the various urban planning schools of thought described in this chapter. What was the basic view that each held with regard to the role of transportation in the urban milieu?

19. What is the purpose of the Highway Trust Fund?

20. Give several specific examples of the way in which technological development has affected the structure of cities.

21. What do you think was the rationale behind placing provisions for federal aid to urban mass transit systems in housing laws?

22. What were the objectives of the TOPICS program?

23. How can you adapt the methodology of needs study to the planning of a system of parks and playgrounds within a city?

24. Describe how the methodology discussed in Section 7.4.5 can be adapted to help plan a regional water supply system.

25. What are some major components of travel demand?

26. You have been assigned the task of developing materials for the conduct of public information and citizen involvement programs as specified in the request for consulting services quoted in Section 7.5.4. Research the topic and write a report not to exceed ten double-spaced pages.

REFERENCES

7.1 FEDERAL HIGHWAY ADMINISTRATION, *America's Highways 1776–1976: A History of the Federal-Aid Program,* U.S. Department of Transportation, U.S. Government Printing Office, Stock No. 050-001-00123-3, Washington, DC, 1976.

7.2 NATIONAL TRANSPORTATION POLICY STUDY COMMISSION, *National Transportation Policies through the Year 2000,* Final Report, U.S. Government Printing Office, Washington, DC, June 1979.

7.3 TRANSPORTATION RESEARCH BOARD, *Urban Transportation Alternatives: Evolution of Federal Policy,* Special Report 177, National Research Council, Washington, DC, 1977.

7.4 HOFFMAN, H. W., *Sagas of Old Western Travel and Transport,* Howell-North Books, San Diego, CA, 1980.

7.5 GRAY, G. E., and L. A. HOEL, Eds., *Public Transportation: Planning, Operations and Management,* Prentice-Hall, Englewood Cliffs, NJ, 1979.

7.6 MARTIN, J., *Mule to MARTA,* Vols. I and II, Atlanta Historical Society, Atlanta, GA, 1977.

7.7 GOODMAN, W. I., and E. C. FREUND. *Principles and Practice of Urban Planning,* International City Manager's Association, Washington, DC, 1968.

7.8 NOLEN, J. (Ed.), *City Planning,* D. Appleton and Company, New York, 1929.

7.9 TAYLOR, G. R., *Satellite Cities: A Study of Industrial Suburbs,* D. Appleton and Company, New York, 1915.

7.10 HOWARD, E., *Garden Cities of To-morrow,* first published in 1898 as *Tomorrow: A Peaceful Path to Real Reform,* Faber and Faber Limited, London, 1902.

7.11 CITY AND COUNTY OF HONOLULU, *General Plan for Urban and Urbanizing Areas,* Planning Department, Honolulu, HI, August 1960.

7.12 MICKELSON, R. P., *Transportation Development Process,* Synthesis of Highway Practice 267, National Cooperative Highway Research Program, Transportation Research Board, National Research Council, Washington, DC, 1998.

7.13 VOORHEES, A. M., and R. MORRIS, *Estimating and Forecasting Travel for Baltimore by Use of a Mathematical Model,* Highway Research Board Bulletin 224, National Research Council, Washington, DC, 1959, pp. 105–114.

7.14 HALL, P. (Ed.), *Von Thunen's Isolated State,* Pergamon Press, Oxford, 1966.

7.15 LOSCH, A., "The Nature of Economic Regions," *Southern Economic Journal, 5* (1938): 71–78.

7.16 U.S. SUBCOMMITTEE ON THE CITY, *New Urban Rail Transit: How Can Its Development and Growth-Shaping Potential Be Realized,* Committee on Banking, Finance and Urban Affairs. U.S. House of Representatives, 96th Congress, First Session, U.S. Government Printing Office, Washington, DC, 1980.

7.17 ENVIRONMENTAL PROTECTION AGENCY, "Air Quality: Transportation Plans, Programs, and Projects; Federal or State Implementation Plan Conformity; Rule," *Federal Register,* Vol. 58, No. 225, pp. 62188–62253, Wednesday, November 24, 1993.

7.18 LOWRY, I. S., A Model of Metropolis, Technical Report RM-4035-RC, the RAND Corporation, Santa Monica, CA, 1964.

7.19 LEE, D. B., Jr., "Requiem for Large Scale Models," *Journal of the American Institute of Planners,* Vol. 39, No. 2 (1973): 163–178.

7.20 WAGENER, M., "Operational Urban Models: State of the Art," *Journal of the American Institute of Planners,* vol. 60, No. 1 (1994): 17–29.

7.21 PARSONS BRINCKERHOFF QUADE & DOUGLAS, *Land Use Impacts of Transportation: A Guidebook,* National Cooperative Highway Research Program Report 423A, Transportation Research Board, National Research Council, Washington, DC, 1999.

7.22 *Honolulu Star-Bulletin,* July 30, 1985.

7.23 FEDERAL HIGHWAY ADMINISTRATION, Guidebook on Statewide Travel Forecasting, Prepared by the Center for Urban Transportation Studies, University of Wisconsin-Milwaukee in Cooperation with the Wisconsin Department of Transportation, 1998.

8

Travel Demand Forecasting

8.1 INTRODUCTION

In Chapter 7 we explained that the purpose of the travel-forecasting phase of the urban transportation planning process is to perform a conditional prediction of travel demand in order to estimate the likely transportation consequences of several transportation alternatives (including the do-nothing alternative) that are being considered for implementation. This prediction is also conditional on a predicted target-year land-use pattern. The major components of travel behavior were identified as:

1. The decision to travel for a given purpose (trip generation)
2. The choice of destination (trip distribution)
3. The choice of travel mode (mode choice)
4. The choice of route or path (network assignment).

Figure 8.1.1 illustrates that travel-demand models can be put together in a sequence. In this *sequential demand-modeling* arrangement, also known as the four-step process, the outputs of each step become inputs to the following step, which also takes relevant inputs from the specification of the alternative plan under study (network description) and from the land-use and socioeconomic projection phase. Also shown on the figure are two ancillary steps, auto/occupancy and time of day. The most commonly used models for each of the four steps of the sequential process are covered in this chapter. For each model the relevant dependent and independent variables are identified and the method of calibration is described. Additionally, the advantages and disadvantages of each model are discussed. Model selection, of course, should be guided by the rules discussed in Chapter 13 in relation to modeling in general.

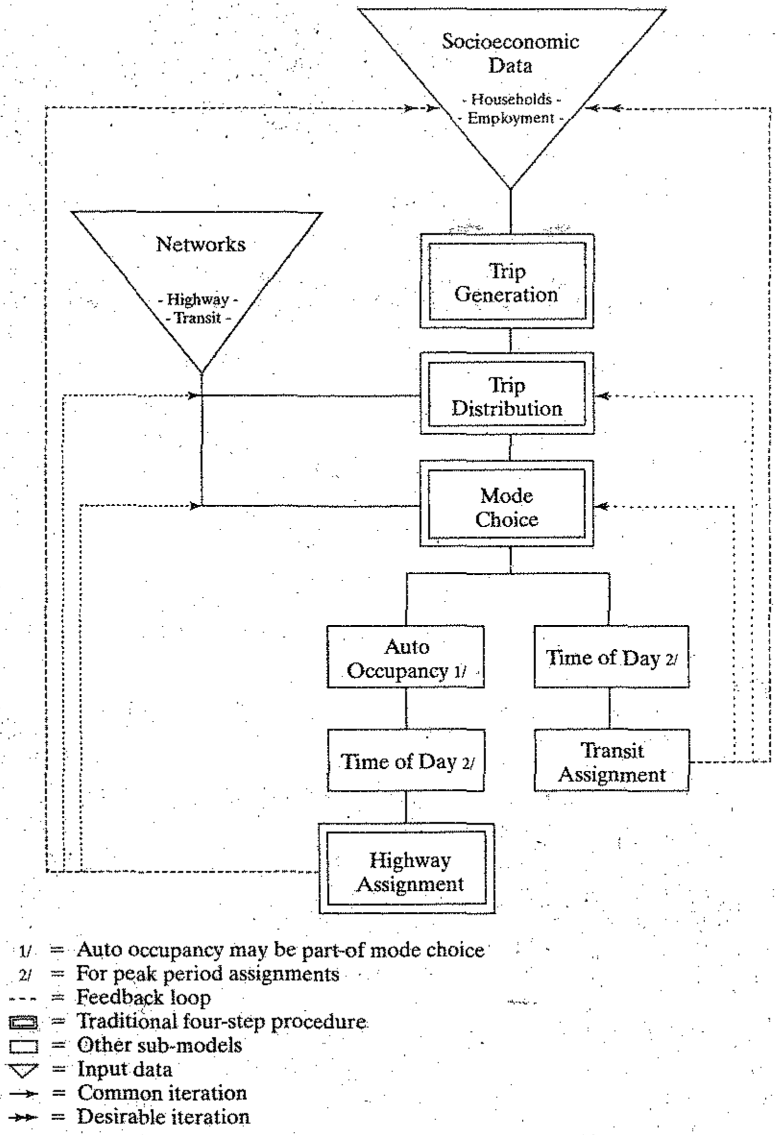


Figure 8.1.1 Travel demand-forecasting process. (*Source:* NCHRP [8.1].)

The four-step modeling process shown in Fig. 8.1.1 was originally developed in connection with the planning of major highway facilities during the 1950s and 1960s. At first glance the process may appear to have remained unchanged since that time. In reality, however, it has undergone significant modification in response to an improved understanding of travel behavior by modelers, the need to address emerging policy questions (e.g., high-occupancy vehicle facilities and congestion pricing), and advances in computational technology. More powerful personal computers allowed the specification of more complex and detailed models. This evolutionary development will undoubtedly continue.

At any given time the state of travel-demand modeling may be described as consisting of:

1. A set of *standard models* that are used by most typical transportation planning organizations, such as Metropolitan Planning Organizations (MPOs) and state Departments of Transportation (DOTs), with relatively minor variations. Reference [8.1] describes the standard models of the late 1990s. These models are often referred to as being *trip-based* because they consider individual trips (as described in Section 8.2) to be the basic unit of travel.
2. A set of the best practice models that represent advanced operational models that are developed and implemented by either progressive transportation planning organizations or organizations that face modeling needs beyond the typical. For example, MPOs in very large areas need more sophisticated models of mode choice than smaller, automobile-oriented areas. Reference [8.2] documents what was considered to be the best practice in 1992.
3. A set of models and modeling approaches that are motivated *by alternate paradigms,* usually based on the findings of advanced research, that either are at the prototype stage or have found limited implementation but are not entirely operational. Such a paradigm shift, which gave rise to what became known as *activity-based approaches,* appeared on the modeling horizon during the late1970s and gained momentum toward the end of the twentieth century (e.g., Ref. [8.3.]). As of that time, elements of the activity-based perspective were being incrementally introduced into the best practice. For example, some operational model sets began to consider *trip chains* (rather than trips) as elemental units of travel. Other examples include the introduction of traveler "lifestyle" characteristics (see Section 8.6) as explanatory variables and a shift from purely empirically calibrated models to probabilistic behavioral travel choice models.
4. A set of *special purpose* models that are developed to address specific issues in a limited context.

The major thrust of this chapter is to describe the standard and some of the best practice models in the context of travel-demand forecasting for a large urban region. Section 8.6 briefly covers the rationale and motivation of activity-based approaches and Section 8.7 describes a special purpose model based on the economic concept of *demand elasticity.* Small-area, site-specific analysis techniques are presented in Chapter 9.

8.2 TRIP GENERATION

8.2.1 Background

The objective of a trip-generation model is to forecast the number of person-trips that will begin from or end in each travel-analysis zone within the region for a typical day of the target year. Prior to its application, a trip-generation model must be estimated and calibrated using observations taken during the base year by means of a variety of travel surveys (see

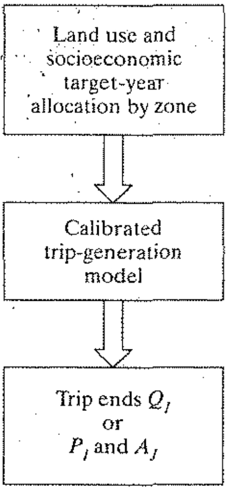

Land use and socioeconomic target-year allocation by zone
Calibrated trip-generation model
Trip ends Q_I or P_I and A_J

Figure 8.2.1 Trip-generation inputs and outputs.

Chapter 7). The total number of person-trips generated constitutes the dependent variable of the model. The independent or explanatory variables include land-use and socioeconomic factors that have been shown to bear a relationship with trip making. When applying a calibrated trip-generation model for predictive purposes, the numerical values of the independent variables must be supplied by the analyst. These values are obtained from the areawide land use and socioeconomic projection phase, which precedes the trip-generation step. As Fig. 8.2.1 illustrates, the output of a trip-generation model consists of the amount of trip making or the *trip ends* Q_I of each zone I within the region [8.4].

8.2.2 Trip Purpose

In contemporary transportation planning the zonal trip making Q_I is estimated separately for each of a number of trip purposes, typically including work trips, school trips, shopping trips, and social or recreational trips. In certain special context studies other categories are considered appropriate as well. For example, a study that examined the travel behavior of users of a special service for elderly and handicapped persons in Honolulu, HI, considered travel for medical and rehabilitational purposes to be relevant categories to that analysis [8.5].

The reason separate trip-generation models are usually developed for each trip purpose is that the travel behavior of trip-makers depends on the trip purpose. For example, work trips are undertaken with daily regularity, mostly during the morning and afternoon period of peak traffic, and overwhelmingly from the same origins to the same destinations. This is also true in the case of school trips. Social and recreational trips, on the other hand, are clearly of a different character. Figure 8.2.2 illustrates that the time-of-day distribution of trips also varies among purposes [8.6].

8.2.3 Zonal-Based versus Household-Based Models

A transportation planning study cannot possibly trace the travel patterns of every individual residing within a region. As a result, the geographical patterns of trip making are summarized by dividing the region into smaller travel-analysis zones and by associating the estimated trips

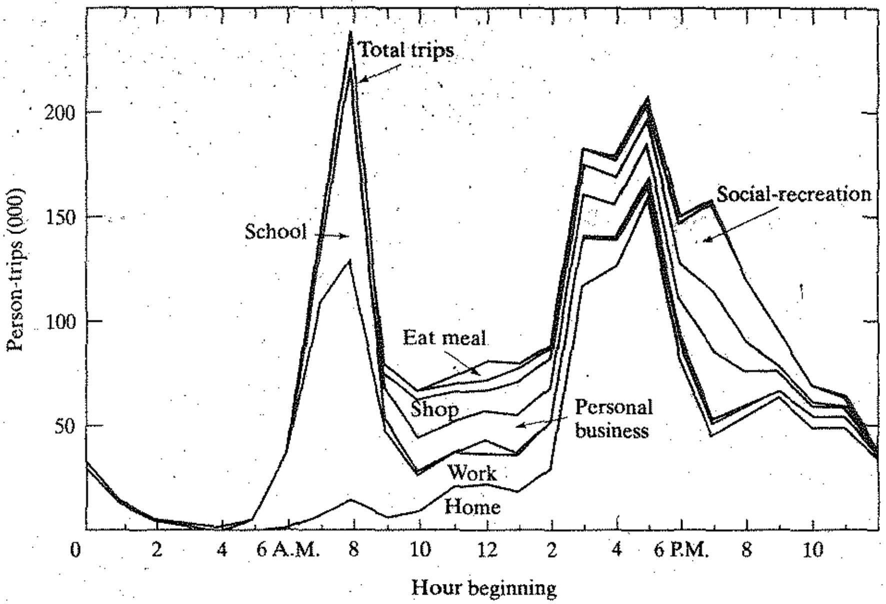


Figure 8.2.2 Hourly distribution of internal person-trips by trip purpose. (From Keefer [8.6].)

with these zones. Early models of trip generation considered the zone to be the smallest entity of interest as far as trip making was concerned. Consequently these models are calibrated on a *zonal basis,* meaning that the overall zonal characteristics were used as independent or explanatory variables. These zonal attributes included variables such as the zonal population, the average zonal income, the average vehicle ownership, and the like. Using zonal averages, however, tends to mask internal (or intrazonal) variability and affects the accuracy of the estimated trip levels. For example, two zones may have the same average income (e.g., in the middle-income range), but one may be composed of a homogenous group of households with respect to income, whereas the second may be composed of two heterogeneous groups, one at high and the other at low income. If income is not linearly related to trip generation, a *zone-based* (or *aggregate*) model will not be sensitive to the intrazonal income differences. *Household-based* (or *disaggregate*) models of trip generation are also available. More advanced models consider the individual rather than the household and the elemental decision-making unit.

The rationale of household-based models is that households with similar characteristics tend to have similar travel propensities irrespective of their geographical location within the region. The calibration of household-based models employs a sample of households rather than a sample of zones. These models are known as disaggregate models because they decompose (or disaggregate) each zone into smaller units. However, this disaggregation is not geographical, as households of the various types may be interspersed throughout a zone. Consequently this decomposition is not necessarily equivalent to delineating smaller zones. In order to arrive at the required estimates of zonal trip generation (thus retaining the spatial characteristics of travel within the region), it is necessary to recombine the contribution of each group of similar households found within the zone into a zonal total. For this reason the land-use projections that provide the inputs to a trip-generation model must specifically forecast the number of households by type. The data requirements of this *market segmentation* approach to aggregation increase rapidly with the number of household classes used. For example, classification by 3 levels of income (e.g., low, medium, high), and 4 levels of household size (e.g., 1, 2, 3, 4+) leads to 12 household types. Adding 3 strata of a third characteristic results in 36 classes. With more household types, larger sample sizes are needed for model estimation. Moreover, the target-year household composition within the region must be projected in terms of the larger number of household classes prior to applying the model. *Sample enumeration* is an alternate way by which aggregate predictions can be obtained from models estimated at the disaggregate level. This technique uses a representative sample of the relevant population (e.g., households within a zone), and the predicted behavior of the sample is taken as an estimate of aggregate behavior. Often Monte Carlo simulation (see Chapter 14) is used along with the sample enumeration method to reduce the computational requirements of this approach. Disaggregate models are further discussed in Section 8.6.

8.2.4 Productions and Attractions

The trips that are predicted by a trip-generation model for each zone are often referred to as the trip ends associated with that zone. Trip ends may be classified as either origins and destinations (O-D) or productions and attractions (P-A). As used in trip-generation studies, the terms *origin* and *production* on one hand and *destination* and *attraction* on the other are not identical. To understand this difference, consider the two zones *I* and *J* of Fig. 8.2.3. Typically each of these zones will contain residences as well as nonresidential land uses, such as places of business, schools, and commercial establishments. The figure captures this fact by showing a portion of each zone as residential and a portion as nonresidential, even though the two types of activity within each zone may be intermingled. Now consider a single worker whose residence is located in zone *I* and whose place of employment is in zone *J*. On a typical workday this trip-maker will travel from zone *I* to zone *J* in the morning and back from zone *J* to zone *I* in the evening. In the morning zone *I* is the trip-maker's *origin* and zone *J* is the trip-maker's *destination*. In the evening, zone *J* becomes the origin and zone *I* the destination. Thus origins and destinations are defined in terms of the direction of a given interzonal trip. In this example each of the two zones experienced two trip ends during the day: one origin and one destination.

The terms *production* and *attraction*, on the other hand, are *not* defined in terms of the directions of trips but in terms of the land use associated with each trip end. A *trip production* is defined as a trip end connected with a residential land use in a zone, and a *trip attraction* is defined as a trip end connected to a nonresidential land use in a zone. On the basis of these definitions, zone *I* of Fig. 8.2.3 has *produced* two trips, whereas zone *J* has *attracted* two trips. This distinction is made because the zonal trip productions can be more easily estimated from the socioeconomic characteristics of the zone's population and the related travel needs of the population for various purposes, whereas the zonal trip attractions depend on the availability and intensity of nonresidential opportunities found within the zone. For example, if a significant portion of the population of a zone consisted of working-age adults, that zone would produce a high number of work trips. On the other

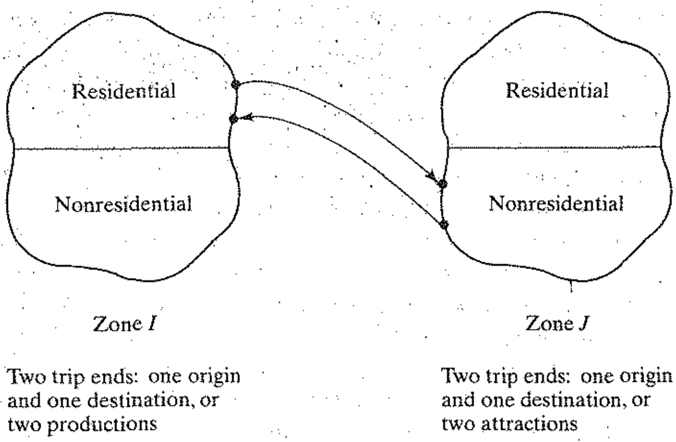

Residential
Nonresidential
Zone I
Two trip ends: one origin and one destination, or two productions
Residential
Nonresidential
Zone J
Two trip ends: one origin and one destination, or two attractions

Figure 8.2.3 Trip-end definitions.

hand, if a zone were predominantly nonresidential (e.g., a downtown employment zone), it would be likely to attract many work trips produced by zones that are dispersed throughout the region.

Thus a typical trip-generation study involves the application of residential trip-production and nonresidential trip-attraction models. The former contain a set of explanatory variables that describe the demographic makeup of the zone's population. The latter rely on a set of explanatory variables that capture the type and intensity of nonresidential activities within the zone. In the general case each zone I will have a number of productions P_I and a number of attractions A_I.

While the vast majority of the trips occurring within urban areas have a production and an attraction trip end, there are trips for which the definition is not directly applicable, for example, trips that take place between two nonresidential activities, such as a trip from the place of employment to a shopping area. Trips can also be classified as *home-based* (HB) or as *non-home-based* (NHB). The former category consists of trips that either begin or end at a residence, whereas the latter neither begin nor end at a residence. This leaves a small percentage of trips usually occurring during the noncritical off-peak periods of the day that have both their origins and their destinations in a residence (e.g., a trip to a friend's house). To account for NHB trips in a production-attraction format, their zone of origin is assumed to be the producing zone and the zone of destination is considered to be the attracting zone. The three most common mathematical formulations of trip generation are regression models, trip-rate analysis models, and cross-classification models.

8.2.5 Regression Models

Chapter 13 presents the underlying theory of least squares regression and shows regression models as linear or nonlinear on one hand and as simple or multiple on the other. All these types of regression models can be employed in connection with trip-generation studies. The selection of the most appropriate form in a particular case is usually based on experience and preliminary investigations into the matter. A frequently used regression model is the linear multiple-regression model.

In a trip-production multiple regression model the dependent variable can estimate either the total trips produced by a zone P_i if it is an aggregate model or the household trip-production rate if it is a household-based model. The independent variables included in a zone-based model are characteristics of the zone as a whole, whereas the independent variables employed by a disaggregate model are household characteristics. The calibration of the former is based on a set of observations for a number of zones, each observation corresponding to a zone; the calibration of a disaggregate model employs a number of base-year observations, each corresponding to an individual household in a sample of households drawn randomly from the region. In the case of multiple regression models of trip attractiveness the independent variables consist of nonresidential attributes.

In all cases each term of the equation can be interpreted as the contribution of the corresponding independent variable to the magnitude of the dependent variable. That is, a unit change in an independent variable is seen to result in a change in the dependent variable, which equals the magnitude of the coefficient of the independent variable. The constant term a_0 captures effects that are not explicitly included in the model.

8.2.6 Trip-Rate Analysis

Trip-rate analysis refers to several models that are based on the determination of the average trip-production or trip-attraction rates associated with the important trip generators within the region. Table 8.2.1, for example, displays the trip-generation rates associated with various land-use categories in downtown Pittsburgh, which were obtained by one of the very first major urban transportation studies [8.6].

This table includes production rates by residential land uses and attraction rates by several nonresidential land uses. Care must be exercised to apply trip-rate models in the same context that they were calibrated. In this case the rates represent person-trips (rather than vehicle-trips) per thousand square feet of each land use. Also, it is almost always true

TABLE 8.2.1 Golden Triangle Floor-Space Trip-Generation Rates Grouped by Generalized Land-Use Categories

Land-use category	Thousands of square feet	Person-trips	Trips per thousand square feet
Residential	2744	6574	2.4
Commercial			
Retail	6732	54,833	8.1
Services	13,506	70,014	5.2
Wholesale	2599	3162	1.2
Manufacturing	1392	1335	1.0
Transportation	1394	5630	4.0
Public buildings	2977	11,746	3.9
Total[a]	31,344	153,294	
Average[a]			4.9

[a] Includes trips to public open spaces.

Source: Keefer [8.6].

that pre-1985 data do not include short trips made by nonmotorized modes such as bicycling and walking.

The hypothetical trip-attraction rates of Table 8.2.2 are expressed in terms of the number of trips attracted per employee for the case of retail and nonretail land uses and in terms of school trips attracted per student enrolled in each of the three types of educational institutions [8.4, 8.7].

Trip rates for a large number of commercial developments and other land uses are supplied in the *Trip Generation* manual of the Institute of Transportation Engineers. These rates are more appropriate for site impact analysis. They are presented in Chapter 9.

8.2.7 Cross-Classification Models

Cross-classification (or category analysis) models may be thought of as extensions of the simple trip-rate models discussed previously. Although they can be calibrated as area- or zone-based models, in trip-generation studies they are almost exclusively used as disaggregate models. In the residential-generation context, household types are classified according to a set of categories that are highly correlated with trip making. Three to four explanatory variables, each broken into about three discrete levels, are usually sufficient. Typically household size, automobile ownership, household income, and some measure of land development intensity are used to classify household types.

The trip rates associated with each type of household are estimated by statistical methods, and these rates are assumed to remain stable over time. Table 8.2.3 presents a cross-classification table that shows the calibrated nonwork home-based trip-production rates for various types of households defined by (1) four levels of household size (i.e., number of persons per household), (2) three levels of car ownership (i.e., vehicles available per household), and (3) three levels of residential density (i.e., dwelling units per acre), a surrogate for accessibility to nonwork activities (e.g., shopping, entertainment).

TABLE 8.2.2 Example of Procedure for Trip-Attraction Estimates Person-Trip Attractions[a]

		Trips per employee						
			Retail					
Trip purpose	Trips per household	Nonretail	CBD	Shop center	Other	University	High school	Other
Home-based work		1.70	1.70	1.70	1.70			
Home-based shop			2.00	9.00	4.00			
Home-based school						0.90	1.60	1.20
Home-based other	0.70	0.60	1.10	4.00	2.30			
Non-home-based[b]	0.30	0.40	1.00	4.60	2.30			

[a] Illustration data only, not to be used directly.

[b] Non-home-based productions and attractions have the same rate and are used to allocate to zones and areawide control total developed in the trip-production model.

Source: Federal Highway Administration [8.4].

TABLE 8.2.3 Example: Total Home-Based-Non-work Trip Rates

Area type	Vehicles available per household	Persons per household 1	2, 3	4	5 +
1. Urban: high density	0	0.57	2.07	4.57	6.95
	1	1.45	3.02	5.52	7.90
	2+	1.82	3.39	5.89	8.27
2. Suburban: medium density	0	0.97	2.54	5.04	7.42
	1	1.92	3.49	5.99	8.37
	2+	2.29	3.86	6.36	8.74
3. Rural: low density	0	0.54	1.94	4.44	6.82
	1	1.32	2.89	5.39	7.77
	2+	1.69	3.26	5.76	8.14

Source: Oahu Metropolitan Planning Organization [8.8].

The documentation of the study that produced this table [8.8] specifies points of demarcation between the levels of density denoted as high (urban), medium (suburban), and low (rural). Each cell of the table contains the calibrated daily trip-production rate per household expressed in terms of person-trips per household per day. Given projections relating to the target-year household composition of a zone, the application of this calibrated model for predictive purposes is straightforward, as the following example illustrates.

Example 8.1: Cross Classification

An urban zone contains 200 acres of residential land, 50 acres devoted to commercial uses, and 10 acres of park land. The following table presents the zone's expected household composition at some future (target) year.

Vehicles per household	Persons per household 1	2, 3	4	5
0	100	200	150	20
1	300	500	210	50
2+	150	100	60	0

Using the calibrated cross-classification table of Table 8.2.3, estimate the total nonwork home-based trips that the zone will produce during a typical target-year day. The rates are given as trips per household per day.

Solution The total productions are estimated by summing the contribution of each household type:

$$P_I = \sum_h N_h R_h$$

where N_h and R_h are the number of households of type h and their corresponding production rate. For example, the 300 single-person one-car households contribute (300)(1.45) = 435 nonwork home-based trips per day. Summing over all household types gives

$$P_I = 5760 \text{ trips per day}$$

TABLE 8.2.4 Average Daily Person-Trips per Household by Household Size and Income

Urban area population	Income in 1990 U.S. ($)	Persons per household					Weighted average
		1	2	3	4	5+	
50,000	<$20,000	3.6	6.5	9.1	11.5	13.8	6.0
to	$20,000–39,000	3.9	7.3	10.0	13.1	15.9	9.3
199,999	>$40,000	4.5	9.2	12.2	14.8	18.2	12.7
Weighted average		*3.7*	*7.6*	*10.6*	*13.6*	*16.6*	*9.2*
200,000	<$20,000	3.1	6.3	9.4	12.5	14.7	6.0
to	$20,000–39,000	4.8	7.2	10.1	13.3	15.5	9.4
499,999	>$40,000	4.9	7.7	12.5	13.8	16.7	11.8
Weighted average		*3.7*	*7.1*	*10.8*	*13.4*	*15.9*	*9.0*
500,000	<$20,000	3.6	7.1	9.0	12.0	14.0	6.0
to	$20,000–39,000	4.8	7.1	9.8	12.7	14.6	8.9
999,999	>$40,000	4.8	7.8	11.5	13.6	16.6	11.5
Weighted average		*4.0*	*7.3*	*10.2*	*13.0*	*15.4*	*8.7*
1,000,000	<$20,000	3.7	6.3	8.1	10.0	11.8	5.7
or	$20,000–39,000	4.9	7.6	9.1	12.3	15.1	9.0
more	>$40,000	5.4	7.9	10.3	12.4	15.3	10.8
Weighted average		*4.2*	*7.3*	*9.3*	*12.0*	*14.8*	*8.5*

Source: NCHRP 365, 1998 [8.1].

Discussion Only the residential land-use sector of the zone entered into the solution because trip productions are associated with the residential characteristics of the zone. The commercial and recreational characteristics of the zone would be relevant to the estimation of the attractiveness of the zone for these purposes. In that case properly calibrated attractiveness models would be required. Table 8.2.4, taken from reference [8.1], presents a typical cross-classification table for trip productions in terms of daily person-trips per household. In this example households are characterized by household size (persons per household), income (low, medium, high) and urbanized area population (four levels). This table was compiled using data derived from the 1990 U.S. Census and other national sources.

8.2.8 The FHWA-Simplified Trip-Production Procedure

Figure 8.2.4 presents a hypothetical example of a residential trip-generation procedure developed by the FHWA [8.4]. This procedure combines several of the concepts discussed in this section. Curve A represents the distribution of households by income and auto ownership. In the example shown the auto ownership of a group of households with an annual income of $12,000 is distributed as follows: 2% own no autos, 32% own one auto, 52% own two autos, and 14% own three or more autos. Incidentally, depending on the way in which curve A is calibrated, it may represent a zonal (aggregate) distribution, in which case the income variable would be a zonal average, or it may represent a household-based (disaggregate) distribution, in which case the group of households illustrated may correspond to a subset of all the households in a zone. In the latter case the percentages obtained can be interpreted as probabilities; for example, the probability that any $12,000 per year household will own two cars is 0.52. (*Note:* the $12,000 in Figure 8.2.4 roughly correspond to

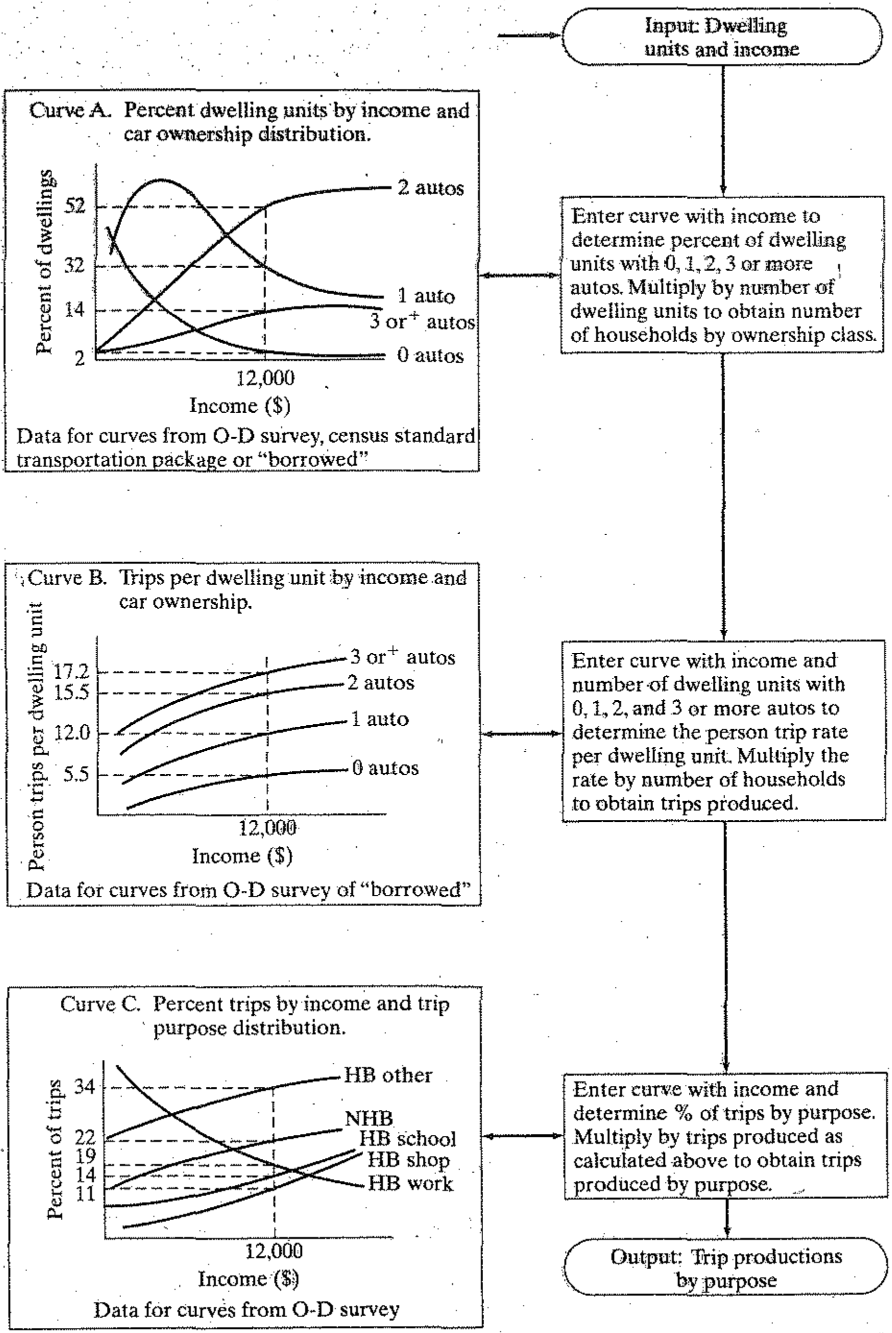

Input: Dwelling units and income
Curve A. Percent dwelling units by income and car ownership distribution.
Percent of dwellings
52
32
14
2
2 autos
1 auto
3 or+ autos
0 autos
12,000
Income ($)
Data for curves from O-D survey, census standard transportation package or "borrowed"
Enter curve with income to determine percent of dwelling units with 0, 1, 2, 3 or more autos. Multiply by number of dwelling units to obtain number of households by ownership class.
Curve B. Trips per dwelling unit by income and car ownership.
Person trips per dwelling unit
17.2
15.5
12.0
5.5
3 or+ autos
2 autos
1 auto
0 autos
12,000
Income ($)
Data for curves from O-D survey of "borrowed"
Enter curve with income and number of dwelling units with 0, 1, 2, and 3 or more autos to determine the person trip rate per dwelling unit. Multiply the rate by number of households to obtain trips produced.
Curve C. Percent trips by income and trip purpose distribution.
Percent of trips
34
22
19
14
11
HB other
NHB
HB school
HB shop
HB work
12,000
Income ($)
Data for curves from O-D survey
Enter curve with income and determine % of trips by purpose. Multiply by trips produced as calculated above to obtain trips produced by purpose.
Output: Trip productions by purpose

Figure 8.2.4 Example of urban trip-production procedure. (From Federal Highway Administration [8.4].)

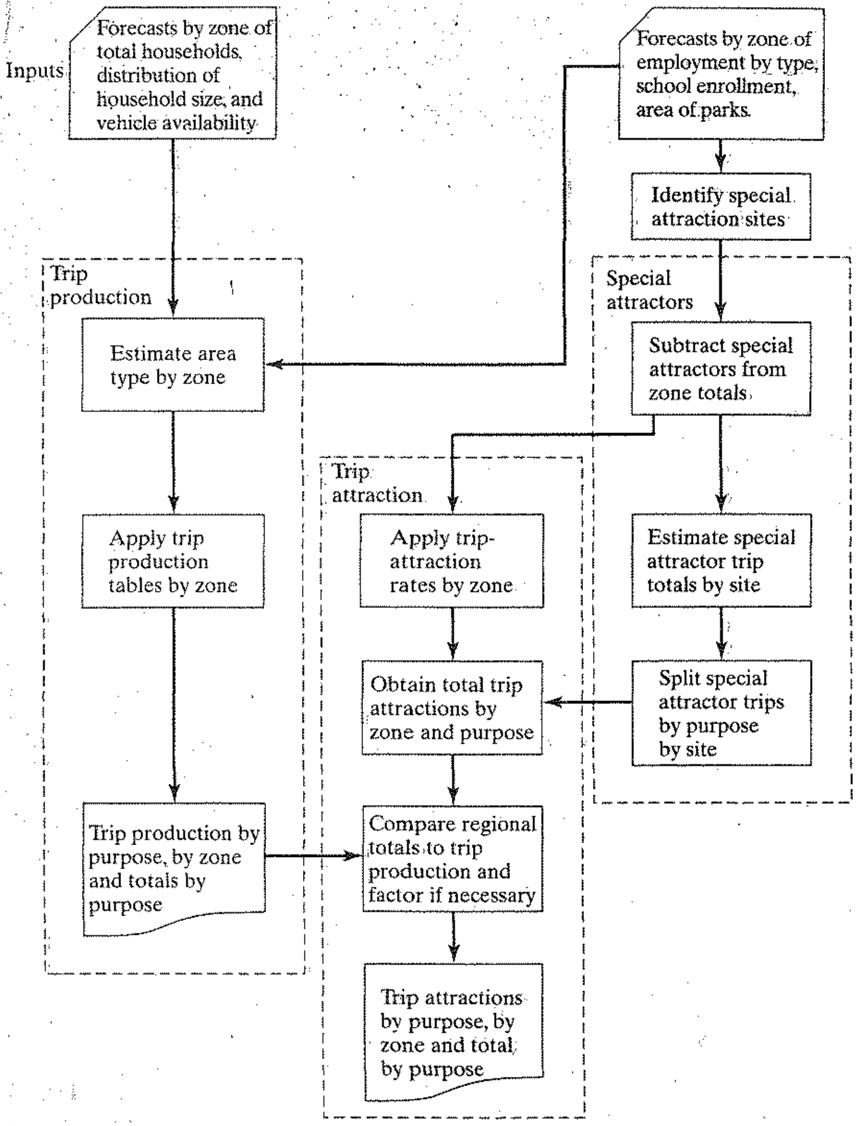

Inputs
Forecasts by zone of total households, distribution of household size, and vehicle availability
Forecasts by zone of employment by type, school enrollment, area of parks.
Identify special attraction sites
Trip production
Estimate area type by zone
Apply trip production tables by zone
Trip production by purpose, by zone and totals by purpose
Special attractors
Subtract special attractors from zone totals
Estimate special attractor trip totals by site
Split special attractor trips by purpose by site
Trip attraction
Apply trip-attraction rates by zone
Obtain total trip attractions by zone and purpose
Compare regional totals to trip production and factor if necessary
Trip attractions by purpose, by zone and total by purpose

Figure 8.2.5 Application of long-range trip-generation procedure. (From Oahu Metropolitan Planning Organization [8.8].)

$52,000 in 2000.) The family of curves designated as *Curve A* illustrates one method by which the automobile-ownership level of households may be determined when income is known. As the inserted note explains, these lines can be derived from census data, a local origin-destination (or, it may be added, any household interview survey) or, in the absence of local data, they may be borrowed from a similar urban area. When developing these curves, care must be exercised to ensure that the sum of the percentages corresponding to a particular income always equals 100%.

Similar curves of automobile ownership may be derived by using "average zonal automobile ownership" on the horizontal axis. Such curves are known as *aggregate share models* and are used extensively in the United States.

A major drawback of these curves is that they do not capture the behavioral variables that affect the decision to purchase automobiles. Examples of a behavioral model of automobile ownership and trip-generation models are given in Section 8.6.

The regression lines of curve B provide the person-trip rates for household types defined by income and auto ownership, and curve C divides these trips among several trip purposes. Adherence to the instructions that accompany the figure leads to the target-year estimate of trip productions by purpose.

8.2.9 Summary

The purpose of trip generation is to estimate the target-year trip ends by travel purpose for each zone within the region. Commonly, these trips are expressed as residential trip productions and nonresidential trip attractions. The most common mathematical forms of trip-generation models are multiple regression equations, trip-rate models, cross-classification models, and their combinations.

Figure 8.2.5 illustrates the trip-generation procedure used to obtain long-range forecasts in a particular urban area. This figure shows that two sets of inputs, residential and nonresidential characteristics, were first obtained from zonal socioeconomic and land-use projections. The specific variables used are listed among the inputs. Residential projections were used by the trip-production models, which took the form of household-based cross-classification tables to estimate the target-year zonal trip productions by purpose. The nonresidential land-use projections were used primarily in relation to the trip-attraction model, which was of the form of multiple regression. Several special attractors, including airports, major shopping centers, and universities, have been given special treatment because of their unique trip-attraction characteristics. The final outputs were the zonal productions and attractions by trip purpose.

8.3 TRIP DISTRIBUTION

8.3.1 Background

The next step in the sequential forecasting model system is concerned with the estimation of the target-year trip-volumes Q_{IJ} that interchange between all pairs of zones I and J, where I is the trip-producing zone and J is the trip-attracting zone of the pair. The rationale of trip distribution is as follows: All trip-attracting zones J in the region are in competition with each other to attract trips produced by each zone I. Everything else being equal, more trips will be attracted by zones that have higher levels of "attractiveness." However, other intervening factors affect the choice of J as well. Consider, for example, the case of two identical shopping centers (i.e., of equal attractiveness) competing for the shopping trips produced by a given zone I. If the distances between zone I and each of the two centers are different, shoppers residing in zone I will show a preference for the closer of the two identical centers. Thus the intervening difficulty of travel between the producing zone I and each of the competing zones J has a definite effect on the choice of attraction zone. In the shopping center example

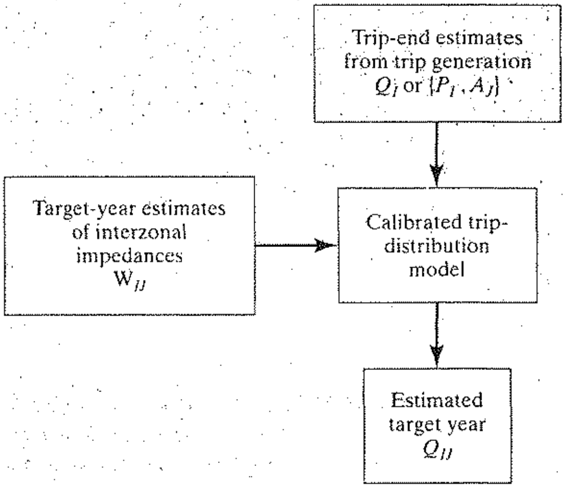

Trip-end estimates from trip generation Q_I or $\{P_I, A_J\}$
Target-year estimates of interzonal impedances W_{IJ}
Calibrated trip-distribution model
Estimated target year Q_{IJ}

Figure 8.3.1 Trip-distribution inputs and outputs.

distance is cited as a measure of this difficulty of travel, but other measures of this effect may be used, such as travel time or some generalized cost that includes travel time, out-of-pocket cost, and the like. The notation W_{IJ} is used for this generalized cost, which is also known as travel *impedance*. When applying a specific model for predictive purposes, care must be taken to use the same measure of impedance that was employed to calibrate the model.

Figure 8.3.1 conceptually illustrates that a trip-distribution model estimates the interzonal person-trip volumes Q_{IJ} based on the productions of each zone I, the attractiveness of zone J, and the interzonal impedance W_{IJ}. The production and attraction inputs are obtained from the preceding trip-generation phase, and estimates of the target-year interzonal impedances are obtained from the specification of the alternative transportation networks under investigation. A table showing the interzonal impedances is known as a *skim table.*

The most common mathematical formulations of trip distribution include various growth-factor models, the gravity model, and a number of *opportunities* models. The following sections discuss the gravity model and one growth-factor model (the Fratar model). The application of a class of models known as discrete choice models (e.g., logit) based on the economic principle of utility maximization constitutes the best practice. The logit model and its variants are discussed in Section 8.4.

8.3.2 The Gravity Model

The gravity model gets its name from the fact that it is conceptually based on Newton's law of gravitation, which states that the force of attraction between two bodies is directly proportional to the product of the masses of the two bodies and inversely proportional to the square of the distance between them, or

$$F = k\frac{M_1M_2}{r^2} \tag{8.3.1}$$

Variations of this formula have been applied to many situations involving human interaction. For example, the volume of long-distance telephone calls between cities may be modeled in this manner, with the population sizes of the cities replacing the masses of particles and the distance between cities or the cost of telephone calls taking the place of r. The exponent of the impedance term in the denominator, however, does not need to be exactly equal to 2 but may be replaced by a model parameter c.

The application of this concept to trip distribution takes the form

$$Q_{IJ} = k\frac{P_I A_J}{W_{IJ}^c} \tag{8.3.2}$$

Equation 8.3.2 states that the interchange volume between a trip-producing zone I and a trip-attracting zone J is directly proportional to the magnitude of the trip productions of zone I and the trip-attractiveness of zone J and is inversely proportional to a function of the impedance W_{IJ} between the two zones.

Using the usual mathematical modeling terminology, the interzonal volume is the dependent variable; the productions, attractions, and impedances are the independent variables; and the constants k and c are the parameters of the model that must be estimated through calibration using base-year data.

The parameter k can be eliminated from Eq. 8.3.2 by applying the *trip-production balance* constraint, which states that the sum over all trip-attracting zones J of the interchange volumes that share I as the trip-producing zone must equal the total productions of zone I, or

$$P_I = \sum_x Q_{Ix} \tag{8.3.3}$$

Equation 8.3.3 ensures that the model will distribute to the competing zones J exactly as many trips as are produced by zone I.

Substituting Eq. 8.3.2 into Eq. 8.3.3 and taking the terms not involving the index x outside the summation, we obtain

$$P_I = kP_I \sum_x \frac{A_x}{W_{Ix}^c} \tag{8.3.4}$$

Solving for k yields

$$k = \left[\sum_x \frac{A_x}{W_{Ix}^c}\right]^{-1} \tag{8.3.5}$$

which is the expression for k that ensures that the trip balance Eq. 8.3.3 is satisfied.

Substituting Eq. 8.3.5 into Eq. 8.3.2 leads to the classical form of the gravity model:

$$Q_{IJ} = P_I \left[\frac{A_J/W_{IJ}^c}{\sum_x (A_x/W_{Ix}^c)}\right] \tag{8.3.6}$$

The bracketed term is the proportion of the trips produced by zone I that will be attracted by zone J in competition with all trip-attracting zones x. Note that the numerical value of this fraction would not be affected if all attraction terms were multiplied by a constant. This implies that the attraction terms can measure the relative attractiveness of zones.

For example, one employment zone may be said to be twice as attractive as another, based on the number of employment opportunities available. In this context the estimated target-year trip attractions of a zone J (denoted by A_J^* to distinguish them from the relative attractiveness term used earlier) may be computed by applying the following *trip-attraction balance* equation to the results of the model:

$$A_J^* = \sum_x Q_{xJ} \tag{8.3.7}$$

The gravity formula is often written alternatively as

$$Q_{IJ} = P_I \left(\frac{A_J F_{IJ}}{\sum_x A_x F_{Ix}} \right) \tag{8.3.8}$$

where

$$F_{IJ} = \frac{1}{W_{IJ}^c} \tag{8.3.9}$$

is known as the *travel-time* (or *friction*) *factor.* Note that the calibration constant c is now implicit in the friction factor.

Finally, a set of interzonal *socioeconomic adjustment factors* K_{IJ} are introduced during calibration to incorporate effects that are not captured by the limited number of independent variables included in the model. The resulting gravity formula becomes

$$Q_{IJ} = P_I \frac{A_J F_{IJ} K_{IJ}}{\sum_J A_J F_{IJ} K_{IJ}} = P_I p_{IJ} \tag{8.3.10}$$

where p_{IJ} is the probability that a trip generated by zone I will be attracted by zone J. As mentioned earlier, a table that contains the interzonal impedances W_{IJ} is known as a *skim table.*

Example 8.2: Application of the Gravity Model

The target-year productions and relative attractiveness of the four-zone city have been estimated as follows:

Zone	Productions	Attractiveness
1	1500	0
2	0	3
3	2600	2
4	0	5

The calibration of the gravity model for this city estimated the parameter c to be 2.0 and all socioeconomic adjustment factors to be equal to unity. Apply the gravity model to estimate all target interchanges Q_{IJ} and to estimate the total target-year attractions of each zone given that the target-year interzonal impedances W_{IJ} will be as shown in the following skim table.

I \ J	1	2	3	4
1	5	10	15	20
2	10	5	10	15
3	15	10	5	10
4	20	15	10	5

Solution The gravity model calculations of the interchange volumes are shown in tabular form for the two trip-producing zones ($I = 1$ and $I = 3$). For $I = 1$, $P_1 = 1500$:

J	A_J	F_{1J}	K_{1J}	$A_J F_{1J} K_{1J}$	p_{1J}	Q_{1J}
1	0	0.0400	1.0	0	0	0
2	3	0.0100	1.0	0.0300	0.584	875
3	2	0.0044	1.0	0.0089	0.173	260
4	5	0.0025	1.0	0.0125	0.243	365
				0.0514	1.000	$1500 = P_1$

For $I = 3$, $P_3 = 2600$:

J	A_J	F_{3J}	K_{3J}	$A_J F_{3J} K_{3J}$	p_{3J}	Q_{3J}
1	0	0.0044	1.0	0.0	0	0
2	3	0.0100	1.0	0.03	0.188	488
3	2	0.0400	1.0	0.08	0.500	1300
4	5	0.0100	1.0	0.05	0.312	812
				0.16	1.000	$2600 = P_3$

To find the total target-year trip attractions of the nonresidential zones ($J = 2$, $J = 3$, and $J = 4$), apply the trip-attraction balance (Eq. 8.3.7) to get

$$A_2^* = 875 + 488 = 1363$$

$$A_3^* = 260 + 1300 = 1560$$

$$A_4^* = 365 + 812 = 1177$$

The solution is summarized by the following *trip table:*

I \ J	1	2	3	4	Sum
1	0	875	260	365	1500
2	0	0	0	0	0
3	0	488	1300	812	2600
4	0	0	0	0	0
Sum	0	1363	1560	1177	4100

Discussion The trip-generation data indicate that there are three types of zones in this city: Zone 1 is purely residential because it is shown to have productions only, zones 2 and 4 are purely nonresidential because they produce no trips, and zone 3 is a mixed land-use zone because it has both productions and attractions. The impedance matrix represents an estimate of interzonal impedances for the target year. The diagonal elements of this matrix represent intrazonal impedances, that is, the impedances associated with trips that begin *and* end within each zone. It is possible, of course, that trips produced by the mixed land-use zone 3 could be attracted by the nonresidential sector of the same zone. The sum of each row of the trip table produces the total productions of the corresponding zone *I*, whereas the sum of each column represents the total attractions of each zone *J*. Again, note that the purely residential zone has no attractions and the purely nonresidential zones have no productions. The mixed zone has both.

Example 8.3: The Generation-Distribution Sequence

You are a planning consultant to a trading firm that is considering the construction of a major shopping center in the city of Trinity. At present the city consists of three residential zones and the central business district (CBD), where all shopping activity is concentrated. Your clients can acquire land for the proposed center at the location shown and are interested in your prediction of the patronage that the center will attract if built to compete with the CBD. The following data have been made available to you:

1. *Daily shopping trip production (trips per person):*

X_1 \ X_2	0	1	2
≤2	0.2	0.3	0.4
3	0.1	0.2	0.3
≥4	0.1	0.2	0.3
		$X_3 = \text{I}$	

X_1 \ X_2	0	1	2
≤2	0.3	0.4	0.5
3	0.2	0.2	0.4
≥4	0.2	0.2	0.5
		$X_3 = \text{II}$	

where

X_1 = household size, in persons/household

X_2 = auto ownership, in cars/household

X_3 = household income level (I or II)

2. *Relative shopping attractiveness:* The relative shopping attractiveness of commercial zones has been found to be given by the following multiple regression equation:

$$A = 5X_a + 3X_b$$

where

X_a = area of shopping floor space provided, in acres

X_b = available parking area, in acres

3. *Land-use and socioeconomic projections:*

Residential Zones

				Number of households	
Zone	X_1	X_2	X_3	Base year	Target year
1	2	0	I	300	500
	2	1	I	300	400
	3	1	I	200	300
	2	2	II	0	50
2	2	1	I	400	500
	2	1	II	300	200
	3	2	I	200	300
	3	0	I	100	400
3	1	1	II	200	200
	2	2	II	300	400
	3	2	II	400	300
	4	2	II	200	400

Commercial Zones

	Base year		Target year	
Zone	X_a	X_b	X_a	X_b
4 (CBD)	3.0	2.0	3.0	2.5
5	0.0	0.0	2.0	3.0

4. *Gravity model parameters:*

(a) $\ln F = -\ln W$, where W is the interzonal impedance, in minutes (see Fig. 8.3.2).

(b) K_{IJ}

J I	4 (CBD)	5 (center)
1	1.0	0.9
2	0.9	1.2
3	1.0	1.0

You are asked to calculate all target-year interchange volumes and the target-year patronage of the two commercial zones.

Solution First, apply the calibrated trip-generation models and the available land-use and socioeconomic projections to find the target-year productions and relative attractiveness of the five zones. The shopping trip-production model is a disaggregated cross-classification model. Considering the units of the calibrated production rate, the contribution of each household type to the total zonal productions is

(number of households)(household size)(trips per person)

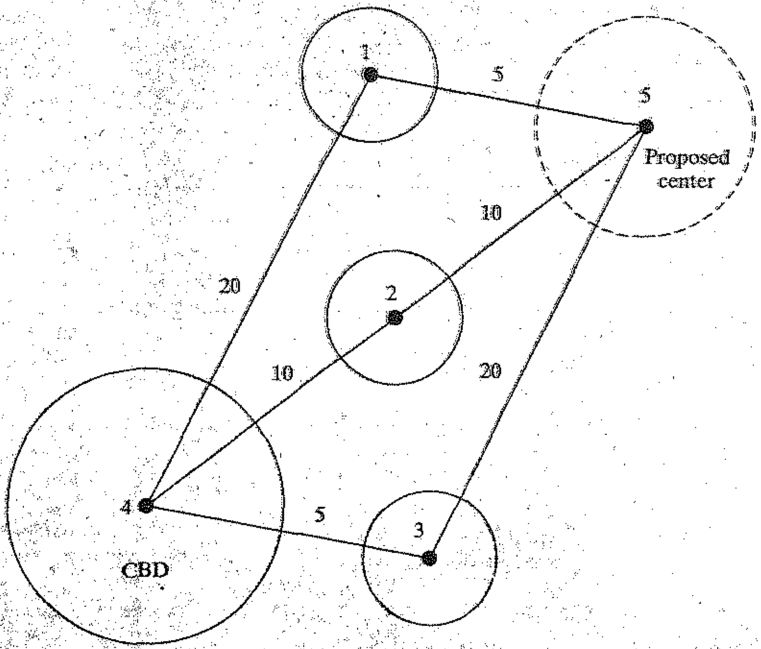


Figure 8.3.2 Interzonal impedances.

Hence for each of the trip-producing zones:

Zone 1	Zone 2	Zone 3
500 × 2 × 0.2 = 200	500 × 2 × 0.3 = 300	200 × 1 × 0.4 = 80
400 × 2 × 0.3 = 240	200 × 2 × 0.4 = 160	400 × 2 × 0.5 = 400
300 × 3 × 0.2 = 180	300 × 3 × 0.3 = 270	300 × 3 × 0.4 = 360
50 × 2 × 0.5 = 50	400 × 3 × 0.1 = 120	400 × 4 × 0.5 = 800
P_1 = 670	P_2 = 850	P_3 = 1640

The target-year attractiveness of the two competing commercial zones is calculated via the calibrated trip-attractiveness equation and the relevant land-use projections as follows:

$$A_4 = 5 \times 3 + 3 \times 2.5 = 22.5$$

$$A_5 = 5 \times 2 + 3 \times 3.0 = 19.0$$

The target-year interchange volumes are computed using the gravity model with the given $c = 1$ and the given K_{IJ} factors. Proceeding as in Example 8.4, the following trip table results:

I \ J	4 (CBD)	5 (center)	P_I
1	166	504	670
2	400	450	850
3	1354	286	1640
A_J^*	1920	1240	

Thus 1240 of the estimated 3160 daily shopping trips (or 39% of the total) will be attracted by the proposed shopping center if built.

Discussion This example illustrates the application of the demand-forecasting models discussed so far and shows how the steps of the sequential forecasting procedure are linked together. The prerequisite selection and calibration of the given models had already been carried out using base-year data. Also, the target-year land-use and socioeconomic projections are given.

The production model is of the cross-classification type, and the production rates are given as trips per person. This is reflected in the calculations where this rate multiplies the total number of persons belonging to each socioeconomic category. The attraction model is a zonal (aggregate) multiple regression equation using shopping floor space and parking availability as the determinants of attractiveness. The dependent variable is relative attractiveness and not trip attractions.

The gravity model of trip distribution incorporates the effect of interzonal impedance, which is clearly seen in the results. Since the productions and attractions are defined irrespective of direction, the actual patronage of the two centers will be half of the attractions just calculated. Note, however, that the same result would be obtained by using half of the productions of each zone in the gravity model.

8.3.3 Calibration of the Gravity Model

The calibration of the gravity model in the form of Eq. 8.3.6 involves the determination of the numerical value of the parameter c that fixes the model to the one that reproduces the base-year observations. Equation 8.3.8 is simply another way of expressing Eq. 8.3.6 by substituting Eq. 8.3.9 in the latter. Hence knowledge of the proper value of c fixes the relationship between the travel-time factor and the interzonal impedance.

Unlike the calibration of a simple linear regression model where the parameters can be solved for by a relatively easy minimization of the sum of squared deviations (see Chapter 13), the calibration of the gravity formula is accomplished through an iterative procedure: An initial value of c is assumed and Eq. 8.3.6 is applied using the known base-year productions, attractiveness, and impedances to compute the interzonal volumes Q_{IJ}. These results are then compared with those observed during the base year. If the computed volumes are sufficiently close to the observed volumes, the current value of c is retained as the calibrated value. Otherwise an adjustment to c is made and the procedure is continued until an acceptable degree of convergence is reached. Most commonly, the friction-factor function F rather than the parameter c is used in the calibration procedure. In that case Eq. 8.3.8 is employed in the place of Eq. 8.3.6.

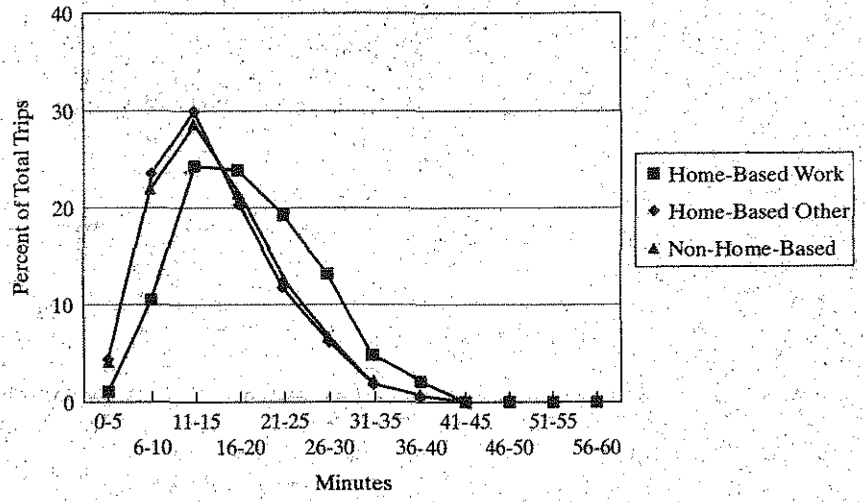


Figure 8.3.3 Typical trip-length frequency distributions by trip purpose. (From Martin and McGuskin [8.1].)

The results of calibration are then expressed in terms of the appropriate equation relating the friction factor and the interzonal impedance. Example 8.4 illustrates this procedure and clarifies the role of the socioeconomic adjustment factors K_{IJ} as well [8.7].

The comparison between the computed and the observed values of Q_{IJ} is accomplished by using the *trip-length frequency distribution*.This distribution consists of a plot of the percentages of the regionwide trips versus their interzonal impedance and has the general shape illustrated in Fig. 8.3.3; The frequency of trips eventually decreases with increasing impedance, as should be expected [8.1].

The base-year trip-length frequency distribution may be compared with that resulting from applying the model during each iteration of the calibration procedure until the latter distribution sufficiently conforms to the former. The following example illustrates the major thrust of the gravity-model calibration procedure (Fig. 8.3.4).

Example 8.4: Calibration of the Gravity Model

Consider the five-zone city shown by Fig. 8.3.5(a). Two of the zones are purely residential, and the remaining three are purely nonresidential. The base-year interzonal impedances are specified in terms of travel time in minutes and are shown in parentheses on the arcs joining pairs of zones. The observed base-year productions, attractiveness, and trip-interchange volumes are inserted in the figure. It is required to find the value of c and the values of K_{IJ} that cause Eq. 8.3.8 to reproduce the observed base-year data.

Solution By taking the natural logarithm of both sides, Eq. 8.3.9 may be rewritten as

$$\ln F = -c \ln W \tag{8.3.11}$$

In other words the negative of the parameter c is the slope of a straight line relating the logarithmic transformations of the friction factor and the interzonal impedance.

Figure 8.3.6 plots the base-year trip-length frequency distribution using the base-year observations. Impedance is shown in 5-min increments, and the ordinate represents the percent of the total trips that travel at the corresponding impedance level.

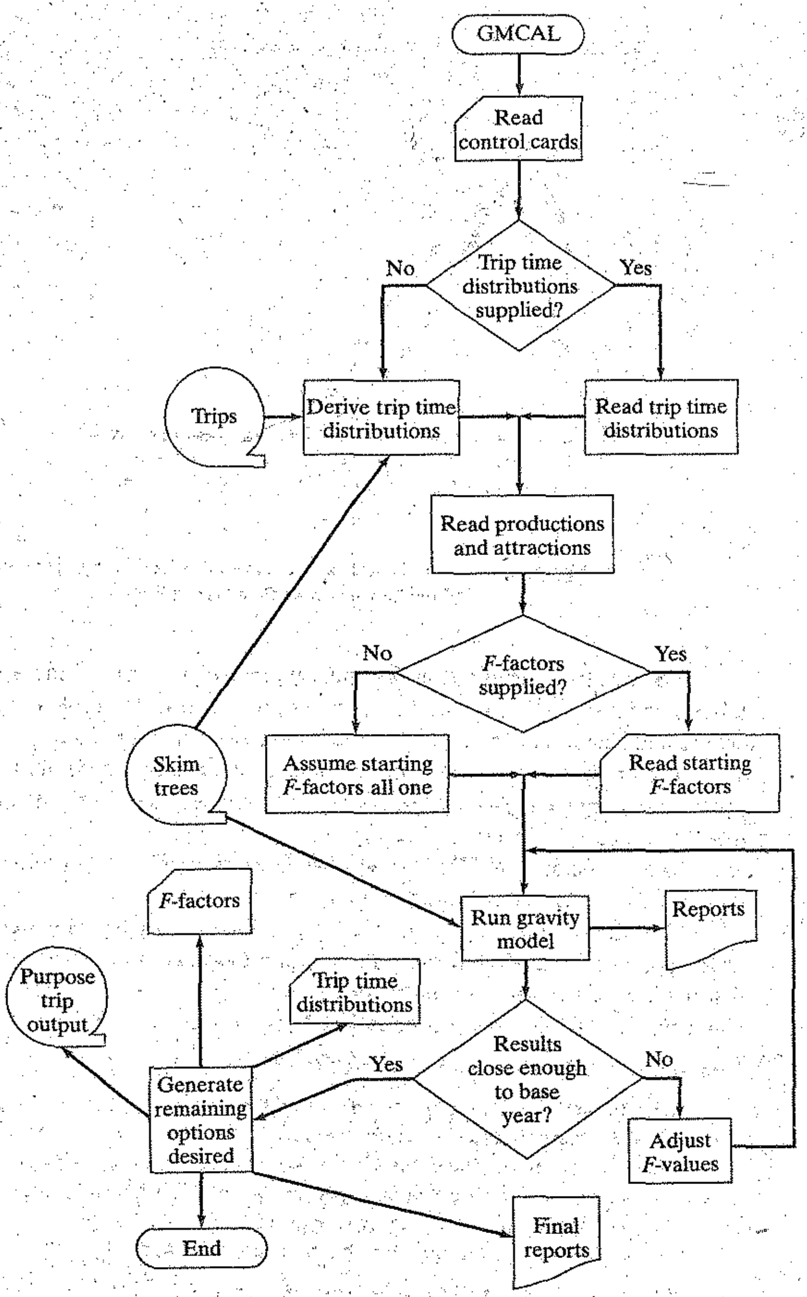
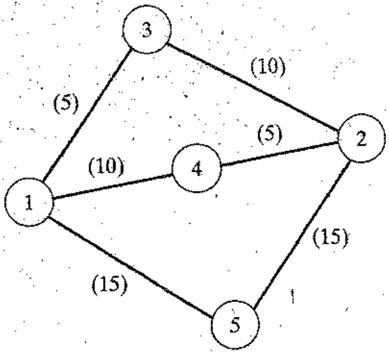

GMCAL
Read control cards
Trip time distributions supplied?
No
Yes
Trips
Derive trip time distributions
Read trip time distributions
Read productions and attractions
F-factors supplied?
No
Yes
Skim trees
Assume starting F-factors all one
Read starting F-factors
F-factors
Run gravity model
Reports
Purpose trip output
Trip time distributions
Results close enough to base year?
Yes
No
Generate remaining options desired
Adjust F-values
Final reports
End

Figure 8.3.4 Gravity-model calibration procedure. (From Federal Highway Administration [8.7].)



(a) Five-zone city

Zone I	P_I	A_I
1	500	0
2	1000	0
3	0	2
4	0	3
5	0	5

(b) Base-year generation

I \ J	3	4	5
1	300	150	50
2	180	600	220

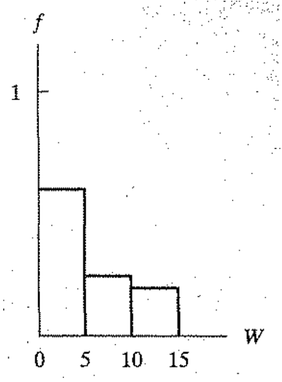
(c) Base-year distribution

Figure 8.3.5 Base-year data for Example 8.4.

Frequency computation :

W	ΣQ_{IJ}	$f = (2)/\text{sum}$
5	$300 + 600 = 900$	0.60
10	$150 + 180 = 330$	0.22
15	$50 + 220 = \underline{270}$	$\underline{0.18}$
	Sum = 1500	1.00

(a) Frequency calculation



(b) Base-year trip length frequency distribution

Figure 8.3.6 Trip-length frequency distribution.

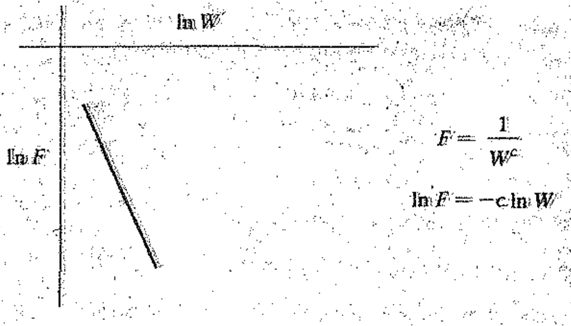
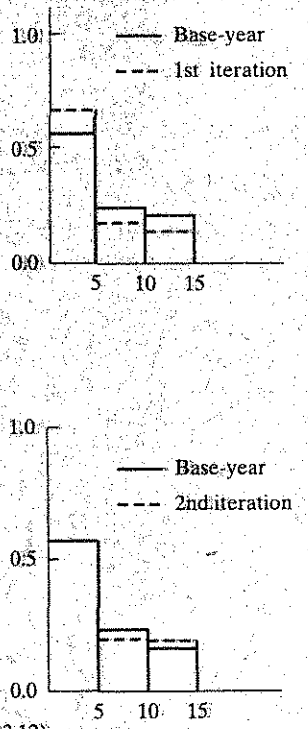


Figure 8.3.7 ln F versus ln W for $c = 2$.

The calibration procedure begins by assuming an initial estimate for c, say, 2.0. This assumption is reflected in the plot of Eq. 8.3.11 shown in Fig. 8.3.7, which plots the initially assumed relationship between F and W.

Application of the gravity formula using the assumed value of c leads to the interzonal volume estimates shown in Fig. 8.3.8 along with the *calculated* trip-length frequency distribution superposed on the *observed* trip-length frequency distribution to illustrate the discrepancy between the two. The assumption that $c = 2.0$ is seen to overestimate the percentage of trips at low impedances and to underestimate the percentage at the high-impedance end. To rectify



I \ J	3	4	5
1	303	114	83
2	123	741	136

W	ΣQ_{IJ}	f
5	1044	0.70
10	237	0.16
15	219	0.14

(a) First iteration ($c = 2$)

I \ J	3	4	5
1	251	145	104
2	176	654	170

W	ΣQ_{IJ}	f
5	905	0.60
10	321	0.21
15	274	0.19

(b) Second iteration (F_{IJ} from Eq. 8.3.12)

Figure 8.3.8 Results of first two iterations.

this situation, the F-factors are adjusted to cause a shift of the calculated distribution toward the observed distribution. A commonly used formula for this adjustment is

$$F^* = F\,\frac{\text{observed}}{\text{calculated}} \tag{8.3.12}$$

where

F^* = adjusted friction factor at a given impedance

Observed = corresponding base-year percentage of trips

Calculated = current estimate of the percentage of trips at that impedance level

For example, the assumed friction factor corresponding to an impedance level of 5 min is equal to 5^{-2}, or 0.04. After the first iteration, the adjusted factor becomes

$$F_5^* = 0.04\left(\frac{0.6}{0.7}\right) = 0.034$$

The adjusted friction factors obtained in this manner are plotted against the interzonal impedances on Fig. 8.3.9. If a single value of c were desired, simple linear regression could be used

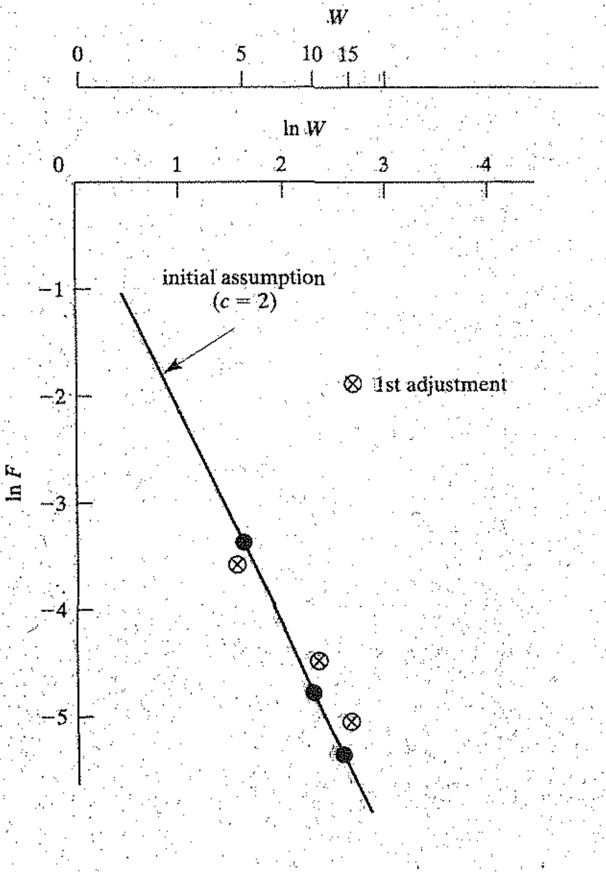


Figure 8.3.9 Friction-factor adjustment.

to fit the best-fitting straight line through this scatter diagram. However, at this point it becomes clear that the friction factor function need not be linear; it could be allowed to take the form that best describes the scatter diagram. The following general gamma function relating F and W has been suggested by the FHWA [8.7]:

$$F = aW^b e^{cW} \tag{8.3.13}$$

where e is the base of natural logarithms and a, b, and c are calibration constants. Figure 8.3.10 shows the range of shapes that this smoothing function can yield depending on the magnitude of parameter b. (*Note:* For $a = 1$ and $c = 0$ this function reduces to Eq. 8.3.9.) The adjusted friction-factor function is used in the next iteration and the calibration procedure continues until the computed distribution is sufficiently close to the observed distribution. The friction-factor function used last provides the desired calibration parameters.

Figure 8.3.8 also includes the results of the second iteration of the simple example being described. For simplicity, the new friction factors were applied directly, as computed by Eq. 8.3.12. The new trip-length frequency distribution is now closer to the observed base-year distribution.

Even though the regional trip-length frequency distribution is now close to the observed distribution, certain pronounced discrepancies remain at the interchange level. To adjust for

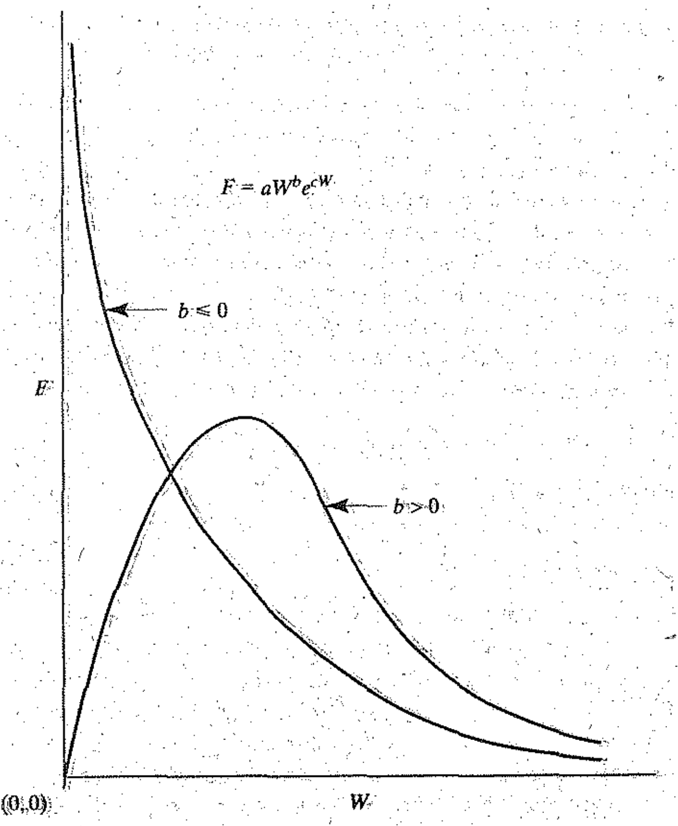


Figure 8.3.10 Shapes assumed by the F-factor.
(From Federal Highway Administration [8.7].)

these, the calibration procedure fine tunes the model by introducing a set of *zone-to-zone* socioeconomic adjustment factors

$$K_{IJ} = R_{IJ} \frac{1 - X_I}{1 - X_I R_{IJ}} \tag{8.3.14}$$

where

R_{IJ} = ratio of observed to calculated Q_{IJ}

X_I = ratio of the base-year Q_{IJ} to P_I, the total productions of zone I

The following values of K_{IJ} would result if Eq. 8.3.14 were applied at the end of the second iteration of the example problem.

I \ J	3	4	5
1	1.7	1.0	0.5
2	1.0	0.8	1.4

Application of the gravity formula with the final friction and socioeconomic factors will result in a closer fit between observed and calculated distributions.

The distinction between *attractions* and *attractiveness* merits a brief comment at this point. Some calibration procedures assume that base-year attractiveness is the same as base-year attractions. This is evident in the flowchart of Fig. 8.3.4 where productions and attractions (but not attractiveness) are specified as inputs to the calibration procedure. As explained earlier, the number of attractions of a particular zone depends on both the zone's relative attractiveness vis-à-vis all competing zones *and* the subject zone's separation from trip-producing zones. Thus two identical shopping zones (having the same attractiveness) may actually attract different volumes of trips not because of differences in their attractiveness but because one of the two zones may be at a more remote location than the other. As a result, substitution of base-year attractions for base-year attractiveness can lead to improper results.

8.3.4 Limitations of the Gravity Model

Despite its rational analogy to the law of gravitation, the gravity model of trip distribution has (as any other model of the real world) certain limitations. The major criticisms of the model usually focus on the simplistic nature of impedance (or zonal separation), its apparent lack of a behavioral basis to explain how individuals make choices among potential destinations, and its reliance on K-factors for adjustment.

In its original application the interzonal impedance was measured solely in terms of highway travel time and thus failed to capture explicitly the effect of the presence of other modes of travel, particularly transit services in transit-intensive urban areas. Developers of operational models responded to this criticism by incorporating additional variables (such as parking and toll costs) in the expression of impedance. The term "generalized cost" is often used in place of impedance to reflect this practice. Moreover, measures of *composite impedance* have been considered that incorporate travel times and costs associated with all modes providing services between pairs of zones, including transit and, more recently, nonmotorized modes (i.e., walking and bicycling). In some instances the generalized costs of the various modes would be weighted by the expected shares of trips that each attracted. Another approach entailed the use of *composite utility* (in the form of the *logsum* variable), which is computed by certain mode choice models. This concept is explained further in Section 8.4.

The second major criticism of the gravity model is the absence of any variables that reflect the characteristics of the individuals or households who decide which destinations to choose in order to satisfy their activity needs. The major response of practitioners to this drawback has been to stratify households into several relatively homogeneous groups of decision-makers and to develop separate models for each group. Household income and automobile ownership are two common segmentation criteria. A more recent practice in some metropolitan areas has been to abandon the gravity model altogether in favor of more behaviorally based *destination choice models*. These are discussed further in Section 8.4. Incidentally, it has been shown by Anas [8.9] that there is a mathematical equivalency between the gravity model and the discrete choice logit model discussed in the next section.

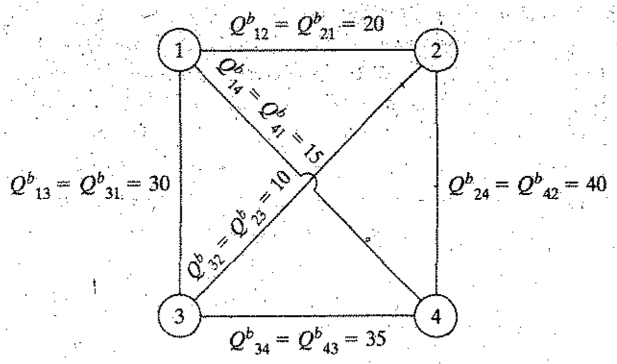
The use of K-factors to adjust for discrepancies between the observed base-year trip-length frequency distribution and that resulting from the use of the final friction factors alone has been a concern for two reasons. The first reason is related to difficulties arising from attempts to interpret the effects captured by the K-factors and the second has to do with the question of whether these effects would hold true between the base and target years. The need for the K-factors has been explained partially as capturing special conditions between some zonal pairs such as the need to cross a river. Other findings showed that K-factors were needed to rectify a mismatch between the types of jobs in which residents of producing zones were engaged and the type of employment available in the trip-attracting zones. For example, blue-collar workers in zone I could be sent by the gravity model to white-collar jobs in zone J because the latter is closer to I than a third zone containing blue-collar jobs. To minimize this difficulty, some gravity-model applications resort to stratifying jobs by industry and employment type or income at the cost of added computational burden. Experience has also shown that the causes of the problem may be rooted in historical and cultural factors that are unique to the local area. A good understanding of local conditions and their likelihood to persist over time can provide invaluable insights that can potentially aid the modeler in interpreting and applying K-factors with good judgment.

8.3.5 The Fratar Model

Several naive trend or simple growth-factor models have also been developed for use in special situations. Among these the Fratar model [8.10] is often used to estimate *external* trips, that is, trips that are either produced and/or are attracted outside the boundaries of the region under study from outlying areas whose character is not explicitly analyzed.

The Fratar model begins with the base-year trip-interchange data as illustrated in Fig. 8.3.11(a). Usually this model does not distinguish between productions and attractions and considers the interzonal trips irrespective of their direction. Consequently the values shown represent the total interchange volumes between two zones, and $Q_{IJ} = Q_{JI}$. Since no distinction is made between productions and attractions, the trip generation of each zone is denoted by Q_I; the following trip balance equation provides the necessary relationship between the trip generation of a zone I and the trip interchanges that involve zone I:

(8.3.15)


1
2
3
4
$Q^b_{12} = Q^b_{21} = 20$
$Q^b_{14} = Q^b_{41} = 15$
$Q^b_{13} = Q^b_{31} = 30$
$Q^b_{32} = Q^b_{23} = 10$
$Q^b_{24} = Q^b_{42} = 40$
$Q^b_{34} = Q^b_{43} = 35$

(a) Base-year trip-interchange data

Zone I	G_I
1	2
2	2
3	3
4	1

(b) Estimated growth factors

Figure 8.3.11 Fratar model inputs.

The estimate of the target-year trip generation $Q_I(t)$, which precedes the trip-distribution phase, is computed by multiplying the base-year trip generation, $Q_I(b)$, by a simple *growth factor, G_I*. This growth factor is based on the anticipated land-use changes that are expected to occur within the zone between the base year and the target year. Thus

$$Q_I(t) = G_I[Q_I(b)] \tag{8.3.16}$$

Subsequently the Fratar model estimates the target-year trip distribution $Q_{IJ}(t)$ that satisfies the trip balance (Eq. 8.3.15) for that year. Mathematically, the model consists of successive approximations and a test of convergence in an iterative procedure. During each iteration the target-year trip-interchange volumes are computed based on the anticipated growth of the two zones at either end of each interchange. The implied *estimated* target-year trip generation of each zone is then computed according to Eq. 8.3.15 and compared to the *expected* target-year trip generation (Eq. 8.3.16). A set of *adjustment factors, R_I*, are then computed by

$$R_I = \frac{Q_I(t)}{Q_I(\text{current})} \tag{8.3.17}$$

If the adjustment factors are all sufficiently close to unity, the trip balance constraint is satisfied and the procedure is terminated. Otherwise the adjustment factors are used along with the current estimate of trip distribution Q_{IJ} (current) to improve the approximation. A comparison of Eqs. 8.3.16 and 8.3.17 shows that the adjustment factors used in all but the first iteration and the original growth factors applied during the first iteration play the same mathematical role. Their interpretation, however, is not the same: The growth factors constitute a prediction of the actual growth of each zone between the base year and the target year, but the subsequent adjustment factors are merely mathematical adjustments that facilitate the convergence of the solution to the predicted zonal trip generation.

The basic equation employed by the Fratar model to calculate the portion of the target-year generation of zone I that will interchange with zone J is

$$Q_{IJ}(\text{new}) = \frac{[Q_{IJ}(\text{current})]R_J}{\sum_x [Q_{Ix}(\text{current})]R_x} Q_I(t) \tag{8.3.18}$$

This equation is similar to that of the gravity model. The expected trip generation of zone I is distributed among all zones so that a specific zone J receives its share according to a zone-specific term divided by the sum of these terms for all "competing" zones x. When Eq. 8.3.18 is applied to all zones, two estimated values result for each pair of zones: The first represents the portion of the generation of zone I allotted to the interchange due to the influence of zone J (or Q_{IJ}), and the second is the portion of the generation of zone J allotted to the interchange due to the influence of zone I (or Q_{JI}). As the following example shows, these two values are not necessarily equal. Since the Fratar model employs only one interzonal volume estimate $Q_{IJ} = Q_{JI}$, the two values are simply averaged; that is,

$$Q_{IJ}(\text{current}) = Q_{JI}(\text{current}) = \frac{Q_{IJ}(\text{new}) + Q_{JI}(\text{new})}{2} \tag{8.3.19}$$

and these values are used to calculate the new adjustment factors as explained previously.

An asymmetric form of the Fratar model begins with a base-year trip table in the production-attraction format. In this case the sum of each row represents the base-year productions, whereas the sum of each column represents the base-year attractions of the corresponding zone. Each zone is given two growth factors: one associated with the expected growth in residential activity (and therefore productions), whereas the second captures the zone's nonresidential growth (i.e., attractions).

A wider class of models employing procedures similar to the Fratar model known as iterative proportional fitting (IPF) are in common use for various transportation applications. In essence, IPF models begin with a *seed matrix* (the base-year trip table in the case of Fratar) and *target marginal distributions* (the target-year sums of rows and columns in the case of Fratar). An iterative procedure is then applied until the updated matrix (the target-year trip table in the case of Fratar) satisfies the marginal constraints. One common application of IPF involves the estimation of vehicle volumes interchanging between on- and off-ramps along a freeway segment based on the marginal sums of on- and off-ramp volumes and an initial seed matrix. Another common application is used to synthesize the classification of households by two or more variables (say, income and automobile ownership) when only the marginal distributions of these variables are known. It has been shown that the choice of the seed matrix has a significant effect on the resulting solution.

Example 8.5: Application of the Fratar Model

Consider the base-year trip distribution of the simple four-zone system of Fig. 8.3.11. Assuming that the growth factors for the four zones are as shown, find the target-year trip distribution.

Solution The accompanying trip table summarizes the base-year data. Note that $Q_{IJ} = Q_{JI}$, as required by the Fratar model:

J I	1	2	3	4	$Q_I(b)$	×	G_I	=	$Q_I(t)$
1	0	20	30	15	65		2		130
2	20	0	10	40	70		2		140
3	30	10	0	35	75		3		225
4	15	40	35	0	90		1		90
$Q_J(b)$	65	70	75	90					

Step 1: Use the trip balance Eq. 8.3.15 to compute the base-year trip generation for each of the four zones and multiply this total by the corresponding growth factor to calculate the target-year trip generation of each zone.

The marginal sums of each row or column of the trip table represent the base-year trip generation for the respective zones. The computation of the target-year generation using the row sums is also shown.

Step 2: For the first iteration, equate the adjustment factors to the growth factors and the current interchange flows to the base-year interchange volumes.

Step 3: Apply Eq. 8.3.18 to all pairs of zones *I, J* to get

J *I*	1	2	3	4
1	0	36	81	13
2	51	0	38	51
3	117	39	0	69
4	13	33	44	0

Step 4: Apply Eq. 8.3.19 to arrive at single volume estimates for each interchange:

J *I*	1	2	3	4	Q_I(current)
1	0	43.5	99.0	13.0	155.5
2	43.5	0	38.5	42.0	124.0
3	99.0	38.5	0	56.5	194.0
4	13.0	42.0	56.5	0	111.5
Q_I(current)	155.5	124.0	194.0	111.5	

Step 5: Apply Eq. 8.3.15 to calculate the target-year trip generation of each zone that is implied in the results of step 4. The results of this step are shown in the marginal column and row of the preceding table.

Step 6: Apply Eq. 8.3.17 to compute the adjustment factors:

$$R_1 = \frac{130}{155.5} = 0.84 \qquad R_2 = 1.13 \qquad R_3 = 1.16 \qquad R_4 = 0.81$$

These adjustment factors show that the current solution overestimates the target-year generation of zones 1 and 4 (i.e., the adjustment factors are less than unity) and underestimates it for zones 2 and 3. If a better approximation is desired, the procedure returns to step 3, using the given adjustment factors and the contents of the trip table of step 4 as the current interchange volumes.

8.3.6 Limitations of the Fratar Model

The Fratar model suffers from three major drawbacks: (1) It breaks down mathematically when a new zone (e.g., a new housing development) is built after the base year since all base-year interchange volumes involving such a zone would be equal to zero; (2) convergence to the target-year generation totals is not always possible; and (3) the model is not sensitive to the impedance W_{IJ}, which has been shown to affect significantly the interzonal distribution of trips. For these reasons the application of the Fratar model is restricted to situations where the more sophisticated models (such as the gravity model, and, more recently, the destination choice approach) cannot be used.

8.3.7 Summary

The aim of trip distribution is to estimate the target-year interchange volumes between all pairs of zones. The trip productions of each zone I obtained from the earlier trip-generation phase are distributed among the trip-attracting zones J. The trip volume that a zone J would attract depends on its relative attractiveness (i.e., the availability of nonresidential activities vis-à-vis all competing zones of attraction) and the relative impedance between the producing zone I and the subject zone J. Estimates of the interzonal impedances are obtained from the specification of the transportation alternative plan under consideration.

The most common formulation of trip distribution is the gravity model, which is conceptually based on Newton's law of gravitation. This section presented the mathematical development of the model, described how it can be calibrated, and illustrated its application. The gravity model can be calibrated separately for each of several trip purposes if the outputs of the antecedent trip-generation phase permit it. Also, it may be calibrated for total daily volumes (i.e., person-trips per day) or for smaller time periods of the day (e.g., person-trips per peak period).

A simple growth-factor model, the Fratar model, was described. Although insensitive to interzonal impedance, this model can be useful in special situations where the detailed data required by more sophisticated models are not available.

A recent trend toward more behaviorally based models, such as the destination choice model, has been noted.

8.4 MODE CHOICE

8.4.1 Background

In a typical travel situation trip-makers can select between several travel modes. These may include driving, riding with someone else, taking the bus, walking, riding a motorcycle, and so forth. A *mode choice,* or *mode split,* model is concerned with the trip-maker's behavior regarding the selection of travel mode. The reasons underlying this choice vary among individuals, trip type, and the relative level of service and cost associated with the available modes. If readers were to contemplate the reasons behind their choice of travel mode to and from school or work, they would have a tangible example of what these factors mean. Additionally, it is likely that readers have established a pattern of mode choice that remains relatively constant as long as these conditions remain the same. When significant changes in these conditions occur, trip-makers respond to varying degrees by shifting from one mode to another. For example, a significant increase in the parking fees charged at a destination may induce some people to shift from driving a car to riding a bus.

The characteristics of the trip also have an effect on the choice of mode. It seems more likely, for example, that a person would choose to travel to work or school by a mass transit system but prefer the private automobile, if available, for social trips. As discussed in relation to trip generation and trip distribution, it is not unusual for a regional transportation study to decompose trip making into trip purpose categories and to model each component separately. This practice could then be extended to the mode choice phase as well.

In addition to the attributes of the available modes and the trip type, the socioeconomic status of the trip-maker affects the choice of travel mode. Thus trip-makers may also be classified into finer categories, such as income or age, and separate estimates may be obtained for each of these socioeconomic subgroups. In many early transportation planning studies a particular subgroup (referred to as the *transit-captive* subgroup) has been singled out for special treatment. As this group's name implies, it consists of people who for various reasons do not have ready access to private transportation, and hence whose mobility is almost exclusively dependent on the public transit system. Included in this group are many of the elderly, the poor, the very young, and even the second primary individual of one-car households. Because this group is of considerable size, public transportation policy at the federal, state, and local levels has specifically addressed the needs of the members of this group.

To summarize, the mode choice behavior of trip-makers can be explained by three categories of factors: the characteristics of the available modes; the socioeconomic status of the trip-maker; and the characteristics of the trip. These are the categories of independent variables that would be included in the mathematical models of mode choice. The dependent variable would be the market share or the percent of travelers that are expected to use each of the available modes.

One of the simplest modal split models employs simple *diversion curves*, such as the one illustrated in Fig. 7.4.1. Elaborate empirical *diversion-type models* stratified by the trip-maker, mode, and trip attributes have been supplanted by probabilistic *discrete choice* models based on the principle of utility maximization.

Although simple in concept, the diversion models were awkward to calibrate and use, especially if more than two competing travel modes were included. In its full form, which involved a number of trip purposes, the Washington, DC, model consisted of 160 different curves. More computationally efficient probability-based models of modal choice have been developed, including discriminant analysis models, probit analysis models, and the most popular logit analysis models [e.g., 8.11]. These models of human choice have been applied to many situations to explain how people select between competing alternatives. Each alternative is described by a utility (or disutility) function, and the probability associated with an individual's choosing of each of the competing alternatives is expressed mathematically in terms of these utilities. Extended to groups of individuals via the theory of probability, these models estimate the proportion of the group that is likely to choose each of the competing alternatives. The development of each model involves two steps: the selection of its mathematical form and the calibration of appropriate utility functions that render the selected model capable of reproducing the available base-year data.

8.4.2 Utility and Disutility Functions

A *utility* function measures the degree of satisfaction that people derive from their choices. A *disutility* function represents the generalized cost (akin to the concept of impedance) that is associated with each choice. The magnitude of either depends on the characteristics (or *attributes*) of each choice and on the characteristics (or socioeconomic status) of the individual making that choice. In the case of modal choice the characteristics of the trip (e.g., trip purpose) also bear a relationship to the utility associated with choosing a particular mode of travel. To specify a utility function, it is necessary to select both the relevant variables from this list and the particular functional form relating the selected variables.

The utility (or disutility) function is typically expressed as the linear weighted sum of the independent variables of their transformation; that is,

$$U = a_0 + a_1X_1 + a_2X_2 + \cdots + a_rX_r \tag{8.4.1}$$

where U is the utility derived from a choice defined by the magnitudes of the attributes X that are present in that choice and weighted by the model parameters a. In the context of mode choice U is a disutility and is negative. This is because typical independent variables include travel times and costs that are perceived as losses (i.e., negatively).

Early attempts to describe the utility associated with travel modes calibrated a *separate* utility function for each mode, as illustrated by the following hypothetical three-mode case:

$$U_1 = 6.2 + 2.4X_1 + 3.5X_2 \tag{8.4.2a}$$

$$U_2 = 3.4 + 3.1X_1 + 2.9X_3 \tag{8.4.2b}$$

$$U_3 = 4.3 + 2.9X_1 + 3.2X_3 \tag{8.4.2c}$$

The three modes in this hypothetical example may be the private auto, a local bus system, an express bus system, respectively, and the independent variables (or attributes) may represent the cost, level of service, and convenience associated with a mode. This type of formulation is known as a *mode-specific* (and, in the general case, *choice-specific*) model because the same attributes are assigned different weights for different modes. It is not even necessary to include the same variables in the utility equations of different modes. This, of course, is similar to saying that some attributes are either absent or given zero weights in certain modes. Although there may be some validity in this hypothesis, it causes a problem when a new mode is introduced. In that case it would be next to impossible to estimate the utility associated with the new mode because ~~the~~ necessary base-year data required for the calibration of its utility function would be unavailable. This new product problem has haunted consumer behavior analysts as well. A way to resolve it was proposed by Lancaster [8.12] as a new approach to consumer theory, where he postulated the idea of a *choice-abstract* (or *attribute-specific*) approach. This theory is based on the hypothesis that when making choices, people perceive goods and services indirectly in terms of their attributes, each of which is weighted identically across choices. Thus trip-makers perceive two distinct modes offering the same cost, level of service, and convenience as being identical. Continuing with the three-mode example, a *mode-abstract* model of modal choice would use a single equation to measure utility; for example,

$$U = 3.1 + 2.8X_1 + 1.2X_2 + 0.9X_3 \tag{8.4.3}$$

Differences in the utilities U associated with each of the competing modes arise because of differences in the magnitudes of the attributes X of these modes. For example, one mode may be faster but costlier than another, and this fact is reflected in their calculated utilities. The attribute-specific approach has a strong conceptual foundation. However, in practical applications, it is not possible to enumerate all the relevant attributes involved in the choice of mode. The first constant term in Eq. 8.4.3 is meant to capture the effect of variables that are not explicitly included in the model. Since it is unlikely that a given set of competing modes will be identical in these excluded attributes, it is reasonable to attempt to capture these unexpressed differences by calibrating for alternative-specific constants by weighting the explicitly identified attributes equally across modes by utilizing any of the modes in the choice set as the base mode. Thus, in equation 8.4.2, instead of having $a_1 = 6.2$, $a_2 = 3.4$, and $a_3 = 4.3$, the model would likely be estimated with mode 2 as the base and the alternative-specific coefficients would be $a_1 = 2.8$, $a_2 = 0$, and $a_3 = 0.9$. The calibrated utility function in the case of the three-mode example may then become

$$U_K = a_K + 2.5X_1 + 1.5X_2 + 0.8X_3 \tag{8.4.4}$$

where U_K is the utility of mode K and a_K is the calibrated mode-specific constant for the same mode, which represents the fixed advantage or disadvantage of mode K vis-à-vis the base mode. The new product problem resurfaces but in a milder form since the selection of a mode-specific constant for a new mode is much more amenable to professional judgment vis-à-vis the mode-specific models, where none of the coefficients is known. The estimation process requires that one of the mode-specific constants be known. Otherwise a unique solution is not possible. One mode-specific constant is usually set equal to zero for the base mode. Note that if the values of the attributes X included in an attribute-specific utility expression are equal for two modes, differences in the shares of the two modes (due to excluded variables) would be captured by the relative values of the mode-specific constants. For this reason the constants are sometimes referred to as *mode-bias* coefficients.

Although the attribute-specific utility function shown in Eq. 8.4.4 is conceptually convenient, practical problems often call for a mixed form that includes both attribute-specific and choice-specific terms. The need to capture the added utility of using transit for travel oriented toward the central business district (CBD) is one example. The effect of this attribute is usually positive (i.e., adds utility or reduces disutility) for transit trips because of the limited availability and high cost of parking at the CBD destination. A binary variable taking the values of 0 for non-CBD orientation and 1 for CBD-destined trips may be included along with its coefficient in the transit utility equation.

The inclusion of variables describing the demographic and socioeconomic characteristics of trip-makers in a utility function brings to bear another consideration. Since these attributes describe the trip-maker, they are the same for all choices (e.g., modes) in the trip-maker's choice set. Thus they do not differentiate in any way between choices. For such characteristics to be sensitive to alternative choices, they must be included in the same term as a modal attribute. An example of this is the inclusion of an explanatory variable that represents the *ratio* of travel cost (a modal attribute) to the trip-maker's income level (a decision-maker's attribute).

A utility-based modal choice model estimates the market share of each mode based on the utility associated with it. In a *deterministic* model it would seem reasonable that all travelers (if they know what is good for them) will select the mode with the highest utility. However, the models being discussed are not deterministic but rather, *probabilistic.** In

*The probabilistic derivation of the logit model assumes that random errors associated with the specification of utilities are independent and identically distributed according to the Gumbel distribution with mean 0 and standard deviation σ.

other words the calculated modal utilities are related to the likelihood that a given mode will be selected or, when dealing with groups of travelers, the proportion or fraction that will select each mode of travel. The relationship between this fraction and the utilities of competing modes has been cast in various forms, the most popular of which is the logit model, which can be applied to the case of two [8.13] or multiple [8.11] competing modes. When applied to a discrete number of alternatives, these *random utility* models are called *discrete choice models* [8.14, 8.15].

8.4.3 The Multinomial Logit (MNL) Model

The multinomial logit model calculates the probability of choosing mode K if disaggregate or the proportion of travelers in the aggregate case that will select a specific mode K according to the following relationship:

$$p(K) = \frac{e^{U_K}}{\sum_x e^{U_x}} \tag{8.4.5}$$

The general form of this equation resembles the fractional term employed by the gravity model of trip distribution: A term relating to the subject mode K appears as the numerator and the summation of the similar terms corresponding to all competing modes is placed in the denominator. This specification ensures that all trips that have been estimated to occur on a specific interchange are assigned to the available modes; that is, the following trip balance equation is satisfied:

$$Q_{IJ} = \sum_x Q_{IJx} \tag{8.4.6}$$

Equation 8.4.6 would still be satisfied by writing the proportion attracted by each mode as

$$p(K) = \frac{U_K}{\sum_x U_x} \tag{8.4.7}$$

For reasons that lie beyond the scope of this book the logistic transformation of the utilities (Eq. 8.4.5) is preferred.

Done **Example 8.6 Application of the Logit Model**

A calibration study resulted in the following utility equation:

$$U_K = a_K - 0.025X_1 - 0.032X_2 - 0.015X_3 - 0.002X_4$$

where

X_1 = access plus egress time, in min

X_2 = waiting time, in min

X_3 = line-haul time, in min

X_4 = out-of-pocket cost, in cents

The trip-distribution forecast for a particular interchange was a target-year volume of $Q_{IJ} = 5000$ person-trips per day. During the target year trip-makers on this particular interchange will have a choice between the private automobile (A) and a local bus system (B). The target-year service attributes of the two competing modes have been estimated to be:

Attribute	X_1	X_2	X_3	X_4
Automobile	5	0	20	100
Local bus	10	15	40	50

Assuming that the calibrated mode-specific constants are 0.00 for the automobile mode (i.e., base mode) and −0.10 for the bus mode, apply the logit model to estimate the target-year market share of the two modes and the resulting fare-box revenue of the bus system.

Solution The utility equation yields

$$U(\text{A}) = -0.625 \qquad \text{and} \qquad U(\text{B}) = -1.530$$

According to the logit equation (Eq. 8.4.5),

$$p(\text{A}) = 0.71 \qquad \text{and} \qquad p(\text{B}) = 0.29$$

Therefore the market share of each mode is

$$Q_{IJ}(\text{A}) = (0.71)(5000) = 3550 \text{ trips/day}$$

$$Q_{IJ}(\text{B}) = (0.29)(5000) = 1450 \text{ trips/day}$$

The fare-box revenue estimate is

$$(1450 \text{ trips/day})(\$0.50/\text{trip}) = \$725 \text{ per day}$$

Discussion The terms of the utility function used in this example are negative. As negative quantities they represent cost (i.e., disutility) components. The more negative this quantity is, the less attractive the mode will be. Because of the exponential transformation of the utilities, the market shares are not directly proportional to the magnitudes of utility. Division of the numerator and denominator of Eq. 8.4.5 by $e^{U(\text{A})}$ results in the following form:

$$p(\text{B}) = \frac{e^{U^*}}{1 + e^{U^*}} \qquad \text{and} \qquad p(\text{A}) = \frac{1}{1 + e^{U^*}} \tag{8.4.8}$$

where U^* is the difference in the utilities of the two modes. This form resembles that advanced by Stopher [8.13]. In the bimodal case the logistic transformation results in a sigmoidal curve, as illustrated in Fig. 8.4.1.

Done

Example 8.7 Introduction of a New Mode

It is desired to examine the effect of introducing a rapid transit (RT) system in the city of Example 8.6. A related study has projected that the service attributes of the proposed system for the interchange under consideration will be

$$X_1(\text{RT}) = 10 \qquad X_2(\text{RT}) = 5 \qquad X_3(\text{RT}) = 30 \qquad X_4(\text{RT}) = 75$$

Based on professional experience, the mode-specific constant for the new mode is somewhere between the other two but closer to the bus system, say, −0.06. Find the market shares of the three modes that will result from implementing the rapid-transit proposal and the effect on the

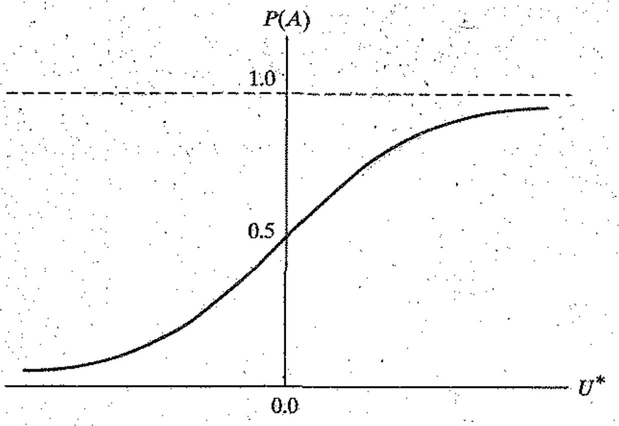


Figure 8.4.1 Binomial logit model.

revenues of the public transportation authority, which operates both the local bus and the rapid-transit systems.

Solution Assuming that the attributes of the existing modes will not be affected by the introduction of the new mode, the utilities of the three alternatives will be

$$U(\text{A}) = -0.625 \qquad U(\text{B}) = -1.530 \qquad U(\text{RT}) = -1.070$$

Proceeding as before, we obtain

$$p(\text{A}) = 0.489 \quad \text{and} \quad Q_{IJ}(\text{A}) = 2445$$
$$p(\text{B}) = 0.198 \quad \text{and} \quad Q_{IJ}(\text{B}) = 990$$
$$p(\text{RT}) = 0.313 \quad \text{and} \quad Q_{IJ}(\text{RT}) = 1565$$

The revenue will be $990 \times 0.5 + 1565 \times 0.75 = \$2582$ per day.

Discussion One of the attractive characteristics of the logit model is the fact that it could be easily extended to this situation. In this example the proposed system is seen to attract 31.3% of the interchange volume, or 1565 person-trips, reducing the auto usage by 1105 trips and the local bus patronage by 460 daily trips. In this connection it is appropriate to mention that the specification of competing modes does not have to be restricted to the generic categories illustrated in this example. Depending on special concerns, finer categories of modes or submodes may be considered for calibration. For example, a study in Honolulu, HI [8.8], which addressed the question of carpooling, has calibrated mode-specific constants for the following modes:

1. Driving alone
2. Auto with 2 occupants
3. Auto with 3+ occupants
4. Regular bus
5. Express bus

The fare-box revenue on this interchange increased from $725 to $2582 per day. This, however, is not necessarily a sufficient reason to implement the proposed system. This decision must also consider the costs of constructing and operating the system as well as other impacts.

Example 8.8: Sensitivity to Other Policies

A city council is contemplating a proposal to charge a rapid-transit fare of \$1.50 rather than \$0.75. Determine the effect of this policy on the utilization of the three modes and on the public transportation authority's revenues.

Solution The proposed policy will cause a change in the utility of the rapid-transit system to −1.220 and will affect the patronage of all modes. Proceed as before.

Mode	Proportion	Market share
Auto	0.511	2555
Local bus	0.207	1035
Rapid transit	0.282	1410

The revenue will be 850 × 0.5 + 1250 × 1.50 = \$2633 per day.

Discussion This example shows how the response of trip-makers to various public policies and combinations of policies can be examined. The fare increase induced some patrons to shift back to the auto and bus modes. The public transit share decreased, the revenues increased. In other words the extra fare revenue collected outweighed the losses that resulted from patronage losses. This is not always the case, however (see the discussion of price elasticity in Section 8.7.5). The foregoing discussion of the MNL model relied on examples related to the choice of mode. It is easy to see, however, that the model can be applied to other traveler choices. Several metropolitan areas, for example, have estimated MNL *destination choice models*. A particular modeling effort in Honolulu, HI, has estimated a destination choice model (rather than the traditional gravity model) for the tourist segment of the market [8.16]. The utility expression for competing destinations included destination attributes, degree of accessibility, and cost by various modes and traveler characteristics. A relatively simple MNL destination choice model was implemented in Portland, OR [8.3]. In that case the utility function contained a polynomial of travel time (travel time plus travel time squared) and the natural log of the attractions of each potential destination as a measure of size.

8.4.4 The Incremental (or Pivot-Point) Logit Model

Equation 8.4.5 gives the probability of choosing alternative K given the utilities of alternatives belonging in the choice set. Consider the general case where (as a consequence of a combination of policies) the utilities of one or more alternatives are changed. Let ΔUx represent the change in the utility of alternative x. Applying Eq. 8.4.5 to calculate the new shares of each alternative yields

$$P'(K) = \frac{e^{(U_k + \Delta U_k)}}{\sum_x e^{(U_x + \Delta U_x)}} = \frac{e^{U_K} e^{\Delta U_K}}{\sum_x e^{U_K} e^{\Delta U_K}} \tag{8.4.9}$$

Dividing the numerator and denominator by the denominator of Eq. 8.4.5 gives

$$P'(K) = \frac{\dfrac{\exp(U_K)}{\sum_x \exp(Ux)} \times \exp(\Delta U_K)}{\sum_x \dfrac{\exp(Ux)}{\sum_x \exp(Ux)} \times \exp(\Delta Ux)} = \frac{P(K) \times e^{\Delta U_K}}{\sum_x P(x) \times e^{\Delta U_x}} \qquad (8.4.10)$$

Thus the revised probability $P'(K)$ of choosing K due to changes in the utilities of one or more alternatives can be incrementally computed by pivoting about the baseline probabilities $P(x)$. The baseline probabilities do not need to be calculated by the MNL model; they may be obtained from surveys of existing conditions.

Example 8.9

Assume that the shares of the three modes of example 8.7 had been obtained from a base-year survey. Calculate the effect that the policy of Example 8.8 will have on the patronage of the local bus.

Solution The baseline shares are $P(A) = 0.489$, $P(B) = 0.198$, and $P(RT) = 0.313$. The only change modifies the utility of rapid transit by $\Delta U(RT) = -0.002 \times 75$ or -0.15 in accordance with the given utility expression. The incremental utilities of the other two modes equal zero.

Applying the incremental logit formula to calculate the new share of the local bus yields

$$P'(B) = \frac{0.198 \times \exp(0)}{0.489 \times \exp(0) + 0.198 \times \exp(0) + 0.313 \times \exp(-0.15)} = 0.207$$

This is identical to the answer obtained via the MNL model of Example 8.8.

Discussion The incremental logit model is also known as the *pivot-point* model. Note that if the values of any attribute change, in this case X_4, then the change in utility is strictly a function of the terms that include these attributes. Also, when computing the incremental utility, the mode-specific constant and any terms corresponding to unmodified attributes are eliminated.

8.4.5 Independence of Irrelevant Alternatives (IIA) Property

Consider a multinomial logit (MNL) model that includes alternatives A and B in the choice set. Applying Eq. 8.4.5 to compute the probabilities associated with the two alternatives and computing their *ratio*, we obtain

$$\frac{P(A)}{P(B)} = \frac{\exp(U_A)}{\exp(U_B)} = \exp(U_A - U_B) \qquad (8.4.11)$$

This equation states that the ratio of the two probabilities is a function of the difference in the utilities of the two alternatives and is not affected by the utility of any other alternative in the choice set. In other words this ratio is not influenced by any change in the utility of a third ("irrelevant") alternative. Another way of stating this condition is that any change in the utility of the third alternative will affect the shares of A and B by the same proportion. This condition is known as the *independence of irrelevant alternatives* (*IIA*) property. To illustrate the IIA property quantitatively, consider the results of Examples 8.6, 8.7, and 8.8. The ratio of $P(A)/P(B)$ in Example 8.6 is $0.71/0.29 = 2.45$. After the introduction of the rapid-transit mode (Example 8.7) there was a reduction in the shares of both the automobile (A) and the local bus (B) so that the ratio of the probabilities remained $0.489/0.198 = 2.47$. The difference is due to rounding. Moreover, increasing the cost of rapid transit (Example 8.8) had a proportional effect on the share of the other two modes with no change in their ratio; that is, $0.511/0.207 = 2.47$.

In many instances this result is counterintuitive. For example, it would be reasonable to expect that changes in the cost of rapid transit would have a proportionally greater impact on the local bus (i.e., another transit mode) than on the automobile. The often-cited example of the blue bus/red bus makes this point dramatically clear.

Suppose that two modes, the automobile and a bus service using blue buses, serve a zonal interchange. Further assume that the utilities of the two modes are identical. Given these conditions, the MNL formula will yield equal shares for the two modes. Now assume that half of the buses are painted red and considered to be a third mode having the same utility as the other two. With three equal utility modes, the MNL equation would yield equal shares, each mode attracting one-third of the market. This would be a relatively inexpensive way for bus transit companies to increase their patronage. In reality, however, the automobile would retain its 50% share and the remaining 50% would be split equally between the blue and red buses. The IIA property is the culprit as it ensures that the ratio of the probabilities of the original "modes" remains the same. A nested logit structure can reduce this problem.

8.4.6 The Nested Logit Model

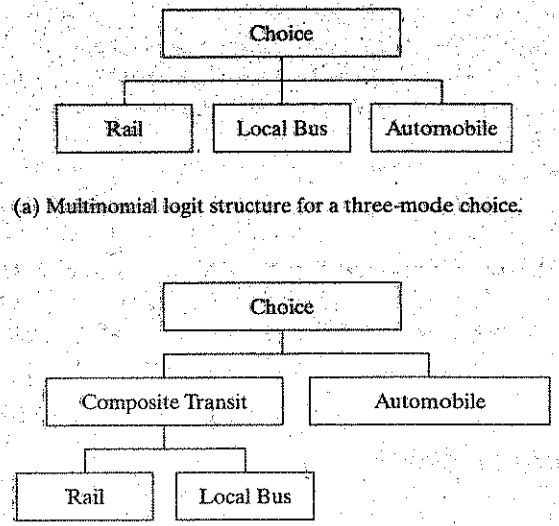
The best practice approach to rectifying (or at least minimizing) the counterintuitive implication of the IIA property of the MNL model is to employ a nested (or hierarchical) structure where similar alternatives are clustered together [(e.g., Refs. [8.14, 8.15]). Figure 8.4.2 compares the MNL and a nested structure involving three mode choices: the automobile, a local bus service, and rail transit. The MNL places these modes on a single level resulting in the usually undesirable IIA condition. By contrast, the nested structure groups the bus and rail together as subchoices of the *composite* transit mode. This structure permits a change in the utility of one of the transit modes (say, the local bus) to affect the share of the other transit mode (i.e., rail) to a greater degree than a mode (in this case the automobile) that does not belong to the transit nest. In other words a greater degree of choice substitution is allowed *within* nests than *between* nests.

Examining the top-level decision of whether to travel by the automobile (A) or transit (T) gives

$$P(A) = \frac{\exp(U_A)}{\exp(U_T) + \exp(U_A)} \tag{8.4.12a}$$

$$P(T) = \frac{\exp(U_T)}{\exp(U_T) + \exp(U_A)} \tag{8.4.12b}$$

where the *composite transit utility* $U_T = f(U_B, U_R)$.



(a) Multinomial logit structure for a three-mode choice.

(b) Nested Logit Structure

Figure 8.4.2 Comparison of MNL and nested logit models.

By moving to the lower transit level, the *conditional probabilities** of choosing the bus (B) or the rail (R), given the decision to travel by transit, become

$$P(B|T) = \frac{\exp(U_B)}{\exp(U_B) + \exp(U_R)} \tag{8.4.13a}$$

$$P(R|T) = \frac{\exp(U_R)}{\exp(U_B) + \exp(U_R)} \tag{8.4.13b}$$

To calculate the unconditional probabilities of choosing bus or rail, use the following equations:

$$P(B) = P(B|T) \times P(T) \tag{8.4.14a}$$

$$P(R) = P(R|T) \times P(T) \tag{8.4.14b}$$

The utility of the composite transit mode needs to capture the characteristics of all transit submodes (i.e., bus and rail). This is normally accomplished by including in the transit utility expression the *Logsum* variable (defined as the natural log of the denominator of Eq. 8.4.13) multiplied by its calibration coefficient:

$$Logsum = \ln\{\exp(U_B) + \exp(U_R)\} \tag{8.4.15}$$

The transit utility expression takes the form

$$U_T = a_T + \cdots + a_n \times X_n + \theta \times Logsum \tag{8.4.16}$$

The numerical value of the *Logsum* coefficient resulting from estimating the model provides information about the appropriateness of the selected nesting structure.

*Conditional probabilities are discussed in Chapter 13.

If the estimated value of θ is 0, the transit utility of Eq. 8.4.16 is independent of the utilities of the submodes. Consequently the primary choice between transit and auto is not affected by changes in the utilities of the submodes. Any such change redistributes the market shares of the submodes solely between them. In this case the submodes are said to be *perfect substitutes* of each other.

If the estimated value of θ turns out to be greater than 0 but less than 1, the selected structure is acceptable. A value of θ exactly equal to 1 implies that there exists an equivalent MNL model that is equally appropriate (i.e., the IIA property holds), whereas a value greater than 1 indicates that the selected nesting structure is inappropriate and other structures need to be investigated. It is fairly easy to show mathematically that when θ equals 1, the equivalent MNL equation is obtained by modifying the utilities of the subchoices as follows:

$$U^{\text{MNL}}_{\text{subchoice}} = a_T + \cdots + a_n \times X_n + U^{\text{nested}}_{\text{subchoice}} \qquad (8.4.17)$$

The following three examples illustrate the effect of θ.

Example 8.10

An estimation procedure for a mode choice model of the nested logit structure shown in Fig. 8.4.2 found that $U_T = a_T + \theta \times \textit{Logsum}$ with $a_T = -0.52$ and $\theta = 0$. For a particular zonal interchange the following modal utilities were calculated in accordance with the estimated nested logit model:

$$U_A = -0.26 \qquad U_B = -0.92 \qquad U_R = -0.82$$

Calculate

(a) the corresponding mode shares

(b) the effect of a policy that is expected to cause a change $\Delta U_B = -0.20$.

Solution

Part a: Baseline conditions

Nest level

Mode *m*	U_m	$\exp(U_m)$	$P(m\mid T)$
B	−0.92	0.399	0.476
R	−0.82	0.440	0.524
		Σ = 0.839	1.000

$$U_T = -0.52 + 0 \times \ln(0.839) = -0.52$$

Primary choice level

Mode *m*	U_m	$\exp(U_m)$	$P(m)$
A	−0.26	0.771	0.564
T	−0.52	0.595	0.436
		Σ = 1.366	1.000

By Eqs. 8.4.14

$$P(B) = P(B|T) \times P(T) = \underline{0.476 \times 0.436} = 0.208$$
$$P(R) = P(R|T) \times P(T) = \underline{0.524 \times 0.436} = 0.228$$
whereas $P(A) = \underline{0.564}$
$\Sigma = 1.000$

Part b: After change

Nest level

Mode m	U_m	$\exp(U_m)$	$P(m\|T)$
B	-1.12	0.326	0.426
R	-0.82	0.440	0.574
		$\Sigma = 0.766$	1.000

$$U_T = -0.52 + 0 \times \ln(0.0.766) = -0.52$$

Primary choice level

Same as above:

$$P(A) = 0.564 \quad \text{and} \quad P(T) = 0.436$$

By Eqs. 8.4.14 imp

$$P(B) = P(B|T) \times P(T) = 0.476 \times 0.426 = 0.186$$
$$P(R) = P(R|T) \times P(T) = 0.524 \times 0.574 = 0.250$$
whereas $P(A) = \underline{0.564}$
$\Sigma = 1.000$

Discussion Since $\theta = 0$, the two transit submodes are perfect substitutes of each other. Consequently the share lost by the local bus was entirely gained by the rail transit, whereas the auto share remained the same.

Example 8.11

An estimation procedure similar to that of Example 8.10 in another metropolitan area found that $U_T = a_T + \theta \times Logsum$ with $a_T = -0.42$ and $\theta = 1.0$. For a particular zonal interchange the following modal utilities were calculated in accordance with the estimated nested logit model:

$$U_A = -0.36 \qquad U_B = -0.88 \qquad U_R = -0.78$$

Calculate

(a) the corresponding mode shares

(b) the effect of a policy that is expected to cause a change $\Delta U_R = -0.10$.

Solution

Part a: Baseline conditions

Nest level

Mode m	U_m	$\exp(U_m)$	$P(m\vert T)$
B	-0.88	0.415	0.475
R	-0.78	0.458	0.525
		$\Sigma = 0.873$	1.000

$$U_T = -0.52 + 1.0 \times \ln(0.839) = -0.56$$

Primary choice level

Mode m	U_m	$\exp(U_m)$	$P(m)$
A	-0.36	0.698	0.550
T	-0.56	0.571	0.450
		$\Sigma = 1.366$	1.000

By Eqs. 8.4.14

$$P(B) = P(B|T) \times P(T) = \underline{0.475 \times 0.450} = 0.214$$

$$P(R) = P(R|T) \times P(T) = \underline{0.525 \times 0.450} = 0.236$$

whereas $P(A) = \underline{\cancel{0.450}}$ 0.556

$$\Sigma = 1.000$$

Part b: After change

Nest level

Mode m	U_m	$\exp(U_m)$	$P(m\vert T)$
B	-0.88	0.415	0.500
R	-0.88	0.415	0.500
		$\Sigma = 0.830$	1.000

$$U_T = -0.42 + 1.0 \times \ln(0.830) = -0.61$$

Primary choice level

Mode m	U_m	$\exp(U_m)$	$P(m)$
A	-0.36	0.698	0.462
T	-0.61	0.543	0.438
		$\Sigma = 1.241$	1.000

imp

By Eqs. 8.4.14

$$P(B) = P(B|T) \times P(T) = 0.438 \times 0.500 = 0.219$$

$$P(R) = P(R|T) \times P(T) = 0.438 \times 0.500 = 0.219$$

whereas

$$P(A) = \underline{0.562}$$

$$\Sigma = 1.000$$

Discussion The auto to bus share ratio before the change in the utility of the third mode (rail) was 0.550/0.214 = 2.57 and after the change 0.562/0.219 = 2.57. In other words, when $\theta = 1.0$, the IIA property holds true and an equivalent MNL model exists. According to Eq. 8.4.17, the corresponding MNL utilities become

$$U_A = -0.36 \qquad U_B = a_T - 0.88 = -1.30 \qquad U_R = a_T - 0.78 = -1.20$$

Applying the MNL to the baseline conditions using these modified utilities yields

Mode *m*	U_m	exp(U_m)	*P*(*m*)
A	−0.36	0.698	0.550
B	−1.30	0.273	0.214
R	−1.20	0.301	0.236
		Σ = 1.272	1.000

This result is identical to that obtained via the equivalent nested model.

Example 8.12

A third city undertook a similar nested logit model estimation study and found that $U_T = a_T + \theta \times Logsum$ with $a_T = -0.41$ and $\theta = 0.2$. For a particular zonal interchange the following modal utilities were calculated in accordance with the estimated nested logit model:

$$U_A = -0.41 \qquad U_B = -1.05 \qquad U_R = -0.95$$

Calculate

(a) the corresponding mode shares

(b) the effect of a policy that is expected to cause a change $\Delta U_B = -0.30$.

Solution

Part a: Baseline conditions

Nest level

Mode *m*	U_m	exp(U_m)	*P*(*m*|*T*)
B	−1.05	0.350	0.475
R	−0.95	0.387	0.525
		Σ = 0.737	1.000

$$U_T = -0.41 + 0.2 \times \ln(0.737) = -0.47$$

Primary choice level

Mode m	U_m	$\exp(U_m)$	$P(m)$
A	−0.41	0.644	0.515
T	−0.47	0.625	0.485
		Σ = 1.289	1.000

By Eqs. 8.4.14

$$P(B) = P(B|T) \times P(T) = 0.485 \times 0.475 = 0.230$$

$$P(R) = P(R|T) \times P(T) = 0.485 \times 0.425 = 0.255$$

whereas

$$P(A) = \underline{0.515}$$

$$\Sigma = 1.000$$

Part b: After change

Nest level

Mode m	U_m	$\exp(U_m)$	$P(m\|T)$
B	−1.35	0.259	0.401
R	−0.95	0.387	0.599
		Σ = 0.646	1.000

$$U_T = -0.41 + 0.2 \times \ln(0.646) = -0.50$$

Primary choice level

Mode m	U_m	$\exp(U_m)$	$P(m)$
A	−0.41	0.664	0.522
T	−0.61	0.608	0.478
		Σ = 1.272	1.000

By Eqs. 8.4.14

$$P(B) = P(B|T) \times P(T) = 0.478 \times 0.401 = 0.192$$

$$P(R) = P(R|T) \times P(T) = 0.478 \times 0.599 = 0.286$$

whereas

$$P(A) = \underline{0.522}$$

$$\Sigma = 1.000$$

Discussion In this case the before and after share ratios of auto and rail are, respectively, 0.515/0.255 = 0.202 and 0.522/0.286 = 1.83. The change in the utility of the third (bus) alternative did not affect the other two modes proportionately. With a value of θ between 0 and 1, the IIA property of the MNL model does not apply to the nested structure.

It is also worth noting that in all of the three preceding examples the relative shares of bus and rail under the baseline conditions were identical (except for rounding) even though the utilities of the two modes varied among the examples. This result is due to the fact that the differences between the modal utilities happened to be identical and Eqs. 8.4.8 and 8.4.11 hold true. This leads to the conclusion that adding a *scaling constant* (not to be confused with the *choice-specific constant*) to all utilities preserves the integrity of the model. In fact, some analysts make it a practice to add a positive scaling constant to disutilities to express them as positive utilities.

A relatively simple nesting structure was postulated in the discussion of nested logit models so far in order to convey the basic properties of the structure. More complex arrangements are possible as illustrated in Fig. 8.4.3, which has been (with variations) estimated in a number of metropolitan areas (e.g., Ref [8.16]). The model uses five MNL clusters in a nested arrangement. The primary choice is a binary logit between the composite auto and the composite transit modes. The auto mode is composed of the subchoices of driving alone or carpooling by 2 or 3+ trip-makers. The choice between access modes (walk or drive) is included on the transit side. The last two nests capture the choice between local (bus) and premium transit service and the choice between drive access to the premium transit mode (park-and-ride versus being dropped off at the station). This particular structure does not provide for drive access to the local transit service.

Figure 8.4.4 illustrates an alternate structure that allows for drop-offs at both local and premium transit stations but does not provide for park-and-ride access to either of the two. In addition to such relatively sophisticated mode choice formulations, nesting schemes can be specified and estimated for a wider combination of choices. Figure 8.4.5 illustrates a structure that incorporates trip generation (i.e., to make a trip or not), the time of day for trips, the choice of destination, and the choice of mode. Although conceptually specified in the early 1970s [8.15], nested structures of this level of complexity were not attempted until much later.

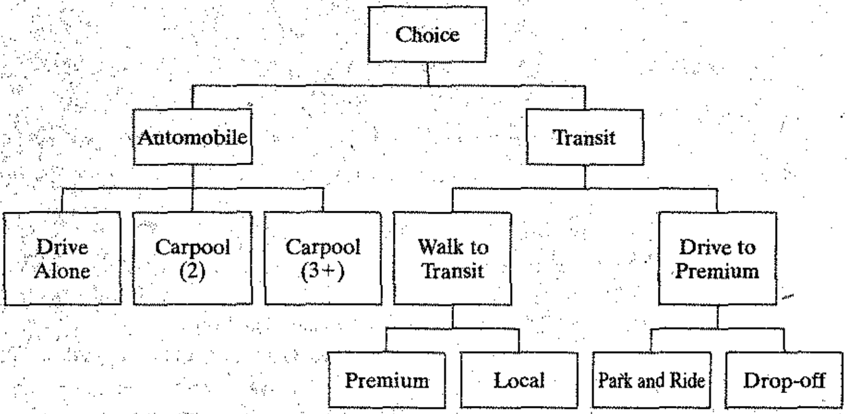


Figure 8.4.3 A three-level nested logit mode choice model.

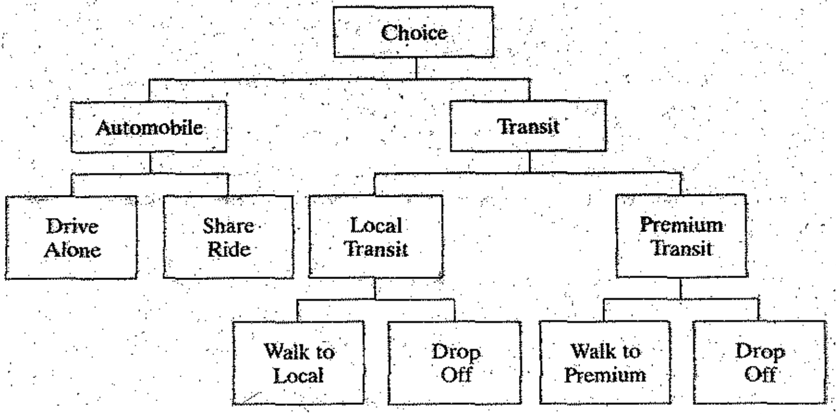


Figure 8.4.4 Alternate three-level nested logit mode choice structure.

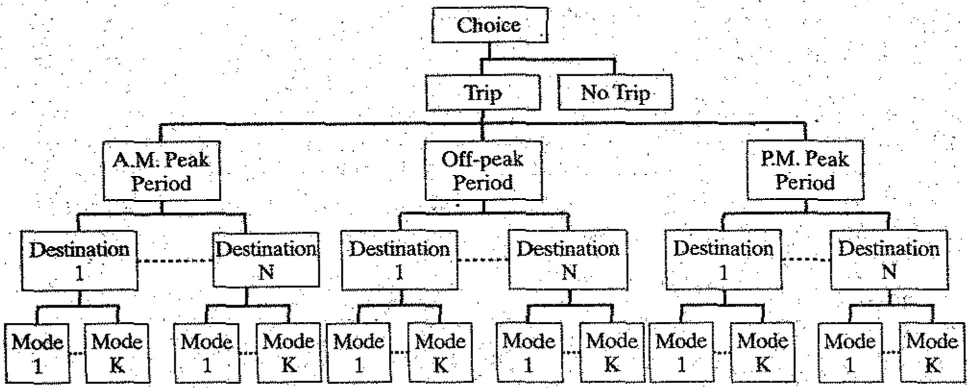


Figure 8.4.5 A combined trip-generation, time-of-day, destination, and mode choice nested logit structure.

8.4.7 Estimation of Logit Models

Estimation of a logit model entails the selection of attributes, attribute coefficients, and mode-specific constants that maximize the probability of replicating the observed mode choices of individuals (or other decision-making units such as households) as revealed in a base-year sample drawn from the population under investigation. The most common technique used is known as the *maximum likelihood (ML) method.* Another is based on the principle of *entropy maximization.*

The ML technique expresses the likelihood that the probability (or likelihood L) that the model is capable of replicating the observed conditions as the product of the probabilities (according to the postulated model) that each member of the sample would make the observed choice:

$$L = \prod_{i=1}^{n} P_i(K_i) \tag{8.4.18}$$

The estimation procedure seeks the combination of model parameters that maximize this function or equivalently the natural logarithm of L, known as the *log likelihood* and designated as *log L*.

The appropriateness of the estimated model is judged on the basis of reasonableness (e.g., correct relative values and signs of attribute coefficients) and by formal statistical tests. The *t*-test (see Chapter 13), for example, is used to determine whether each coefficient is statistically different from zero. Most estimation software report additional statistics. These include the values of the likelihood function associated with several potential models, such as:

1. The estimated model that normally contains a subset of all the available variables
2. A model that includes all available variables ("best" model)
3. A model that expresses utilities in terms of choice-specific constants only (i.e., has no explanatory variables)
4. A no-information model that assumes equal mode probabilities with its likelihood designated as $L(0)$

A goodness-of-fit measure known as the likelihood ratio index, ρ^2, which ranges from 0 to 1, is used to compare the likelihood L associated with the postulated model to $L(0)$,

$$\rho^2 = 1 - \{L/L(0)\} \qquad (8.4.19)$$

A value of 0, implying $L = L(0)$, means that the estimated model does not have any explanatory power beyond the "no-information" model. The closer the index comes to 1, the better the model is. An adjusted form of ρ^2 is commonly reported as well. Another test statistic, known as the likelihood ratio test statistic *LR*, can be used to ascertain whether the exclusion of certain variables (vis-à-vis the "best" model) is appropriate. *LR* is distributed according to the χ^2 (chi-square) distribution with degrees of freedom equal to the sample size.* It is given by

$$LR = -2 \times (\log L_{\text{best}} - \log L_{\text{model}}) \qquad (8.4.20)$$

Logit model estimation may be done at the individual (disaggregate) level or at higher levels of aggregation. Models estimated with disaggregate data can be used to obtain aggregate predictions through the use of the market segmentation and sample enumeration techniques discussed in Section 8.2.3.

In the case of nested logit models the estimation can proceed sequentially from the lowest nesting level to the highest. Alternately, it can be done in a single step in what is known as estimation using *full information*. Specialized software packages are available for this purpose.

8.4.8 Summary

The purpose of a mode choice model is to predict the trip-maker's choice of travel mode. The factors that explain this behavior include:

1. The characteristics of the trip-maker
2. The characteristics of the trip
3. The attributes of the available modes of travel

*Refer to Chapter 13 for a discussion of hypothesis testing.

Mode split models may be aggregate or disaggregate, depending on the level at which they are calibrated.

Purely empirical diversion-curve methods have been replaced by the probability-based multinomial logit and nested logit formulations that were presented in this section. The concepts of utility and disutility employed by the latter were explained. The difference between choice-specific and attribute-specific models of consumer behavior was drawn and illustrated.

8.5 TRIP ASSIGNMENT

8.5.1 Background

The last phase of the four-step transportation-forecasting process is concerned with the trip-maker's choice of path between pairs of zones by travel mode and with the resulting vehicular flows on the multimodal transportation network. This step may be viewed as the equilibration model between the demand for travel (Q_{IJK}) estimated earlier in the process and the supply of transportation in terms of the physical facilities and, in the case of the various possible mass transit modes, the frequency of service provided. Incidentally, this conceptual framework of economic theory is applicable to earlier steps of the process as well and has been so treated by many authors. Examples 8.6 through 8.12 illustrate how people respond to changes in the availability and price of transportation services. If the price of one mode increases relative to another, its market share will decrease.

Returning to the topic of network assignment, the question of interest is, given Q_{IJK}, that is, the estimate of interzonal demand by mode, determine the trip-maker's likely choice of paths between all zones I and J along the network of each mode K and predict the resulting flows q on the individual links that make up the network of that mode (Fig. 8.5.1). The estimates of link utilization can be used to assess the likely level of service and to anticipate potential capacity problems.

The number of available paths between any pair of zones depends on the mode of travel. In the case of private transportation modes a driver has a relatively large set of possible paths and path variations and also a good deal of freedom in selecting between them.

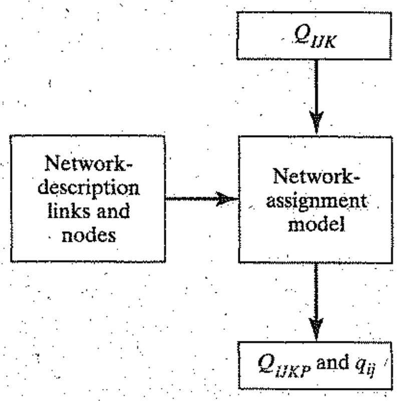


Figure 8.5.1 Trip assignment inputs and outputs.

On the other hand, typical mass transit modes offer a limited number of path (or route) choices.

Three preliminary questions must be dealt with prior to the performance of network assignment. The first is related to the difference between interzonal person-trips and interzonal vehicle-trips, the second is related to the difference between daily trips (i.e., the estimate of the 24-h demand) versus the diurnal distribution of this demand, and the third is concerned with the direction of travel of the trips to be assigned on the transportation network.

8.5.2 Person-Trips and Vehicle-Trips

The forecasts of the person-trip and vehicle-trip flows that are expected to use the transportation system are both relevant to the assessment of its performance. The estimate of person-trips that desire to use a highway, for example, provides an indication of the passenger throughput that will be accommodated. On the other hand, the level of service (see Chapter 4) that the trip-makers experience when traveling on a highway is related to the *vehicular* flow (e.g., vehicles per hour) that desires to use the highway. For this reason the estimated interzonal person-trips must be translated into vehicular-trips prior to performing the highway trip assignment (also known as *traffic assignment*). *Car occupancies* (i.e., persons per car) vary between cities and also between trip types. Reference 8.1 provides summaries of average daily car-occupancy rates by trip purpose and urban area size and presents default adjustment factors by time of day and trip purpose. These rates and adjustment factors were compiled from a 1990 nationwide survey [8.17].

Advanced mode choice models in use by some transportation agencies predict trip tables by car-occupancy level directly in a form that can be fairly easily converted to vehicle-trips. The family of mode choice models exemplified by Fig. 8.4.3 illustrates this possibility. The utility functions of the subchoice automobile modes (i.e., driving alone and carpooling) typically include a wide range of explanatory variables. These include trip purpose and orientation (CBD or other), parking availability and cost at the destination end, the availability of faster high-occupancy vehicle (HOV) facilities, and traveler attributes such as income, age, and so on. Thus they have a stronger behavioral basis than simple rate models calibrated for the conditions prevailing during a base year. In other words they are sensitive to changes in the factors, which motivate decisions relating to driving alone versus carpooling.

Mass transit (or *transit assignment*) must address another issue as well. In this case the specification of an alternative system consists not only of the fixed facilities that constitute the modal network, but also the scheduling of transit services. This means that the analysis of a particular transit alternative must address the question of whether a proposed fleet size and operating schedule and the related vehicular frequencies (i.e., flows) provide sufficient capacity to meet the anticipated interzonal person-trip demand.

8.5.3 Diurnal (Time-of-Day) Patterns of Demand

The highway flows and intersection approach volumes that are used to calculate the prevailing level of service are expressed in vehicles per hour (Chapter 4). On the other hand, the estimates of interzonal flows that are obtained by the trip-generation distribution-mode choice sequence are often based on a 24-h period. As Fig. 8.2.2 illustrates, the demand for transportation exhibits a highly peaked pattern with a sharp peak period in the morning and a generally longer but less pronounced peak period in the evening. It is appropriate, therefore, to investigate the performance of the transportation system under peak-demand conditions when capacity limitations become most critical. The time variation of demand is most relevant to mass transit planning because the scheduling of service is typically tailored to the variation of demand over the 24-h period.

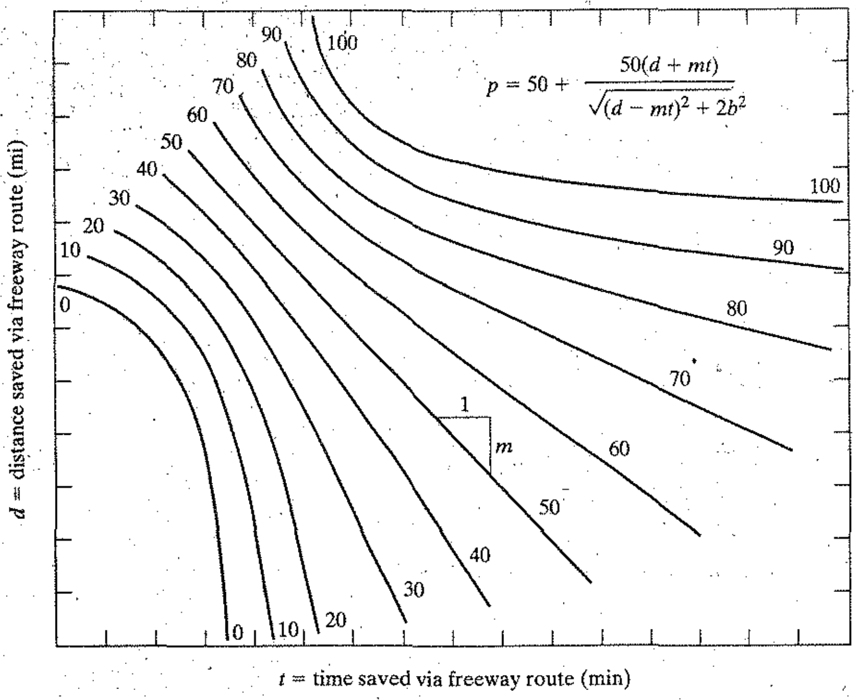
The diurnal distribution of demand may be estimated through the use of factors taken from observations during the base year, or it may be explicitly modeled in the preceding steps of the demand-forecasting process (see e.g., Fig. 8.4.3). Typically the morning peak-period demand is in the range of 10 to 20% of the total daily demand. Standard practice (see Fig. 8.1.1) entails the performance of three separate assignments by the time of day, the morning (A.M.) peak period, the afternoon (P.M.) peak period, and the off-peak balance of the day.

8.5.4 Trip Direction

In the discussion of trip generation a distinction was drawn between productions and attractions on one hand and origins and destinations on the other. It was also explained why most trip-generation models estimate productions rather than origins. However, it is desirable that the assignment of trips (especially by the time of day) retains the direction of these trips. The predominant direction of travel during the morning peak period is toward major activity centers (i.e, CBDs or schools), and the reverse is true during the evening peak period. The experience and knowledge accumulated through studies of the travel patterns within the region aid in the accomplishment of this task. Directionality factors by time of day and trip purpose are typically used to convert production-attraction tables to origin-destination (*O-D*) tables.

8.5.5 Historical Context

The origin of traffic assignment can be traced to the 1950s and early 1960s, when the majority of urban freeways were constructed in U.S. cities. Typically highway engineers wanted to know how many drivers would be diverted from arterial streets to a proposed freeway in order to make decisions related to the geometric design and capacity of proposed urban freeways. The *diversion-curve model* was developed to answer this question. This model employs empirically derived curves to compute the percentage of trips that would use the freeway route between two points on some measure of relative impedance between the freeway route and the fastest arterial route between the two points. Figure 8.5.2 shows that the California diversion curves [8.18] used travel-time and travel-distance differences between the two alternative paths to estimate the percentage of trips that would use the freeway. Figure 8.5.3 illustrates a diversion curve developed by the BPR [8.19], where the ratio of travel times via the two routes serves as the impedance measure. It is interesting to note that when the travel times are equal, less than half of the travelers tend to use the freeway. This may represent unfamiliarity with freeway conditions when the curve was calibrated.



Where p = percent usage,
d = distance saved in miles,
t = time saved in minutes,
m = a coefficient relating the value of a mile saved to a minute lost; in other words, a scale value for the x ordinate for a given scale on the y ordinate,
and b = a coefficient determining how far the vertices of the 100 percent and 0 percent boundaries are from the origin.

Figure 8.5.2 California diversion curves. (From Moskowitz [8.18].)

A shortcoming of the diversion-curve method is the fact that drivers between two points have path options that contain both freeway segments and arterial street segments rather than two distinct all-freeway and all-arterial paths. Moreover, these combinations become computationally complex as the number of zonal pairs and the size of the transportation network increase. Developments in computer technology have made it possible to expand the traffic assignment procedure to large networks. A network assignment procedure requires:

1. A way of coding the modal network for computer processing
2. An understanding of the factors affecting the trip-maker's path preferences
3. A computer algorithm that is capable of producing the trip-maker's preferred paths

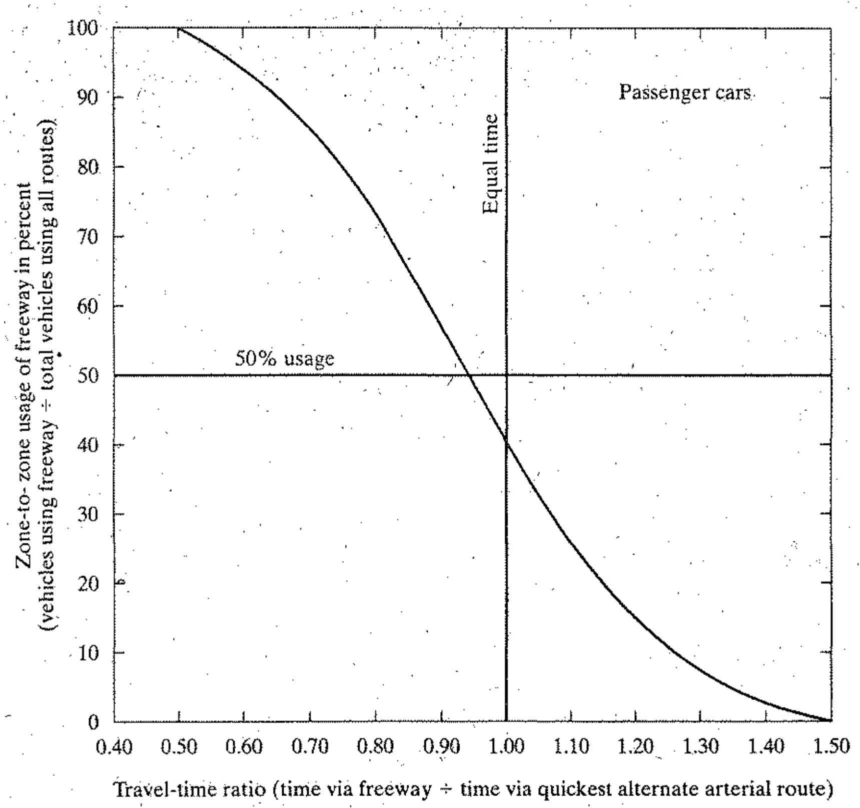


Figure 8.5.3 Bureau of Public Roads diversion curve. (From Bureau of Public Roads [8.19].)

8.5.6 Highway Network Description

Fixed facilities (i.e., major arterials, expressways, and freeways) are specified by a set of *nodes* (i.e., intersections and interchanges) and *links;* usually local and minor streets are not included in the coded highway network. Each node is specified by a numerical code and each link is described by its end nodes. Important characteristics of each link (such as its capacity, free-flow speed, or travel time) are also specified. It is often advantageous to select the coding scheme judiciously to reflect other link attributes as well. For example, nodes that lie exclusively on arterial streets may be denoted by one range of numerical codes (say, between 100 and 1000), whereas nodes that lie on higher-type facilities may be coded with numbers in another range (say, greater than 1000). Thus a link connecting nodes 525 and 666 is clearly a segment of an arterial street, whereas link 1212–1213 is a segment of freeway. Moreover, links 729–1432 and 1198–888 represent an on-ramp (i.e., connecting an arterial to a freeway) and an off-ramp (i.e., connecting a freeway to an arterial).

The advent of geographic information system (GIS) technology (see Section 15.2) has greatly enhanced the efficiency and accuracy of network specification and coding. Figure 8.5.4 shows a portion of the highway assignment network in Honolulu, HI, at the stage of "cleaning." Network facilities can be displayed in a variety of ways, including the

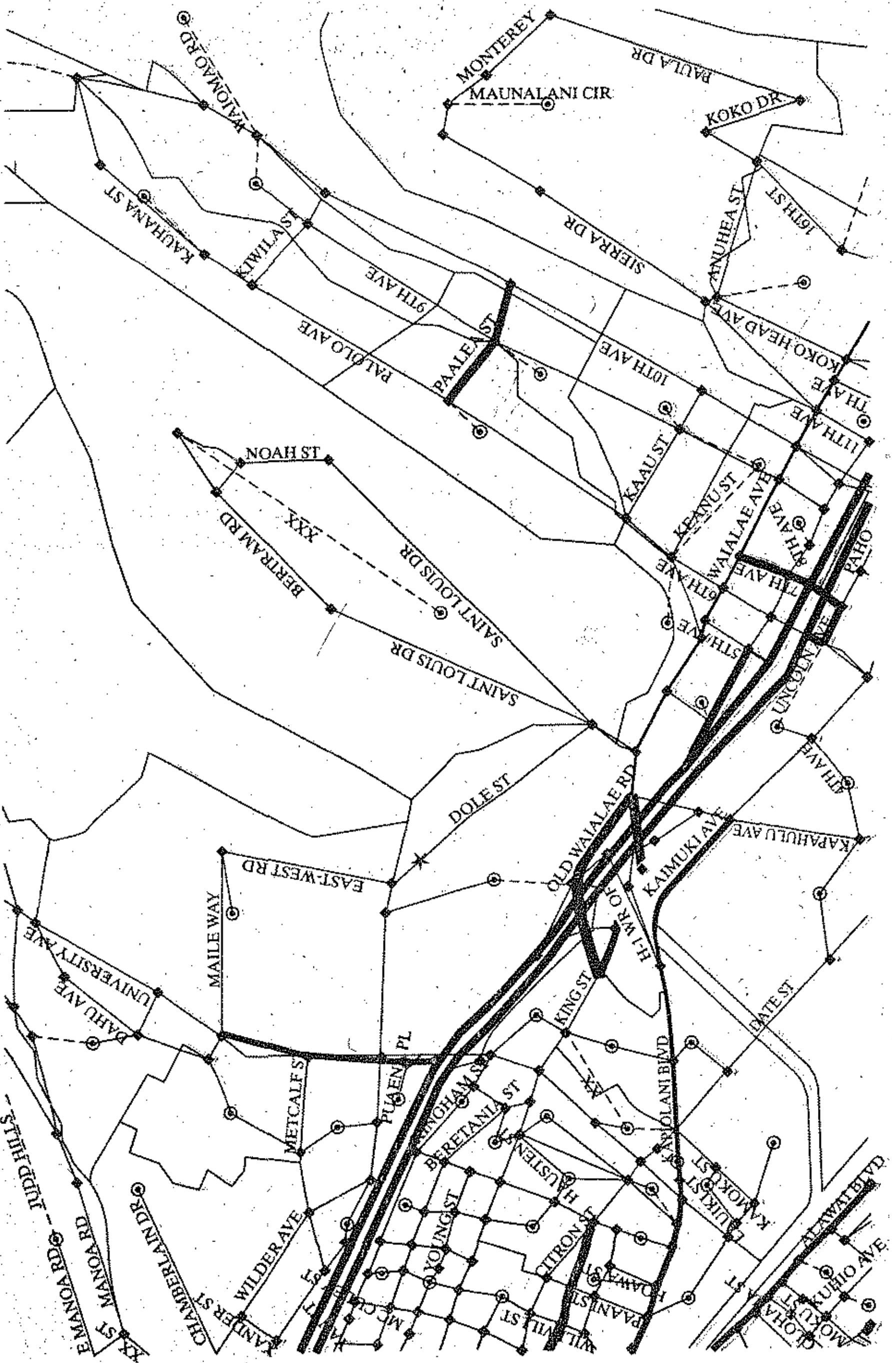


Figure 8.5.4 Part of Honolulu's coded network.
(From Oahu Metropolitan Planning Organization)

one shown where lines with distinct color (not shown) and weight combinations can clearly differentiate between facility type. At this stage it is graphically clear that some roadways have been misspecified. For example, Paalea Street at the right and middle part of the figure appears to have been coded as a freeway segment when in fact it is a collector road. Also, the symbols XX and XXX indicate that an attempt was made to assign street names to the corresponding centroidal connectors, and in the middle of the figure Saint Louis Drive appears to have been associated with two distinct streets. These and other coding errors, of course, were subsequently corrected. Honolulu's 1998 highway assignment network consisted of approximately 3700 links.

In Fig. 8.5.4, the irregular lines that have no distinct nodes at vertices represent zone boundaries. As seen in the figure, zone boundaries often coincide with roadway segments. Travel-analysis zones are coded as a set of imaginary nodes that are referred to as *zonal centroids*. To distinguish them from the actual network nodes, they are usually designated by numerical codes at the lower range of positive integers. Their geographical location is often taken to coincide with the activity or population centroid of the zones they represent, hence their name. Finally, a set of imaginary links known as *centroidal connectors* are introduced to connect the zonal centroids to the assignment network. Although not real links, they are typically given link attributes corresponding to the average conditions that trip-makers

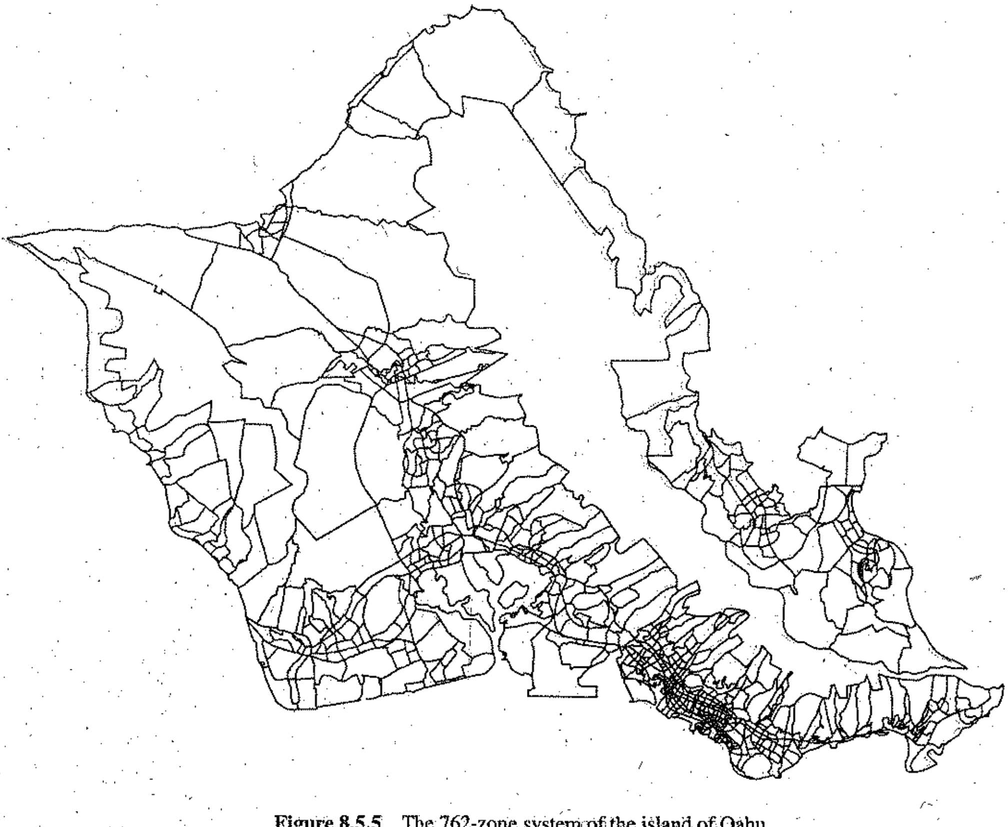


Figure 8.5.5 The 762-zone system of the island of Oahu. (From Oahu Metropolitan Planning Organization)

experience on the noncoded local and minor street system. Figure 8.5.5 shows the travel-analysis zone system (c. 1998) developed for the island of Oahu, HI where the city of Honolulu is located. The 600-square-mile island was partitioned into 762 zones, excluding two large areas of inaccessible mountainous terrain. The zone size ranges from as small as a city block within the CBD and "primary urban corridor" to much larger sizes in outlying areas. Population, employment densities, land use, natural features, and relation to the transportation network serve as guides to zone delineation.

It should be clear to the reader that specifying, coding, and cleaning the zone system and the transportation network is a very tedious task. Well-defined zones and networks are critical. The simple network of Fig. 8.5.6 consists of five zonal centroids (i.e., nodes 1 to 5), six centroidal connectors, nine street intersections (i.e., nodes 6 to 14), and 13 arterial street links. Although not followed in this simple network, proper practice is to connect centroidal

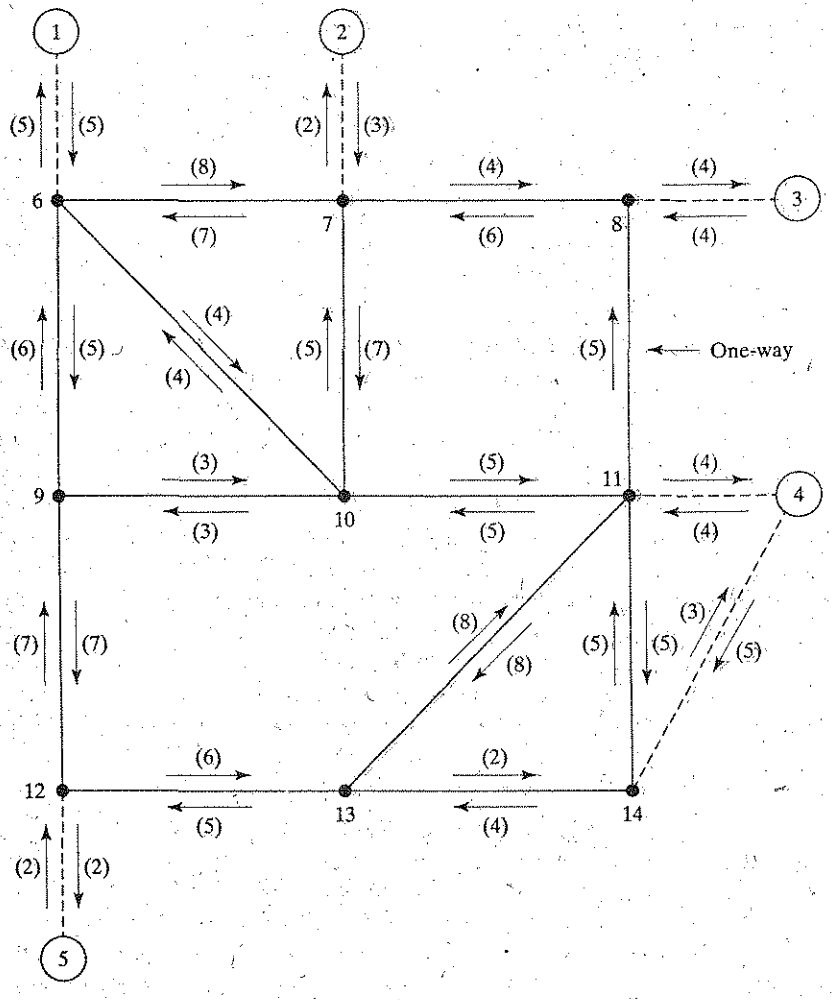


Figure 8.5.6 Hypothetical network.

TABLE 8.5.1 Link Array

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1						5								
2							3							
3								4						
4											4			5
5												3		
6	5						8		5	4				
7		2				7		4		7				
8			4				6							
9						6				3		7		
10						4	5		3		5			
11				4				5		5			8	5
12					2				7				6	
13											8	5		2
14				3							5		4	

links to "dummy" nodes along network links rather than directly to intersections. The reason for this practice is to ensure that traffic flows on the centroidal connectors do not unrealistically load the intersections from nonexisting approaches and thus adversely affect subsequent level-of-service calculations. By allowing these flows to move in their natural direction (i.e., from noncoded minor streets to coded links and then into major intersections) subsequent level-of-service analyses would represent actual conditions more realistically. The numerical values in parentheses correspond to the link impedances in the direction shown. This network is described by the *link array* (Table 8.5.1), each cell of which represents a possible direct link between the row and column nodes. A numerical entry in a cell means that there is in fact such a link, the cell value being, say, the link's impedance. The dimensions of the link array may be increased to include other link attributes as well, such as free-flow speed, length, and capacity. An alternate way to describe the network is by the use of a relational database that can incorporate a large number of link attributes.

8.5.7 Link Flows and Interzonal Flows

A careful distinction must be made between the terms *interzonal flows* (Q_{IJ}) and *link flows* (q_{ij}). The former refer to the demand for travel between a pair of zones. The latter is the flow that occurs on a specific link (i, j) of the transportation network and is the sum of all interzonal flows that happen to include that particular link on their preferred paths. For the sake of clarity, uppercase letters are employed in this book to denote *zones* (I and J) and *interzonal flow* (Q), and lowercase letters are used to denote traffic assignment network *nodes* (i and j) and link flows (q).

8.5.8 Route Choice Behavior

The key to assigning users on the network is the underlying behavioral assumption of route choice. In 1952 Wardrop established two mutually independent principles of route choice. According to the first principle, users choose the route that minimizes their own travel time. According to the second principle, users distribute themselves on the network in such a way that the average travel time for all users is equal (on each route leading from an origin to a destination) [8.20].

The first rule defines the *user equilibrium*, whereupon each user goes on the shortest path. The second rule defines the *system equilibrium*, whereupon the total cost of using the system is minimized. The terms "shortest" and "cost" typically refer to travel time, but elaborate equilibrium formulations account for *generalized costs*, which include travel time, fuel consumption or fare price, average speed, number of stops, and so forth.

A later development in equilibrium principles recognizes the fact that users have only limited information about the network and their transportation options (mode and route) for going from an origin to a destination. Thus it is more logical to base the equilibrium on the perceptions of users. This way each user assigns himself/herself on a path that he or she thinks is the shortest. This is called *stochastic equilibrium*.

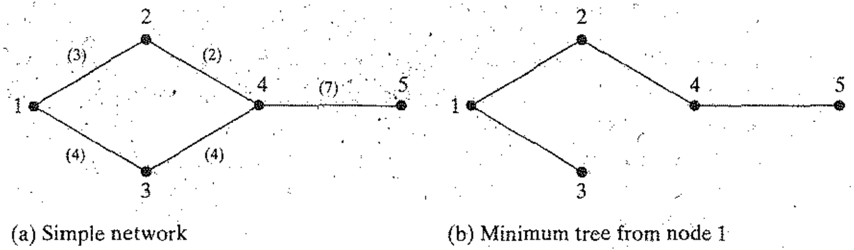
The effect of these three types of equilibria is most notable on networks. According to the user equilibrium, all used paths between the origin and destination require the same travel time (or cost), whereas all unused paths have times that are greater than the shortest time. It is likely that several paths between the origin and destination will not have any flow.

According to system equilibrium, all possible paths are evaluated and users are assigned in a way to minimize the networkwide travel time or cost. This equilibrium rule is useful during the planning stage of large traffic studies: Signal timing, channelization, lane allocations, and other traffic elements can be used to encourage or discourage particular routes so that the networkwide travel time, pollution, or congestion level is kept at a minimum. According to the stochastic equilibrium, all reasonable paths (i.e., paths that logically go from origin to destination) between an origin and a destination will have flow.

8.5.9 Minimum Path Algorithms

Assume that it is required to find the minimum (impedance) path between zones 3 and 5 on the network of Fig. 8.5.6. This task may be accomplished by identifying all possible paths between the two zones, computing their impedances, and choosing the path with the lowest impedance. But even in the case of this extremely simple network, the path enumeration procedure is time-consuming and inefficient. More efficient minimum path algorithms have been developed as variations on a theme advanced by Moore [8.21]. Some determine the minimum path between a pair of zones, whereas others compute the *minimum tree*, which contains all of the interzonal minimum paths that emanate from a zone of origin.

The basic minimum tree algorithm begins at the node of origin and proceeds outward, successively eliminating links that clearly do not belong on any minimum path emanating from the origin. Figure 8.5.7 illustrates this concept. Suppose that the minimum tree emanating from node 1 and terminating in all other nodes of the network is being sought. The minimum path to node 5 passes through node 4. But there are two possible paths to node 4: one via node 2 and the other via node 3. The first takes 5 units of impedance and the second



(a) Simple network (b) Minimum tree from node 1

Figure 8.5.7 Link elimination.

takes 8 units. Therefore the first is the minimum path to node 4 and to any subsequent node whose path passes through node 4. The last link of the longer path (in this case link 3–4) is eliminated from the minimum tree shown by Fig. 8.5.7(b).

The minimum tree may be described numerically by a *tree table*, as shown in Table 8.5.2. The first column of the tree table contains all network nodes *j* including the origin. The second column contains the total impedance of the minimum path from the origin to each node *j*. The last column specifies the node *i* that immediately precedes node *j* on the minimum path from the origin to node *j*. In other words the pair of nodes (*i*, *j*) defines the last link on the minimum path from the origin to node *j*. Thus the fourth row of the table says that the minimum path from node 1 to node 4 takes five units of impedance and that node 4 is immediately preceded by node 2.

TABLE 8.5.2 Simple Tree Table

Node (*j*)	Total impedance to node *j*	Node preceding *j*
1	0	—
2	3	1
3	4	1
4	5	2
5	12	4

The tree table describes a specific path, say, to node 5, as follows: Node 5 is preceded by node 4 (last column of row 5), node 4 is preceded by node 2 (last column of row 4), and node 2 is preceded by node 1 (last column of row 2), which is the origin (last column of row 1). Reversing this order, the path from node 1 to node 5 consists of the following sequence of links: 1–2, 2–4, and 4–5.

Most transportation planning packages (see Section 15.3) incorporate assignment algorithms. The more sophisticated packages incorporate *vine-building* algorithms. A vine-building algorithm, when seeking minimum paths, takes into account both delays at intersections (nodes) and turn prohibitions.

8.5.10 A Minimum Tree-Seeking Procedure

The following procedure produces the tree table that contains every minimum path emanating from the node of origin:

Step 1: Initialize the path impedances of the tree table at zero for the node of origin and a very large number for all other nodes. This large number ensures that the first encountered actual path to a node will be chosen.

Step 2: Enter into a list the links (i, j) that emanate directly from any node i just added to the tree.

Step 3: For each node j included in the list, add the impedance of link (i, j) to the tree table's current total impedance to node i. This quantity represents the total impedance to node j via node i. If this value is smaller than the current tree table entry for node j, replace the current total impedance to j with the new total impedance and enter node i as the node that immediately precedes j. This operation replaces the longer path to node j with the shorter one just discovered. If the new total impedance is greater than the current tree table entry, proceed to the next link in the list.

Step 4: Return to step 2, unless the list is empty, in which case the tree table contains the solution.

Example 8.13: Minimum Tree Algorithm

Find the minimum tree emanating from node 1 for the network described by the link array of Table 8.5.1.

Solution The graphical solution to this problem is summarized in Fig. 8.5.8. The related calculations are shown in Tables 8.5.3 and 8.5.4. Table 8.5.3 shows the changes performed on the tree table as the tree is built outward from the origin (node 1). The second and third columns of the table have been expanded to show these changes as they occur during the procedure. The initial condition (stage I) contains only the node of origin. All links emanating from node 1 are next entered in the list (Table 8.5.4) and are also shown by dashed lines on the graph of the partial tree (Fig. 8.5.8). These links and their link impedances are found in row 1 of the link array (Table 8.5.1). In this case there is only one entry, link 1-6 with a link impedance of five units. The calculations of stage II are shown in Table 8.5.4. The impedance of the new path is computed by adding the impedance of link (1, 6) to the current tree table entry for node $i = 1$

TABLE 8.5.3 Tree Table Changes at the End of Each Stage

Node	Total impedance to node j							Node preceding j					
(j)	I	II	III	IV	V	VI		I	II	III	IV	V	VI
1	0							—					
2	∞			15							7		N
3	∞				21							8	o
4	∞				18							11	
5	∞				19							12	
6	∞	5							1				C
7	∞		13							6			h
8	∞			17							7		a
9	∞		10							6			n
10	∞		9							6			g
11	∞			14							10		e
12	∞			17							9		
13	∞				22							11	
14	∞				19							11	

TABLE 8.5.4 List Changes and Related Calculations

Stage	Links		Compute	Compare to tree table	
N	i	j	new path impedance	stage $N-1$	Decision
II	1	6	$0+5=5$	$5<\infty$	Accept
III	6	7	$5+8=13$	$13<\infty$	Accept
		9	$5+5=10$	$10<\infty$	Accept
		10	$5+4=9$	$9<\infty$	Accept
IV	7	2	$13+2=15$	$15<\infty$	Accept
		8	$13+4=17$	$17<\infty$	Accept
		10	$13+7=20$ Reject		Reject
	9	10	$10+3=13$	$13>9$	Reject
		12	$10+7=17$	$17<\infty$	Accept
	10	7	$9+5=14$	$14>13$	Reject
		9	$9+3=12$	$12>10$	Reject
		11	$9+5=14$	$14<\infty$	Accept
V	8	3	$17+4=21$	$21<\infty$	Accept
	11	4	$14+4=18$	$18<\infty$	Accept
		8	$14+5=19$	$19>17$	Reject
		13	$14+8=22$	$22<\infty$	Accept
		14	$14+5=19$	$19<\infty$	Accept
	12	5	$17+2=19$	$19<\infty$	Accept
		13	$17+6=23$ Reject		Reject
VI	All links emanating from nodes 3, 4, 5, 13, and 14 are rejected; the list is now empty and the procedure ends.				

(i.e., $0+5=5$). This value is compared to the current tree table entry for node $j=6$ (i.e., infinity). Since the new path to node 6 is shorter than the current value, the new path is accepted. Stage II of the tree table and the tree diagram reflect this modification. All links emanating from the newly added node (node 6) are placed in the list to be considered at the next stage. They are also shown by dashed lines on the partial tree diagram. At the end of stage III nodes 7, 9, and 10 are added to the tree, and the links emanating from these nodes (i.e., the links found in rows 7, 9, and 10 of the link array) are placed in the list. Note that the stage IV entries to the list contain two alternative paths to node 10: One via node 7 and one via node 9. Of these two, the second is shorter, so the first may be rejected immediately. The path to node 10 via node 9 is also rejected in favor of the current path to 10, which the tree table shows to be via node 6. The procedure continues until stage VI, when all of the list entries are rejected, that is, when the list is empty. The final tree table emanating from node 1 and the corresponding diagram are shown in Table 8.5.5 and in Fig. 8.5.8.

Discussion This example found the minimum tree emanating from node 1 and terminating in all other nodes of the network. It does not contain any other paths. Thus the sequence of links shown on the tree of Fig. 8.5.8 joining node 3 and node 14 does not represent the minimum path between these nodes. In order to find the minimum tree emanating from node 3, the procedure must be repeated, starting with the appropriate initialization of the tree table.

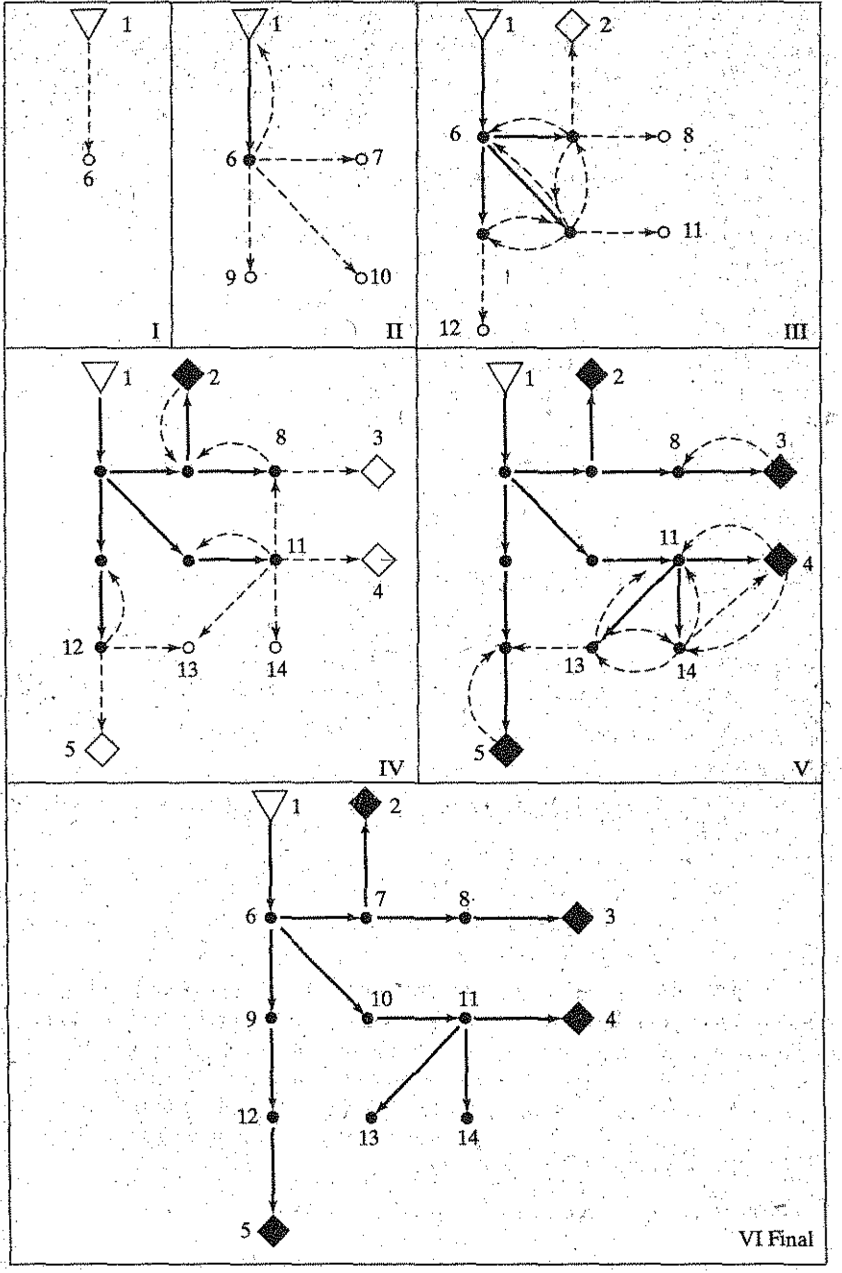


Figure 8.5.8 Tree stages.

TABLE 8.5.5 Final Tree Table

Node (j)	Total impedance to node j	Node preceding j
1	0	—
2	15	7
3	20	8
4	18	11
5	19	12
6	5	1
7	13	6
8	17	7
9	10	6
10	9	6
11	14	10
12	17	9
13	22	11
14	19	11

8.5.11 Free/All-or-Nothing Traffic Assignment

The free/all-or-nothing assignment technique allocates the entire volume interchanging between pairs of zones to the minimum path calculated on the basis of free-flow link impedances. After all interchange volumes are assigned the flow on a particular link is computed by summing all interzonal flows that happen to include that link on their minimum paths.

Example 8.14

Assign the following interzonal vehicular-trips emanating from zone 1 to the network of Example 8.13.

J	2	3	4	5
Q_{1J}	800	500	600	200

Solution The minimum tree emanating from zone 1 is reproduced in Fig. 8.5.9. The interzonal flows using each link of the tree are summed to compute the total contribution of the given flows to these links. Thus link 7–2 takes only the total interchange between zones 1 and 2, and link 7–8 takes the flow from zone 1 to zone 3. Link 6–7 takes the sum of the flows from zone 1 to zone 2 and from zone 1 to zone 3 because it belongs to both minimum paths. These links may also be assigned additional flows if they happen to be part of minimum paths that originate from zones other than zone 1.

8.5.12 Free/Multipath Traffic Assignment

In essence, a free/all-or-nothing assignment assumes that all trip-makers traveling between a specific pair of zones actually select the same path. In reality, interchange volumes are divided among a number of paths, and algorithms that are capable of determining several paths between each pair of zones in order of increasing impedance are available. Therefore it is possible to apportion the interchange volume between these paths according to some realistic rule. The diversion-curve method described earlier is a case where the interzonal

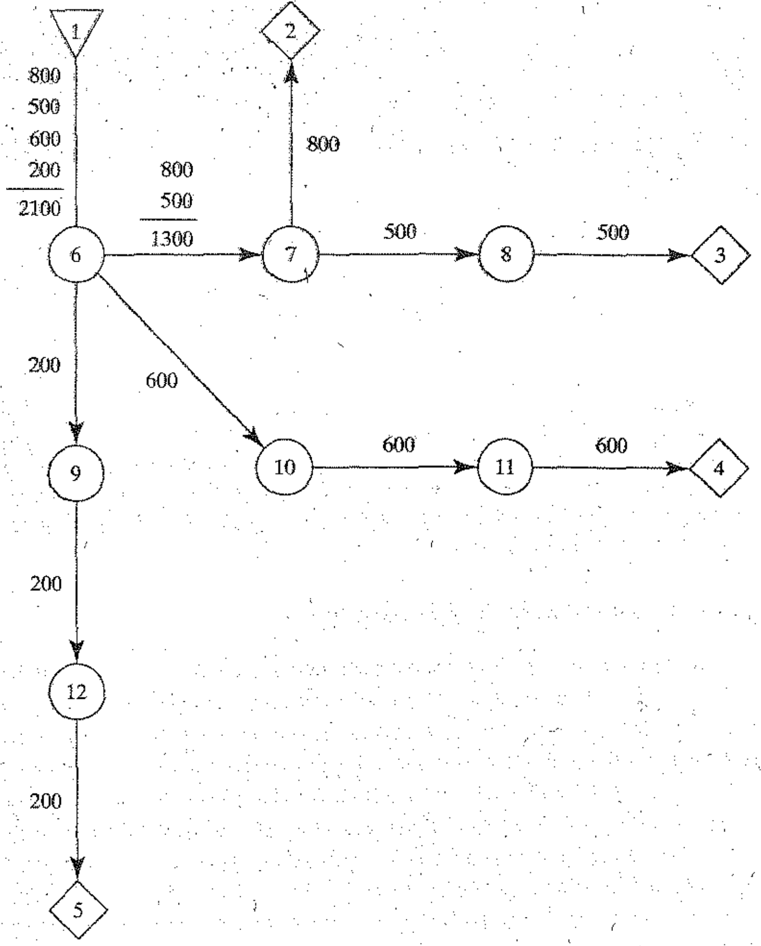


Figure 8.5.9 Minimum tree.

flows are allocated to two competing paths. Other allocation rules are also possible. For example, Irwin and von Cube [8.22] suggested the following inverse-proportion function to compute the fraction to be assigned to each of a number of interzonal routes:

$$p(r) = \frac{W_{IJr}^{-1}}{\sum_{x} W_{IJx}^{-1}} \tag{8.5.1}$$

where W_{IJr} is the impedance of route r from I to J. As the following example illustrates, the use of the multinomial logit (MNL) model (see Section 8.4.3) with disutilities based on path impedances is another possibility.

Example 8.15

A multipath algorithm found the interzonal impedances of the four shorter paths between a pair of zones to be 1.0, 1.5, 2.0, and 3.0 units of disutility. Estimate the percentage of trips to be assigned to each of the four routes according to the MNL model.

Solution Applying Eq 8.4.5 with the negative of the path disutilities in place of the utility terms, we obtain

$$p(1) = 0.47$$

$$p(2) = 0.29$$

$$p(3) = 0.17$$

$$p(4) = 0.07$$

Discussion This example merely illustrates how the MNL model may be applied to the unrestrained multipath assignment problem. Whether this model is adequate for a particular planning study must also be the subject of inquiry. The computational complexity of these models should not escape the reader's attention, but computerized algorithms that deal effectively with the repetitive nature of the calculations are available. Of greater complexity are *capacity-restrained* algorithms that incorporate the effect of traffic flow on link impedance.

8.5.13 Capacity-Restrained Traffic Assignment

In Chapter 3 we showed that as the flow increases toward capacity, the average stream speed decreases from the free-flow speed (u_f) to the speed at maximum flow (u_m). Beyond this point the internal friction between vehicles in the stream becomes severe, the traffic conditions worsen (i.e., levels-of-service E and F), and severe shock waves and slow-moving platoons develop.

The implication of this phenomenon on the results of free-traffic assignment presents the following paradox: The interzonal flows are assigned to the minimum paths computed on the basis of free-flow link impedances (usually travel times). But if the link flows were at the levels dictated by the assignment, the link speeds would be lower and the link travel times (i.e., impedances) would be higher than those corresponding to free-flow conditions. As a result, the minimum paths computed *prior* to trip assignment may not be the minimum paths *after* the trips are assigned. Several iterative assignment techniques address the convergence between the link impedances assumed prior to assignment and the link impedances that are implied by the resulting link volumes. These techniques are known as *capacity-restrained* methods or techniques that employ *capacity restraints.*

The relationship between link flow and link impedance is described as the *link-capacity function.* Several such functions are found in the technical literature. Figure 8.5.10 presents the form developed by the BPR, which is expressed mathematically as

$$w = \bar{w}\left[1 + 0.15\left(\frac{q}{q_{\max}}\right)^4\right] \tag{8.5.2}$$

where

w = impedance of a given link at flow q

$\bar{w}$ = free-flow impedance of the link

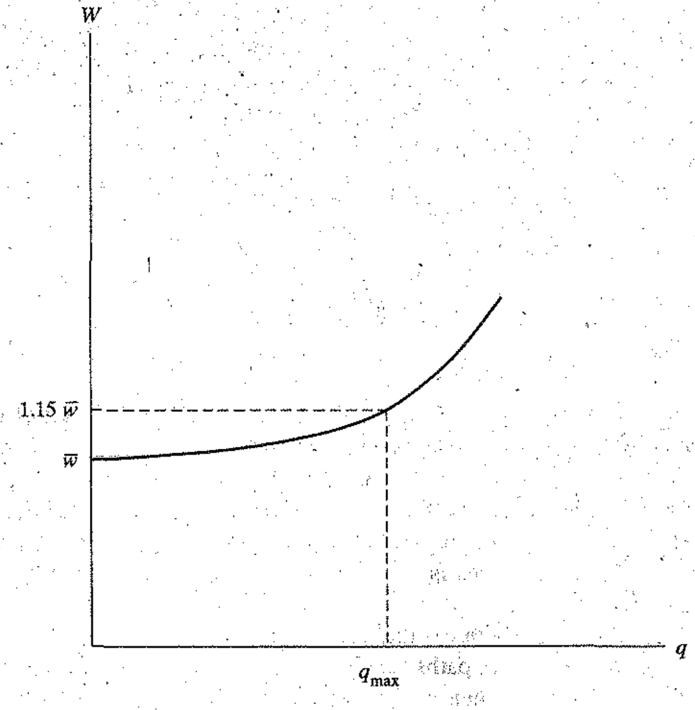


Figure 8.5.10 Bureau of Public Roads link capacity function.

$$q = \text{link flow}$$
$$q_{\max} = \text{link's capacity}$$

This function states that at capacity the lin'k impedance is 15% higher than the free-flow impedance. If the demand were to exceed the capacity of the link, the resulting shock waves and their dissipation times (see Chapter 3) would cause a rapid deterioration in the link flow conditions. The original BPR capacity function was based on the relationships found in the 1965 Highway Capacity Manual (HCM). Since that time the equation has been modified to be consistent with more recent data. One modification was to substitute parameters α and β in place of the constants 0.15 and 4 and to allow these parameters to take different values depending on facility type, design speed, the number of signals and other interruptions per mile, and other characteristics. Table 8.5.6 presents values for the two parameters for freeways and multilane highways based on the 1985 edition of the HCM. Although the modified BPR equation constitutes standard practice, several model sets make use of alternate mathematical forms, including complex conical-section functions.

Capacity-restrained algorithms incorporate link-capacity functions in their search for convergence to an equilibrium state. They may be either all-or-nothing or multipath. An example of the former is the algorithm developed by CATS [8.23], where the following assignment procedure is applied: An interchange is chosen at random, the minimum path is

TABLE 8.5.6 Modified BPR Coefficients

Facility type	Speed (mi/h)	α	β
Freeway	50	0.56	3.6
	60	0.83	5.5
	70	0.88	9.8
Multilane	50	0.71	2.1
	60	0.83	2.7
	70	1	5.4

Source: NCHRP 365, 1998 [8.1].

determined using the free-flow impedances, and the entire interchange volume (i.e., all-or-nothing) is assigned to this minimum path. The impedances of the links that make up this path are updated according to the assigned flows, and another interchange is randomly chosen for similar treatment. The procedure ends when all interchanges are considered. Although not realistically reproducing particular interchange flows, the incremental updating of link impedances is expected to result in realistic estimates of the equilibrium link flows.

A multipath extension of the CATS method begins with an uncongested network, finds all of the minimum paths, and assigns a portion (say, 20%) of each interchange volume to these paths. It then updates the link impedances according to the resulting partial link flows, recomputes the minimum paths, and assigns the next increment of the interchange volumes. The procedure continues until 100% of the interzonal flows are assigned.

The method originally developed by the BPR [8.19] first performs an all-or-nothing assignment based on free-flow link impedances. It then updates all loaded links and repeats the all-or-nothing assignment using the new impedances. After several iterations paths between pairs of zones are assigned a portion of the interzonal flow in proportion to the number of times that each appeared to be the minimum path. Figure 8.5.11 illustrates the results of this procedure on a small hypothetical network after four iterations [8.24].

The Urban Transportation Planning System (UTPS), a battery of computer programs developed by the FHWA, includes a method that is based on the linear programming optimization technique [8.25]. Figure 8.5.12 illustrates this method, which begins with a free/all-or-nothing assignment (iteration 0). The simple two-zone diagram shows that 100% of the interzonal flow is assigned to the minimum path. The BPR function (Eq. 8.5.2) is then applied to update the impedances of the loaded links. Iteration 1 repeats the all-or-nothing assignment based on the updated link impedances. At this stage the linear programming algorithm is applied to calculate the value of the variable λ, which is used to divide the interzonal volume between the first two paths. The link impedances are updated accordingly, and the procedure continues until λ is close to zero, a sign that an equilibrium stage has been reached.

Other capacity-restrained assignments have been reported in the literature, including a two-pass Markov model, which allocates traffic to links based on reasonable transition probabilities [8.26]. Standard practice involves the use of more sophisticated equilibrium assignments. The simpler techniques are still in use in the context of site-specific analyses (see e.g., Chapter 9).

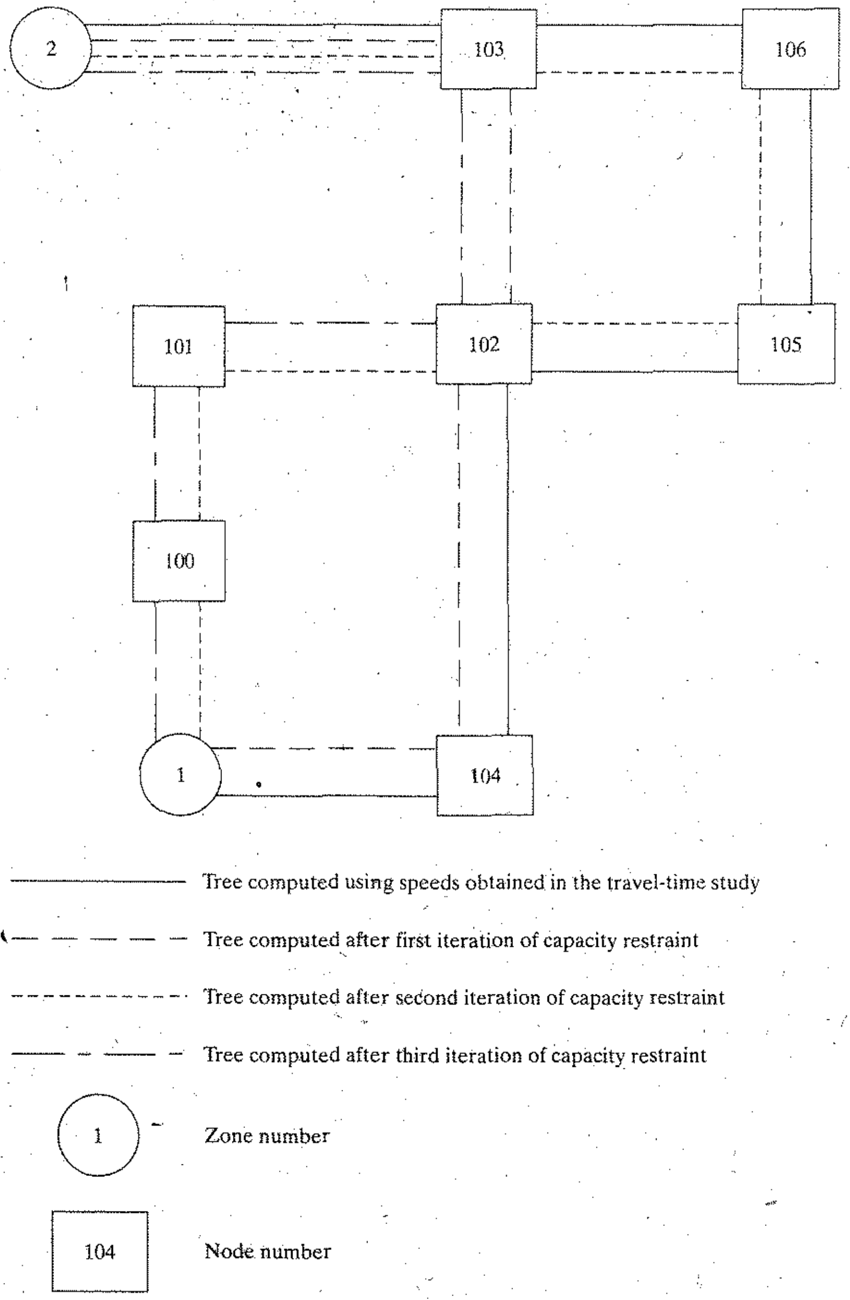

2
103
106
101
102
105
100
1
104
Tree computed using speeds obtained in the travel-time study
Tree computed after first iteration of capacity restraint
Tree computed after second iteration of capacity restraint
Tree computed after third iteration of capacity restraint
1
Zone number
104
Node number

Figure 8.5.11 Illustration of the Bureau of Public Roads multipath assignment procedure.
(From Humphrey [8.24].)

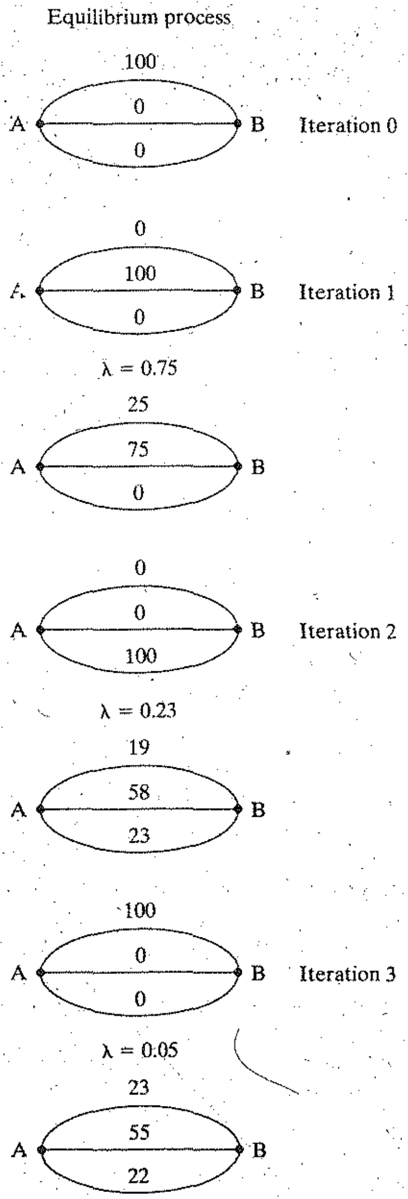


Figure 8.5.12 Urban transportation planning system capacity-restrained assignment procedure. (From Levinsohn et al. [8.25].)

8.5.14 Transit Assignment

The assignment of interzonal trips to a transit network presents certain complications that are not encountered in traffic assignment. In addition to the transit network of links and nodes, transit operations involve the identification of transit routes and schedules. The temporal pattern and directional orientation of demand coupled with resource limitations (e.g.,

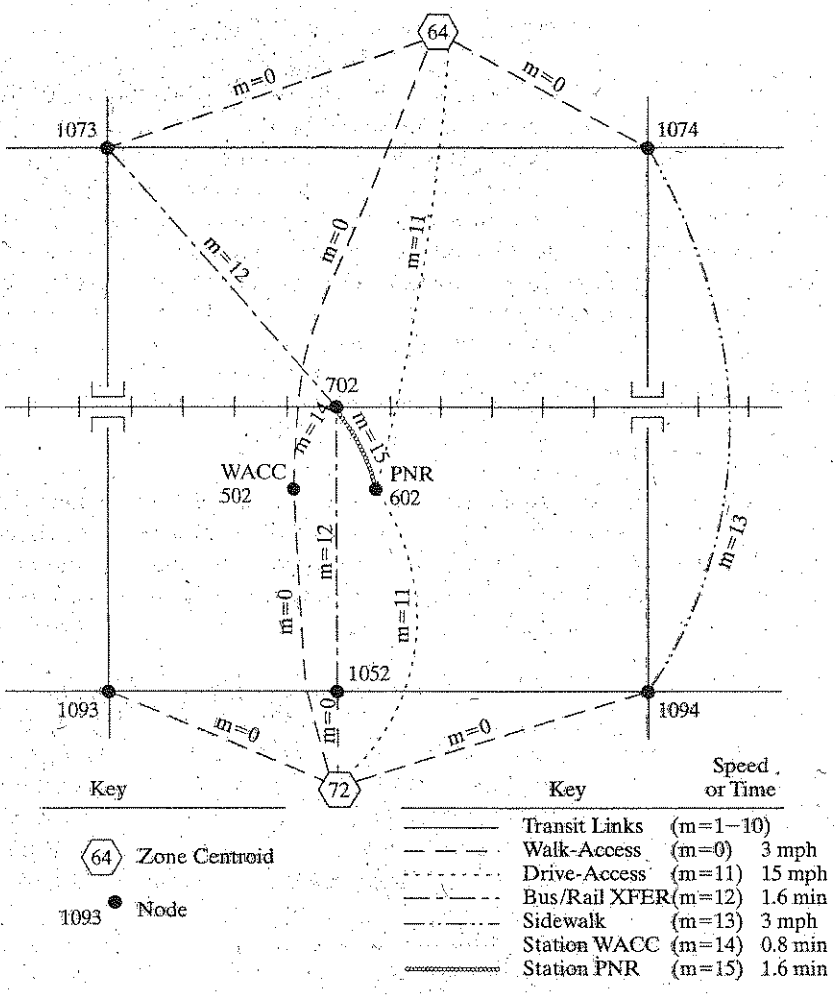


Figure 8.5.13 Sample coding for transit assignment with MINUTP.

available fleet size, operating costs) always dictate a service-coverage that is not ubiquitous. Overlapping routes, the need to transfer between routes, differences between exclusive right-of-way lines and mixed traffic operations, and variabilities of service in time and space add to the challenges associated with transit assignment. Figure 8.5.13 illustrates the level of detail required when specifying a transit network. The methods of transit network analysis, however, are beyond the scope of this book.

8.5.15 Summary

Trip assignment simulates the way in which trip-makers select their paths between zones. Traffic assignment estimates the expected flows that the links of the highway network are likely to experience to help anticipate potential capacity problems and to plan accordingly. It requires a behavioral hypothesis of route choice, a method of describing the highway network for computer processing, a way of selecting the appropriate interzonal paths, and a way of realistically allocating (i.e., assigning) the interzonal volumes on these paths. Several traffic assigr, nent models were described in this section, including a simple two-path diversion-curve model. In addition, traffic assignment models that are more appropriate for the analysis of large networks were described. These models were classified in two ways. First, they were identified as either capacity-restrained or free assignment models, depending on whether or not they explicitly account for the effect of congestion. Second, they were categorized as either all-or-nothing or multipath models, depending on whether they allocate the interzonal demand on a single or multiple paths. In addition to path (or route) allocation, transit assignment must contend with complexities that are not present in traffic assignment.

8.6 TRANSPORT BEHAVIOR OF INDIVIDUALS AND HOUSEHOLDS

8.6.1 Background

This section concentrates on behavioral aspects of transportation demand. The prevailing trend in demand analyses is toward focusing on the household and its individual members in an effort to achieve a better understanding of transport demand. This focus reveals that a large number of factors (i.e., from most tangible factors, such as gender, age, and income, to most elusive ones, such as personality and lifestyle) affect transport behavior and consequently the utilization of transportation systems and modes.

It is well known that transportation is by and large a derived demand. Individual and household needs generate the demand for transportation. The transport behavior of individuals and households results from their transport-related decisions, which may be classified into long- and short-term decisions. The long-term decisions are referred to as *mobility choices* and include decision relations to residential location, employment location, automobile ownership, and mode to work. The short-term decisions are referred to as *travel choices* and include frequency of travel for various purposes, such as mode, destination, route, and time of day of trips.

In disaggregate transport behavior analyses one important element is the unit of analysis (i.e., the entity upon which the analysis is focused): It could be the individual or the household. Individuals can be viewed independently, with their own unique aspirations, goals, and idiosyncratic personalities. However, the behavior of individuals belonging to households is constrained and will necessarily conform to certain role assignments usually defined to meet household as well as individual goals with reasonable efficiency. Therefore it is important to recognize the effects of household characteristic while trying to account for the personal attributes of each person in the household.

8.6.2 Conceptual Models

Several conceptual models describing the process generating the transport behavior of people have been developed. The process generating the transport behavior of people founded upon work by Hartgen and Tanner [8.27], Field et al. [8.27], Salomon [8.28], Ben-Akiva and Lerman [8.29], and Prevedouros [8.30] is presented in Fig. 8.6.1.

The environment within which households and individuals exist may be represented by a set of activities, defined by location, time, and money requirements as well as constraints

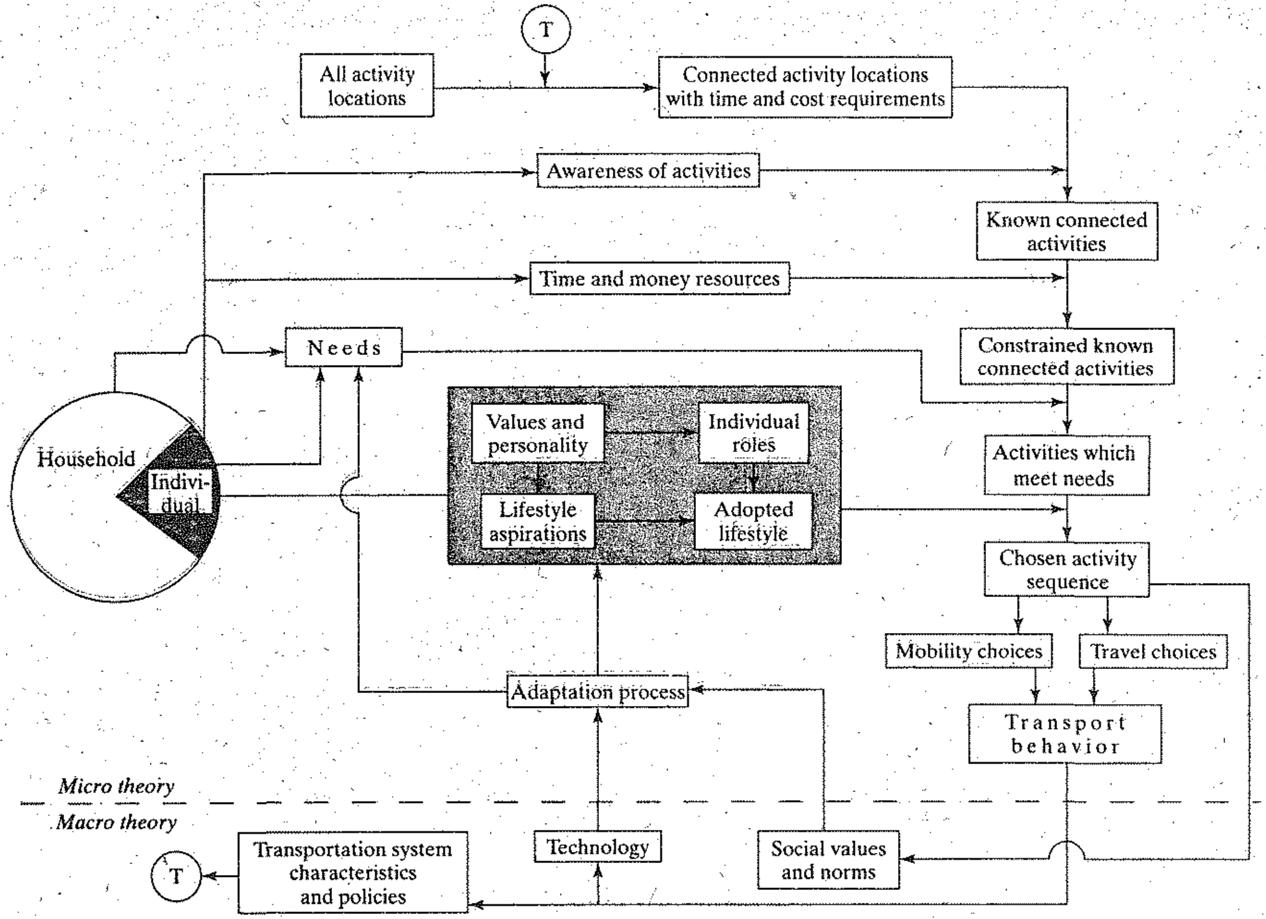


Figure 8.6.1 Conceptual model of factors affecting transport behavior.

(e.g., food store closes at 9 P.M., favorite restaurant is closed on Thursdays, the movie show costs $5 before 3:00 P.M. and $8 after 3:00 P.M., etc.). The activity locations are connected with a transportation system that has distinct characteristics (e.g., network structure, modes, performance, costs to the traveler) and policies (e.g., parking regulations, transit hours of operation, time-variant fare or toll pricing, etc.)

The personal and joint needs of individuals create the set of household needs. Some needs must be fulfilled (i.e., work, school, maintenance) at certain time periods, whereas some others are optional (i.e., recreational activities). When the set of needs that can or must be fulfilled and the resulting set of activities that will meet those needs have been identified, the personal characteristics and responsibilities of each person determine the chosen activity sequence for each household member (which will result in the fulfillment of needs). This is a dynamic process with much variation around an "average" activity pattern as well as longer-term changes of the average pattern.

Focus on the personal characteristics and responsibilities of each individual reveals that each person has a set of values and a distinct personality. These elements interact with the surrounding world generating their lifestyle. Lifestyle, according to Salomon [8.28], is a person's orientation of life through three major decisions: the decision to form a household, the decision to participate in the labor force, and decisions about spending free time (leisure). Many more detailed components can be attributed to lifestyle, which are complements or parts of this broad definition (i.e., preference for residence location, degree of career orientation, preference for specific types of automobiles, entertainment preferences, etc.)

Two major underlying components of a person's lifestyle are his or her personality and occupational status, both of which are responsible to a large degree for the orientation of each person in life. Lifestyle is important because it underlies the decision-making process of individuals. Certainly the chosen lifestyle of each person is not a straight, narrow line connecting points in the person's life. It is better to be viewed as a broad path through life. In other words individuals make decisions and perform activities that broadly fulfill their lifestyle aspirations.

Another important element for each individual household member is that of the role and role commitments (i.e., provision of shelter, food gathering, schooling, working, etc.) The role is tied to a function or a set of functions; by performing a set of activities that fulfills those functions, a person assumes a role. Functions are defined by the needs of individuals and household members. Thus the role constitutes a link between needs and patterns of fulfilling needs. The needs cluster in four broad groups: household/familial, work/career, interpersonal/social, and leisure/recreation [8.27]. The responsibility for the fulfillment of roles within each group is divided among members in the household. The basis for this division is a result of a mix of factors, such as social norms, past experience, time and opportunity constraints, sociocultural and personality factors, and negotiation. There are short- and long-term role commitments. Over time roles change for household members as the household evolves through the life-cycle stages* [8.31].

All of these result in the activity sequence chosen by the individual household member. This sequence allows fulfillment of the needs of the household and achievement of the person's lifestyle aspirations. The chosen activity sequence then serves as the input

*The basic life-cycle stages of a household are the following, in chronological order: single person, couple without children, couple or single parent with dependent children, couple or single parent with independent children (children at driving age or older living at the parental household), couple of seniors, senior single person.

for the mobility and travel choices of the household members, which in turn define the transport behavior (see Fig. 8.6.1).

The sum of the chosen activity sequence and the transport behavior across all individuals in the population feeds back the set of values and norms of the society, as well as the technology and the transportation system and policies to the household (macro theory, Fig. 8.6.1). This dynamic interaction modifies the environment in which individuals live. An adaptation process then takes the place and the household needs and standards and aspirations for each person are adjusted.

The critical elements of this framework are the individual and household needs as well as the lifestyles of the individual household members. There are some basic elements of needs and lifestyles that can be found in most individuals at most places (i.e., work, shelter, household formation, transportation). A number of factors (i.e., culture, environment, available technology, etc.) shape and differentiate needs and lifestyles across individuals, households, or places. The process in Fig. 8.6.1 suggests that people's transport behavior is affected by such factors as household structure, availability and cost of activities, personality and lifestyle, technology available to users, location patterns, personal or household income, social values and norms, and transportation system characteristics and policies.

8.6.3 Demand Models with Behavioral Content

Detailed knowledge of the factors affecting transport behavior enables the estimation of realistic models for transport-demand forecasting, including models of automobile ownership, trip generation, mode choice, residence location choice, and diversion choice (i.e., response of drivers to real-time information on traffic conditions). The latter is critical in the rapidly approaching ITS (intelligent transportation systems era, discussed in Chapter 6).

In addition to the factors mentioned earlier, perception of quality of service and attitudes toward transportation modes can be included in quantitative representations of transport behavior (models) to enrich the coverage of factors affecting transport behavior and to obtain a better understanding of the impact of each factor on transport behavior [8.32, 8.33].

Knowledge of transportation systems and services varies among people, as does the degree of accuracy with which characteristics of transportation systems and services are known to people. Two major characteristics affecting the choice of mode are travel time and travel cost. One's perception may be that the travel time from *A* to *B* by bus is 25 min and 10 to 12 min by private car. Although the actual travel time by bus is 15 to 18 min, this objective fact is not the input utilized by the person considered when deciding his or her mode of transportation from *A* to *B*. Misperceptions also apply to travel cost. Typically people tend to count only the out-of-pocket cost of using an automobile (i.e., gasoline, parking, and tolls) and forget the substantial costs of vehicle depreciation, insurance, and maintenance. The aforementioned misperception of characteristics may bias the judgment of the merits of private auto and public transportation, which would result in a higher than normal proportion of people selecting the private auto as their means of transportation.*

It may take an extensive and expensive marketing campaign by a transit authority to partially correct public misperceptions. It is critical, however, that public transit agencies

*Similar misperceptions are evident in intercity transportation as well. In this context modes are often selected based on their terminal-to-terminal performance, thereby ignoring substantial access and waiting times (e.g., bias toward fast traveling modes; also see Fig. 5.3.1).

and private travel industries utilize models and planning tools that include both their customers' perceptions and objective system characteristics.

The feelings or attitudes toward a transportation mode also affect travel choices in important ways. People tend to have a less favorable attitude toward public transportation (i.e., in terms of scheduling, privacy, convenience, and safety) than toward the private auto [8.32], especially in the United States.

In the remainder of this section we discuss two basic models developed on the basis of household travel behavior: an automobile ownership and a trip-generation model. *Automobile ownership* is considered as one of the key determinants of transport behavior and is included in virtually every model of trip generation and mode choice [8.34]. A basic classification of automobile-ownership models results from the unit of analysis. A model is disaggregate when the unit of analysis is the individual or the individual household. When characteristics of individuals or households are aggregated in any way (i.e., in tracts, zones, etc.) for model estimation the resulting model is aggregate. The data also characterize the geographic area of application of a model. There are models built to fit specific urban environs, states, and nations.

Regression models are used to estimate levels of automobile ownership (e.g., the number of automobiles owned by a household). Poisson regression and logit models are used to estimate a household's probability of having 0, 1, 2, . . . automobiles. Studies show that the number of household drivers, the number of dependent children (i.e., before the age of driving), and location (i.e., high- and low-density locations) play an important role in automobile ownership and usage (refer to Example 8.16 later in this section) [8.30, 8.35].

The classification of *trip-generation* models (i.e., aggregate, disaggregate, local, regional, etc.) is similar to that for automobile-ownership models. A usual method for model estimation is regression. Trip-generation models are often segmented by either gender or level of automobile ownership. Common factors used in trip-generation models are the number of automobiles available (i.e., owned plus rented or company-provided), household structure (i.e., size, life-cycle stage, and number of workers), residential location, and income.

Recent advancements in travel behavior focus on activity analysis [8.36]. Activity analysis involves the integration of concepts from psychological theory, sociological theory, economic theory, and geography, and intends to develop a clear understanding of the people who make the trips (i.e., motivational structure and psychological profile), the interdependencies among people both within and outside the household, the people's opportunity set, transport services, and options. (See the TRANSIMS model in Chapter 15.)

Examples of disaggregate automobile-ownership and trip-generation models are presented in Table 8.6.1. The automobile-ownership model includes the following variables, in the order listed in Table 8.6.1:

1. Number of household members who are eligible to drive (i.e., at age 16 or older)
2. Number of household full-time workers
3. Number of household workers who are employed in the suburbs
4. A binary variable that identifies households with dependent children (i.e., variable is equal to 1 if there are children younger than 16 years of age)
5. A binary variable that identifies households with senior members (i.e., no children present and at least one member at the age of 65 or older)
6. Income in 1989 U.S. dollars

TABLE 8.6.1 Examples of Automobile-Ownership and Trip-Generation Models

Automobile-ownership model [8.30]		Trip-generation model [8.37]	
Dependent variable: Number of household automobiles		Dependent variable: Household trips to work	
Independent variables	Parameter	Independent variables	Parameter
Number of drivers	0.58	Number of workers	3.933
Number of full-time workers	0.11	Number of adult females	0.416
Number working in the suburbs	0.11	Number of adult males	0.918
Household with dependent children	0.11	Household size	0.282
Household with senior members	−0.13	Number of nondrivers	−0.457
Income (thousand 1989 U.S. dollars)	0.04	(High education) (workers)	0.23[a]
Use of mass transit to work	−0.28	(Low education) (workers)	−0.35[a]
Low-density residence location	0.14	$\sqrt{\text{Income (in DFI)/100}}$ (workers)	0.22[a]
Constant	0.23	(High-density) (workers)	−0.11[a]
		(Low-density) (workers)	−0.43[a]
		Constant	−0.36[a]
R^2	0.54	R^2	0.56
Number of cases	1372	Number of cases	1739
Source of data: Survey in Chicago, Illinois		*Source of data:* Survey in the Netherlands	

[a] Parameter not statistically significant; otherwise significant at $\alpha = 0.05$.

7. A binary variable that accounts for public transit use (i.e., it is equal to 1 if one household worker used mass transit to commute to work)
8. A binary variable that is equal to 1 if the household resides at a low-density location (it is equal to 0 otherwise, same as the other binary variables in the model)

Example 8.16

Given a suburban (low-density) zone with 250 households with the following average characteristics: 2.2 drivers, 1.4 full-time workers, 70% employed in the suburbs, 33% of all households have dependent children, 8% of all households are senior households, income is $43,000 and 6.5% use mass transit to work, apply the model in Table 8.6.1 to estimate the average automobile ownership per household.

Solution Using the model in Table 8.6.1 and substituting the given inputs, the average automobile ownership per household is estimated as

$$\begin{aligned}
&+ 0.58(2.2)\\
&+ 0.11(1.1)\\
&+ 0.11(1)(0.70)\\
&+ 0.11(1)(0.33)\\
&- 0.13(1)(0.08)\\
&+ 0.04(43{,}000 \div 10{,}000)\\
&- 0.28(1)(0.065)\\
&+ 0.14(1)\\
&+ 0.23 = 2.06 \quad \text{automobiles per household}
\end{aligned}$$

Discussion This estimate translates into a total of 515 automobiles in the zone examined. If this population was at a high-density location and 15% were using mass transit to work, the average household automobile ownership would be equal to 1.92 automobiles, which would translate into 475 automobiles in the zone examined, a 7.8% reduction in the number of automobiles owned.

The latter part reflects another utility of such models: ability to assess the effect of each factor upon the element of transport behavior examined. In addition, sensitivity analysis (i.e., how much a Δx change of factor x affects the dependent variable) may assist in policies aimed at encouraging or discouraging specific transportation decisions of people. Also, the effects on transport behavior of dynamic changes in the society, such as aging, increases in the workforce population, and economic downturns that affect people's incomes, can be assessed.

The trip-generation model includes the following variables, in the order listed in Table 8.6.1:

1. Number of workers in the household
2. Number of adult females in the household
3. Number of adult males in the household
4. Number of persons in the household
5. Number of household persons at age 12 or older who do not drive

6,7. Level of education, which is defined by two binary variables: high education is equal to 1 if a household member has a college degree (it is 0 otherwise), and low education is equal to 1 if the household member with the highest education has an elementary school degree (it is 0 otherwise)

8. Income is used after the square root of the annual gross income in DFI is taken and then divided by 100

9,10. Density, which is represented by two binary variables: High density is equal to 1 if the household resides in a large metropolitan area served by transit systems (it is 0 otherwise), and low density is equal to 1 if the household resides in a community without transit service (it is 0 otherwise)

Example 8.17

Consider a zone with 250 households with the following average characteristics: household size is 3.2, the number of workers per household is 1.1, 1.4 adult females, and 1.3 adult males per household, 0.9 nondrivers per household, 35% of households have a member with a college degree, 12% of households belong to the low education category, income is 23,000 DFI, and the entire zone is in a high-density area. These data are similar to the types of data that can be taken from census reports by tract or zip code. Apply the model in Table 8.6.1 to estimate the expected number of trips in the zone.

Solution By applying the trip-generation model in Table 8.6.1, the expected work trips generated in the zone are

$$\begin{aligned}&250[(3.993)(1.1)\\&+0.416(1.4)\\&+0.918(1.3)\\&+0.282(3.2)\end{aligned}$$

$$
\begin{aligned}
&- 0.457(0.9)\\
&+ 0.230(1)(0.35)(2.1)\\
&- 0.350(1)(0.12)(2.1)\\
&+ 0.220(\sqrt{23{,}000 \div 100})\\
&- 0.110(1)(2.1)\\
&- 0.430(0)(2.1)\\
&- 0.360] = 250(6.5) = 1622 \quad \text{work trips produced from the households residing in the examined zone}
\end{aligned}
$$

The trip-generation model presented in the example results in the total number of trips produced by a group of households. An older practice consists of the separate estimation of household trips using two models: a model estimating home-based trips and a model estimating non-home-based trips. Then the separate estimates are combined to yield the total number of trips for a zone or area. The implicit assumption in the combination of home-based and non-home-based trips is that these two types of trips are mutually independent. This convenient assumption is hardly realistic [8.36].

The weakness of the combination of dependent trip rates is absent in *trip-chaining* models. These models comprise a theoretically sound trip-generation modeling basis that is gaining acceptance. Trip-chaining modeling employs a procedure that links trips originating from home according to purpose and other characteristics. Separate models for work, school, social activities, household maintenance, and so forth, are developed. These models are connected to a trip-chaining model that identifies applicable chains [8.37, 8.38]. This combination of models provides a more realistic representation of the household trip-making behavior and it maintains the ability to produce home- and non-home-based trip rates for application in older model structures.

The models presented in Table 8.6.1 are not intended to be applied to single households. These models produce meaningful estimates when applied to zones with fairly homogeneous populations. For example, it is likely that if applied to a single household, the models may estimate 6.88 trips or 1.65 automobiles, which, of course, have little physical meaning. On the other hand, these numbers may well represent household averages for a number of households in an area.

Caution should be exercised in the use of behavioral models, particularly when they have been derived from small sample sizes and/or from a specific population group (i.e., survey in affluent suburbs), which may not be representative of the whole population in an area. Considerable attention should be exercised in the application of such models in areas other than the one where the data used for the estimation of the model were collected. The issue of model transferability is important and simple conversion of units (i.e., present worth of past incomes, pounds to dollars, kilometers to miles, etc.) is hardly sufficient for application of a model in another area [8.39]. A brief description of the four common transfer methods can be found in reference [8.40]. This is due largely to the substantial cultural, spatial, economic, and transportation differences between locations, even within the same nation. Similar concerns undermine the validity of empirically derived models over time; caution should be exercised when applying a 1970s behavioral model in the 1990s. As explained earlier, activity-based travel forecasting represents several approaches that attempt to explicitly incorporate the fact that decisions relating to activities take precedence over decisions relating to travel. Facing a myriad of activity options and given their lifestyles, aspirations, and perceived needs, *individuals* make complex decisions about their activity schedules within budget and time constraints. Decisions regarding travel choices are part of this overarching activity context.

At a 1997 major conference on the subject, Goulias [8.3] encapsulated the activity-based modeling paradigm as follows:

> An activity-based travel forecasting system is a system that uses as inputs sociodemographic information of potential travelers and land use information to create schedules followed by people in their everyday life providing as output, for a given day, detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity (including travel time, mode used and so forth). This output is very much like a "day timer" for each person in a given region. A complete operational activity-based forecasting system does not exist yet [8.3].

At the same conference Bowan and Ben-Akiva echoed this emphasis on the generation of activity patterns at the individual (i.e., disaggregate) level. They classified activity-based approaches into *econometric model systems* and *hybrid simulation systems*. The former approaches are usually rooted in the economic concept of utility maximization by deliberative, "rational" decision-makers. In one way or another these models need to account for all (or at least a set of important choices available to individuals and households) and rely on mathematical equations (including, but not limited, to the logit formulations discussed in Section 8.4) that are based on the utility maximization principle. Hybrid simulation approaches, on the other hand, are motivated by theories of cognitive psychology and related behavioral sciences that attempt to explain the rules (or *heuristics*) that individuals use to make decisions when faced with complex situations where exhaustive enumeration or full knowledge of the available choices is lacking. These models rely on Monte Carlo simulation (see Chapter 14) rather than on optimization techniques. They are often called *microsimulation* models because they consider individual travelers as the elemental entities in the system. Within the two major approaches to activity-based travel forecasting there exists a multitude of potential model specifications ranging in terms of the postulated scope and desired degree of detail. Questions related to data availability and computational limitations usually constrain practical applications.

To illustrate these concepts, consider Fig. 8.6.2. The figure shows a worker's *activity schedule* during a 24-h period. This person (from a one-car household) leaves home (H) for work (W) in the morning and drops off a child at school (S) on the way. At noon he or she walks to a business lunch at location (B) and returns to work activities until the end of the workday. On the return home the worker stops at a drugstore (D) to fill a prescription. In the evening the worker with some members of the household go to a restaurant (R) for dinner, watch a movie (M), and return home. The choice of this particular activity schedule by this particular person is the result of the complex long-term and short-term kinds of decisions made individually (by the worker) and collectively (by household members, coworkers, etc.). Among the long-term decisions are household formation and choice of residential and employment locations, the purchase of an automobile and so on. Shorter-term decisions may include decisions about which member of the household should use the car that day, who would drop off the child at school and who would pick up the child at the end of the school

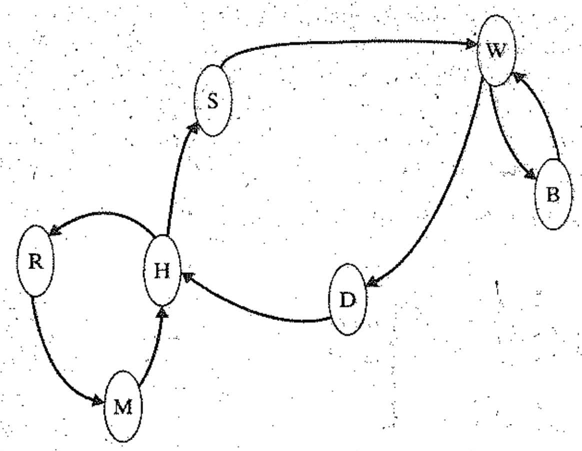


Figure 8.6.2 A worker's daily activity pattern.

day, whether to visit a restaurant and a movie theater or to eat at home and watch television. Other decisions may be made on the spur of the moment (e.g., stopping at the drugstore on the way home because time is available and the worker is using the car). Clearly no travel-demand forecasting model should be expected to capture (and be able to predict) such complex activity patterns for an entire urban area. Different model specifications can result, however, depending on how this complexity is reduced for modeling purposes.

Several researchers have attempted to develop models intended to capture jointly activity scheduling and (embedded) travel patterns, but as of 1999, these attempts have not yielded any operational applications. For example, Ettema, et al. [8.41] tested a hybrid simulation model using "hypothetical spatiotemporal settings" but did not reach the stage of estimating it using real-world data. A simpler modeling approach may classify activities into *mandatory* (or *committed*) and *discretionary*, depending on socioeconomic and other characteristics, allowing the mandatory activities to "anchor" and thus limit the number of admissible activity patterns to be considered explicitly. Simpler models are also possible.

8.6.4 Trip-, Journey-, and Tour-Based Models

Several important points can be seen with regard to the travel aspects of the activity pattern of Fig. 8.6.2. On a 24-h basis the worker's travel pattern consists of a single itinerary beginning in the morning and ending with the return from the theater. At the other extreme, which constitutes the standard practice associated with the four-step modeling process, the itinerary is decomposed into nine distinct trips, each having a producing and an attracting zone. As mentioned in Section 8.6.3, this decomposition results in the loss of important information relating to trip chaining. Assuming that trip purposes are defined as home-based work (HBW), home-based school (HBS), home-based other (HBO) and non-home-based (NHB), the trip from home to the school drop-off location in the morning is no longer understood to be associated with travel to work with an intermediate stop. The worker's itinerary segment that begins at home and ends at work in the morning is simply (from the perspective of the worker, not the child) split into two trips, one classified as HBO and the other as NHB. The first would be reported as a ride-share, whereas the second would be characterized as drive-alone. The mandatory nature of the work activity that gives rise to the home to drop off to work *trip chain* is totally lost. Moreover, when a mode choice model is subsequently applied, it will treat the two trips independently of each other and independently of the work activity using the mode choice equations estimated for all HBO and NHB trips, respectively. The HBO trip will be placed in the same pigeonhole as the trip from home to the restaurant and the trip from the theater to home in the evening, as well as the trip from the drugstore to home after work. This is because all these trips would be classified as HBO. Similarly the NHB trip from the drop-off location to work will be grouped with both trips between the work location and the business lunch site, as well as the trip from work to the drugstore.

One way to retain some knowledge of the difference in the character of these trips is to subdivide them into finer categories. For example, in developing an operational model for Reading, PA, Schultz and Allen [8.42] subdivided NHB trips into non-home-based trips associated with *journey* to or from work (NHBJTW), non-home-based-at-work (NHBWRK) trips such as the two trips to and from the business lunch, and non-home-based-non-work (NHBNWK) trips such as the trip from the restaurant to the movie theater. Analysis of the characteristics of these three NHB trip categories showed that they exhibit very different trip length, mode choice, and time-of-day characteristics. Consequently the resulting Reading models were judged to be more accurate than would be the case if the traditional trip definitions were used.

An alternate approach to being more sensitive to trip chaining and its implications on destination, mode, and time-of-day choices is to subdivide daily itineraries into tours. A *tour* is typically defined as a chain beginning and ending at the home. Thus the itinerary of Fig. 8.6.2 consists of two tours. One beginning with the morning trip and ending with the arrival home after the shopping stop and another from home to the restaurant, to the theater, and back home in the evening. The combinatorial complexity of tour-based models has prevented them from reaching the operational stage in the United States.

An intermediate approach between the Reading trip-stratification scheme and tour-based models was considered in Honolulu in connection with a major update of the land-use and travel forecasting model set used by the Oahu Metropolitan Planning Organization (OMPO). The project first developed a trip-based model similar to the Reading model but with finer trip-type categories. It then proceeded to propose an extended (*journey-based*) model set. The trip-based model was constructed so as to ensure that a fully functional model set would be delivered, considering the risk associated with attempting to construct a more advanced journey-based model [8.43]. The extended models borrowed the idea of *journeys* from the Reading application but considered them as the elemental travel unit instead of subdividing them into trips. Journeys were defined as one-way travel linkages involving primary locations (e.g., job site for workers, school site for students, home). Journey types included *work journeys* between home and work. Two such journeys are shown in Fig. 8.6.2, both having an intermediate stop. *Journeys at work* were defined akin the Reading's trips at work but allowing for intermediate stops and *other journeys* between the home, on one end, and nonwork or nonschool locations at the other. Two such journeys are included in Fig. 8.6.2, each consisting of half the evening trip chain to the restaurant and movie theater. A decision was made to split such loops at the most distant point from home. *School journeys* and *journeys at school* (undertaken by students) were defined in a manner similar to work-related travel. The proposed OMPO journey-based model [8.44] employs nested logit structures and consists of two components, a travel pattern component and a travel details component. The first predicts primary destination, frequency, and (for work and school journeys) time of day, whereas the second addresses intermediate stops, mode of travel, and (for other journeys) time of day.

8.7 OTHER DEMAND-FORECASTING MODELS

8.7.1 Background

Efforts to improve the modeling of transportation demand have followed two main paths. The first represents a move away from purely empirical models and toward models that are grounded on an improved theoretical understanding of travel behavior. The ultimate test of the appropriateness of these theoretical constructs, however, is not their theoretical eloquence but their ability realistically to describe and predict the real world. The second major thrust of modeling has been toward the discovery of simple models that can facilitate decision making by providing useful information quickly and inexpensively. Of course, the availability of parsimonious models that are also soundly based on a realistic theory of travel behavior is ideal, but theoretical difficulties and practical constraints often prevent the attainment of this ideal.

In earlier sections we surveyed alternative formulations for each of the four components of travel behavior, for example, trip generation, trip distribution, modal choice, and trip assignment. In addition, two alternative estimation approaches (aggregate or disaggregate) and two alternative choice theories (choice-specific and choice-abstract) were discussed. Additional modeling options that are available to the transportation planner are presented next.

8.7.2 Demand-Model Consistency

Model consistency is an important consideration relating to modeling the demand for transportation. Consider, for example, the difference between free and capacity-restrained traffic assignment models. As explained earlier, a free assignment allocates the interzonal flows based on the free-flow interzonal impedances associated with alternative routes between zones. Because the impedances implied by the assigned flows could be significantly different from their assumed values, various capacity-restrained algorithms have been developed to ensure internal consistency between the two sets of interzonal impedances.

The question of consistency between the four steps of the sequential process was also addressed. For instance, it was argued that a conformance must be sought between the interzonal impedances used by the trip-distribution phase and those that result from trip assignment; as a matter of fact, several studies have introduced a feedback link between the two models for this purpose. Feedback loops are also clearly evident in Fig. 8.1.1, which represents standard practice. Figure 8.7.1 illustrates that iterating among all four steps in search of overall model consistency is more difficult and resource-intensive than it may seem [8.45].

8.7.3 Simultaneous or Direct Demand Formulations

A related travel-demand theory states that an individual makes travel choices simultaneously rather than in a sequence of discrete steps and that a demand model should be calibrated to reflect this behavior. An often-cited example of *simultaneous* models is the

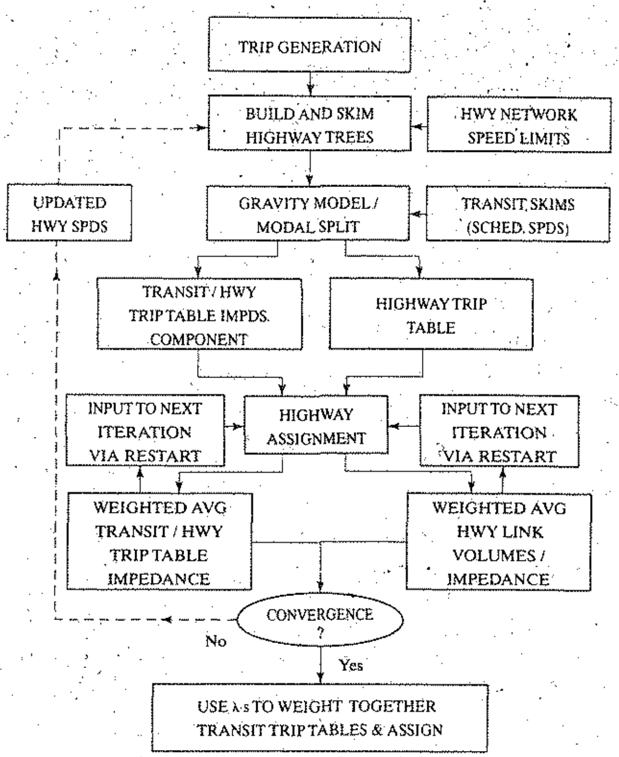

TRIP GENERATION
BUILD AND SKIM HIGHWAY TREES
HWY NETWORK SPEED LIMITS
UPDATED HWY SPDS
GRAVITY MODEL / MODAL SPLIT
TRANSIT SKIMS (SCHED. SPDS)
TRANSIT / HWY TRIP TABLE IMPDS. COMPONENT
HIGHWAY TRIP TABLE
INPUT TO NEXT ITERATION VIA RESTART
HIGHWAY ASSIGNMENT
INPUT TO NEXT ITERATION VIA RESTART
WEIGHTED AVG TRANSIT / HWY TRIP TABLE IMPEDANCE
WEIGHTED AVG HWY LINK VOLUMES / IMPEDANCE
CONVERGENCE ?
No
Yes
USE λs TO WEIGHT TOGETHER TRANSIT TRIP TABLES & ASSIGN

Figure 8.7.1 Iterative procedure among the four steps.
Source: Walker and Peng [8.45].

Quandt and Baumol [8.46] formulation of intercity travel demand, which, using the notation of this book, takes the general form

$$Q_{IJK} = a_0(P_I)^{a_1}(P_J)^{a_2}(C_{IJ*})^{a_3}\left(\frac{C_{IJK}}{C_{IJ*}}\right)^{a_4}(H_{IJ*})^{a_5}\left(\frac{H_{IJK}}{H_{IJ*}}\right)^{a_6}\left(\frac{D_{IJK}}{D_{IJ*}}\right)^{a_7}(Y_{IJ})^{a_8} \qquad (8.7.1)$$

where

Q_{IJK} = travel flow between cities I and J via mode K

P_I, P_J = populations of I and J

C_{IJ}* = least cost of travel between I and J

C_{IJK} = cost via mode K

H_{IJ} = shortest travel time between I and J

H_{IJK} = travel time via mode K

D_{IJ}* = departure frequency of the most frequent mode

D_{IJK} = departure frequency of mode K

Y_{IJ} = weighted average incomes of I and J

$a_0, \ldots, a_8$ = calibration parameters

This model is a simultaneous trip-generation/trip-distribution mode choice equation employing land-use variables (populations), socioeconomic characteristics (income levels), and interzonal impedances by mode (costs, travel times, and frequency of service) to estimate the interzonal demands by mode (Q_{IJK}). In keeping with the purpose of the demand-estimating methodology, these interzonal flows would presumably be assigned to the networks of the modes K serving the region to find the equilibrium link flows. The question of consistency raised earlier between the assumed levels of some of the explanatory variables (e.g., travel times) and the levels implied in the results of the assignment phase resurfaces.

In urban situations the calibration and application of such large models is, to say the least, cumbersome. However, they may be useful for rather coarse estimates at the regional level if the number of zones and the degree of detail in specifying the transportation network are kept to a minimum.

8.7.4 Combined Modeling Strategies

Between the two extremes of sequential model arrangements and large simultaneous models there exist a plethora of options that are partly sequential and partly simultaneous. The following excerpt from a modeling undertaking in Canberra, Australia [8.47], which was intended to be sensitive to the short-term effects of various TSM-type options, illustrates this point:

> If all travel choice decisions for all purpose groups were to be modelled, the scope of modelling required would be extensive and consequently very expensive. It is however possible to reduce the scale and range of models by making some *a priori* assumptions as to the travel choice processes exercised by individual travellers or potential travellers. For example it seems reasonable to assume, certainly for the short term, that workplace and schooling location and the frequency of work and school trips are relatively stable and do not vary significantly under the range of practical conditions which are likely to occur. In these particular cases travel mode choice would therefore appear to be the most important travel decision. Using these and similar reasonings for other purpose categories, a listing of the necessary models was prepared. Specifically these models were:
>
> **(a)** work, mode choice;
> **(b)** school, mode choice;
> **(c)** shopping, mode choice;
> **(d)** shopping, frequency;
> **(e)** shopping, destination;
> **(f)** shopping, mode choice/destination;
> **(g)** social/recreation, mode choice;
> **(h)** social/recreation, destination; and
> **(i)** social/recreation, mode choice/destination.
>
> Other purpose groups were reasoned to be either less significant in terms of scale than the above purpose groups or they were to be less responsive to the specific policy measures, which can be manipulated by Canberra transport planners ([8.47, p. 62]).

Models (d) and (e) are trip-generation and trip-distribution models, respectively, calibrated for shopping trips that could be chained in the sequence d-e-c. Model (f) is a simultaneous modal choice and trip-distribution model that may be applied after estimating trip generation [i.e., model (d)]. Similarly (g) and (h) are modal choice and trip-distribution models, respectively, and (i) is a simultaneous model of these two choices.

Careful contemplation of the quoted terse statement will reveal the wide variety of modeling choices that are available to the contemporary transportation planner, the need to tailor specific models and model arrangements to particular levels of planning and policy issues, and the importance of professional judgment in this most important phase of transportation planning.

One common form of simultaneous models of trip distribution and modal choice is the following *share model:*

$$Q_{IJK} = P_I \frac{A_J e^{U_{IJK}}}{\sum_{X,Y} A_X e^{U_{IXY}}} \tag{8.7.2}$$

Note that the inputs to this model are the exogenously estimated zonal trip productions P_I, zonal trip attractiveness A_J, and interzonal modal utilities U_{IJK}. The output of this simultaneous logit model consists of interzonal flows by mode. Any change in the values U_{IJK} is permitted to affect both the trip distribution and the modal shares of the interzonal demand at the same time. As discussed in Section 8.5, nested logit models can be used to capture simultaneously the combined effects of many travel decisions.

8.7.5 Models of Demand Elasticity

Many planning situations are concerned with immediate or short-term actions or with relatively small changes to the system that do not warrant an elaborate and detailed analytical treatment. Several simplified methods have been developed for this purpose. Simple models are also appropriate to planning studies for rural areas, small- and medium-size urban areas, and certain elements of planning for larger cities.

A transportation-demand model that can be used to provide broad predictions of the response of trip-makers to changes in the transportation system is founded in the economic concept of price elasticity of demand.

Definition of price elasticity of demand. In economic theory the law of demand states that, everything else being constant, the quantity Q of goods or services that consumers demand decreases as their price P increases, and conversely, when the price is reduced, the quantity demanded rises. Figure 8.7.2 illustrates this law. The parallel to transportation viewed as a service that is subject to market forces is inescapable. For example, a patronage drop should be expected to occur following an increase in transit fares. Similarly lowering the downtown parking fees should encourage an increase in automobile use.

The price elasticity of demand, E, is defined as

$$E = \frac{dQ/Q}{dP/P} = \frac{dQ}{dP}\frac{P}{Q} \tag{8.7.3}$$

In words, it is the ratio of the *relative* change in the quantity demanded to the *relative* change in price.

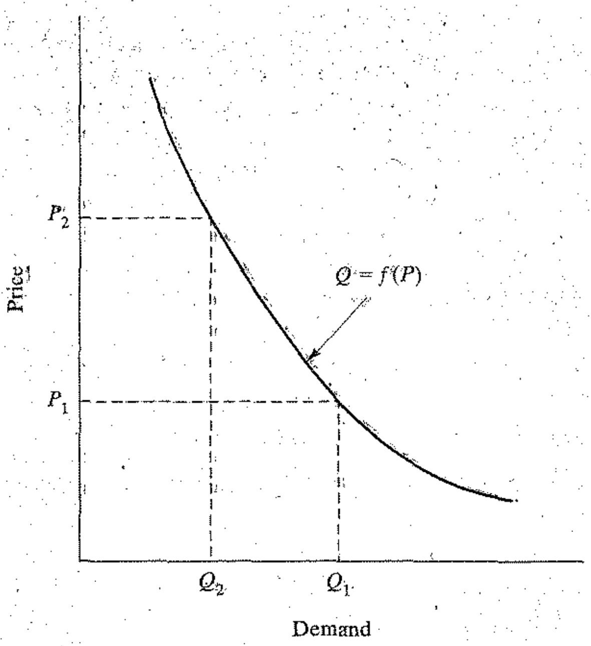


Figure 8.7.2. A hypothetical demand curve.

Example 8.18: Linear Demand Functions

Given a demand function of the form

$$Q = a - bP \tag{8.7.4}$$

express the price elasticity of demand as a function of price.

Solution Applying Eq. 8.7.3, we obtain

$$E = \frac{-Pb}{Q}$$

Substituting Eq 8.7.4 in this gives

$$E = \frac{-Pb}{a - bP}$$

Discussion The negative sign of elasticity reflects the fact that a percentage increase in P will cause a percentage decrease in Q. The solution illustrates that, depending on the demand function, the price elasticity of demand is not constant for all points on the curve. In addition, the value of the price elasticity of demand reflects the implication of a price change on the total revenue (PQ) of the supplier. For example, when $E < -1$, the percent decrease in Q (i.e., the numerator of Eq. 8.7.3) is larger than the percent increase in P (i.e., the denominator of Eq. 8.7.3). In that case the demand is said to be *elastic* and the total revenue after the price increase decreases because the loss of sales volume outweighs the extra revenue obtained per unit sold. When $E > -1$, the demand is said to be *inelastic* and the total revenue after raising P increases. When $E = -1$, the demand is *unitarily elastic* and the revenue derived from selling less units at a higher price is equal to the total revenue prior to raising the price, for example, more units at a lower price. Thus an upward or downward price change may result in an increase, a decrease, or a constancy of revenue. The value of the price elasticity of demand reflects this fact.

Example 8.19: Product Forms

The demand for a particular transit service has been assessed to be a function of fare F and travel time T as follows:

$$Q = aF^bT^c \tag{8.7.5}$$

Calculate the elasticity of demand with respect to (a) fare and (b) travel time.

Solution The two elasticities can be computed via Eq. 8.7.3, except that partial derivatives should be taken.

The fare elasticity of demand is

$$E_f \frac{\partial Q}{\partial F} \frac{F}{Q} = \left(\frac{F}{Q}\right) abF^{b-1}T^c$$

Substituting Eq. 8.7.5 in this relationship yields

$$E_f = b$$

Similarly the travel-time elasticity of demand is

$$E_t = c$$

Discussion This example illustrates that if the demand function is of the product form 8.7.5, the exponents of the price components represent the elasticity of demand with respect to each component. This is the basic reason that Quandt and Baumol have selected this particular functional form for their simultaneous demand equation (8.7.1). A disadvantage of this form is that it assumes that the elasticities are constant. This may be reasonable for price-level changes near the base-data conditions.

Direct and cross elasticities. So far the discussion of elasticity was confined to the effect of changes in the price of a product on the demand for the *same* product. This type of elasticity is called a *direct* elasticity. On the other hand, a price change in one product often affects the demand for *another* product. Price elasticities reflecting this effect are called *cross* elasticities. Cross effects may be positive or negative depending on whether the two products are complementary or substitutes for each other. An increase in the cost of automobile travel would be expected to cause a decrease in automobile use (direct effect) and an increase in transit patronage (cross effect). Another possible cross effect is a decrease in the demand for automobile tires.

Measurement of elasticities. The discussion of Example 8.19 implies that one way of obtaining elasticity estimates is to apply Eq. 8.7.3 to calibrated demand models such as those described earlier and Eqs. 8.7.1 and 8.7.2. Note that if certain important cost or level-of-service variables are not included in a particular model, their associated effect on demand cannot be derived. It would be distributed among the variables included in the model. Also, if a calibrated model is not readily available, this method of elasticity estimation may be very expensive and time-consuming.

Another method of obtaining estimates of elasticities is by observing the effects of actual price changes in the system under study. When undertaking this task, special care must be taken to account properly for demand changes due to factors other than price, such as secular trends attributable to population and demographic changes. When the observed situation involves changes in more than one cost component, it is important to separate the overall demand response to its corresponding individual components. An excellent treatment of these issues may be found in the work reported by Parody and Brand [8.48] relating to a study of the transit system of the city of Jacksonville, FL.

In that study various elasticities for several user subgroups were derived and used to predict the transit demand effects of alternative fare structures. Incidentally, the practice of calibrating demand models specifically for individual market subgroups (known as *market segmentation*) is a matter of increasing interest to planners, especially as it relates to both tailoring transportation services to the needs of these groups and the investigation of how transportation benefits and costs are distributed among them.

Given an actual price change, say, from P_1 to P_2 as shown in Fig. 8.7.2, and an observed demand change from Q_1 to Q_2, the implicit price elasticity of demand (Eq. 8.7.3) may be approximated in several ways:

1. The *shrinkage ratio*, defined as

$$E_{\text{shr}} = \frac{(Q_2 - Q_1)/Q_1}{(P_2 - P_1)/P_1} \tag{8.7.6}$$

2. The midpoint (or linear) *arc elasticity*, computed as

$$E_{\text{arc}} = \frac{(Q_2 - Q_1)/(Q_1 + Q_2)}{(P_2 - P_1)/(P_1 + P_2)} \tag{8.7.7}$$

3. The *log-arc elasticity*, calculated as

$$E_{\text{logarc}} = \frac{\log Q_2 - \log Q_1}{\log P_2 - \log P_1} \tag{8.7.8}$$

These three measures of elasticity yield approximately equal values for relatively small price changes. For larger differences the shrinkage ratio begins to deviate significantly from the other two.

Example 8.20: Application of Elasticities

Given that the log-arc elasticity of demand is −0.28, calculate the effect of an increase in transit fares from 50 to 80¢ given that the patronage prior to the price increase is 20,000 riders per day.

Solution Equation 8.7.8 yields

$$-0.28 = \frac{\log Q_2 - \log 20{,}000}{\log 80 - \log 50}$$

Solving for Q_2 gives

$$Q_2 = 17{,}534 \text{ riders per day}$$

Discussion Since the given direct elasticity was of the log-arc form, Eq. 8.7.8 was applied. To illustrate the differences between the three measures of elasticity, the shrinkage ratio and the arc elasticity implied in the preceding results may be computed. Equation 8.7.6 with $Q_2 = 17{,}534$ yields $E_{\text{shr}} = -0.21$, and Eq. 8.7.7 gives $E_{\text{arc}} = -0.29$. As stated previously, the shrinkage ratio tends to deviate from the other two measures.

Example 8.21: Multiple Price Changes

The shares of the automobile and a transit mode along a corridor are 4500 and 1000 persons per peak period, respectively. The prevailing out-of-pocket costs and travel times associated with the two modes are as follows:

	Time (min)	Cost
Auto	35	$2.00
Transit	50	1.00

The shrinkage ratios with respect to transit prices have been estimated as:

	Time (min)	Cost
Auto	0.05	0.04
Transit	−0.52	−0.30

In other words the direct elasticity of transit demand with respect to transit travel time is −0.52, the cross elasticity of auto demand with respect to transit cost is 0.04, and so forth. The shrinkage ratios with respect to auto prices are:

	Time (min)	Cost
Auto	−0.58	−0.20
Transit	0.12	0.03

The city department of transportation services is contemplating the opening of an exclusive bus lane that would save an average of 10 min per trip. At the same time the city council is holding public hearings on a proposal to raise downtown parking rates and thus cause an increase in automobile travel costs to $2.60. Estimate the likely effects on peak-hour travel demand of both actions combined.

Solution The direct effect of reducing the transit travel time on transit patronage may be predicted by Eq. 8.7.6, which may be rewritten as

$$Q_2 - Q_1 = \frac{E_{\text{shr}} Q_1 (T_2 - T_1)}{T_1}$$

$$= \frac{(-0.52)(1000)(40 - 50)}{50}$$

$$= 104$$

Thus the transit patronage would increase by 104 peak-hour trips, or by 10.4% in relation to the initial demand of 1000, as a result of a 20% decrease in travel time. Because of the way in which shrinkage ratios are defined, the same answer could have been obtained by proportioning the given shrinkage ratio. Specifically, this ratio states that a 1% increase (decrease) in travel time would result in a 0.52% decrease (increase) in transit patronage. Since the contemplated decrease in travel time is 20%, the percentage of increase in patronage would be 20 × 0.52 = 10.4% of the original demand, or 1000 × 0.104 = 104 trips.

The remaining likely direct and cross effects are summarized next:

	Demand changes	
Action	Automobile	Transit system
20% Transit-time reduction	−45	+104
30% Auto cost increase	−270	+9
Combined effects	−315	+113

Discussion In combination the two actions would result in a decrease of 315 auto person-trips per peak and an increase in transit patronage of 113. The available information precludes any definite answers to the question of what would happen to the net loss of 202 peak-hour person-trips, for example, whether they would shift to another time period, would be given up, or would shift to another mode that is not included in the analysis. The question of internal consistency discussed in Section 8.7.2 merits consideration: The predicted decrease in auto trips may in fact cause a reduction in auto travel times, whereas the added transit patronage may induce the scheduling of additional departures, thus decreasing the average waiting times, and consequently travel times. These changes may have further direct and cross effects on the peak-hour demand before a new equilibrium is reached.

8.8 SUMMARY

In this chapter we presented the fundamental elements of the standard four-step demand-forecasting process, the purpose of which is to predict how a proposed regional transportation system will be used at some future time. The process is driven by scenaria describing the distribution of future land uses and socioeconomic characteristics between small analysis zones in the region and a description of a regional multimodal transportation system. Consequently it constitutes a conditional prediction of future demand given these inputs. The process is called *sequential* because it applies a chain of models in sequence to predict the number of trips that each zone will either produce or attract (trip generation), the interchange volumes between pairs of zones (trip distribution), the shares of interzonal volumes that will use each of the available travel modes (mode choice), and the allocation of interzonal trips to the interzonal paths provided by the transportation network (network assignment). The resulting interzonal volumes by mode can be translated to link flows in order to aid in the assessment of the ability of the transportation system to accommodate the demands that will be placed on it.

Given a land-use pattern, several proposed transportation system alternatives (including the do-nothing alternative) are typically examined. The outputs of the demand-forecasting models for each alternative are included among the consequences or impacts (see Chapter 10) that enter the process of system evaluation and selection (see Chapter 11). The demand-forecasting process can also aid in predicting the transportation consequences of land-use changes and in providing guidance to related public policies.

The transport behavior of individuals and households was examined as well. Analyses based on disaggregate units such as individuals and/or households offer insights on the effects of personal and household characteristics on travel behavior and trip patterns. Factors ranging from household structure and income to lifestyles and personality were addressed along with applications of household variables in automobile-ownership and trip-generation modeling.

Principles of simultaneous model structures, demand elasticity models (which are particularly useful in assessing the effects of price or service changes on transport demand), and strategies that combine a wide spectrum of transport-demand-forecasting methodologies were presented as well.

Finally, the fact that large-scale transportation-demand forecasting is a monumental enterprise must not escape the attention of the reader. The task is facilitated by the availability of computer software packages that can be used for the calibration and application of the models described in this chapter; these are covered in Chapter 15.

EXERCISES

1. An origin-destination survey in ten travel-analysis zones provided the following data relating to zonal residential densities (households/acre) and average daily trip productions per household. Calibrate and plot a model of the form $10^Y = AX^{-B}$.

Density X	42	5	25	10	4	15	20	12	14	22
Trip rate Y	1.5	4.0	2.1	2.6	4.8	2.0	2.5	3.3	1.9	2.0

2. Prepare an essay describing the major factors that affect your choice of travel mode for different purposes.

3. A zone in the CBD is projected to contain 1,525,000 ft^2 of residential space; 3,675,000 ft^2 of service establishments; and a total retail activity floor area of 2,100,000 ft^2. Government and other public buildings occupy a total area of 615,000 ft^2. Using the data obtained in Pittsburgh, calculate the trip generation of this zone.

4. An international hotel chain is planning the construction of a motel/office development in a resort town. The preliminary design includes 2100 rooms, a sit-down restaurant having a total floor space of 2500 ft^2, and 5000 ft^2 of office space, which the company is planning to lease to various local firms. Apply the trip rates published by the Institute of Transportation Engineers to estimate the total trip attractions during the afternoon traffic peak hour. (Instructor must furnish tables or formulas.)

5. A high-rise apartment building containing 350 units is planned for a residential area of a city. Because the area is zoned for low-density residential land uses, the developer has applied for a zoning variance. At the legally required public hearing several residents of the area have opposed the zoning change, claiming that the proposal will add to the traffic-congestion problem during the peak hours, but they were unable to substantiate their claim. Calculate the likely peak-hour trip generation of the proposed project. (Instructor must furnish tables or formulas.)

6. Use the cross-classification table (Table 8.2.3) to calculate the total non-work-home-based productions of each of the zones that are expected to contain the following mixtures of households (HH):

Zone 1: Suburban

Veh/HH Persons/HH	0	1	2+
1	50	150	100
2, 3	10	500	300
4	100	400	100

Zone 2: Rural

Veh/HH Persons/HH	0	1	2+
1	300	50	100
2, 3	100	200	100
4	400	300	150

7. A residential zone is expected to have 1500 dwelling units. For a $12,000 average income, calculate (a) the person-trips per dwelling unit for units that own 0, 1, 2, and 3+ autos and (b) the total-trip generation by trip purpose. Use the income–auto-ownership distribution given in Fig. 8.2.5 and assume that curve C applies to all subgroups within the zone.

8. Given

1.

Zone	Productions	Attractiveness
1	1000	2
2	0	5
3	2000	1

2. W_{IJ}:

J I	1	2	3
1	5	20	10
2	20	5	10
3	10	10	5

K_{IJ}:

J I	1	2	3
1	1.1	1.5	0.8
2	0.6	1.2	0.5
3	1.0	1.4	1.3

3. $\ln F = -1.5 \ln W$

Apply the gravity model to calculate all interchange volumes.

9. Complete the following table given that $P_1 = 1000$ trips per day, $C = 2.0$, and all $K_{IJ} = 1.0$.

Zone	A_J	W_{IJ}	F_{IJ}	Q_{IJ}
1	0	2		
2	400	20		
3	300	5		
4	100	5		
5	200	10		

10. Assuming that the relationship between F and W is of the form $F = AW^{-c}$, apply the method of least squares to the following data to estimate the parameters A and c.

F	0.03	0.04	0.02	0.03
W	7	5	12	8

11. The final iteration in a calibration of the gravity model yielded the following friction-factor and impedance values:

F	1.0	4.0	0.5	0.3
W	12	4	15	20

(a) Calibrate a relationship of the form $F = aW^{-b}$.

(b) Apply your results to the following case: Two residential zones (1 and 2) are expected to produce 6500 and 3800 person-trips per day, respectively. Two nonresidential zones (3 and 4) are competing for these trips. The planning commission has received a proposal to improve parts of the transportation system, which, if implemented, would affect certain interzonal impedances as shown:

Do-nothing:

J I	3	4
1	10	14
2	8	14

Proposed plan:

J I	3	4
1	10	10
2	8	8

Given the following additional information, calculate the effect of the proposal on the total trips attracted by the nonresidential zones.

$$A_3 = 10 \qquad A_4 = 15 \qquad \text{all } K_{IJ} = 1.0$$

12. A base-year trip-generation study obtained the data shown relating to the daily person-trip productions per dwelling unit (Y) and residential density (X dwelling units per acre).

Y	3.5	6.5	4.0	2.2
X	30.0	10.0	50.0	70.0

(a) Calibrate and plot the relationship $Y = (a + bX)^{-1}$.

(b) Apply your answers to part (a) to the following situation: A residential zone I has an area of 500 acres and contains 7500 dwelling units. Two zones (J and L) are competing for the trips produced by I. Given the following information, calculate the trip interchange volumes Q_{IJ} and Q_{IL} if $W_{IJ} = 12$, $W_{IL} = 8$, $\ln F = -1.5 \ln W$, $A_J = 0.5\, A_L$, and all $K_{IJ} = 1.0$.

13. A gravity-model calibration study obtained the following final values for the travel-time factors and the interzonal impedances:

F	0.19	0.10	0.07	0.05
W	4	7	9	12

Using F as the dependent variable, calculate the parameter c of Eq. 8.3.9.

14. Given the following data:

I	P	A
1	2000	4
2	0	10
3	0	12

W_{IJ}

J I	1	2	3
1	5	10	15
2	10	4	10
3	15	10	15

Estimate all interchange volumes assuming that $c = 1.9$ and that all socioeconomic adjustment factors are equal to unity.

15. After calibrating Eq. 8.3.13 the following interzonal information became available. The observed and calculated interzonal flow interchanging between I and J were 2500 and 2100, respectively. The data corresponding to the interchange between zones I and L were 1960 and 2060. Given that the total production of zone I was 12,000 trips, calculate the socioeconomic adjustment factors for the two interchanges.

16. Computerize the application of the gravity model of trip distribution.

17. Perform a second Fratar model iteration using the results of Example 8.5.

18. Computerize the Fratar model, making sure to place an upper limit on the number of iterations in order to avoid infinite looping.

19. Given the utility equation

$$U_K = a_K - 0.003X_1 - 0.04X_2$$

where X_1 is the travel cost in cents and X_2 is the travel time in minutes.

(a) Calculate the market shares of the following travel modes:

Mode K	a_K	X_1	X_2
Automobile	−0.20	120	30
Express bus	−0.40	60	45
Regular bus	−0.60	30	55

(b) Estimate the effect that a 50% increase in the cost of all three modes will have on modal split.

20. Given the utility expression

$$U_K = A_K - 0.05T_a - 0.04T_w - 0.02T_r - 0.01C$$

where

T_a = access time

T_w = waiting time

T_r = riding time

C = out-of-pocket cost

(a) Apply the logit model to calculate the shares of the automobile mode ($A_K = -0.005$) and a mass transit mode ($A_K = -0.05$) if

Mode	T_a	T_w	T_r	C
Auto	5	0	30	100
Transit	10	10	45	50

(b) Use the incremental logit model to estimate the patronage shift that would result from doubling the bus out-of-pocket cost.

21. Prove Eq. 8.4.17.

22. The application of a route-building algorithm results in the following final tree table:

Node to	Total time	Node from
1	0	—
2	8	3
3	3	1
4	7	3
5	17	6
6	10	2
7	8	4
8	19	5
9	13	6
10	14	9
11	17	10
12	18	9
13	18	10
14	25	12
15	23	13
16	22	13

Sketch the minimum tree, and specify the link travel times.

23. Complete the tree table that describes the minimum tree shown in Fig. E8.23.

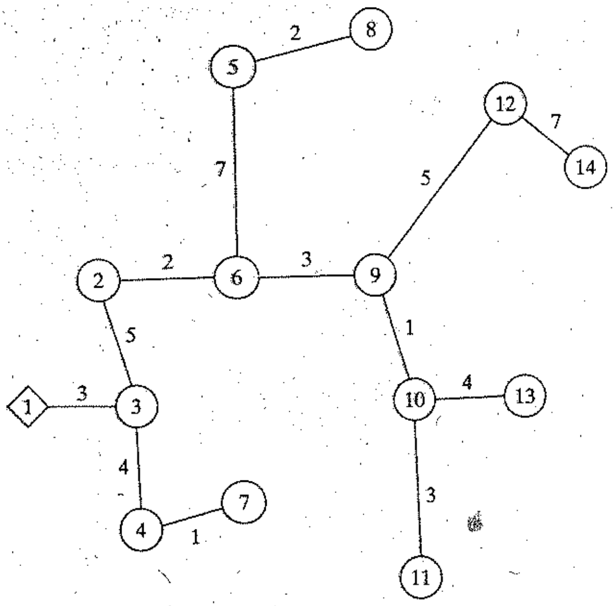


Figure E8.23

24. Without drawing the entire network described by the accompanying link table, find and sketch the minimum tree emanating from node 1.

i	j	w_{ij}
1	4	2
2	5	4
2	6	3
3	8	5
4	1	2
4	5	6
4	8	10
5	2	4
5	4	6
5	6	4
6	2	3
6	5	4
6	7	9
7	6	9
7	8	7
8	3	5
8	4	10
8	7	7

25. Given the link table of Exercise 8.24

(a) Find graphically the minimum tree emanating from node 1.

(b) Using your answer to part (a), calculate Q_{1-2} and Q_{1-3} if $P_1 = 2500$ trips per day, $A_2 = 1.5$, $A_3 = 3.5$, all other P_I and $A_J = 0$, and all $K_{IJ} = 1.0$.

(c) Perform an all-or-nothing assignment to allocate the interzonal volumes calculated in part (b) to the links of the network.

26. Write a computer program for the minimum tree–seeking procedure described in this chapter.

27. Allocate the following peak-hour interchange volumes produced by zone 10 to the network described.

J	11	12	13	14	15	16	17
Q_{IJ}	200	150	190	270	320	110	540

Link Table

i	j	w_{ij}	i	j	w_{ij}
10	14	6	14	13	15
10	15	5	14	15	7
10	16	12	15	10	5
11	14	5	15	11	3
11	15	3	15	14	7
11	17	8	15	16	8
12	14	21	16	10	12
12	16	4	16	12	4
12	17	3	16	13	7
13	14	15	16	15	8
13	16	7	16	17	2
13	17	6	17	11	8
14	10	6	17	12	3
14	11	5	17	13	6
14	12	21	17	16	2

If the capacity of link 10–16 is 1000 veh/h (vehicles per hour) calculate this link's travel time as implied by your answer.

28. A roadway connection is planned between cities A and B, which at present have no direct connection. The cities are approximately 800 mi apart. There is no passenger rail connection between A and B, but there are four commercial flights a day from each city to the other. City A has a population of 250,000, and B has 350,000. The median household income is \$22,000 and \$23,500 in A and B, respectively.

The following model provides estimates on the expected daily volume (DV) of passenger traffic between two cities. Assuming that (1) the average vehicle occupancy will be 1.5 passengers per vehicle, (2) each lane can serve up to 1700 veh/h, and (3) 40% of the expected daily volume will occur in 1 h (peak, or design hour), estimate and decide how many lanes are required (even number):

$DV = 2000 + 30$ population in the largest of the two cities, in thousands

$+ 25$ population in the smallest of the two cities, in thousands

$+ 240$ median household income in the largest of the two cities, in thousands

+ 200 median household income in the smallest of the two cities, in thousands

− 1.5 distance between the two cities, in mi

− 0.005 distance between the two cities squared

− 500 passenger rail connection, 1 if it exists, 0 otherwise

− 150 number of daily commercial flights between the two cities

The standard error of estimate (SEE) for the dependent variable is ±1000.

29. Focus on the model in Exercise 28 and answer the following questions:
(a) What type of model is this?
(b) Would you be confident in using this model for a real application? Why?
(c) Why do population and income have positive contributions to daily passenger traffic, and why do distance and commercial airline flights have a negative contribution? (It may be necessary to consult Chapter 13 to answer this part.)

30. Identify the characteristics of the following three transportation-demand models and indicate their role in the demand-forecasting process.

$$Q_{IJ} = aP_I A_J W_{IJ}^b$$
$$Q_{IJ} = aY_I^b X_I^c W_{IJ}^d$$
$$Q_{IJK} = aX_I^b Y_I^c Z_J^d W_{IJK}^e$$

where

P_I = productions of zone I

A_J = attractions of zone J

W_{IJ} = travel impedance from I to J

Y_I = average income in zone I

X_I = population of zone I

Z_J = total employment in zone J

W_{IJK} = impedance from I to J via mode K

31. Specify and discuss the forecasting model structures that are possible given the models calibrated for Canberra, Australia (see Section 8.7.4).

32. Calculate and interpret the income elasticity of demand in model 2 of Exercise 30 assuming that *a, b, c, d,* and *e* are constants.

33. An increase of transit fares from 40 to 60¢ has resulted in a decrease in transit patronage from 500,000 to 450,000 trips per day. Calculate the shrinkage ratio, the linear-arc elasticity, and the log-arc elasticity.

34. An increase in gasoline prices from $1.00 to $1.30 per gallon resulted in a decrease of automobile use from 1,000,000 to 960,000 trips. Estimate the likely impact of an increase in gasoline prices from $1.30 to $1.50. Solve this problem using the three alternative measures of elasticity.

35. A special service for the elderly and handicapped currently serves 2500 persons per day. Given that the current fare is 50¢ and that the linear-arc fare elasticity of demand is −0.45, calculate

(a) the loss of patronage that would result from doubling the fare, (b) the effect on fare-box revenues, (c) the implied shrinkage ratio, and (d) the implied log-arc elasticity.

36. A 20% increase in automobile costs has been observed to cause a 5% increase in transit patronage relative to the patronage prior to the increase and a 10% decrease in auto usage. Calculate the implied direct and cross elasticities of demand as measured by E_{shr}, E_{arc}, and $E_{\text{log-arc}}$.

37. A zonal interchange is served by a local bus route and an express bus route. The current travel times and fares associated with the two types of service are:

	Travel time (min)	Fare
Local	50	\$0.50
Express	30	1.00

Given the following linear-arc elasticities of demand and that the current transit patronage of 4000 trips per peak period is split 40–60 between the express and local bus services, calculate the effect of raising the express bus fare to \$1.50:

	Local		Express	
	Time	Fare	Time	Fare
Local	−0.02	−0.03	+0.01	+0.02
Express	+0.09	+0.62	−0.08	−0.15

38. For the system of Exercise 37, estimate the effect of expanding the number of express buses and thus reducing the express bus travel time to 25 min.

39. Examine the effect that a 10-min reduction in the travel time offered by the local bus service would have on the total peak-hour transit usage between the two zones described in Exercise 37.

40. Assuming that the elasticities given in Exercise 37 are shrinkage ratios, estimate the combined effect of raising express bus fares to \$1.30 *and* lowering the local bus fare to \$0.40.

41. Repeat Exercise 37 assuming that the given elasticities are log-arc elasticities.

42. Repeat Exercise 37 assuming that the given elasticities are shrinkage ratios.

REFERENCES

8.1 MARTIN, W. A., and N. A. MCGUSKIN, *Travel Estimation Techniques for Urban Planning,* National Cooperative Research Program Report 365, Transportation Research Board, National Research Council, Washington, DC, 1998.

8.2 PARSONS BRINCKERHOFF QUADE & DOUGLAS, *Model Development Review of Best Practices,* Washington, DC, 1992.

8.3 TEXAS TRANSPORTATION INSTITUTE, *Activity-Based Travel Forecasting Conference,* prepared for the U.S. Department of Transportation and the U.S. Environmental Protection Agency, 1997.

8.4 FEDERAL HIGHWAY ADMINISTRATION, *Trip Generation Analysis,* U.S. Department of Transportation, U.S. Government Printing Office, Stock No. 050-001-00101-2, Washington, DC, 1975.

8.5 PAPACOSTAS, C. S., "Honolulu's Handi-Van: Use and Implications," *Traffic Quarterly,* 34, 3 (July 1980): 429–440.

8.6 KEEFER, L. E., Director, *Pittsburgh Area Transportation Study,* vol. I, Study Findings, November 1961.

8.7 FEDERAL HIGHWAY ADMINISTRATION, *Computer Programs for Urban Transportation Planning: PLANPAC/BACK-PAC General Information,* U.S. Department of Transportation, U.S. Government Printing Office, Stock No. 050-001-00125-0, Washington, DC, April 1977.

8.8 OAHU METROPOLITAN PLANNING ORGANIZATION, *Oahu Model Update Study: User's Manual and Training Information,* Honolulu, HI, December 1982.

8.9 ANAS, A., "Discrete Choice Theory, Information Theory, and the Multinomial Logit and Gravity Models," *Transportation Research,* 17B (1983): 13–23.

8.10 FRATAR, T. J., "Forecasting the Distribution of Interzonal Vehicular Trips by Successive Approximations," *Proceedings of the 33rd Annual Meeting,* Highway Research Board, National Research Council, Washington, DC, 1954.

8.11 KANAFANI, A., *Transportation Demand Analysis,* McGraw-Hill, New York, 1983.

8.12 LANCASTER, K. J., "A New Approach to Consumer Theory," *Journal of Political Economy,* 64 (1966): 132–157.

8.13 STOPHER, P. R., *A Probability Model of Travel Mode Choice for the Work Journey,* Highway Research Record 283, Highway Research Board, National Research Council, Washington, DC, 1969, pp. 57–65.

8.14 OPPENHEIM, N., Urban Travel Demand Modeling: *From Individual Choices to General Equilibrium,* John Wiley, New York, 1995.

8.15 HIGHWAY RESEARCH BOARD, *Urban Travel Demand Forecasting,* Special Report 143, Highway Research Board, National Research Council, Washington, DC, 1973.

8.16 HONOLULU RAPID TRANSIT PROGRAM, *Service and Patronage Forecasting Methodology, Final Report,* prepared by Barton-Aschman Associates, Inc. and Parsons Brinckerhoff Quade & Douglas, Inc. for the Department of Transportation Services, City and County of Honolulu, HI, March 1992.

8.17 FEDERAL HIGHWAY ADMINISTRATION, *1990 Nationwide Personal Transportation Survey: Summary and Travel Trends,* Report FHWA-PL-92-027, U.S. Department of Transportation, Washington, DC, 1992.

8.18 MOSKOWITZ, K., *California Model of Assigning Diverted Traffic to Proposed Freeways,* Highway Research Record 130, Highway Research Board, National Research Council, Washington, DC, 1956, pp. 1–26.

8.19 BUREAU OF PUBLIC ROADS, *Traffic Assignment Manual,* U.S. Department of Commerce, U.S. Government Printing Office, Washington, DC, June 1964.

8.20 WARDROP, J. G., *Some Theoretical Aspects of Road Traffic Research,* Proceedings of the Institution of Traffic Engineers, Vol. 1, Part II, London, 1952.

8.21 MOORE, E. F., "The Shortest Path through a Maze," *Proceedings of the International Symposium on the Theory of Switching,* Harvard University, Cambridge, MA, 1957.

8.22 IRWIN, N. A., and H. G. VON CUBE, *Capacity Restraint in Multi-travel Mode Assignment Programs,* Highway Research Board, Bulletin 347, National Research Council, Washington, DC, 1962, pp. 258–289.

8.23 SCHNEIDER, M., *A Direct Approach to Traffic Assignment,* Highway Research Record 6, Highway Research Board, National Research Council, Washington, DC, 1963, pp. 71–75.

8.24 HUMPHREY, T. F., *A Report on the Accuracy of Traffic Assignment When Using Capacity Restraint,* Highway Research Record 191, Highway Research Board, National Research Council, Washington, DC, 1967, pp. 53–75.

8.25 LEVINSOHN, D., et al. *UTPS Highway Network Development Guide,* Federal Highway Administration, U.S. Department of Transportation, Washington, DC, January 1983.

8.26 DIAL, R. A., "A Probabilistic Traffic Assignment Model Which Obviates Path Enumeration," *Transportation Research,* 5, 2 (1971): 83–222.

8.27 HAVENS, J. J., "New Approaches to Understanding Travel Behavior: Role, Life-Style and Adaptation," in *New Horizons in Travel Behavior Research,* edited by P.; TNR. Stropher, A. H. Meyburg, and W. Brög, Lexington Books, Lexington, MA, 1981.

8.28 SALOMON, I., "Life-Styles: A Broader Perspective on Travel Behaviour," in *Recent Advances in Travel Demand Analysis,* Gower, Aldershot, Hampshire, England, 1983.

8.29 BEN-AKIVA, M., and S. LERMAN, *Discrete Choice Analysis: Theory and Application to Travel Demand,* MIT Press, Cambridge, MA, 1985.

8.30 PREVEDOUROS, P. D., *Demographic, Social, Economic and Personality Factors Affecting Suburban Transport Behavior,* Ph.D. dissertation, Department of Civil Engineering, Northwestern University, Evanston, IL, 1990.

8.31 TOWNSEND, T. A., *The Effects of Household Characteristics on the Multi-day Time Allocations and Travel Activity Patterns of Households and Their Members,* Ph.D. dissertation, Department of Civil Engineering, Northwestern University, Evanston, IL, 1987.

8.32 KOPPELMAN, F. S., and E. I. PAS, *Travel-Choice Behavior: Models of Perceptions, Feelings, Preference and Choice,* Transportation Research Record 765, Transportation Research Board, National Research Council, Washington, DC, 1980, pp. 26–33.

8.33 KOPPELMAN, F. S., and P. K. LYON, "Attitudinal Analysis of Work/School Travel," *Transportation Science,* 15, 3 (1981).

8.34 KITAMURA, R., "A Panel Analysis of Household Car Ownership and Mobility," *Proceedings of Japan Society of Civil Engineers,* No 383/IV-7 (1987): 13–27.

8.35 GOLOB, T. F., "The Dynamics of Household Travel Time Expenditures and Car Ownership Decisions," presented at the *International Conference on Dynamic Travel Behavior Analysis,* Kyoto University, Kyoto, Japan, July 18–19, 1989.

8.36 JONES, P. M., F. S. KOPPELMAN, and J. P. ORFEUIL, "Activity Analysis: State-of-the-Art and Future Directions," in *Developments in Dynamic and Activity-Based Approaches to Travel Analysis,* Avenbury, London, 1990.

8.37 GOULIAS, K. G., and R. KITAMURA, *Recursive Model System for Trip Generation and Trip Chaining,* Transportation Research Record, 1236, Transportation Research Board, National Research Council, Washington, DC, 1989, pp. 59–66.

8.38 GOULIAS, K. G., R. M. PENDYALA, and R. KITAMURA, *Practical Method for the Estimation of Trip Generation and Trip Chaining,* Transportation Research Record, 1285, Transportation Research Board, National Research Council, Washington, DC, 1990, pp. 47–56.

8.39 KOPPELMAN, F. S., and G. ROSE, "Geographic Transfer of Travel Choice Models: Evaluation and Procedures," *Proceedings of the International Symposium on New Directions in Urban Modelling,* University of Waterloo, Canada, July 1983.

8.40 KARASMAA, N., and M. PURSULA, "Empirical Studies of Transferability of Helsinki Metropolitan Area Travel Forecasting Models," *Transportation Research Record 1607,* National Research Council, Washington, DC, 1997, pp. 38–54.

8.41 ETTEMA, D., A. BORGERS, and H. TIMMERMANS, "Simulation Model of Activity Scheduling Behavior," *Transportation Research Record 1413,* National Research Council, Washington, DC, 1993, pp. 1–11.

8.42 SCHULTZ, G. W., and W. G. ALLEN, "Improved Modeling of Non-Home-Based Trips," *Transportation Research Record 1556,* National Research Council, Washington, DC, 1996, pp. 22–26.

8.43 PAPACOSTAS, C. S., and G. G. W. LUM, "Honolulu's Land Use and Travel Demand Model Update Project," Proceedings of the ITE Regional Conference, Melbourne, Australia, Institute of Transportation Engineers, 1995.

8.44 RYAN, J., "Journey-Based Travel Models for Honolulu," Invited Presentation, 78th National Meeting of the Transportation Research Board, National Research Council, Washington, DC, 1999.

8.45 WALKER, W. T., and H. PENG, "Alternative Methods to Iterate a Regional Travel Simulation Model: Computational Practicality and Accuracy," *Transportation Research Record 1556,* National Research Council, Washington, DC, 1995.

8.46 QUANDT, R. E., and W. J. BAUMOL, "The Demand for Abstract Modes—Theory and Measurement," *Journal of Regional Science,* 6, 2 (1966): 13–26.

8.47 WIGAN, M. R., (Ed.), *New Techniques for Transport Systems Analysis,* Special Report 10, Australian Road Research Board and Bureau of Transport Economics, Vermont, Victoria, 1977.

8.48 PARODY, T. E., and D. BRAND, *Forecasting Demand and Revenue for Transit Prepaid Pass and Fare Alternatives,* Transportation Research Record 719, Transportation Research Board, National Research Council, 1979, pp. 35–41.

PART 3

Transportation Impacts

9

Traffic Impact and Parking Studies

9.1 INTRODUCTION

In this chapter we provide an overview of two common types of traffic engineering studies: traffic impact studies and parking studies. The former assess the impact of proposed new developments or expansions of existing developments on traffic networks (i.e., impact of anticipated traffic on existing road networks surrounding the site). The latter focus on the analysis of needs and the principles of the design of parking facilities.

9.2 TRAFFIC IMPACT STUDIES

9.2.1 Background

Few people realize the connection between traffic and proposed developments. Why should transportation analysis be necessary in the design of a shopping mall, a large apartment building, a new employment center, and so forth? Perhaps these developments require some internal circulation design and parking lots or parking structures. The role of transportation analysts, however, goes far beyond traffic circulation and parking analysis.

New or expanded developments generate new or additional traffic: new shoppers, new residents, and new employees in the examples mentioned earlier. In general, developments attract people because certain needs of people can be fulfilled in a new development. For example, the new development may include a supermarket that is larger or closer to the one previously used (the underlying need is household maintenance), or the new development may host offices, restaurants, and movie theaters (the underlying needs are work and entertainment) that would attract people. Users of these facilities would travel to the new development by various travel modes (i.e., driving, sharing a ride, using public transportation, biking, walking, or a combination of them).

To quantify the transportation implications of new developments, the new or additional trips have to be estimated; their origins and destinations must be determined and the modes and routes selected have to be established. Furthermore, traffic is dynamic. Directional flows vary by the time of day and with time. New transportation facilities and services are implemented or terminated over time. Thus the characteristics of the generated trips and their impact on traffic must be forecast and assessed. This requires specialized traffic engineering knowledge.

9.2.2 Basic Characteristics

Traffic impact studies (TIS) are necessitated by the increasing levels of congestion in growing areas, particularly those that are located within the boundaries of large urban areas. In an attempt to control unplanned growth and unmanageable loads of traffic, traffic impact studies became a requirement to examine whether the road network surrounding a proposed development will be able to handle the additional traffic while still offering acceptable levels of service (e.g., performance at level C or better). TIS are required by municipalities or counties that individually determine the acceptable levels of intersection performance. For example, in some locations LOS D is acceptable for peak periods. Typically the TIS is a part of the environmental impact statement (EIS) of the proposed development. Public projects such as an extensive roadway widening, a new airport terminal, and so on, may also be required to submit an EIS.

Traffic impact studies provide answers to the following questions:

1. What are the existing traffic conditions on the network surrounding the proposed development?
2. How much additional traffic will be generated by the proposed development?
3. How will additional traffic affect existing conditions?
4. What roadway improvements or changes in the site plan would be necessary to minimize the traffic impact of the proposed development?

Traffic impact studies are not required for all developments. A traffic impact study may be necessary if at least 100 new inbound (i.e., entering the site) or outbound (i.e., exiting from the site) trips are generated along the peak direction of traffic during the peak hour of the existing traffic. This corresponds to developments of substantial size: 160 single family houses, or 220 multifamily units, or 10,000 ft^2 of retail space, or 60,000 ft^2 of office space. Smaller developments that (1) may generate traffic safety hazards or (2) are located in traffic congested areas may necessitate a traffic impact study. Municipalities may request a TIS for reasons other than the ones mentioned before. The roles of developers, traffic consultants, and government agencies vary between localities.

Several assumptions are put forth in traffic impact studies, particularly for large developments, because proposed developments may be fully operational in the near future, whereas their impact will be felt over a longer period.

The following are some of the issues that need resolution:

- Determination of the appropriate size of the study area surrounding the proposed development
- Peak periods to be analyzed (i.e., morning, noon, evening, weekends, etc.)

- Time frame of analysis, which includes the base year when the proposed facility opens and the target or horizon year for assessment of impacts
- Anticipated background growth (e.g., overall growth of the surrounding area)
- Identification of committed transportation improvements or changes until the horizon year
- Study methodology (i.e., analysis of intersections using a specific version of HCM (e.g., [9.1, 9.2] or other method, utilization of nationwide or area-specific rates)
- Requirements for additional analyses (e.g., accidents, sight distance, gap availability, weaving, queuing, etc.)

The size of the development largely dictates the number of intersections to be analyzed (i.e., two to four intersections for a 300,000 ft^2 strip mall), which, in turn suggests the equipment to be utilized. Small developments can be analyzed "by hand"; computers are used for simple bookkeeping tasks and calculations. Large developments are typically analyzed with the aid of personal computers. Traffic network flow models are utilized for the proper assignment of flows on the network and the faster repetition of calculations for future scenarios. Also, for large developments traffic impact studies are required during various phases of the design and issuance of requisite permits because the final site plan is likely to be different from the original plan (e.g., redesign to respond to new economic or demand trends, public reaction, etc.) Computer processing of the traffic impact study vastly improves responsiveness and efficiency.

The basic inputs and outputs of TIS are as follows: *inputs:* traffic volumes by movement and direction, link volumes, network characteristics (street system: link lengths, capacities, signal timings), land-use patterns, trip-generation rates, and distribution by direction (i.e., percent of traffic entering and leaving the site); *outputs:* link and intersection volumes, capacity and performance analysis of intersections (usually with the HCM method; see Chapter 4), and recommendations for improvements within the site and of the surrounding network.

9.2.3 Overview of Steps

The methodology of a traffic impact study includes the following steps [9.3]:

1. Meetings with the municipality to discuss the scope and extent of the study and the assumptions and horizon years to be used for analysis. At this stage the municipality may agree to provide data on traffic volumes, signal timings, accidents, and planned transportation improvements.
2. Field surveys, which usually include the following observations: detailed reconnaissance of the project site, roadway network in the area, traffic control devices, signal phasings and timings at signalized intersections, roadway geometrics, parking regulations, transit routes and stops, adjacent land uses, and driveway locations.
3. Traffic counts and surveys. The standard traffic counts (detailed in Chapter 4) are conducted during the agreed upon time periods, days, and seasons. Surveys of pass-by motorists, neighboring residents, and employees may be collected for guidance in modal split, trip distribution, and traffic assignment.

4. Analysis, which consists of the following tasks listed in the order in which they are often conducted:
   - a. Trip generation: estimation of trip rates and application of modal split based on experience (historical data in the area) or from local surveys
   - b. Trip distribution
   - c. Estimation of nonsite traffic. The impact of such traffic should not be attributed to the site studied; otherwise unreasonable demands for roadway improvements or downsizing of the development may be placed on the developer of the project under study
   - d. Assignment of site and nonsite trips on the roadway network based on a network equilibrium rule or an empirical technique*
   - e. Capacity and performance analysis of signalized and unsignalized intersections
   - f. Evaluation of results and recommendations for improvements
5. Revisions of the site plan and incorporation of selected improvements
6. Production of reports for the client and the municipality

9.2.4 Major Components of Traffic Impact Studies

In this section we focus on components of a traffic impact study, which require specific analyses. These components are trip-generation, modal split, trip-distribution, traffic assignment, and intersection analyses. Methodologies of these fundamental steps of transportation analysis are detailed in Chapters 4 and 8. In this section we present some basic examples of the application of these transportation analysis techniques to conduct traffic impact studies.

Trip generation. Trip-generation estimation is conducted separately for site and nonsite traffic. Nonsite traffic includes all through traffic that has neither origin nor destination at the site as well as the traffic generated by developments within the study area, but outside the specific site under analysis. In the absence of locally derived rates ITE's *Trip Generation* manual [9.4] is utilized for assessing site and nonsite traffic. An example of the form of the data in this manual is given in Fig. 9.2.1.

The *Trip Generation* manual employs simple additive or multiplicative models of the form

$$T = a + bX \qquad \text{(additive)} \tag{9.2.1}$$

$$\ln T = a + b \ln X \quad \text{(multiplicative; original form } T = a'X^b\text{)}$$

where

T = total number of generated trips

X = total GFA or GLA or another characteristic of the site

a, b = given model parameters

GFA = gross floor area

GLA = gross leasable area

*New transportation infrastructure, planned or under construction, is accounted for here and in task 4e.

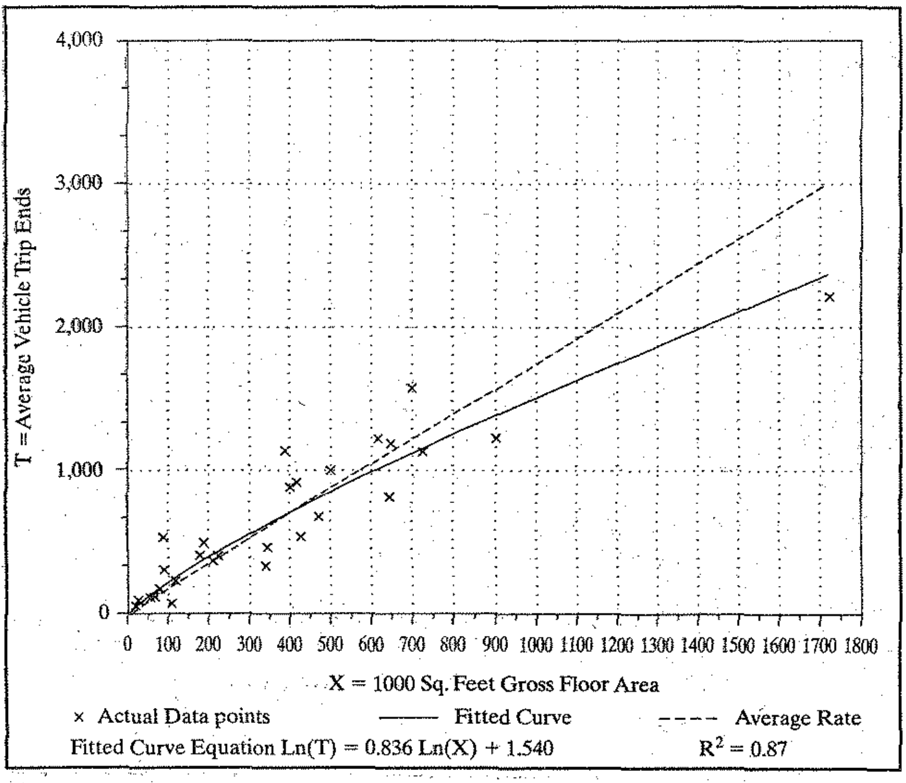
Office Park
(750)

Average Vehicle Trip Ends vs:	1000 Sq. Feet Gross Floor Area
On a:	Weekday, A.M. Peak Hour
Number of Studies:	29
Average 1000 Sq. Feet GFA:	372
Directional Distribution:	89% entering, 11% exiting

Trip Generation Per 1000 Sq. Feet Gross Floor Area

Average Rate	Range of Rates	Standard Deviation
1.74	0.72 - 5.89	1.46

Data Plot and Equation



Fitted Curve Equation Ln(T) = 0.836 Ln(X) + 1.540 $R^2 = 0.87$

Figure 9.2.1 Example of trip rates (office park development). (Reprinted with permission from *Trip Generation*, 6th ed., © 1997 Institute of Transportation Engineers, Washington, DC.)

Office Park
(750)

Average Vehicle Trip Ends vs:	1000 Sq. Feet Gross Floor Area
On a:	Weekday, P.M. Peak Hour
Number of Studies:	31
Average 1000 Sq. Feet GFA:	370
Directional Distribution:	14% entering, 86% exiting

Trip Generation per 1000 Sq. Feet Gross Floor Area

Average Rate	Range of Rates	Standard Deviation
1.50	0.73 - 4.50	1.32

Data Plot and Equation

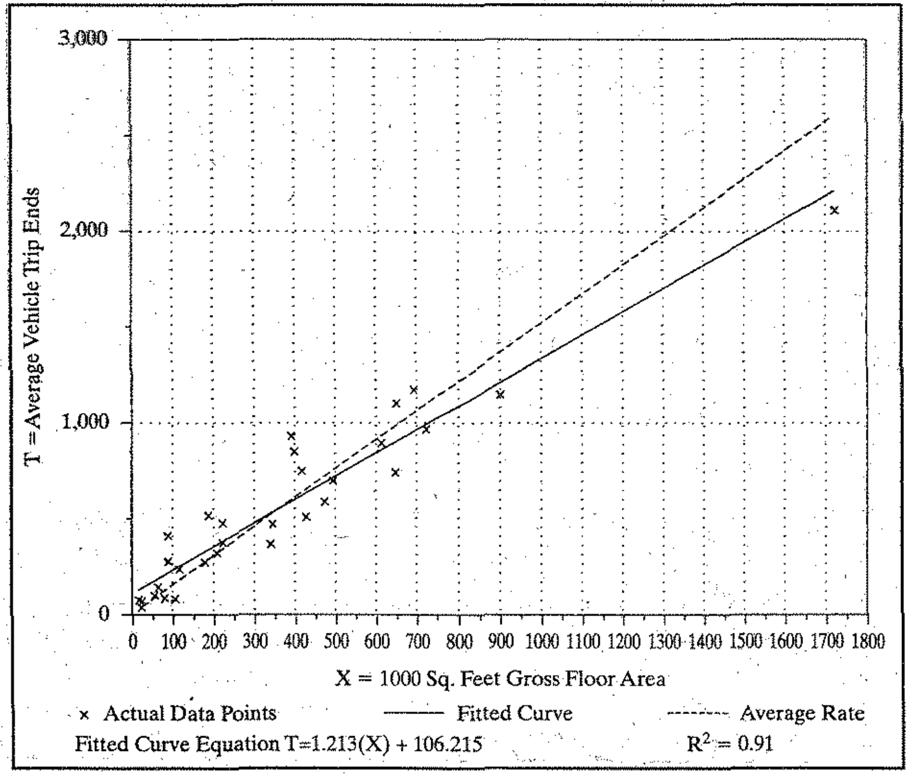


Figure 9.2.1 ***Continued***

Example 9.1

Estimate the peak-hour number of trips generated by an office park with 190,000 ft^2 GFA for 1 h during the morning and evening peak periods.

Solution The appropriate formulas from the *Trip Generation* manual are (see Fig. 9.2.1):

$$\text{A.M. peak:} \quad \ln(T) = 0.836 \ln(X) + 1.540 \qquad (9.2.2a)$$

$$\text{P.M. peak:} \quad T = 1.213(X) + 106.215 \qquad (9.2.2b)$$

Simple substitution of $X = 190$, since inputs should be in thousands, results in 375 vehicle-trips in the morning peak hour (Eq. 9.2.2a), and 337 in the evening peak hour (Eq. 9.2.2b).

The present traffic in the area (assessed with traffic counts) needs to be projected to the horizon year. This is done by augmenting traffic using an annual growth rate (one of the assumptions mentioned earlier) or by using historical growth rates for the general area. More sophisticated forecasting tools incorporating demographic and socioeconomic parameters may be employed for large-scale developments (see Chapter 8). The simpler the forecasting method is (i.e., constant annual growth rate) and the farther the horizon year is, the more uncertain the forecasts are. In such cases it may be wise to employ alternative growth scenarios, such as constant growth, accelerated growth, tapering growth, and so forth. Examples are given next.

	Growth (%)		
Years from present	Constant	Accelerated	Tapering
1–5	3	2	8
5–10	3	3	5
10–15	3	4	2

A simple method in common use is a compounded growth equation (see Chapter 12). For example, 1000 vehicle-trips in 2000 will correspond to 1344 trips in 2010 compounded annually, assuming a compounded 3% annual growth rate. As the reader may calculate, the same number of trips in 2000 will correspond to 1396 trips in 2010, assuming a 5% growth rate for the first 4 years, 3% growth rate for the next 4 years, and 1% growth rate for the last 2 years (high growth that tapers off).

Modal split. The estimated trips need to be adjusted to reflect public transit use and ride sharing. Usually the prevailing modal split in similar developments in the general area is adopted if no locally calibrated modal split is available. Modal split is essential because it varies substantially from place to place. The *Trip Generation* data are from suburban locations with nearly 100% auto modes share.

Example 9.2

For 1000 generated trips, estimate the number of vehicle-trips for the following scenarios: (1) 100% automobile use with average occupancy equal to 1.2 persons per car and (2) 65% automobile use with average occupancy equal to 1.2, 25% ride sharing with average occupancy equal to 2.5, and 10% public transit (in this case this corresponds to 29 buses per hour).

Solution Scenario 1 results in 833 vehicle-trips (1000/1.2 = 833). Scenario 2 results in 671 vehicle-trips, which reflects a 19% reduction of the trips estimated in scenario 1 [1000(0.65)/1.2 + 1000(0.25)/2.5 + 29 = 671].

Note that the rates resulting from the *Trip Generation* manual implicitly account for average occupancy (i.e., 1.2 persons per car); thus the occupancy adjustment illustrated earlier should be avoided when rates from that source are used. In general, the analyst should pay attention to trip rates; some may be in terms of person-trips and others in terms of vehicle-trips.

Trip distribution. Trip generation results in the total number of trips generated by the site analyzed. Some of these trips are entering the site while others are leaving the site. Knowledge of the destination of exiting trips and the origin of entering trips is necessary so that the routes followed and the impact on intersections can be assessed in the steps of traffic assignment and intersection performance analysis. Trip distribution enables identification of the general direction of origin or destination of the trips (i.e., 12% of traffic is coming from zone *x*) and trip assignment enables the assignment of trips along specific routes.

The split of trips in entering and exiting can be taken by surveying neighboring similar sites or from the *Trip Generation* manual, as shown in the following example.

Example 9.3

Based on Example 9.1, estimate the number of trips entering and exiting the site for the morning and evening peak hours.

Solution For the particular kind of site development the *Trip Generation* manual specifies the following average values (see Fig. 9.2.1):

A.M. peak hour: 89% entering and 11% exiting

P.M. peak hour: 14% entering and 86% exiting

Given the 375 and 337 trips estimated for the A.M. and P.M. peak hours, respectively, the directional distribution is as follows:

A.M. peak hour: 334 entering and 41 exiting

P.M. peak hour: 47 entering and 290 exiting

The directions from which the traffic will access the site, or depart from the site, depend on:

- The type of proposed development
- Competing developments in the surrounding area
- The size of the proposed development
- The land uses in the area and its population
- The flow conditions and the characteristics of the surrounding street system

So far the location of the site as well as the study area have been established. The study area is defined by the intersections to be analyzed. To apply trip distribution, an influence area (a perimeter surrounding the study area) needs to be established. This area should include at least 80% of the trips ending at or departing from the site. One way to define the influence area is by setting a reasonable upper bound travel time between the site and the limits of the influence area. Some examples of travel times are the shopping mall, 15 to 30 min; the office park, 30 to 45 min [9.4]. Then the influence area is divided into zones (i.e., census tracts, physical or legal boundaries, etc.). A basic method for distributing trips from many origins to one (the site in this case) or several destinations is the gravity model, which is presented in Chapter 8.

Note that the smaller the zones and the more accurate the distances or travel times are utilized, the more accurate are the trip-distribution results. Large zones are inhibited with large aggregation errors, and inaccurate travel characteristics result in a biased trip distribution (the distribution favors shorter distances or travel times). Distances and travel times are discussed further in the next section.

Traffic assignment. The traffic assignment step determines the amount of traffic that will use certain routes of the roadway network between the site and the surrounding zones (within the influence area). Links of the network will be loaded differently, depending on the origins and destinations, as well as on the traffic conditions on each link. As a result, some links or network segments may receive the bulk of the site-generated traffic, whereas others may receive no additional traffic.

As discussed in Chapter 8, the fundamental methodology for traffic assignment is based on some principle of network equilibrium. There are three mutually exclusive principles: *user* equilibrium (assign users on the shortest path from their origin to their destination), *system* equilibrium (assign flows to minimize the total travel time spent by all users in the network), and *stochastic* equilibrium (assign users on paths that *they think* are the shortest). The latter incorporates the notion of perception of travel times while it implicitly assumes that all reasonable paths between the origin and destination will have some flow; realistically, this is a better equilibrium principle [9.5].

A measure of travel cost is necessary to perform traffic assignment: usually travel time or distance. The use of travel time instead of distance is preferred because it represents actual flow conditions on the network. Distance remains constant over time, whereas travel time fluctuates by the *time of day* (e.g., it is shortest during off-peak hours and much longer during peak hours; thus travel time accounts for the level of congestion) or by the *type of the facility* (e.g., under uncongested conditions, 5 mi on an expressway can be traversed in shorter time than on an urban arterial, which typically has a lower speed limit as well as traffic signals and other interruptions). Thus travel time accounts for facility-specific characteristics as well.

The traffic assignment phase can be hand-solved only for very small networks (i.e., up to eight nodes and/or 12 links). Usually computer programs are employed for conducting the traffic assignment. The process is often iterative in an attempt to estimate realistic traffic loads. Two simple assignment methods that can be applied quickly with a spreadsheet program are the *FHWA* and the *incremental* methods. Their algorithms follow these steps for each origin-destination (O-D) pair:

FHWA

1. Compute the travel times based on existing flows, t'.
2. Assign all flow to the minimum path (all-or-nothing).

3. Compute travel times $t_a(V_a)$ as a function of the existing and assigned flows.
4. Compute revised times $t'' = 0.75t' + 0.25\, t_a(V_a)$.
5. Take t'' as the new base time and repeat steps 2 to 4 three more times.
6. The final assignment of flows on each path is the average of the four flows.

INCREMENTAL

1. Compute the travel times based on existing flows.
2. Assign an increment of the flow to the minimum path (usually the smaller the increment is, the better are the results).
3. Update link travel times based on the existing and assigned flow.
4. Repeat until all the flow for the O-D pair has been assigned.

Example 9.4

A volume of 524 vehicles must be assigned between an O-D pair connected with three routes A, B, and C. The three routes have existing volumes, $Q_{\max}$ and base-travel times as shown in the first four columns in the following table. Use the FHWA and the incremental methods with four equal increments to perform this assignment.

Solution First, the travel time with the existing traffic is estimated by applying Eq. 8.5.2, which for link A is modified as follows:

$$t' = 22\left[1 + 0.15\left(\frac{550}{1000}\right)^4\right] = 22.3$$

The spreadsheet on the next page illustrates the two methods.

FHWA Method

Link B has the lowest travel time and receives the volume of 524 veh/h. This increases its travel time from 20.3 to 25.9 min. The travel times of links A and C do not change. Then the weighting formula is applied: $t'' = 0.75\,t' + 0.25\,t_a(V_a) = 0.75 \times 20.3 + 0.25 \times 25.9 = 21.7$. At this point the second iteration can commence based on the t'' of iteration 1. After four iterations links A, B, and C receive 0, 262, and 262 veh/h, respectively.
An exact solution is easy to produce in this case. The assignment that yields the best travel-time equilibrium is 0, 222, and 302 veh/h for links A, B, and C, respectively. Based on this, the "% error" column was estimated.

Incremental Method

Four equal increments of 25% or 131 veh/h are made as shown in the table. Observe that each increment is added to the previous one, so that no summation and averaging is needed at the end of the process. After four iterations links A, B, and C receive 0, 262, and 262 veh/h, respectively.

Discussion Although in this simple example both methods produced identical results (which are considerably far from the optimum solution), the incremental method tends to produce better results than the FHWA method if a large number of increments of diminishing size are used, for example, eight increments of 30, 20, 15, 10, 10, 5, 5, and 5%.

Intersection analysis. After the trip assignment is complete, all additional (site and nonsite) volumes are known for the base as well as for the horizon year. The calculation of volumes by movement (i.e., through, right, and left) is the next step; it is illustrated in the detailed example in Section 9.2.6. At this point all site-generated volumes at each intersection

FHWA

					1			2			3			4				
Link	Existing traffic	$Q_{\max}$	Base II	t'	V_a	$t_a(V_a)$	t''	V_a	$t_a(V_a)$	t''	V_a	$t_a(V_a)$	t''	V_a	$t_a(V_a)$	t''	V_{avg}	% error
A	550	1000	22	22.3	0	22.3	22.3	0	22.3	22.3	0	22.3	22.3	0	22.3	22.3	0	0%
B	1440	1500	18	20.3	524	25.9	21.7	0	20.3	21.4	524	25.9	22.5	0	20.3	21.9	262	18%
C	3700	4000	19	21.1	0	21.1	21.1	524	22.5	21.5	0	21.1	21.4	524	22.5	21.7	262	−13%

Incremental

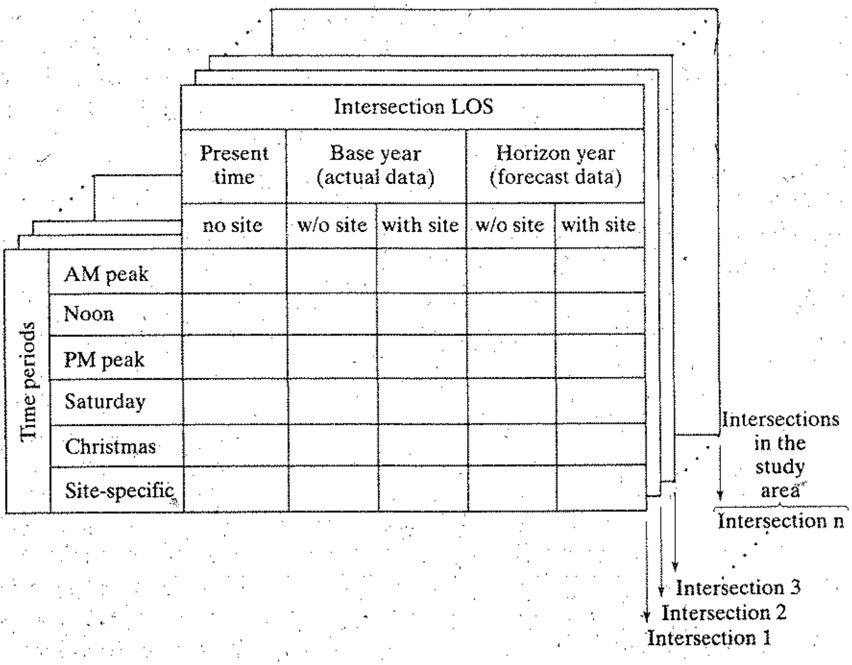
					1		2		3		4		
Link	Existing traffic	$Q_{\max}$	Base II	t'_a	$V_{25\%}$	$t_a(V_a)$	$V_{50\%}$	$t_a(V_a)$	$V_{75\%}$	$t_a(V_a)$	$V_{100\%}$	$t_a(V_a)$	% error
A	550	1000	22	22.3	0	22.3	0	22.3	0	22.3	0	22.3	0%
B	1440	1500	18	20.3	131	21.2	131	21.2	262	22.5	262	22.5	18%
C	3700	4000	19	21.1	0	21.1	131	21.4	131	21.4	262	21.7	−13%

are available by movement. Simple superposition of these volumes to the preexisting ones results in the final volumes to be used for intersection analysis. Thus the final traffic volumes for intersection analysis are the summation of (1) existing traffic, (2) growth of existing traffic at the horizon year, (3) site-generated traffic, and (4) anticipated nonsite traffic (i.e., from neighboring proposed development sites).

The intersection capacity and performance analysis is conducted as illustrated in Chapter 4. The total number of intersection analyses required are given in Fig. 9.2.2. For example, if five intersections were to be analyzed for morning and evening peak periods, a total of 5 × 2 × 5 = 50 intersection analyses would be necessary. This is because present time, base, and horizon year analyses are required and traffic conditions with and without the site are analyzed so that present, base-year, and future year conditions, as well as the impact of the subject project, can be evaluated.

In the horizon year analyses planned roadway improvements, such as widened approaches, installation of signal control, and so forth, should be included. Other actions, such as the elimination of curb parking and the scheduling of new transit service, may be included if sufficient information is available.

The analysis of intersections has the potential to reveal service deficiencies, such as approaches or whole intersections performing at unacceptable levels of service. These deficiencies have to be dealt with before seeking approval from the municipality. Potential deficiencies and remedies are discussed in Section 9.2.5.



		Intersection LOS				
		Present time	Base year (actual data)		Horizon year (forecast data)	
		no site	w/o site	with site	w/o site	with site
Time periods	AM peak					
	Noon					
	PM peak					
	Saturday					
	Christmas					
	Site-specific					

Figure 9.2.2 Intersection analysis reporting of results. (A selected number of time periods is usually considered (e.g., A.M. and P.M. peak hours only).)

Queue lengths as well as weaving areas at interchanges may need investigation to assure that safe and efficient traffic operations are maintained. Traffic simulation packages are usually employed for this type of analyses. Small or large signalized-intersection networks can be analyzed efficiently and accurately with several traffic simulation packages (see Chapter 15).

9.2.5 Site and Network Improvement Alternatives

The traffic impact study may reveal a number of traffic-related deficiencies that need to be improved. In some cases the deficiencies are so substantial that large elements of the design or size of the whole project may need to be changed. In this section improvements that may be within the site or in the surrounding roadway network are presented.

There are a number of improvements that can reduce and/or improve the flow of traffic within the site. They are grouped in three categories: (1) access locations, (2) internal circulation, and (3) demand management programs. Most of these improvements pertain to large-scale developments. In some cases, however, the access and internal circulation design of small or medium-sized sites need to be improved before approval from the municipality can be granted.

Access improvements facilitate the flow entering or leaving the site. Potential improvements may be the widening of entrance and exit points, which may include the placement of bays (exclusive lanes) for turning movements. Queues should not impede the internal circulation as well as the outside traffic. Exit and entrance lanes should supply enough capacity and storage so that the performance and safety of operations is not compromised.

Internal circulation improvements facilitate the flow within the site. Proper pavement markings as well as signs must be placed to assure the safety of operations. Major concerns arise from the expected presence of heavy vehicles in the internal traffic of the site (i.e., public transportation buses, employee transportation buses, and pick-up, delivery, or waste removal trucks). Thus turning radii should be designed to accommodate the movement of large vehicles, the parking layout should allow sufficient space for the maneuvering of heavy vehicles, and vertical clearances should account for the potential presence of oversized vehicles, while small bridges and other landscaping elements should be able to withstand the stress from heavy vehicles. Finally, the loading and unloading ramps should be carefully designed: They should be practical, spacious, and concealed from public view, if possible, to enhance aesthetics.

Demand-management programs could also be considered. They aim to reduce the number of vehicles using the road network to go to the site and the internal network and the parking space of the site. Cooperation with the local public transit authority for rerouting buses through the site and programs matching commuters as well as incentives for ride sharing (i.e., bonuses, free parking, privilege to park closer to entrance) have a potential to reduce the number of vehicles on site. In the case of large employment facilities the institution of flexible work hours has the potential to reduce local congestion for accessing and circulation in the facility during peak hours (i.e., peak-demand spreading).

On the surrounding roadway system improvements may take place at intersections, arterial streets, and freeway interchanges. The operation of intersections may be improved easily by altering the signal phasing and timings as well as the progression settings (offsets). Other improvements include the addition of lanes and/or different channelization schemes (i.e., allocation of lanes for turning movements). Often the addition of lanes is infeasible or unaffordable. Also, due to the site-generated traffic, signal control may be warranted for previously unsignalized intersections. Semiactuated controllers at the access points of large sites may facilitate the safe and efficient processing of traffic flows.

Several improvements may take place along arterials. Additional lanes increase capacity, walkways and curbs facilitate the safe processing of pedestrians, and lighting, particularly near access points and busy intersections, improves night driving conditions and safety, whereas new or altered regulations can affect the operational and safety characteristics (i.e., revision of speed limits and parking regulations).

If the site is close to an expressway facility, new access ramps may be installed, or the design of existing ramps, channelization, and weaving areas may need revisions. There are two major concerns at interchanges: the processing of flow moving at considerable speeds (particularly in the merging and weaving areas on the freeway) and the elimination of spilling over queues, which have the potential to clog ramps or parts of arterials. Ramp metering at freeway entrances can control the number of vehicles entering the facility so that uncongested flow conditions may be maintained.

The effectiveness of the selected improvements should be assessed and the least costly or most cost-effective options should be identified. This is an additional reason for having the traffic impacts analysis done on a computer.

In general, states, counties, and municipalities are increasingly hesitant in expending on infrastructure improvements to facilitate the traffic generated by one or several developments in an area. Roadway infrastructure improvements are costly and the small tax base of suburban communities can hardly provide the funds for such projects. Thus in many places *impact fees* have been instituted and schemes for shared funding of network improvements from the public and the private sectors have been established. Certainly both the developer and the municipality try to minimize their share in the contribution for roadway improvements. Comprehensive and accurate information from the transportation analysts will guarantee that each party pays only its fair share of the expenditures.

9.2.6 Comprehensive Example

This example illustrates analyses for a small-scale-site-development traffic impact study. At the corner of East-West Road and Oahu Avenue in Ocean County a site will be developed in the next few years after the county issues a permit. The traffic impact study (TIS) will be used in the application for site development. This site is planned to be developed as an office park. The developer's architect stated that the gross floor area (GFA) of the office park is going to be equal to 157,300 ft^2. The development is expected to be fully operational in year $N + 4$; the present year is N. No other sites have been proposed for development in the neighborhood of the subject site. Access to the site will be provided at the north side of its boundaries; it is planned to be an unsignalized intersection with stop control at the exit from the project, unless the TIS indicates otherwise.

There are two communities close to the project: Northtown with 18,000 residents and Easttown with 25,000 residents at the present time. The county statistics indicate that Northtown grows at an annual rate of 3.2%, whereas Easttown grows by a rate of 1.5%. The county is expected to grow at an average rate of 2.0% for the next 10 years. The county's planners advised that 100% of the site traffic will be coming from or going to the two neighboring communities.

The county required that detailed volume travel-time surveys be conducted to estimate the functions for accurate assignment of traffic. The traffic analysts have agreed with county planners that the evening peak period should be considered at this stage of the proposal.

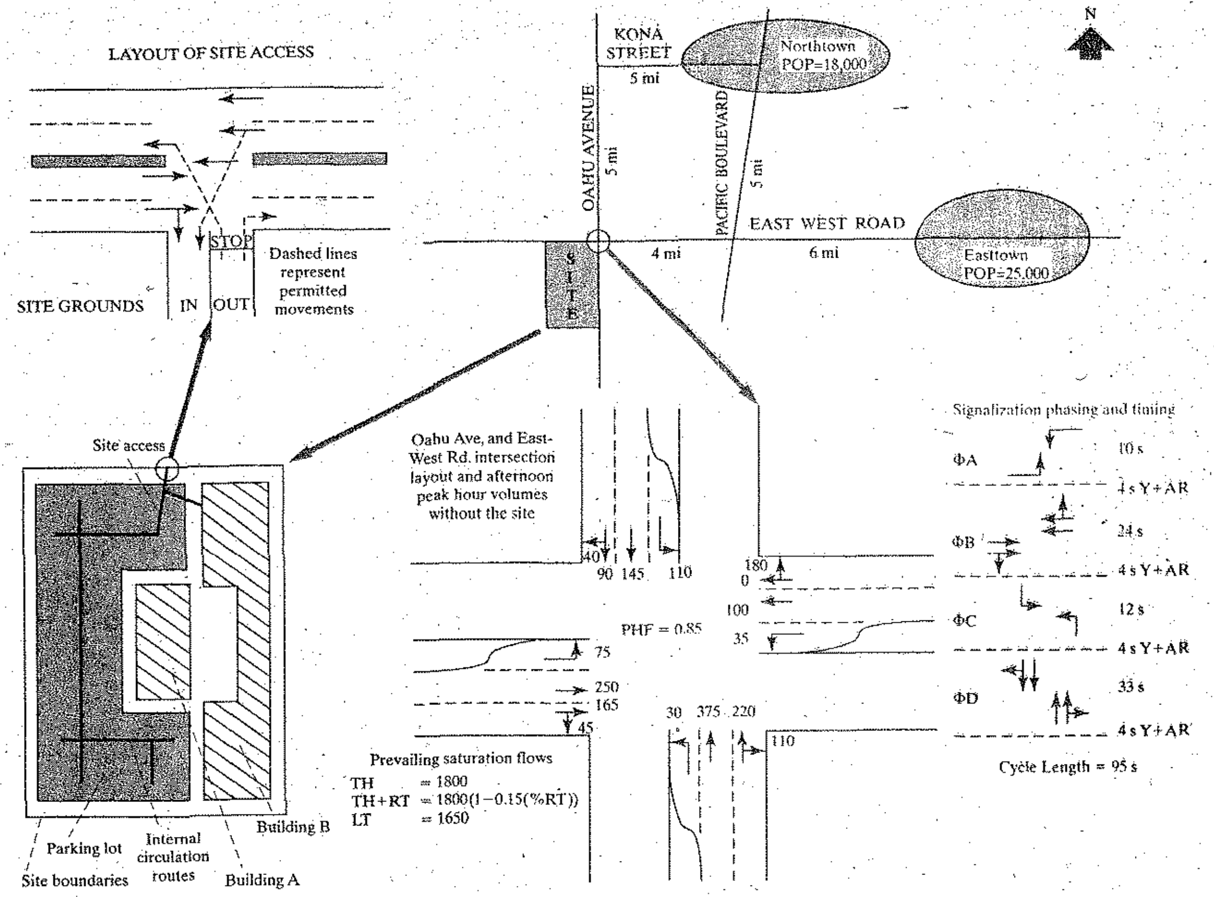


Figure 9.2.3 Description of site and surrounding area.

Figure 9.2.3 contains most of the information required for this preliminary TIS. Ocean County also required that (1) all lanes or lane groups should operate under LOS D or better and (2) the demand for each movement of the unsignalized intersection should be at most 50% of the potential capacity of the corresponding movement.

The analysis proceeds as follows:

Step 1: Evaluation of the performance of the signalized intersection at time N that the project is not in place

Step 2: Site and nonsite traffic estimation

- **Step 2.1:** Estimation of the project-generated traffic
- **Step 2.2:** Traffic distribution
- **Step 2.3:** Traffic assignment on the roadway network
- **Step 2.4:** Imposition of site volumes onto the year $N + 4$ background volumes

Step 3: Evaluation of the performance of the signalized intersection assuming that the project is in full operation

Step 4: Assessment of the capacity of the unsignalized intersection

Step 5: Recommendations for changes so that the requirements set forth by the county are met

The performance of the intersection under present conditions is analyzed and evaluated in Table 9.2.1 using the HCM 2000 procedure presented in Chapter 4. The analysis reveals that with the existing signal timings all left-turning movements operate at level-of-service D. Overall the intersection operates at LOS C (but only 0.3 s/veh below the threshold for LOS D). Table 9.2.2 presents the base future scenario, that is, traffic analysis at year $N + 4$ without the site, assuming an average growth of 2.0% a year, which translates into a growth factor of 1.0824 (2.0% growth for 4 years: $(1 + 0.02)^4 = 1.0824$). All average delays per lane group increase, placing the intersection at LOS D. The SB-LT movement, however, is marginally into LOS E (not acceptable). A small change in green times should improve it to LOS D.

Trips generated by the proposed site are estimated as follows. A model supplied by ITE's *Trip Generation* manual [9.4] is used; it is identical to the one in Eq. 9.2.2b. This model is for the evening peak period (1 h during the evening peak period), and the proportions for entry and exit are 14 and 86%, respectively. Thus

$$T = 1.213(X) + 106.2 \quad \text{for} \quad X = 157.3 \rightarrow T = 297 \text{ trips}$$

and

$$\text{Enter (INs)} = 42 \text{ trips, exit (OUTs)} = 255 \text{ trips}$$

The year $N + 4$ populations of the surrounding communities need to be estimated before the site-generated traffic is distributed.

$$\begin{aligned} \text{Northtown:}\quad & 18{,}000\,(1 + 0.032)^4 = 20{,}417 \\ \text{Easttown:}\quad & 25{,}000\,(1 + 0.015)^4 = \underline{26{,}534} \\ & \qquad\qquad\qquad\qquad\quad 46{,}951 \end{aligned}$$

The description of the site development states that there are no other sites to be developed in the neighborhood of the subject site. Thus there is no nonsite, other than the background traffic to be considered. If there was such traffic, the analysis of nonsite traffic would

TABLE 9.2.1 Present Time: Year *N* without the Site

Approach	Lane group	Volume	Right turn (%)	PHF	Adjusted volume	Satur. flow	Flow ratio	Critical mvmt.	Green	Cycle length	Capacity	X	Delay	LOS
EB	LT	75		0.85	88	1650	0.053	1	10	95	174	0.51	50.4	D
	TH	250		0.85	294	1800	0.163	1	24	95	455	0.65	38.7	D
	TH + RT	210	21	0.85	247	1742	0.142		24	95	440	0.56	36.0	D
WB	LT	35		0.85	41	1650	0.025		10	95	174	0.24	42.2	D
	TH	100		0.85	118	1800	0.065		24	95	455	0.26	29.8	C
	RT	180	100	0.85	212	1530	0.138		24	95	387	0.55	36.3	D
NB	LT	30		0.85	35	1650	0.021		12	95	208	0.17	38.8	D
	TH	375		0.85	441	1800	0.245	1	33	95	625	0.71	33.4	C
	TH + RT	330	33	0.85	388	1710	0.227		33	95	594	0.65	31.7	C
SB	LT	110		0.85	129	1650	0.078	1	12	95	208	0.62	52.5	D
	TH	145		0.85	171	1800	0.095		33	95	625	0.27	23.4	C
	TH + RT	130	31	0.85	153	1717	0.089		33	95	596	0.26	23.2	C
							X_c =	65.0%		Overall intersection =			34.7	C

TABLE 9.2.2 Base Future Scenario: Year $N + 4$ without the Site (2% Annual Growth Rate)

Approach	Lane group	Volume	Right turn (%)	PHF	Adjusted volume	Satur. flow	Flow ratio	Critical mvmt.	Green	Cycle length	Capacity	X	Delay	LOS
EB	LT	81		0.85	96	1650	0.058	1	10	95	174	0.55	52.3	D
	TH	271		0.85	318	1800	0.177	1	24	95	455	0.70	40.9	D
	TH + RT	227	21	0.85	267	1742	0.154		24	95	440	0.61	37.5	D
WB	LT	38		0.85	45	1650	0.027		10	95	174	0.26	42.6	D
	TH	108		0.85	127	1800	0.071		24	95	455	0.28	30.1	C
	RT	195	100	0.85	229	1530	0.150		24	95	387	0.59	37.8	D
NB	LT	32		0.85	38	1650	0.023		12	95	208	0.18	39.0	D
	TH	406		0.85	478	1800	0.265	1	33	95	625	0.76	36.1	D
	TH + RT	357	33	0.85	420	1710	0.246		33	95	594	0.71	33.8	C
SB	LT	119		0.85	140	1650	0.085	1	12	95	208	0.67	55.6	E
	TH	157		0.85	185	1800	0.103		33	95	625	0.30	23.7	C
	TH + RT	141	31	0.85	166	1717	0.096		33	95	596	0.28	23.5	C
							X_c =	70.3%		Overall intersection =			36.5	D

be identical to the analysis of the site-generated traffic. Then both site and nonsite traffic would have to be combined to result in the proper intersection loadings (volumes in year $N + 4$).

The site-generated traffic is distributed using a simplified version of the gravity model. The distances between the communities and the site are excluded because they are nearly identical.

	entering trips		exiting trips
Northtown to site	$42 \dfrac{20{,}417}{46{,}951} = 18$	Site to Northtown	$255 \dfrac{20{,}417}{46{,}951} = 111$
Easttown to site	$42 \dfrac{26{,}534}{46{,}951} = 24$	Site to Easttown	$255 \dfrac{26{,}534}{46{,}951} = 144$

The assignment of traffic on specific routes is next. There is only one practical way for traffic to go between the site and Easttown: along East-West Road. It is not conceivable that such traffic will utilize the route via Northtown, for example, a 21-mi trip instead of a 10-mi trip. Thus the traffic from the site to Easttown and from Easttown to the site can be readily assigned on East-West Road.

The traffic between the site and Northtown has two options. It may follow the 9-mi route (East-West Road and Pacific Boulevard) or the 10-mi route (Oahu Avenue and Kona Street). In order for the assignment to satisfy county requirements, volume travel-time surveys were conducted. The following models were estimated:

East-West/Pacific: $$TT = 2 + 2.5(V/2000)^2 \tag{9.2.3a}$$

Oahu/Kona: $$TT = 2 + (V/2200)^2 \tag{9.2.3b}$$

where

TT = travel time per mile

V = traffic volume on each direction

First, the year $N + 4$ volumes (without the site) along the routes identified previously need to be estimated. Figure 9.2.4(a) presents these calculations. Then the 111 trips from the site to Northtown and the 18 trips from Northtown to the site must be assigned on the network. This problem is amenable to a simple and exact algebraic solution, which should be preferred to heuristic methods, such as the FHWA and incremental methods presented earlier in this chapter. Call x the proportion of the 111 trips to be assigned on the East-West/Pacific route. To reach travel-time equilibrium, the following relationship must be true:

East-West/Pacific route *Oahu/Kona route*

$$9\left[2 + 2.5\left(\frac{687 + 144 + x}{2000}\right)^2\right] = 10\left[2 + \left(\frac{920 + (111 - x)}{2200}\right)^2\right] \tag{9.2.4}$$

Note: The 144 trips from the site to Easttown have already been assigned to East-West Road.

Travel times equate at $x = 23$. The equilibrium travel time is 22.1 min, which corresponds to an average speed of 24.4 mi/h along the East-West/Pacific route and to 27.1 mi/h

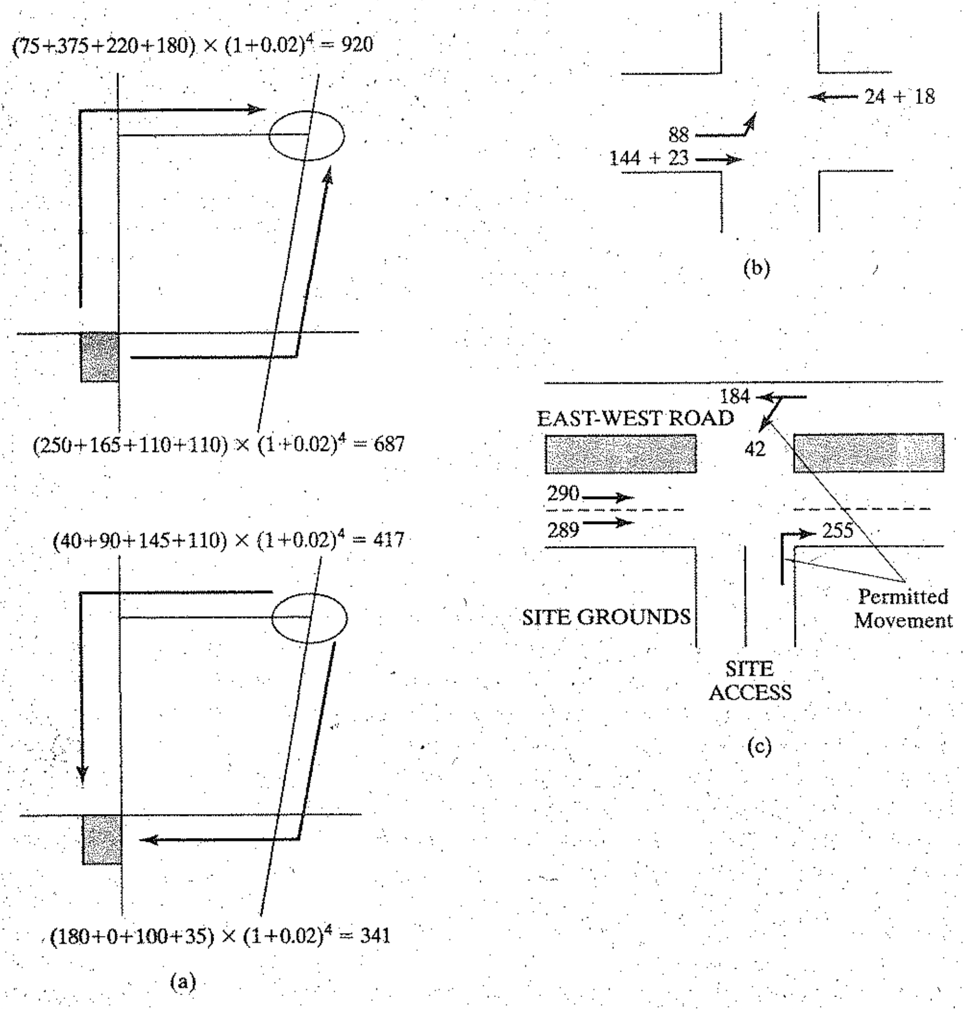


Figure 9.2.4 (a) Existing volumes for traffic assignment; (b) Intersection loadings from site-generated traffic; (c) Traffic volumes per movement of the unsignalized (stop-controlled) intersection at the access point of the site.

on the Oahu/Kona route. The analyst should always check results for reasonableness: Average speeds below 15 mi/h or above 45 mi/h may be unreasonably slow or fast, respectively, for suburban arterial streets.

First, we note that 24 trips from Easttown to the site have already been assigned to East-West Road. The 18 trips from Northtown to site are all assigned on the East-West/Pacific route because after all the site-generated traffic is added, the resulting travel time is shorter than that on the Oahu/Kona route:

East-West/Pacific: Total volume = 341 + 24 + 18 → TT = 18.8 min

Oahu/Kona: Total volume = 417 + 0 → TT = 20.4 min

All traffic assignment calculations were based on the assumption that travelers will choose the route providing the minimum travel time between the origin and destination until equilibrium is reached. Figure 9.2.4(b) presents the final assignment of traffic and the corresponding loading on the intersection under analysis. The site traffic must be allocated on specific lanes. The allocation is as follows:

1. *WB traffic.* The 42 added vehicle-trips (through movement) are all assigned to the center lane because the rightmost lane already carries a higher volume:

$$180(1 + 0.02)^4 > 100(1 + 0.02)^4 + 42$$

2. *EB traffic.* The 167 vehicle-trips (through movement) need to be distributed on the center and rightmost lanes. If x is the portion to be assigned on the center lane, then

$$\overset{\textit{before}}{\frac{45 + 165}{250}} = \overset{\textit{after}}{\frac{(45 + 165) \times (1.02)^4 + (167 - x)}{250 \times (1.02)^4 + x}} \rightarrow \qquad x = 91$$

Thus 91 vehicles are added to the center lane and 76 vehicles are added to the rightmost lane.

At this point the performance of the signalized intersection at the time $N + 4$ with the site in full operation can be assessed. Table 9.2.3 presents this analysis. The site-generated volumes have been added to the year $N + 4$ background traffic; bold type in the volume column highlights the cells containing the sums of background and site-generated traffic. The results indicate that 3 out of 12 lanes will be providing unacceptable LOS. However, since the distribution of traffic has changed (i.e., the EB approach has been loaded with a large amount of additional traffic) and the overall traffic load has increased (i.e., due to the expected areawide growth of 2%), new signal timings are needed. They may alleviate the problem of poor intersection performance.

Webster's formula for the estimation of optimal cycle length is utilized (as in Chapter 4). The total lost time is 20 s ($L = 5$ phases* × 4 s per phase), while the sum of the flow ratios for each critical movement per phase is CS = 0.58. These inputs result in $C_0 = 82.7$ s. An 85-s cycle length is selected. Table 9.2.4 shows that all 12 lanes are estimated to operate under LOS D or better.

The next step is the evaluation of the performance of the unsignalized intersection that provides access to the site. Figure 9.2.4(c) presents the intersection configuration and the traffic volumes present when the site becomes fully operational. Two movements need to be analyzed: the left turn from the major street into the site and the right turn exiting from the site. The potential capacity is estimated separately for each permitted movement:

1. *Left turn from major Street:* $t_c = 4.1$ s, $t_f = 2.2$ s, $V_c = 579$ veh/h

$$\alpha = \frac{579(4.1)}{3600} = 0.6594 \quad \text{and} \quad \beta = \frac{579(2.2)}{3600} = 0.3538$$

*The five phases are EWL, ETL, EWT, NSL, and NST. Two small adjustments were made to the greens: (EWL = −2s, ETL = +2 s) and (NSL = +1.4 s, NST = −1.4 s). These small changes worsened the overall delay from 36.6 s/veh to 37.1 s/veh but produced an acceptable level of service for all lanes.

TABLE 9.2.3 Future Scenario: Year $N + 4$ with the Site

Approach	Lane group	Volume	Right turn (%)	PHF	Adjusted volume	Satur. flow	Flow ratio	Critical mvmt.	Green	Cycle length	Capacity	X	Delay	LOS
EB	LT	**169**		0.85	199	1650	0.121	1	10	95	174	1.15	155.5[a]	F
	TH	**362**		0.85	425	1800	0.236	1	24	95	455	0.94	63.6	E
	TH + RT	**303**	16	0.85	357	1757	0.203		24	95	444	0.80	47.6	D
WB	LT	38		0.85	45	1650	0.027		10	95	174	0.26	42.6	D
	TH	**150**		0.85	177	1800	0.098		24	95	455	0.39	31.9	C
	RT	195	100	0.85	229	1530	0.150		24	95	387	0.59	37.8	D
NB	LT	32		0.85	38	1650	0.023		12	95	208	0.18	39.0	D
	TH	406		0.85	478	1800	0.265	1	33	95	625	0.76	36.1	D
	TH + RT	357	33	0.85	420	1710	0.246		33	95	594	0.71	33.8	C
SB	LT	119		0.85	140	1650	0.085	1	12	95	208	0.67	55.6	E
	TH	157		0.85	185	1800	0.103		33	95	625	0.30	23.7	C
	TH + RT	141	31	0.85	166	1717	0.096		33	95	596	0.28	23.5	C
							$X_c =$	85.0%		Overall intersection =			49.1	D

[a]Must use $X = 1$ in the estimation of the uniform delay (d_1)

TABLE 9.2.4 Future Scenario: Year *N* + 4 with the Site-Improved Signal Timings

Approach	Lane group	Volume	Right turn (%)	PHF	Adjusted volume	Satur. flow	Flow ratio	Critical mvmt.	Green	Cycle length	Capacity	*X*	Delay	LOS
EB	LT	**169**		0.85	199	1650	0.121	1	14.6	85	283	0.70	46.8	D
	TH	**362**		0.85	425	1800	0.236	1	24.5	85	519	0.82	41.7	D
	TH + RT	**303**	16	0.85	357	1757	0.203		24.5	85	506	0.70	35.0	C
WB	LT	38		0.85	45	1650	0.027		5	85	97	0.46	53.5	D
	TH	**150**		0.85	177	1800	0.098		16.9	85	358	0.49	35.1	D
	RT	195	100	0.85	229	1530	0.150		16.9	85	304	0.75	47.9	D
NB	LT	32		0.85	38	1650	0.023		11	85	214	0.18	34.8	C
	TH	406		0.85	478	1800	0.265	1	28.5	85	604	0.79	35.8	D
	TH + RT	357	33	0.85	420	1710	0.246		28.5	85	573	0.73	33.0	C
SB	LT	119		0.85	140	1650	0.085	1	11	85	214	0.66	49.9	D
	TH	157		0.85	185	1800	0.103		28.5	85	604	0.31	22.2	C
	TH + RT	141	31	0.85	166	1717	0.096		28.5	85	576	0.29	22.0	C
							X_c =	87.1%		Overall intersection =			37.1	D

Substitution in Eq. 4.9.3 results in $C_p = 1005$ vehicles. The demand of 42 vehicles is well below the 50% of the potential capacity. Thus for this movement the county's criterion is met.

2. *Right turn from minor street:* $t_c = 6.9$ s, $t_f = 3.3$ s, $V_c = 289$ veh/h

$$\alpha = \frac{289(6.9)}{3600} = 0.5539 \quad \text{and} \quad \beta = \frac{289(3.3)}{3600} = 0.2649$$

Substitution in Eq. 4.9.3 results in $C_p = 714$ vehicles. The demand of 255 vehicles is well below the 50% of the potential capacity. Thus for this movement the county's criterion is met as well.

The study of the unsignalized intersection is concluded at this point but it is not complete. Some potential problems that may require further study are:

1. The left-turn movement from the major street into the site may be overloaded in the morning peak period. This is likely to cause a substantial disruption to the westbound through traffic, whereas the queue may back up all the way to the intersection of East-West Road with Oahu Avenue. A left-turn bay may be required.
2. A number of vehicles exiting from the site (right turn) may need to take a left turn at the intersection of East-West Road with Oahu Avenue. It is not certain that under peak-period conditions such weaving can be accomplished safely. Moving the access point to the west end of the site may mitigate this problem; if not, then semiactuated signal control with detectors at the access site and the lane for left turn into the site may be required.

This traffic impact study indicates that if certain signalization modifications are implemented, the subject road network will be able to handle the additional volumes generated by the site in year $N + 4$. Therefore the site may be developed according to the original plan. However, detailed analysis of morning conditions (i.e., a left-turn bay for accessing the site may be required) is necessary to assure the safety and efficiency of operations when the site becomes operational.

The previous TIS example is substantially simplified compared with a real-world TIS. However, all fundamental elements of a TIS have been preserved. A similar real-world TIS would likely include:

- Analysis of more intersections
- Analysis and evaluation of morning peak traffic conditions
- Inclusion of other proposed developments in the immediate area that are likely to exist
- Distribution of the traffic to a larger number of smaller zones
- Consideration of more routes between the site and each zone (as applicable)
- Analysis and recommendation of specific treatments for identified problems, such as the weaving problem stated earlier

9.3 PARKING STUDIES

9.3.1 Background

Parking is an important urban transportation element. It has various long- and short-term impacts on individuals, communities, and transportation systems. First, parking affects mode choice. Individuals having an automobile available will probably choose to access their destination by automobile if parking is available and conveniently located at the destination, and if the cost of parking is reasonable. In other words inexpensive (or free) and plentiful parking is an incentive for using private automobiles, whereas scarce, inconvenient, and/or expensive parking is a substantial disincentive for using private automobiles [9.6].

Parking also affects the vitality of communities, commercial and business centers, transit systems, and airports as well as the efficiency of traffic circulation in downtown areas. For example, in certain European cities it is estimated that 40% of the total travel time to work is consumed in searching for parking [9.7].

Parking has certain direct economic impacts as well. At a microscale parking costs may result in tax benefits for employers and entrepreneurs. At a macroscale parking generates revenues for both public (i.e., municipality revenue from metered parking and citations) and private institutions (i.e., for profit development of land).

As a result, most individuals and institutions are concerned with parking: automobile users, local governments (i.e., service to residents and visitors, economic vitality of businesses plus revenue for the community), private businesses (i.e., customer attraction, convenience to employees plus for-profit parking development), hospitals, schools and colleges, public services, and so forth.

Around the end of the century nearly 50 cities in the United States, Canada, and Europe allowed developers to pay a fee instead of providing the parking spaces required by zoning ordinances [9.8]. The fee revenue is used to provide new public parking spaces in lieu of the private parking spaces that developers would have provided. *In lieu parking* programs may reduce the cost of development, encourage shared parking, improve urban design, support historic preservation, and encourage developers to reduce parking demand instead of increasing parking supply. For example, the Eco Pass program in California has shown that paying the transit fare for commuters who arrive by bus is cheaper than providing the parking required for commuters who arrive by car [9.8]. In lieu programs, employee parking cash-out (see Section 6.4.2), provision of mass transit passes instead of parking, as well as the analysis of parking demands and needs, capacity, circulation, ventilation, security, and compliance with national and local regulations make comprehensive parking analyses complex.

9.3.2 Types of Parking

There are two broad categories of parking: public and private. *Public* parking may be curb-side (on streets and alleys) or off-street. Curb-side parking may be free or not, and it may be regulated or unregulated (i.e., no parking during rush hours, no parking overnight, etc.) In downtown areas curb-side parking is usually metered and regulated.

Off-street parking is usually in lots, decks (within multipurpose buildings), or in exclusive parking structures. Private firms or public agencies may be operating these facilities, which are open to the public. Some facilities may operate under certain rules (e.g., parking on a long-term/contract basis).

Private parking includes home or apartment building garages, stalls and driveways, or affiliate-specific parking (i.e., permit required).

The arrangement of stalls and pricing schemes are two important characteristics of parking. Stalls may be parallel or angled (varying from 20 to 90°). Pricing schemes usually try to maximize revenue as well as to fulfill certain objectives. Some pricing schemes are designed to encourage short stay (high turnover): \$1 for up to $\frac{1}{2}$ h, \$3 if $\frac{1}{2}$ h is exceeded and \$2 for every hour thereafter, or parking meters with short-duration dials (this hinders convenience because coins must be inserted frequently while the likelihood of getting a citation increases). Other pricing schemes are designed to encourage long stay (low turnover): \$3 for 1 h or less, \$0.50 for each hour thereafter.

9.3.3 Types of Parking Studies

Parking studies include financial feasibility, functional design, structural design, and demand studies. This chapter focuses on demand and functional design studies. There are three major types of parking demand studies: comprehensive, limited, and site-specific [9.9]. *Comprehensive studies* cover an entire area, such as the central business district (CBD).

A major objective is to estimate demand for parking. The status quo reveals utilization forced by existing conditions and does not represent actual demand. Usually surveys are employed to assess the demand for parking: both the need and preferred locations for parking. Careful sampling design and analysis are necessary to compensate for the natural bias toward oversampling short-term parkers [9.9].*

In comprehensive studies the future parking demand is estimated with the use of forecasting models, which include population growth, demographic, social and economic trends, as well as trends of the local economy and use of transportation modes. Analytic and comprehensive inventories of on- and off-street parking are gathered along with detailed information on utilization patterns. From these, current deficiencies of the parking supply are identified (i.e., lack of supply, interference with traffic circulation). Then proposed scenarios for alleviating current deficiencies and fulfilling anticipated demands are developed and evaluated for judgment by officials and/or private interests. The development and evaluation of scenarios or *alternatives analysis* is conducted with a number of criteria, such as (1) encouragement or discouragement of private automobile use, (2) identification of primary recipients of service and ways to screen out nonprimary parkers, (3) derivation of a pricing schedule, (4) issues regarding access distance (i.e., convenience and safety for walking), (5) satisfaction of municipal and private perspectives, (6) zoning requirements, and (7) budget and future costs/income flows.

Limited studies are similar to comprehensive studies but with reduced geographic coverage and fewer requirements. Typically in limited studies only one type of parking may be investigated (i.e., curb-side parking) while the estimation of future demand may not be required.

Site-specific studies are geographically narrow but analytically extensive. Focus sites may include existing, planned, or expanding hospitals, campuses, shopping malls, residential, office, and industrial developments. Detailed inventories of existing supply and utilization are taken and future demands are forecast. In addition, attention is paid in regard to the various types of users of the parking supply: people who do business or work at the site

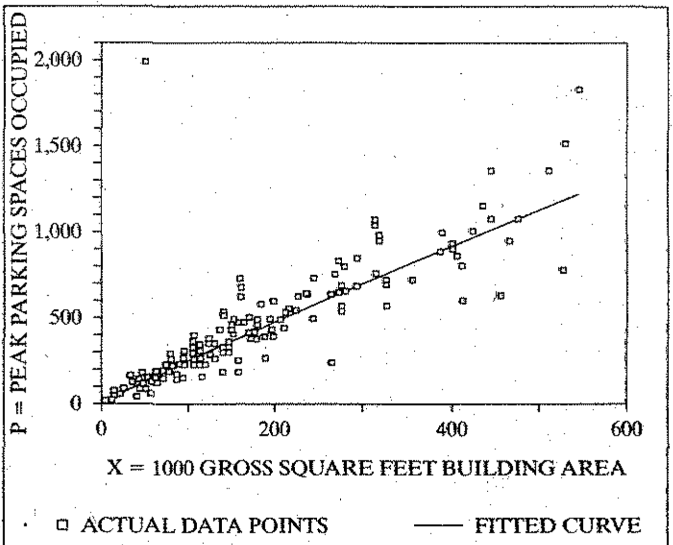
*Assume a 100-stall lot, where 50 stalls are filled with long-term parkers with zero turnover rate during the sampling period and the other 50 stalls are filled with short-term parkers with a turnover rate of five. Assume, for the sake of the argument, that all parkers were sampled. If so, 300 surveys were collected, 50 from long-term parkers and 250 from short-term parkers; the bias toward sampling short-term parkers is quite obvious.

GENERAL OFFICE BUILDING (711-716)
Peak Parking Spaces Occupied vs: 1,000 GROSS SQUARE FEET BUILDING AREA
On a WEEKDAY

PARKING GENERATION RATES

Average Rate	*Range of Rates*	*Standard Deviation*	*Number of Studies*	*Average 1,000 GSF Building Area*
2.79	0.75–32.93	2.25	207	168

DATA PLOT AND EQUATION



Fitted Curve Equation: $\mathrm{Ln}(P) = 0.93\ \mathrm{Ln}(X) + 1.253$
$R^2 = 0.870$

Figure 9.3.1 Examples of parking demand rates for general-purpose office buildings and movie theaters.
(Reprinted with permission from *Parking Generation*, 2nd ed., © 1987 Institute of Transportation Engineers, Washington, DC.)

(i.e., primary recipients of parking service) and people who park at the site to go elsewhere. Information on the users' access mode, the mix of users in regard to their parking occupancy (i.e., in hospitals, visitors stay up to a few hours, doctors and nurses may stay in excess of 16 h, and other staff stays 8 to 9 h) are often measured. Change of shifts and the overlapping parking utilization patterns of land uses such as industrial parks are of critical importance due to concerns in parking availability and turnover, internal circulation, and potential congestion at access points.

Assessment of site parking demand. Parking demand for proposed sites is commonly estimated from ITE's *Parking Generation* manual [9.10], which is conceptually and presentationally identical to the *Trip Generation* manual [9.4]. Figure 9.3.1 presents two graphs, one for general purpose office buildings and one for movie theaters.

MOVIE THEATER (443)

Peak Parking Spaces Occupied vs: SEATS

On a: SATURDAY

PARKING GENERATION RATES

Average Rate	*Range of Rates*	*Standard Deviation*	*Number of Studies*	*Average Number of Seats*
0.26	0.11-0.42	0.11	9	1562

DATA PLOT AND EQUATION

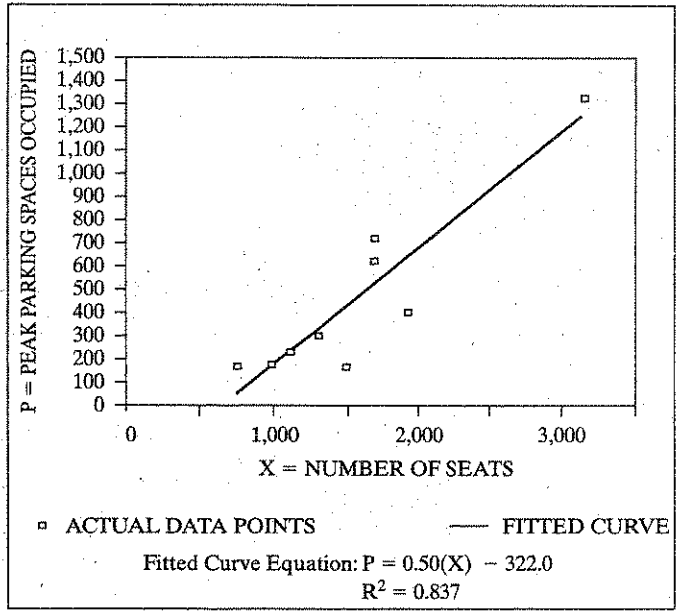


Figure 9.3.1 *Continued*

Parking demand dictates the size of the parking facility. A substantial parcel of land is always required for the establishment of a parking lot or structure. To achieve better utilization of a parking lot, it is preferred to develop land uses with complementary parking use requirements [9.11]. An office building and a movie theater complex sharing the same parking lot is a good example. Typically the office building causes peak parking use between 8 A.M. and 5 P.M. during weekdays, whereas the movie theater complex causes peak parking use between 6 P.M. and midnight during weekends. The following example illustrates this point.

Example 9.5

A joint development of a general purpose office building with a 200,000 ft^2 gross building area and a movie theater complex with 1500 seats is planned. The facilities will be served by a common parking lot. Estimate the number of stalls required and assess whether the efficiency improves with shared use of a parking lot as opposed to separate parking lots for each of the developments. Also, estimate efficiency as the degree of stall utilization over 1 week. Lot utilization on weekdays and Saturdays is specified in the following table (based on data in Table 5–6 of Ref. [9.12]).

	Office		Cinema	
Time	Weekday	Saturday	Weekday	Saturday
5	0%	0%	0%	0%
6	3%	0%	0%	0%
7	20%	3%	0%	0%
8	63%	10%	0%	0%
9	77%	13%	0%	0%
10	100%	13%	0%	0%
11	100%	17%	0%	0%
12	90%	17%	33%	33%
13	90%	13%	50%	67%
14	97%	10%	50%	67%
15	77%	7%	50%	67%
16	77%	7%	50%	67%
17	47%	3%	50%	67%
18	23%	3%	67%	83%
19	7%	3%	67%	83%
20	7%	3%	83%	100%
21	3%	0%	83%	100%
22	3%	0%	83%	100%
23	0%	0%	67%	83%
24	0%	0%	50%	67%

Solution The parking generation equations are as follows:

Office building (weekday): $\ln P = 0.93 \ln X + 1.253$ (Fig. 9.3.1)

Movie theaters (weekday): $P = 0.32X - 174.0$ ([9.10], p. 60)

Movie theaters (weekend): $P = 0.50X - 322.0$ (Fig. 9.3.1)

Based on the inputs supplied, the required parking is

Office building (weekday): 483 stalls

Movie theaters (weekday): 306 stalls

Movie theaters (weekend): 428 stalls

It would be erroneous and excessive to suggest that for the shared use of the parking lot $483 + 306 = 789$ stalls are required in a typical weekday. Better efficiencies may be realized if the parking demand is assessed based on the utilization during weekdays and Saturdays. Columns 2 through 5 in the following table were generated by multiplying the lot utilization by the demand for the parking estimated earlier. Then two columns of totals are created, one for weekdays and one for Saturdays. The last column shows the maximum value for parking demand in each time period. The value of 620 is the design target; it should be increased by 5 to 10% to account for the difficulty in finding parking when a large lot is nearly full. Thus a lot size of about 650 stalls would be sufficient instead of 800 stalls as initially estimated.

	Office		Cinema				
	Parking requirement						
	483	483	306	428	Total		
Time	Weekday	Saturday	Weekday	Saturday	Weekday	Saturday	Max
5	0	0	0	0	0	0	0
6	16	0	0	0	16	0	16
7	97	16	0	0	97	16	97
8	306	48	0	0	306	48	306
9	370	64	0	0	370	64	370
10	483	64	0	0	483	64	483
11	483	81	0	0	483	81	483
12	435	81	102	143	537	223	537
13	435	64	153	285	588	350	588
14	467	48	153	285	620	334	620
15	370	32	153	285	523	318	523
16	370	32	153	285	523	318	523
17	225	16	153	285	378	301	378
18	113	16	204	357	317	373	373
19	32	16	204	357	236	373	373
20	32	16	255	428	287	444	444
21	16	0	255	428	271	428	428
22	16	0	255	428	271	428	428
23	0	0	204	357	204	357	357
24	0	0	153	285	153	285	285
Max	483	81	255	428	620	444	620

Based on the estimates of the preceding table, the increase in lot utilization can be estimated for (1) two separate and independent parking lots and (2) one combined parking lot. We will assume that either lot is empty during the 4 h of the day that are not shown in the table. We need to sum up the contents of each column in the table. By doing so, we estimate that a total of 4267 stall-hours of occupancy from 0:00 to 23:00 are estimated for the parking lot of the office building during a weekday. The weekly parking lot utilization therefore is

$$\{[(4256 \div 24) \times 5 + (596 \div 24) \times 2] \div 7\} \div 483 = 33.3\%$$

$$\{[(2397 \div 24) \times 5 + (4209 \div 24) \times 2] \div 7\} \div 428 = 28.4\%$$

The lot size weighted average utilization for the two separate parking lots (911 spaces total) is 31%. The average utilization for the combined parking lots (620 spaces total) is 41.2%, as estimated here.

$$\{[(6664 \div 24) \times 5 + (4804 \div 24) \times 2] \div 7\} \div 620 = 41.2\%$$

which represents a substantial improvement of the lot utilization over time. In this way (1) a substantial parcel of land was preserved: land for 620 instead of 911 stalls for separate lots,

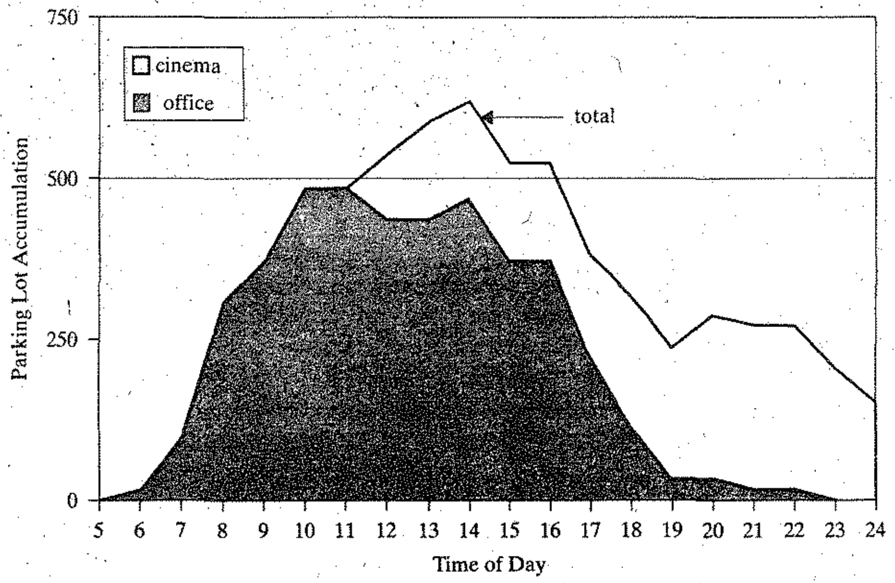


Figure 9.3.2 Parking occupancy profiles of a parking lot shared by an office building and a movie theater complex (profiles during a weekday based on the data of Example 9.5).

which represents a 32% reduction in land consumption, whereas the unused land may be landscaped or put in other revenue/utility-generating uses, and (2) the efficiency of land utilization increased from 31 to 41%.

Figure 9.3.2 presents the accumulation plot of shared parking lot utilization during a weekday for the example illustrated previously. The lot fills quickly in the morning. The office parking occupancy drops around the lunch hour. At the same time patronage for the movie theater starts increasing and the lot reaches its maximum utilization at 3 P.M., followed by a substantial decrease in occupancy due to the departure of most workers of the office building. Around 8 P.M., a spike in occupancy is observed. It is caused by the theater patrons coming for the late show while several patrons of the evening show have not yet departed.

9.3.4 Parking Measurements and Analysis

There are four major indices describing the parking utilization of the area or site of focus: occupancy, accumulation, turnover, and average duration of occupancy. Typically 85 to 95% instead of 100% of the available parking capacity is used in the analysis. This is because levels of utilization higher than 95% are hard to attain due to efficiency losses in turnover and circulation [9.9]. In other words, when the utilization of the capacity exceeds 95%, it becomes very difficult to find an empty parking stall. The supply or capacity results directly from the inventory measurements.

The four indicators are defined as follows:

$$\textit{Occupancy}\ (\%) = 100\,\frac{\text{number of spaces occupied}}{\text{total spaces available}}$$

$$\textit{Accumulation} = \text{number of vehicles parked at a given time}$$

Turnover = number of vehicles utilizing the same stall over a given period of time (four or more during an 8-h period indicates a high turnover rate)

$$\textit{Average duration} = \frac{\text{total vehicle hours}}{\text{total number of vehicles that parked}} = \frac{\sum_{i=1}^{N} t_i}{N}$$

where t_i is the duration of parking occupancy of vehicle i.

There are various methods of collecting the data that are required for the estimation of these four indices. Three of the most common methods are *ins and outs, fixed period,* and *license plate* surveys.

According to *ins and outs,* all vehicles parked in the focus area are counted at the beginning of the survey period. Then vehicles entering and exiting the area are counted (i.e., lot accesses and/or streets carrying traffic into or out of the area of focus are monitored continuously). Another occupancy count is conducted at the end of the survey period to check whether measurements balance. This method can yield the overall accumulation and occupancy only; turnover rates and average duration cannot be estimated.

According to *fixed-period* sampling, all vehicles parked in the focus area are counted at the beginning of the survey period. Then the same area is covered in increments of $\frac{1}{4}$ to 1h; that is, occupancy counts are conducted every $\frac{1}{4}$ to 1h. This method may miss short-term parkers and may be difficult to conduct in areas with private garages.

License plate surveys result in the most accurate and realistic data because essentially every parking stall is monitored at fixed intervals (the shorter the intervals are, the higher the accuracy is; i.e., avoid missing short-term parkers). This method is very labor-intensive and it entails certain liability problems (i.e., close monitoring of private property by copying vehicle license numbers). This can be partly alleviated by omitting the first or last digit of the license plate.

A license plate survey is conducted as follows: Every surveyor is assigned to specific parking stalls (i.e., the floor of a parking structure or a stretch of roadway with curb parking). For each parking stall there is a corresponding box on the survey form. Both standardized and project-specific forms are used. Typically every 30 min the surveyor walks along his or her inspection area marking the license number of each vehicle parked in the corresponding box on the form. For the next 30 min the same is repeated on another row of the form. Usually symbols are defined for denoting empty stalls as well as stalls occupied by the same vehicles as in the previous 30 min. At the end of the survey period the data are manually consolidated for each stall and input to the computer (i.e., spreadsheet software) for analysis.

Table 9.3.1 presents actual license plate survey data, encoding and analysis. The 50-stall parking lot contains 2 handicapped stalls, 20 permit-holder-only stalls, and 28 visitor stalls. A license plate survey every 15 min resulted in the left side of the table listing all of the partial license plate numbers of vehicles in stalls. License numbers in bold correspond to occupancy by a vehicle other than the one in the previous survey period. The occupancy of each stall is encoded as a series of 0 and 1 in the right side of the table. Shading is used to denote the difference of parked vehicles between consecutive periods. This helps to fill in the turnover column correctly; it shows how many different vehicles utilized each specific stall.

TABLE 9.3.1 License Plate Survey Data and Encoding

Stall	Type	12:00	12:15	12:30	12:45	12:00	12:15	12:30	12:45	Turnover
1	♿					0	0	0	0	0
2						0	0	0	0	0
3						0	0	0	0	0
4	PERMIT	GRF9	GRF9	GRF9	GRF9	1	1	1	1	1
5		FVS5	FVS5	FVS5	FVS5	1	1	1	1	1
6		EFE6			GRZ7	1	0	0	1	2
7		GFE0	GFE0	GFE0	GFE0	1	1	1	1	1
8		GTP6			ENG6	1	0	0	1	2
9		GJR5	GJR5	GJR5	**FOS4**	1	1	1	1	2
10		EGX4	EGX4	EGX4	EGX4	1	1	1	1	1
11		GRC1	GRC1	GRC1	GRC1	1	1	1	1	1
12		BYE	BYE		GCG5	1	1	0	1	2
13		GBW7	GBW7	GBW7	GBW7	1	1	1	1	1
14		FZP5	FZP5	**FNN2**	FNN2	1	1	1	1	2
15		GVR4	GVR4	GVR4	GVR4	1	1	1	1	1
16		GWT6	GWT6		ETT5	1	1	0	1	2
17			GWE8	GWE8	GWE8	0	1	1	1	1
18		GAA6	GAA6	GAA6	GAA6	1	1	1	1	1
19		GJS9	GJS9	GJS9	GJS9	1	1	1	1	1
20		GRT1	GRT1	GRT1	GRT1	1	1	1	1	1
21		FWW3			EGZ5	1	0	0	1	2
22		GFW6	GFW6	GFW6	GFW6	1	1	1	1	1
23	VISITOR	ECT6	**EEC2**	EEC2		1	1	1	0	2
24		GEF1	GEF1	GEF1	GEF1	1	1	1	1	1
25		EVA8	EVA8	EVA8	EVA8	1	1	1	1	1
26		FTP2	FTP2	FTP2	FTP2	1	1	1	1	1
27		FRY1	FRY1	**CGG8**		1	1	1	0	2
28		EFJ3	EFJ8	EFJ3	EFJ3	1	1	1	1	1
29		FJF9	FJF9	FJF9	**GPE6**	1	1	1	1	2
30		GDX4	GDX4	GDX4	GDX4	1	1	1	1	1
31		GNT7	**GDJ2**	GDJ2	GDJ2	1	1	1	1	2
32		GPP4	GPP4	GPP4	GPP4	1	1	1	1	1
33		GWZ4	GWZ4			1	1	0	0	1
34		GDR4	GDR4	GDR4	GDR4	1	1	1	1	1
35		DIDO	**FJP8**	FJP8	FJP8	1	1	1	1	2
36		FZJ1	FZJ1	FZJ1	FZJ1	1	1	1	1	1
37		GSU0	GSU0	GSU0	GSU0	1	1	1	1	1
38		FYB4	FYB4		GVZ3	1	1	0	1	2
39		GDY6	GDY6	**ETA9**	ETA9	1	1	1	1	2
40		FWN3	FWN3	FWN3	FWN3	1	1	1	1	1
41		FNF7	FNF7	FNF7	FNF7	1	1	1	1	1
42		BZE1		FZE4	FZE4	1	0	1	1	2
43		FFW5	FFW5	FFW5	**GAH2**	1	1	1	1	2
44		FWX8	FWX8	FWX8		1	1	1	0	1
45		**GWT9**	FDE4		GATT	1	1	0	1	3
46		ERV2	ERV2	ERV2	ERV2	1	1	1	1	1
47			EBW3	EBW3	EBW3	0	1	1	1	2
48		FTP0	FTP0	FTP0	FTP0	1	1	1	1	1
49		FWP8	FWP8	FWP8	FWP8	1	1	1	1	1
50		FFG6	FFG6	**GCG3**	**GRZ5**	1	1	1	1	3
ALL					ACCUMULATION	45	43	39	43	Avg. T/O
					% OCCUPANCY	90%	86%	78%	86%	1.36
PERMIT					ACCUMULATION	18	16	14	19	Avg. T/O
					% OCCUPANCY	90%	80%	70%	95%	1.30
VISITOR					ACCUMULATION	27	27	25	24	Avg. T/O
					% OCCUPANCY	96%	96%	89%	86%	1.50

Source: University of Hawaii, CE 462: Traffic Engineering, Parking Study, Spring 1999; Liza Garcia's data.

The bottom part of Table 9.3.1 presents the summary of measurements and the derivation of three of the four standard factors: accumulation, occupancy, and turnover, for the entire lot, and separately for the visitor and permit-only parts. The permit-only part shows the typical reduction in occupancy due to lunch hour departures and the return to near-full occupancy by 1:00 P.M. As expected, the visitor stalls exhibit a higher turnover than the permit-only stalls (the difference should be higher over an 8-h span.) If one assumes that the visitor stall turnover is typical and extrapolates for 8 h (extrapolation is not a recommended practice), then the turnover is estimated as 1 + 0.5 × 8 = 5 vehicles, which is quite high.

The average duration is estimated to be 37.6, 38.9, and 36.8 min for the total, permit-only and visitor stalls, respectively. Specifically, summing up the turnover column yields that 69 different vehicles were parked during the period of observations. The accumulation for each period multiplied by the sample interval (15 min) gives the vehicle hours of occupancy for the parking area. Then $[44 + 45 + 39 + 43] \times 15 \div 69 = 37.6$ min is the estimate for average duration. At this point all four parking factors have been estimated.

9.3.5 Design, Operation, and Other Considerations

Extensive information on parking lot and structure designs is presented in various manuals and textbooks [9.13, 9.14]. Figure 9.3.3 illustrates a basic lot or deck design. Several principles of design are discussed next.

Parallel on-street parking is preferred to angled parking for reasons of safety. When parked at an angle, backing into traffic may be dangerous due to poor visibility. In addition, the width of one traffic lane may not be sufficient for maneuvering out of the parking stall.

Angling stalls at 60° result in the most efficient utilization of space in lots and decks although other angles may provide a larger number of stalls for specific lot footprints. Unutilized corners should be landscaped or designated for motorbike or bicycle parking.

Due to the continuous downsizing of automobiles (particularly of those designed for urban commuting), a large number of "tighter" stalls with the designation "compact" should be supplied. This increases the overall parking capacity as well as the land-use efficiency.

It is often possible to design a simple parking lot by following standardized design modules that are different for large and small cars and for each angle of parking, typically from 45 to 90° in increments of 5°. Figure 9.3.4 presents the dimensions of typical modules (in ft, where applicable). Explanations as well as values for a 60° design are given next.

	Magnitude		
Dimension	Small car	Large car	Explanation
θ	60°	60°	Parking angle
i	1.42	1.67	Interlock reduction
o	1.75	2.58	Overhang
SL	16.00	16.00	Stall length
W_1	29.67	35.50	Wall-to-wall width, single-loaded aisle
W_2	46.00	55.00	Wall-to-wall width, double-loaded aisle
W_3	44.58	53.33	Wall-to-interlock width, double-loaded aisle
W_4	43.17	51.67	Interlock-to-interlock width, double-loaded aisle
W_5	42.50	49.08	Curb-to-curb width, double-loaded aisle
AW	13.33	16.00	Aisle width
VP	16.33	19.50	Vehicle projection

Source: Ref. [9.21].
(Reprinted with permission from *The Dimensions of Parking*, 3/e, ©1993 Urban Land Institute, Washington, D.C., USA.)

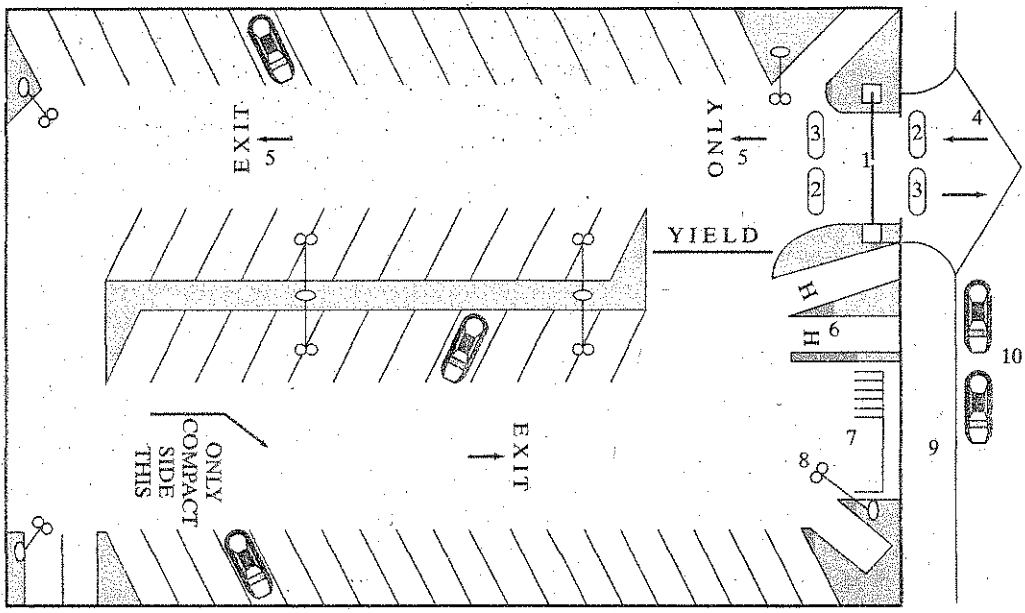


Figure 9.3.3 Parking lot layout (approximate scale).
1, Access gates; 2, vehicle presence detector to open gate; 3, vehicle departure detector to close gate; 4, transition from roadway to pavement; 5, internal circulation directions; 6, dedicated handicapped parking stalls; 7, space provisions for bicycles and motorbikes; 8, lighting; 9, curb; 10, street.

Special attention should be paid to parking for handicapped persons. Local ordinances may require a specific number or percentage of stalls designated for use by handicapped drivers. These stalls should be located the closest to the facility access points and clearly noted for the exclusive use by handicapped drivers. The handicapped parking stall requirements mandated by the Americans with Disabilities Act (ADA) may be summarized as follows:

Parking facility size (stalls)	Handicapped stall requirement
Up to 100	1 for each 25-stall increment
101 to 200	4 plus 1 for each 50-stall increment over 100
201 to 500	6 plus 1 for each 100-stall increment over 200
501 to 1000	2% of stalls
1001 and over	20 plus 1 for each 100-stall increment

For example, the 650-stall parking lot of Example 9.5 would require 13 designated handicapped stalls. The typical handicapped stall is arranged perpendicularly to the curb; it has a width of 8 ft and a clearance of 5 ft on both sides.

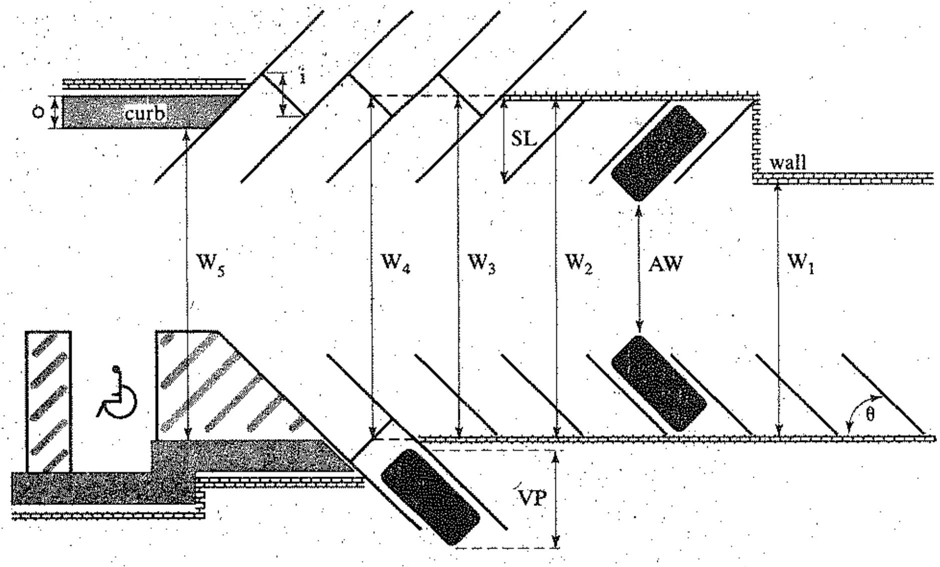


Figure 9.3.4 Typical dimensions for parking layout design.
(Reprinted with permission from *The Dimensions of Parking*, 3/e, ©1993 Urban Land Institute, Washington, D.C., USA.)

Security in parking lots is important for both drivers and passengers and for vehicles. At a minimum, proper illumination should be supplied, while fencing, retractable gates, and security personnel are additional options of security, depending on the form and the size of the lot, the crime risk at the area, and the hours of operation. Often in hospitals and campuses escort services are offered instead of or in addition to parking security.

There are many options in collecting parking fees: manual or electronic (digital) parking meters, collection boxes, coin/token collectors, authorization cards, time-stamped cards, collector (manual or automatic) at the exit, and permits (visually displayed or electronic: signal transmitted to sensor).

Automated parking with computer-controlled robotic dollies, which automatically park and retrieve automobiles, eliminates ramps at multistoried parking structures as well as valet parking operators. For the same capacity smaller structures are required—which are easier to design to look like an ordinary office building, while noise, fumes, internal congestion, accidents, and security problems are almost completely eliminated. Backup computers and electricity generators are essential for reliable operation. Systems with ability to process up to 240 automobiles in 1 h are currently in operation in the United States and elsewhere [9.15].

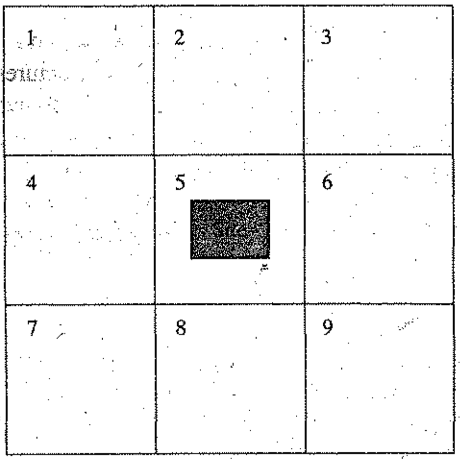
9.4 SUMMARY

This chapter presented practical applications of traffic engineering studies. Traffic impact studies are essentially small-scale (localized) planning studies where the impacts of a proposed development on the existing roadway network are assessed and evaluated. The results of traffic impact studies permit a site to be developed as proposed, or after certain modifications to the site per se and/or the surrounding roadway network and traffic controls. Large-scale traffic analyses, including urban signalized intersection networks, are conducted with powerful traffic simulation and planning packages.

Parking studies analyze parking, which is an indispensable component of urban transportation. Parking studies are either site-specific or area-specific. In a typical parking study the demand for parking is assessed, the present conditions are evaluated, and strategies for alleviating parking problems are suggested. Principles of parking design were also discussed.

EXERCISES

1. Consider a proposed office park with 3250 people employed in it. Estimate the morning and evening flows of vehicle-trips for a weekday and sketch the flows at the single access point of the site, which is at a midblock location. The site is located along an east-west road, and 30% of the site-generated traffic comes from (or is destined to) east. Identify high flow movements and potential operation deficiencies. The following trip-generation models apply: $T = 1/(1.97/X - 0.000053)$ with 92% enter and 8% exit (A.M.), and $\ln T = 0.75 \ln X + 0.87$ with 15% enter and 85% exit (P.M.)
2. Based on Exercise 1, estimate three modal split assignment scenarios. Scenario 1: 100% auto. Scenario 2: 70% auto, 20% carpooling with average occupancy 2.2 persons per car and 10% bus transit (one bus every 10 min enters site from each direction). Scenario 3: 40% auto, 25% carpooling with average occupancy 2.2 persons per car, 10% bus transit (one bus every 10 min enters site from each direction), and 25% rail transit, which stops near the site (walk access). Estimate the vehicle-trips for each scenario as well as the reduction from scenario 1. Is the reduction significant? Sketch A.M. volumes at the access point for all scenarios.
3. The site in Fig. E9.3 is expected to generate 2480 trips of which 90% will be from zones 1 through 9. Estimate the flows from each zone to the site. Assume that the rate of attraction is inversely proportional to the square of the distance between the site and each zone centroid.

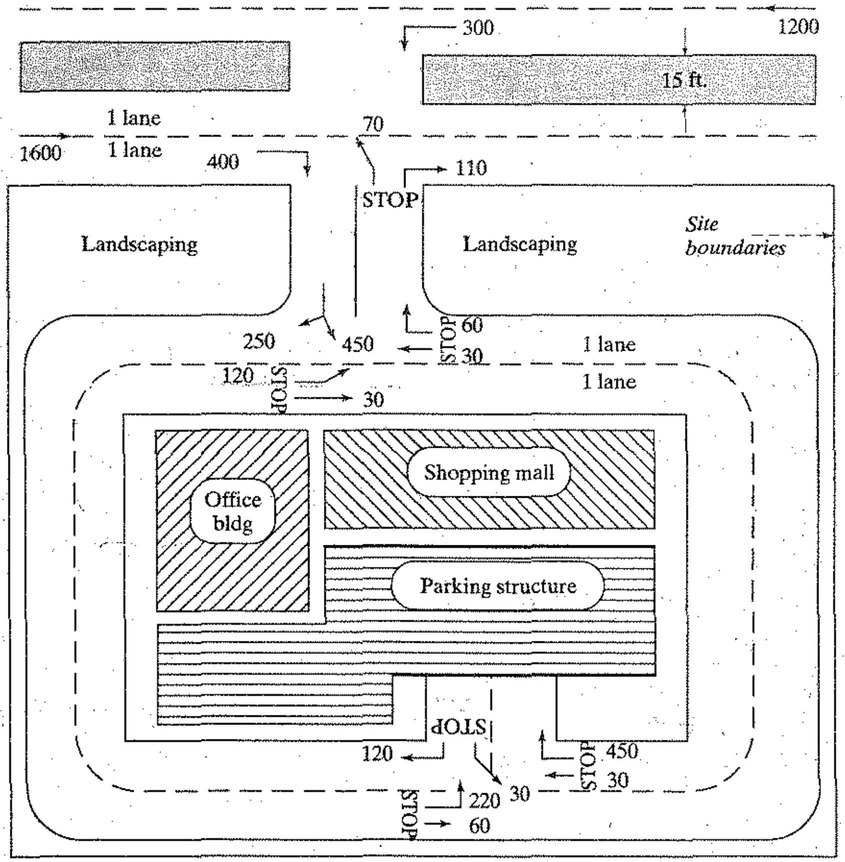


Zone	Population	Average distance
1	1300	10
2	3000	7
3	7000	9
4	800	6
5	3000	3
6	5500	6
7	500	10
8	1800	6
9	4400	10

Figure E9.3

4. Consider Exercise 3 and assume that the rate of attraction is inversely proportional to the square root of the distance between the site and each zone centroid. Do flows change, and how much? What is the underlying difference in people's behavior, which in one case suggests that the square of the distance should be used (Exercise 3), whereas in the other case the square root of the distance should be used?

5. Figure E9.5 presents the proposed layout of a site and the expected flows of traffic. Identify problematic spots at the access point of the site and in the internal circulation. Then design an improved layout that corrects most of the problems. (*Note:* Numbers on the drawing represent traffic volumes, unless noted otherwise.)



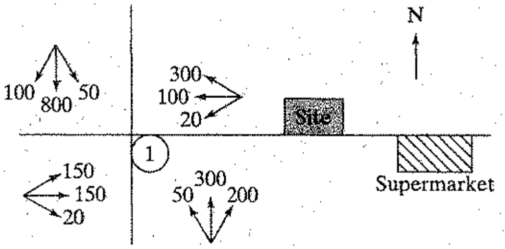
NOTES: 1) Buildings not to scale, 2) A.M. peak flows; reverse flows at P.M. peak

Figure E9.5

6. A proposed movie theater is expected to generate 250 new trips during the evening rush hour (80% enter and 20% exit). The consultant and the municipality have agreed to distribute the trips empirically. To accomplish this, vehicle counts were taken at a nearby supermarket and an intersection during the evening rush hour. Figure E9.6 shows these counts. Estimate the site-generated number of vehicle-trips by direction and movement at the site access point and at intersection *I*.

7. Based on the data of Example 9.4, assign 1500 vehicles to links A, B, and C using (1) the FHWA method, (2) the incremental method with five equal increments, and (3) the incremental method with five diminishing increments of 40, 25, 18, 12, and 5%.

8. Based on the data of Example 9.4, assign 1500 vehicles to links A, B, and C using (1) the incremental method with eight diminishing increments of 30, 20, 15, 10, 10, 5, 5, and 5% and (2) the incremental method with eight diminishing increments of 30, 22, 16, 12, 8, 6, 4, and 2%. Summarize the lessons learned from this and the previous exercise.



Traffic counts at supermarket	
From east, IN	57
From west, IN	49
OUT, eastbound	51
OUT, westbound	55

Figure E9.6

9. A developer plans to install a large movie theater complex with approximately 3000 seats. The developer's site is next to a large office building that has a parking structure attached to it. The developer's site has no room for parking and he is considering whether to accept the offer of the office building management to provide him with up to 300 stalls of parking on weekdays and up to 1500 stalls of parking during weekends for a cost of $4.00 per stall (validated parking to be absorbed by the developer) or to lease part of an adjacent parcel of vacant land at $2.50 per square foot per annum. Approximately 1/3 of the theater's sitting capacity is expected to be utilized during weekdays (100% during weekends) and 250 ft^2 is needed for each parking stall (includes parking space, access, and space for maneuvering). Round parking demand estimates upward to the closest hundred increment. Which option should he choose?

10. Design an open parking lot with space for 120 passenger cars. Use a square lot layout and place the access point in the middle of one of its sides. Angle stalls by 60° and use 9 × 19-ft stall size and 20-ft-wide aisles for access and maneuvering.

11. The following table presents a sample of parking observations of 15 curb stalls. Estimate the occupancy, accumulation, turnover, and average duration for this sample of parking stalls.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
9:00	Y	Y	N	Y	N	N	N	Y	Y	N	N	Y	Y	Y	N
9:30	Y*	N	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y*	Y	Y
10:00	Y	Y	Y*	Y	Y*	Y	N	Y	Y*	Y	Y	Y	Y	N	Y
10:30	N	Y*	N	Y*	Y	Y*	Y	Y	Y*	N	Y	Y	Y*	Y	Y*
11:00	N	Y*	Y	Y	Y	Y	Y	Y	Y*	Y	Y*	Y*	N	Y*	Y*
11:30	Y	Y	Y	N	N	Y*	Y*	Y*	N	Y*	N	Y	Y	Y	N
12:00	Y	N	N	N	Y	N	N	N	N	Y	N	N	N	Y	Y

Y, stall occupied; N, stall empty; * occupancy by a different vehicle from that in the preceding time period.

12. [Class Project] On the map shown in Fig. E9.12 a site will be developed as an office park with a gross leasable area (x) of 418.25 thousand square feet. Conduct a traffic impact study and a parking study for this site. The time points to be considered are (1) the present time, (2) the time when the development becomes fully operational (3 years from the present time; 3% annual growth of background traffic), and (3) a target year set at 7 years from the time when the development becomes operational (1.5% annual growth of background traffic).

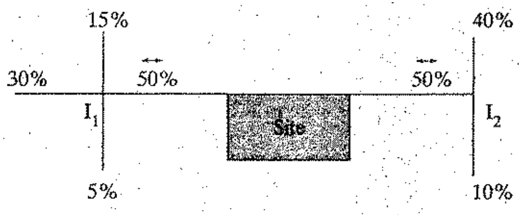


Figure E9.12

The trip-generation equations are

A.M. peak-hour trips (T): $\ln(T) = 0.8 \ln(x) + 2$

with 90% enter and 10% exit

P.M. peak-hour trips (T): $\ln(T) = 0.9 \ln(x) + 1$

with 15% enter and 85% exit

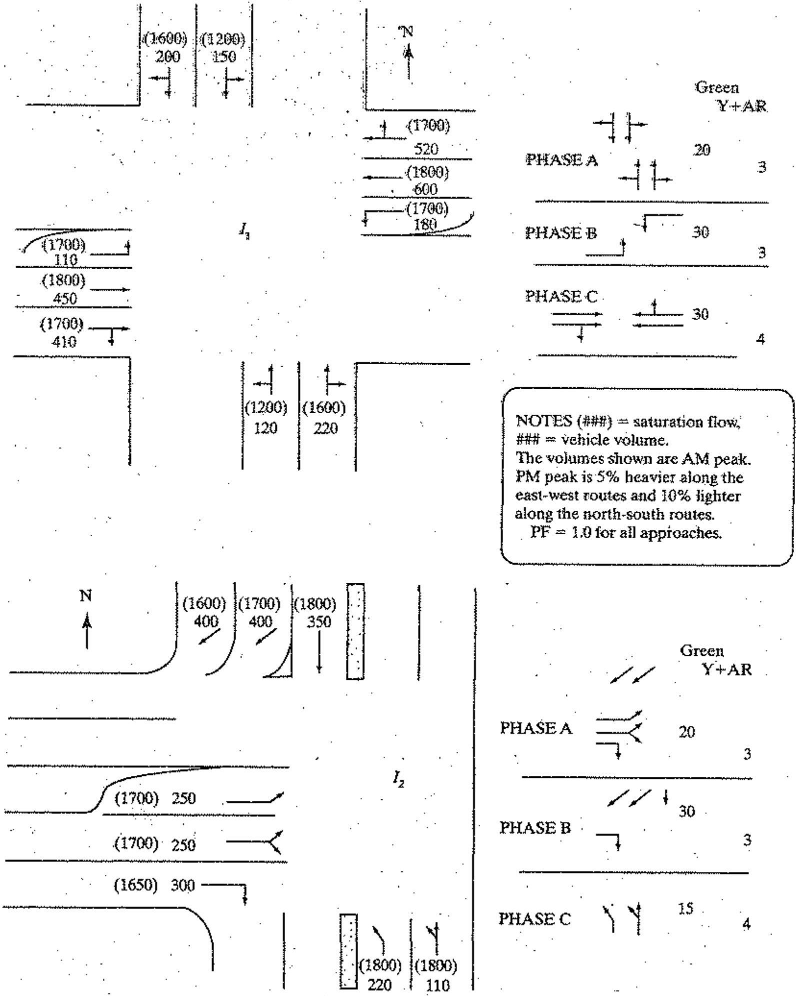
It is likely that signalized control should be placed at the site access point. This T intersection along with intersections I_1 and I_2 must be analyzed for 1 h in the A.M. and P.M. weekday peak period. New signal timings should be estimated and up to one additional lane per approach may be placed if the approach LOS is worse than C. Estimates of through traffic at the access point should result from the volumes of intersection I_2. All three intersections must operate under compatible cycle lengths so that arterial progression is maintained.

The expected modal splits are as follows:

Opening time ($t + 3$):	80% drive alone, 20% car pool (occupancy 2.4)
Target time ($t + 10$):	60% drive alone, 30% car pool (occupancy 2.0) 10% use rail transit

In t+3 and t+10 distribute volumes per lane so that flow ratios are approximately equal.

These model splits should be used in both the traffic impact and the parking study. Figure E9.12 includes the expected distribution of traffic. Existing conditions of intersections I_1 and I_2 are shown in Fig. E9.13.



NOTES (###) = saturation flow,
= vehicle volume.
The volumes shown are AM peak.
PM peak is 5% heavier along the east-west routes and 10% lighter along the north-south routes.
PF = 1.0 for all approaches.

Figure E9.13

REFERENCES

9.1 Transportation Research Board, *Highway Capacity Manual,* Special Report 209, National Research Council, Washington, DC, 1985.

9.2 ———, *Highway Capacity Manual,* Report 209, 3rd ed., National Research Council, Washington, DC, 1998.

9.3 Institute of Transportation Engineers, *Traffic Access and Impact Studies for Site Development: A Recommended Practice,* Final Report by the Transportation Planners Council, ITE, Washington, DC, September 1989.

9.4 ———, *Trip Generation,* 6th ed., ITE. Washington, DC, 1997.

9.5 Dial, R. B., "A Probabilisitic Multipath Traffic Assignment Model Which Obviates Path Enumeration," *Transportation Research,* 5 (1971): 83–104.

9.6 Surber, M., D. Shoup, and M. Wachs, *Effects of Ending Employer-Paid Parking for Solo Drivers,* Transportation Research Record 957; National Research Council, Washington, DC, 1984, pp. 67–70.

9.7 Organization for Economic Co-operation and Development, *Evaluation of Urban Parking Systems,* OECD Publications, Paris, December 1980.

9.8 Shoup, D. C., "In-Lieu Parking Fees," *Urban Mobility Professional,* No. 7 (January 1999).

9.9 Krohn, R. A., *Parking Studies: General Procedures and Case Studies,* Graduate Report UCB-ITS-GR-85-2, Institute of Transportation Studies, University of California at Berkeley, April 1985.

9.10 Institute of Transportation Engineers, *Parking Generation,* 2nd ed., ITE, Washington, DC, 1987.

9.11 Urban Land Institute, *Shared Parking,* a study conducted by Barton-Aschman Associates, Washington, DC, 1983.

9.12 Urban Land Institute and National Parking Association, *The Dimensions of Parking,* 3rd ed., Washington, DC, 1993.

9.13 Highway Research Board, *Parking Principles,* Highway Research Special Report 125, National Academy of Sciences, Washington, DC, 1971.

9.14 Institute of Transportation Engineers, "Employment Center Parking Facilities," a summary report by the ITE Technical Council Committee 6F-24, *Institute of Transportation Engineers Journal* (1988): 29–35.

9.15 Editorial, "Robots Could Take Anxiety Out of Parking," *Civil Engineering* (June 1989): 21–22.

10

Air Quality, Noise, and Energy Impacts

10.1 INTRODUCTION

The main objective of the travel-demand-forecasting models examined in Chapters 8 and 9 is to estimate the impacts of transportation systems that are directly related to travel. These impacts include the amount of trip making, the geographical distribution and orientation of trips, the utilization of the available and proposed modes of travel, and the consequences of these travel choices on the loading of the transportation network in terms of link flows and of the impedances (e.g., travel times) experienced by the users of the system.

Up to the 1960s transportation decisions in the public sector were generally based on the assessment of the capital and operating costs of transportation facilities vis-à-vis the expected direct improvements in the levels of service and travel times experienced by the user. The explicit consideration of indirect and nonuser impacts was generally confined to practical cost-related items such as the appropriate or just compensation of individuals and businesses for right-of-way acquisition and relocation.

As described in Chapter 7, the civil rights and the environmental movements of the 1960s contributed to the evolution of an altered perspective. Civil rights concerns affected the understanding of the role of transportation by addressing issues relating to the rights to mobility and accessibility to employment and other opportunities on the part of various societal subgroups. The environmental movement resulted in an increasing awareness about many indirect socioeconomic and environmental effects of transportation decisions (see Appendix A).

In this chapter we address three of the many transportation-related impacts that have become an integral part of contemporary transportation planning and decision making. These are air quality, noise generation, and energy consumption. For each of these impacts a brief historical note is presented, several mitigation strategies are described, and simple models for estimating the impact are illustrated. These models are simplistic and in some cases based on outdated data. They are included in this chapter to help the reader develop a "feel" of how various factors affect the environmental impacts covered. Contemporary practice involves the use of sophisticated computer software, the intricacies of which are beyond the scope of an introductory transportation textbook.

10.2 AIR POLLUTION

10.2.1 Background

The release of air pollutants in the atmosphere is a concomitant result of human activities. In some instances naturally produced air pollutants outweigh man-made pollution, but the latter tends to be concentrated in urbanized areas where people live and work. The problem of air pollution is not new. In early fourteenth-century London the smoke and odor consequences of coal burning became such a public nuisance that several commissions were appointed to combat them. In his book on the subject of air pollution Perkins [10.1] quotes the following declaration by King Edward I:

> Be it known to all within the sound of my voice, whosoever shall be found guilty of burning coal shall suffer the loss of his head.

More recently connections between air pollution and respiratory disease have been demonstrated and detrimental environmental effects on the global scale have been discerned. Several localized pollution episodes have resulted in documented deaths and highlighted the severity of the problem. A December 1948 episode in Donora, PA, and a December 1956 episode in London are most notable.

The first major law enacted by the U.S. Congress in relation to air pollution was the 1955 Air Pollution Act, which provided federal support for research into the subject. The Clean Air Act of 1963 recognized the contribution of "urbanization, industrial development, and increasing use of motor vehicles" to the problem and encouraged automobile manufacturers to address it. Two years later the 1965 Motor Vehicle Air Pollution Control Act provided for the establishment of vehicle-emission standards and opened the way to a series of amendments that led to the Air Quality Act of 1970, which provided for national ambient air quality standards, for a reduction of vehicle emissions of several pollutants by 90% of their 1970 levels, and for state implementation plans to conform to these provisions. The 1970 Federal-Aid Highway Act explicitly required that highway planning must be consistent with implementation plans to attain and maintain established regional ambient air quality standards. This requirement resulted in an accelerated level of activity regarding the monitoring and modeling of the air quality impacts of transportation systems and in the integration of air quality considerations into the transportation planning process. The subsequent legislative history of the issue has been, to say the least, tumultuous. Nevertheless, the problem has come to the forefront and has claimed a special place in the planning, design, and implementation of transportation projects. The 1990 amendments to the Clean Air Act imposed more stringent requirements on geographical areas that did not comply with tightened pollution standards. These places were designated as *nonattainment* areas. Table 10.2.1 presents the national air quality standards issued by the Environmental Protection Agency (EPA) in 1997. The ozone 1-h standard applied only to nonattainment areas in 1997 as a transitional standard to the 8-h level. Parenthetical values

TABLE 10.2.1 1999 National Air Quality Standards

Pollutant	Standard value		Standard type
Carbon monoxide (CO)			
8-h average	9 ppm	(10 mg/m^3)[b]	Primary
1-h average	35 ppm	(40 mg/m^3)[b]	Primary
Nitrogen dioxide (NO_2)			
Annual arithmetic mean	0.053 ppm	(100 μg/m^3)[b]	Primary & secondary
Ozone (O_3)			
1-h average[a]	0.12 ppm	(235 μg/m^3)[b]	Primary & secondary
8-h average	0.08 ppm	(157 μg/m^3)[b]	Primary & secondary
Lead (Pb)			
Quarterly average		1.5 μg/m^3	Primary & secondary
Particulate < 10 micrometers (PM-10)			
Annual arithmetic mean		50μg/m^3	Primary & secondary
24-h average		150 μg/m^3	Primary & secondary
Particulate < 2.5 micrometers (PM-2.5)			
Annual arithmetic mean		15 μg/m^3	Primary & secondary
24-h average		65 μg/m^3	Primary & secondary
Sulfur dioxide (SO_2)			
Annual arithmetic mean	0.03 ppm	(80 μg/m^3)[b]	Primary
24-h average	0.14 ppm	(365 μg/m^3)[b]	Primary
3-h average	0.50 ppm	(1300 μg/m^3)[b]	Secondary

[a] The ozone 1-h standard applied only to nonattainment areas in 1997 as a transitional standard to the 8-h level.

[b] Parenthetical values are approximate equivalent concentrations.

Source: Environmental Protection Agency [10.].

are approximate equivalent concentrations. *Primary standards* are required to protect public health, including the physiological response of children, the elderly, and people suffering from asthma and other respiratory diseases. *Secondary standards* are intended to protect the public welfare, including annoyance, loss of visibility, and damage to crops and livestock.

10.2.2 Problem Dimensions

The combustion of transportation fuels releases several contaminants into the atmosphere, including carbon monoxide, hydrocarbons, oxides of nitrogen, and lead and other particulate matter. Hydrocarbons, of which more than 200 have been detected in exhaust emissions, are the result of the incomplete combustion of fuel. Particulates are minute solid or liquid particles that are suspended in the atmosphere; they include aerosols, smoke, and dust particles. Photochemical smog is the result of complex chemical reactions of oxides of nitrogen and hydrocarbons in the presence of sunlight.

Once emitted into the atmosphere, air pollutants undergo mixing or diffusion, the degree of which depends on topographic, climatic, and meteorological conditions. These include wind speed and direction, and atmospheric stability.

The assessment of the air pollution effects of transportation may be undertaken at three levels: microscale analysis in the immediate vicinity of a transportation facility such as a highway, mesoscale analysis in areas that are somewhat removed from the facility, which includes the contribution of other mobile and stationary sources of pollution, and macroscale analysis, extending from the regional to the global levels. Available air pollution estimation models range from simple models that provide rough estimates of emission levels to very complex numerical models that trace the diffusion of pollutants in space and time and also simulate the chemical processes that follow.

10.2.3 Emission Levels

Vehicular emissions of air pollutants are usually measured in grams per vehicle-mile of travel and are related to several factors, including vehicle type and age, ambient temperature, and altitude. The operating cycle, which consists of starts and stops, speed changes, and idling time, is also an important factor. A disproportionate fraction of carbon monoxide and hydrocarbons are emitted during cold starts of the engine.

The general relationships between speed and emissions are illustrated by Fig. 10.2.1. Carbon monoxide emissions generally decrease with speed, partly due to the air-to-fuel ratio supplied to the engine at different speeds. Up to about 30 to 40 mi/h a similar relationship occurs in the case of hydrocarbons, but a mild increase in emissions is seen thereafter. The emission of nitrogen oxides exhibits a different pattern; that is, it generally increases with speed.

A number of emission models utilizing these factors have been developed and many traffic simulation models (see Chapter 15) have been supplemented by emission-estimating subroutines. The EPA has developed MOBILE [10.2] a computer program that estimates the emissions resulting from various combinations of traffic flows, vehicle mixes, and other factors. The model's most common version is MOBILE5a. It was released in 1993. MOBILE6 became available in the late 1990s. The latest version attempts to account for the separation of start and running exhaust emissions, roadway facility type, average traffic speeds, and so on. Additional important features (and modeling issues) shared by current models include the following:

- Technology shares. Emissions from highway vehicles are estimated on a fleetwide basis using information on the share of each model year's fleet that use different technologies (e.g., fuel delivery systems, catalytic converter type).
- Aging and the corresponding increase in emissions over time as vehicles accumulate mileage and components, including emission control components, age, and deterioration
- Advanced engine management and diagnostics such as the introduction of second-generation onboard diagnostic systems (OBD-II) to the light duty fleet introduces complications to the modeling of aging.

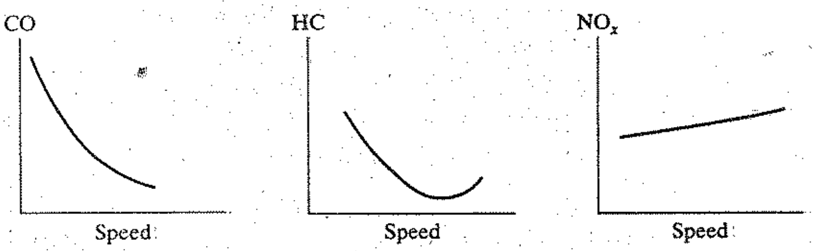


Figure 10.2.1 General relationship between speed and emissions.

- Effects of specific fuel content such as sulfur and oxygenates
- The estimation of emissions from heavy duty vehicles is complicated because their engines are regulated on a mass/work basis (grams per brake horsepower-hour), whereas emission analysis generally requires emissions on a mass/activity basis (i.e., grams per mile), which necessitates the use of complex conversion factors.
- Information on the total numbers of vehicles by vehicle type, the registration distributions by the age of each vehicle type, and the annual mileage accumulation rates by the age of each vehicle type are required to model emission factors for the entire in-use fleet of highway vehicles. Modelers should be able to use local instead of national data.
- Nonexhaust, or evaporative, emissions and leaks

The California Air Resources Board (CARB) estimates on-road motor vehicle emissions with a package of models called the Motor Vehicle Emission Inventory (MVEI) models. The three main computer models that form the MVEI are CALIMFAC, WEIGHT, and EMFAC. The CALIMFAC model produces base emission rates for each model year when a vehicle is new and as it accumulates mileage and the emission controls deteriorate. The WEIGHT model calculates the relative weighting each model year should be given in the total inventory, and each model year's accumulated mileage. The EMFAC model uses these pieces of information, along with the correction factors and other data, to produce fleet composite emission factors. The software and its documentation are free and available from download.

CALINE, developed by the California Department of Transportation (CALTRANS), is a dispersion model for predicting air pollutant levels near highways and arterial streets. The program computes the effect on air quality, measured at several locations (the maximum number of locations is 20 for CALINE3) of a roadway situated on a relatively flat terrain. It is based on the Gaussian plume approximation, but it can also account for deposition and sedimentation in order to compute the concentration of particulate matter. CALINE3 was developed in the late 1970s and is available from NTIS. The CALINE4 model is available from CALTRANS.

A simple model for carbon monoxide based on MOBILE has been proposed by Raus [10.3]. This model uses several nomographs that are based on a typical 1980 vehicle mix for various altitudes and ambient temperatures and is included here for illustrative purposes. The family of curves corresponding to altitudes up to 4000 ft above sea level is shown in Fig. 10.2.2.

Example 10.1

According to a traffic forecast, a proposed 4-mi highway is expected to carry 3500 veh/h during the 2-h peak period of the day at an average travel time of 16 min. Apply the Raus model to estimate the total peak-hour emissions of carbon monoxide on the highway for the typical autumn day (60°F).

Solution The total number of vehicle-miles traveled during the typical peak period is

$$(3500 \text{ veh/h})(2 \text{ h/peak})(4 \text{ mi}) = 28{,}000 \text{ veh-mi/peak period}$$

The estimated average speed is 15 mi/h, and according to Fig. 10.2.2, the emission rate corresponding to this speed is 78 g/veh-mi. As a result, the total emissions of carbon monoxide are estimated to be 2,184,000 g, or 4811 lb, per peak period.

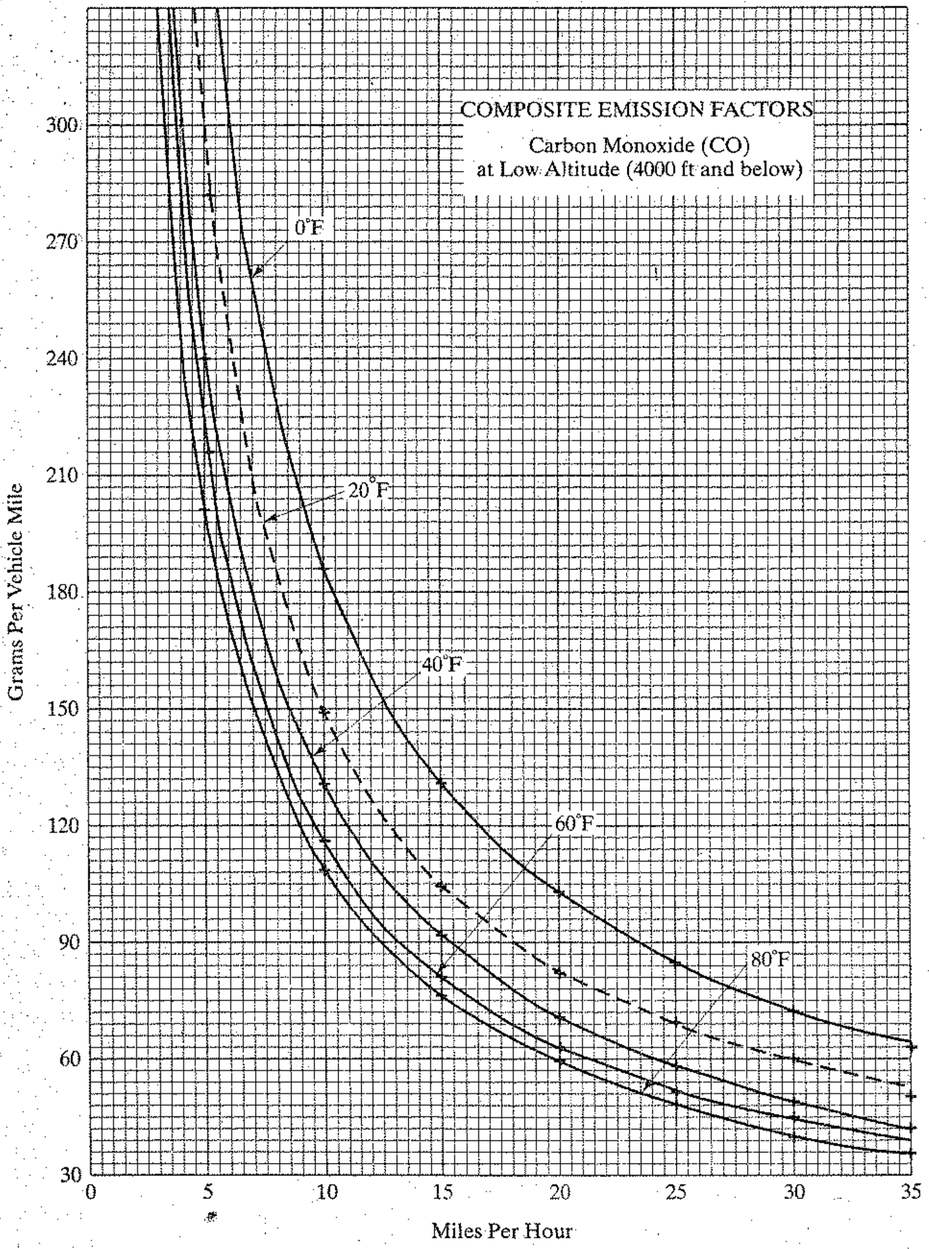


Figure 10.2.2 Carbon monoxide emission factors, 1980 vehicle mix. (From Raus [10.3].)

10.2.4 Air Pollution Dispersion

While the emission level is an important measure of the air pollution impact of various sources, it is the concentration of pollutants in the atmosphere that defines the levels and times of exposure. Following the emission of pollutants, dispersion and chemical oxidation take place in the atmosphere. The dispersion of a pollutant is affected by the strength of the source and topographic and meteorological conditions. The topography of the terrain in the vicinity of the source of pollution affects, among other items, the wind profile near the ground and the generation of turbulence in the form of eddies. Special conditions related to highway facilities also affect the dispersal of highway-generated pollution. For example, the EPA HIWAY model [10.4] analyzes at-grade and depressed highways differently to estimate the concentrations of nonreactive pollutants from highway traffic at various downwind locations. Also, air pollutants released on roadways passing through densely developed urban areas tend to be trapped in the street canyons that are formed by rows of buildings at both sides of the roadway.

One of the most important meteorological conditions that affect the mixing of pollutants is the temperature lapse rate, which is defined as the rate of change of temperature with altitude. This rate is usually referenced to the *adiabatic lapse rate* of −5.4°F per 1000 ft, which corresponds to an atmosphere that is characterized by neutral stability, that is, a situation where air particles tend to maintain their positions. When the temperature drops at a faster rate than the adiabatic (i.e., at a *superadiabatic lapse rate*), the atmosphere is unstable and vigorous mixing takes place. On the other side of the adiabatic lapse rate *subadiabatic lapse rates* tend to inhibit mixing. Due to various meteorological combinations, sometimes certain layers in the atmosphere experience an increase of temperature with altitude. This is known as a *temperature inversion* and is critical especially when it occurs in a layer close to the ground because pollutants are trapped within this layer. The *mixing height* is the height of the atmospheric layer within which mixing occurs. This height varies from locality to locality and also exhibits daily and seasonal variation. The degree of mixing is a function of the atmospheric stability of this layer. Typically the atmosphere near the earth's surface becomes unstable in the morning, allowing for energetic mixing within the mixing layer, which attains its maximum height in the afternoon [10.4].

One of the simplest mathematical models of air pollution diffusion is the box model, which is described in the next subsection. More sophisticated models employing numerical integration of complex Gaussian equations and the modeling of chemical processes require the use of computer algorithms.

10.2.5 The Box Model

The *box model* may be used to approximate the concentration of air pollution within an atmospheric volume defined by a rectangular area and extending to the altitude of the mixing height H, as shown in Fig. 10.2.3. Pollutants emitted into the box at a constant rate E in pollutant weight per unit time are assumed to be mixed instantaneously with the air volume of the box. Clean air is assumed to enter the box at a speed U, and air containing the same concentration as the interior of the box is assumed to exit from the opposite side. The concentration $C(t)$ at any time t inside the box is expressed in pollutant weight per unit volume. Based on these assumptions, the following balance equation applies:

$$E - FC = V\left(\frac{dC}{dt}\right) \tag{10.2.1}$$

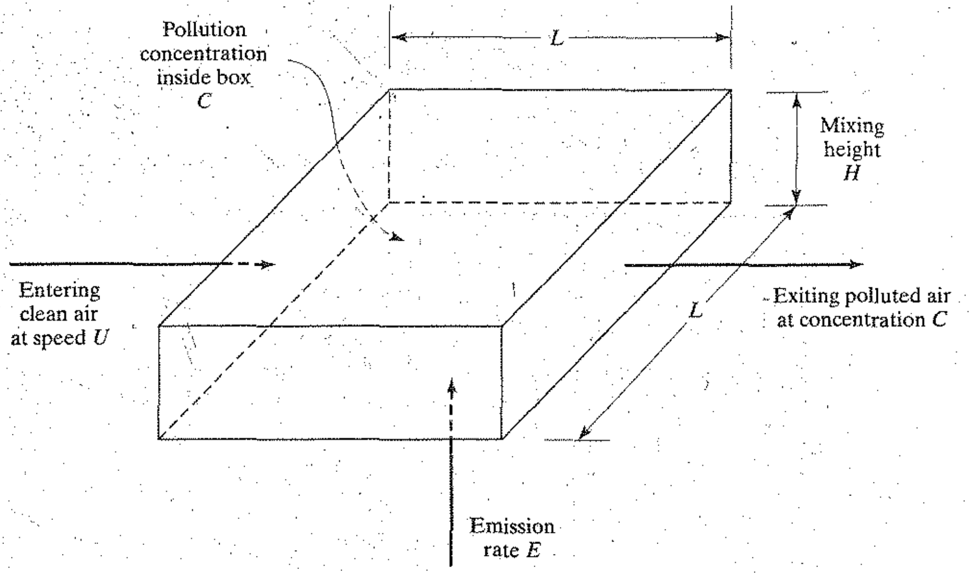

Pollution concentration inside box C
L
Mixing height H
Entering clean air at speed U
L
Exiting polluted air at concentration C
Emission rate E

Figure 10.2.3 Box model.

where

$$F = ULH = \text{airflow (volume per unit time)}$$

$$V = L^2H = \text{box volume}$$

Equation 10.2.1 expresses the rate of change of pollution inside the box as the difference between the amount of pollution entering the box (E) and the amount of pollution exiting the box (FC).

Example 10.2

Solve Eq. 10.2.1 for concentration as a function of time assuming that the air within the box is initially clean (i.e., $C_0 = 0$) and that U, E, and H are constant.

Solution Rewrite Eq. 10.2.1 as

$$\frac{dC}{dt} + \frac{F}{V}C = \frac{E}{V} \tag{10.2.2}$$

Equation 10.2.2 is a first-order linear differential equation of the form

$$\frac{dy}{dx} + p(x)y = g(x) \tag{10.2.3}$$

which can be solved by multiplying both sides by the integrating factor

$$f(x) = e^{\int p(x)\,dx} \tag{10.2.4}$$

thus rendering the left-hand side into the exact differential of the product $yf(x)$. In this case the integrating factor is

$$f(t) = e^{(F/V)t} \tag{10.2.5}$$

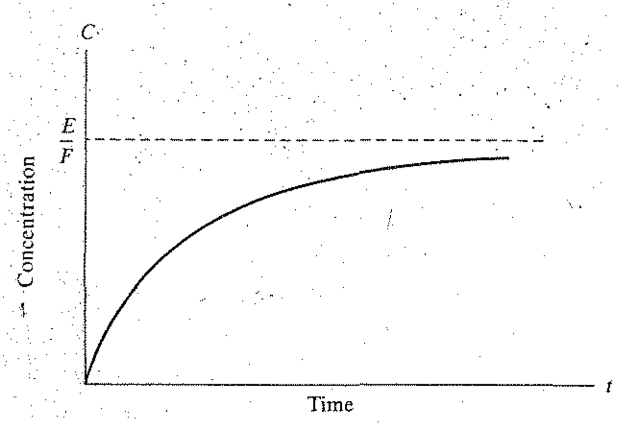


Figure 10.2.4 Pollutant concentration as a function of time.

Hence

$$\frac{d}{dt}(Ce^{(F/V)t}) = \frac{E}{V}e^{(F/V)t}$$

Integrating with respect to t gives us

$$Ce^{(F/V)t} = \frac{VE}{FV}e^{(F/V)t} + K$$

or

$$C = \frac{E}{F} + Ke^{-(F/V)t} \tag{10.2.6}$$

where K is the constant of integration, which can be evaluated at the initial condition $C(0) = 0$ to be

$$K = -\frac{E}{F}$$

As a result, the solution to Eq. 10.2.1 becomes

$$C = \frac{E}{F}[1 - e^{-(F/V)t}] \tag{10.2.7}$$

which is plotted in Fig. 10.2.4. Thus, under the simplifying assumptions of this model, the pollution concentration of the interior of the box tends toward a steady-state level of E/F.

More complex computer-based emission, dispersion, and chemical models are available. Many of these models provide linkages to the travel demand forecasting models covered in Chapter 8 to help estimate the air quality impacts of regional land-use and transportation actions (e.g., Ref. [10.5]).

10.3 NOISE GENERATION

10.3.1 Background

Sound is acoustical energy released into the atmosphere by vibrating or moving bodies. Therefore sound is amenable to objective scientific measurement and investigation. On the other hand *noise* is undesirable or unwanted sound and as such it is cloaked with a certain degree of subjectivity. Extended exposure to excessive sound has been shown to produce physical and psychological damage. Because of its annoyance and disturbance implications, noise adds to mental stress and hence affects the general well-being of those who are exposed to it. Undoubtedly noise has always been a major source of friction between individuals.

Transportation operations are major contributors to noise in the modern urban environment. Noise is generated by the engine and exhaust systems of vehicles, aerodynamic friction, and the interaction between the vehicle and its support system (e.g., tire-pavement and wheel-rail interactions) [10.6]. Because noise diminishes with distance from the source, the most serious transportation-related noise problems are confined to transportation corridors (e.g., highway and railway corridors and aircraft flight paths) and at major transportation hubs (e.g., airports and transit terminals).

Seiff [10.7] reports that in 1970 the Bureau of Motor Carrier Safety issued rules relating to noise levels in the interior of commercial vehicles based on the belief that the safety of these operations would be compromised by the resulting driver fatigue and hearing problems. About the same time various states and local communities began to establish community noise regulations, including motor-vehicle noise standards. The passage of the Noise Control Act of 1972 at the federal level marked the recognition of the problem as a major detriment to urban living of nationwide proportions. Pursuant to the provisions of this act, the FHWA issued the 1973 Policy and Procedure Memorandum [10.8], which promulgated noise standards for various types of land use and stated that the FHWA:

> ... encourages the application of the noise standards at the earliest appropriate stage in the project development process.

Table 10.3.1 presents the noise standards issued by the FHWA in 1973 for several categories of land use, which have remained in effect into the late 1990s. Thus concern about the noise impacts of transportation vehicles and facilities officially entered the calculus of transportation design, planning, and implementation. For requirements on undeveloped lands, the table refers to other sections of the Policy and Procedure Memorandum, PPM [10.8].

10.3.2 Noise Measurement

The quantity of energy or the intensity of a single sound is usually measured on a relative logarithmic scale that employs a unit called a *bel* (B) or in terms of its subdivision, the *decibel* (dB). A bel represents a tenfold increase in energy and is measured in relation to a reference intensity I_0, which is usually taken at the threshold of human hearing. The intensity I of a sound corresponding to L bels is

$$I = 10^L I_0 \qquad (10.3.1)$$

Solving for L yields

$$L = \log_{10}\left(\frac{I}{I_0}\right) \quad \text{B} \qquad (10.3.2)$$

At the threshold of hearing $I = I_0$ and the noise level L is equal to zero. When L is about 14, the sound becomes painful to the human ear.

For finer scaling the bel is divided into 10 dB, and Eq. 10.3.2 becomes

$$L = 10\log_{10}\left(\frac{I}{I_0}\right) \quad \text{dB} \qquad (10.3.3)$$

TABLE 10.3.1 FHWA NOISE STANDARDS

(a) Design noise level/land-use relationships

Land-use category	Design noise level, L_{10}	Description of land-use category
A	60 dBA (exterior)	Tracts of land in which serenity and quiet are of extraordinary significance and serve an important public need, and where the preservation of those qualities is essential if the area is to continue to serve its intended purpose. Such areas could include amphitheaters, particular parks or portions of parks, or open spaces which are dedicated or recognized by appropriate local officials for activities requiring special qualities of serenity and quiet.
B	70 dBA (exterior)	Residences, motels, hotels, public meeting rooms, schools, churches, libraries, hospitals, picnic areas, recreation areas, playgrounds, active sports areas, and parks.
C	75 dBA (exterior)	Developed lands, properties or activities not included in categories A and B.
D	—	For requirements on undeveloped lands see paragraphs 5a(5) and (6), this PPM.
E[a]	55 dBa (interior)	Residences, motels, hotels, public meeting rooms, schools, churches, libraries, hospitals, and auditoriums.

[a] See part (b).

(b) Noise-reduction factors

Building type	Window condition	Noise reduction due to exterior of the structure (dB)	Corresponding highest exterior noise level that would achieve an interior design noise level of 55 dBA (dBA)
All	Open	10	65
Light frame	Ordinary sash		
	Closed	20	75
	With storm windows	25	80
Masonry	Single glazed	25	80
	Double glazed	35	90

Source: Federal Highway Administration [10.8].

An alternative formulation of Eq. 10.3.2 is based on the fact that the sound energy is proportional to the square of the frequency f of the sound; that is,

$$I = af^2 \tag{10.3.4}$$

where a is a proportionality factor. Substitution of this equation in Eq. 10.3.2 yields

$$L = 20 \log_{10}\left(\frac{f}{f_0}\right) \quad \text{dB} \tag{10.3.5}$$

where f is the frequency of the sound being measured and f_0 is the reference frequency at the threshold of hearing. A sound pressure of 20 μPa corresponds to this reference frequency.

Equations 10.3.1 through 10.3.5 apply to sounds that consist of a single frequency. However, typical environmental noises consist of combinations of frequencies of which only those in the approximate range from 500 to 10,000 Hz are detectable by humans. A single noise-level scale in decibels that combines the effect of multifrequency noises in a manner that simulates the sensitivity and response of humans discriminates or weighs against frequencies that lie outside this range. The most common weighting scheme is referred to as the *A-weighted scale* and gives measurements that are measured in A-weighted decibels, or dBA. Figure 10.3.1 presents the A-weighted decibel levels of several common environmental sounds that lie between the threshold of hearing (i.e., dBA = 0) and the level of physiological pain [10.9]. More precisely, the A-weighting scheme de-emphasizes frequencies below 1 kHz and above 6.3 kHz.

10.3.3 Noise Propagation and Mitigation Strategies

Once generated at a source, unshielded noise spreads out spherically as it travels through the air away from the source. Consequently the intensity of the sound diminishes with distance from the source. In addition to these losses in intensity due to spreading, absorption losses also take place as the sound energy is transferred between air particles. When the sound waves encounter natural and manufactured solid objects, they undergo bending or diffraction and reflection, the degree of which depends on the characteristics of the object. Trees and other vegetation, for example, tend to reflect the sound waves in a diffused pattern and are considered to be good interceptors of noise.

The major thrust of noise control strategies is to minimize the noise levels to which the population is exposed. Three categories of transportation noise control strategies are possible: source controls, noise path controls, and receiver-side controls. Potential source controls include vehicle control devices, vehicle maintenance practices, traffic controls, and highway design controls. Noise path controls include the erection of appropriately designed noise barriers that reflect and diffuse noise and the provision of buffer zones between the transportation facility and the population to provide a distance over which noise can be attenuated. Noise control strategies at the receptor site include public awareness programs and building design practices. Figure 10.3.2 illustrates the effects of elevating or depressing the highway.

10.3.4 Noise Measures

Figure 10.3.3, from a U.S. DOT study [10.10], shows that the noise levels generated by transportation facilities are characterized by a good amount of variability with respect to time. It is, therefore, necessary to establish meaningful statistical noise measures that describe the magnitude of the problem while capturing this variability. Commonly used statistical measures include the following:

1. L_p denotes the noise level at a receptor site that is exceeded p percent of the time. Commonly used levels of this measure include the noise level that is exceeded 10, 50, and 90% of the time. L_{10} is a peak noise level used by most highway departments in the United States and endorsed by the FHWA. Noise level L_{90} is a background level that is exceeded most of the time.
2. The equivalent noise level denoted by L_{eq} is defined as

$$L_{\text{eq}} = 10 \log_{10}\left(\frac{1}{T}\int_0^T \frac{f^2}{f_0^2}\,dt\right) \tag{10.3.6}$$

Sound levels and human response		
Common sounds	Noise level (dB)	Effect
Carrier deck jet operation Air raid siren	140	Painfully loud
	130	
Jet takeoff (200 ft) Thunderclap Discotheque Auto horn (3 ft)	120	Maximum vocal effort
Pile drivers	110	
Garbage truck	100	
Heavy truck (50 ft) City traffic	90	Very annoying Hearing damage (8 h)
Alarm clock (2 ft) Hair dryer	80	Annoying
Noisy restaurant Freeway traffic Man's voice (3 ft)	70	Telephone use difficult
Air-conditioning unit (20 ft)	60	Intrusive
Light auto traffic (100 ft)	50	Quiet
Living room Bedroom Quiet office	40	
Library Soft whisper (15 ft)	30	Very quiet
Broadcasting studio	20	
	10	Just audible
	0	Hearing begins

Figure 10.3.1 Sound levels and human response. (From Environmental Protection Agency [10.9].)

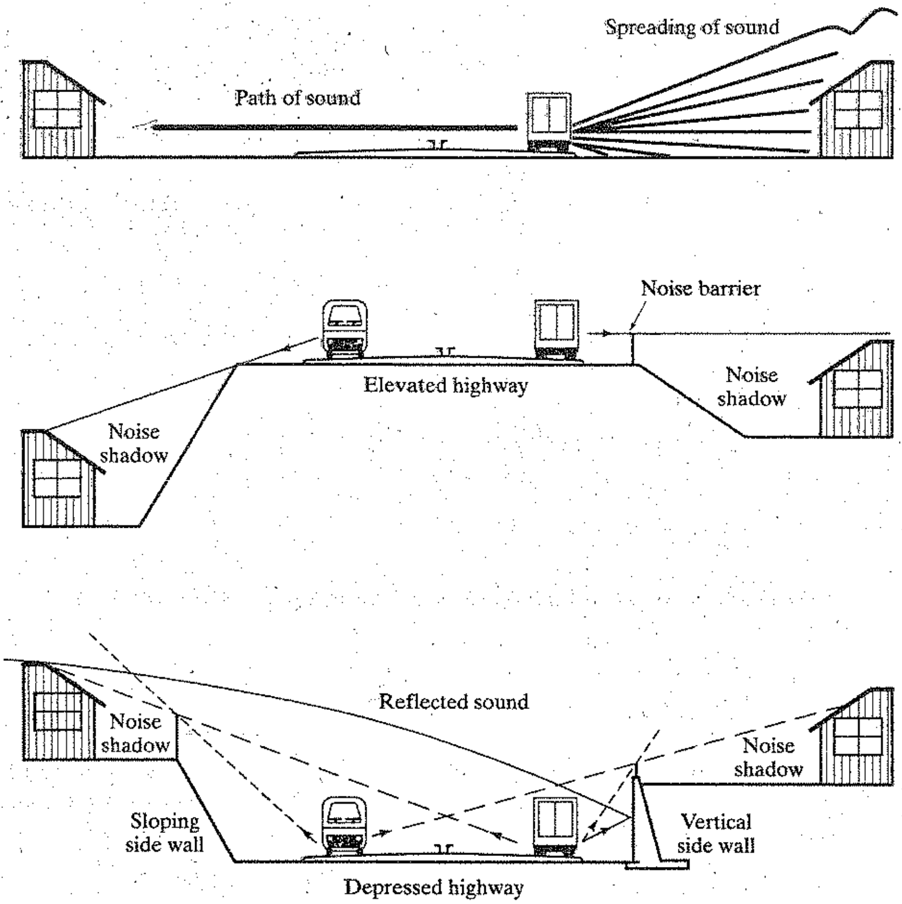


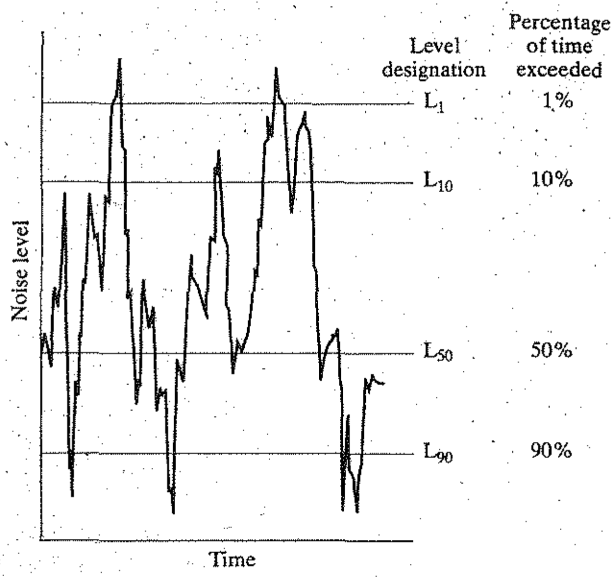
Figure 10.3.2 Effects of elevating and depressing highways, and noise barriers on noise exposure of adjacent land uses.

where T is the period of time over which the measurement is made. The equivalent noise level may be approximated by a series of N discrete measurements as follows:

$$L_{\text{eq}} = 10 \log_{10} \left(\frac{1}{N} \sum_{i} 10^{(L_i/10)} \right) \qquad (10.3.7)$$

where L_i is the average noise level during interval i.

Other noise impact measures for the assessment of transportation noise have been proposed. Some of these combine the preceding measures in various ways that attempt to capture the annoyance caused by the noise [10.11].



Typical of the highest levels occurring, although momentary peak levels from very noisy sources (such as an unmuffled truck or motorcycle) may be 5 to 10 dBA above the L_1

This level appears in ppm 90.2 noise specifications (used for this study)

Also known as the mean

Figure 10.3.3 Noise-level variability. (From U.S. Department of Transportation [10.10].)

10.3.5 Mathematical Models of Transportation Noise

Wesler [10.12] traces the first mathematical formulation of traffic noise to the following empirical equation presented in the 1952 Wright Air Development Center *Handbook of Acoustic Noise Control* [10.13]:

$$L_{50} = 68 + 8.5 \log V - 20 \log D \quad \text{dB} \tag{10.3.8}$$

where

V = traffic volume, in veh/h

D = distance from a traffic line to the observer, in ft

Note that this equation does not recognize the fact that a given volume can occur at two different speeds.

Since then many researchers have attempted to calibrate highway-related noise models for various traffic conditions. Various manual and computer-based models became available for the analysis of noise impacts and the design of noise-amelioration devices such as noise barriers. The FHWA, for example, has published several models, which have been subsequently enhanced and refined [10.14]. These models include several manual and computerized solution procedures. Figure 10.3.4 is a nomograph that can be used to estimate the unshielded noise level (i.e., in the absence of noise barriers) at some distance from a highway. Inputs to this model are the volumes and speeds of automobiles, medium trucks, and heavy trucks using the highway, and the output consists of an estimate of L_{10} at a given distance away. The noise level caused by each component is calculated by the procedure

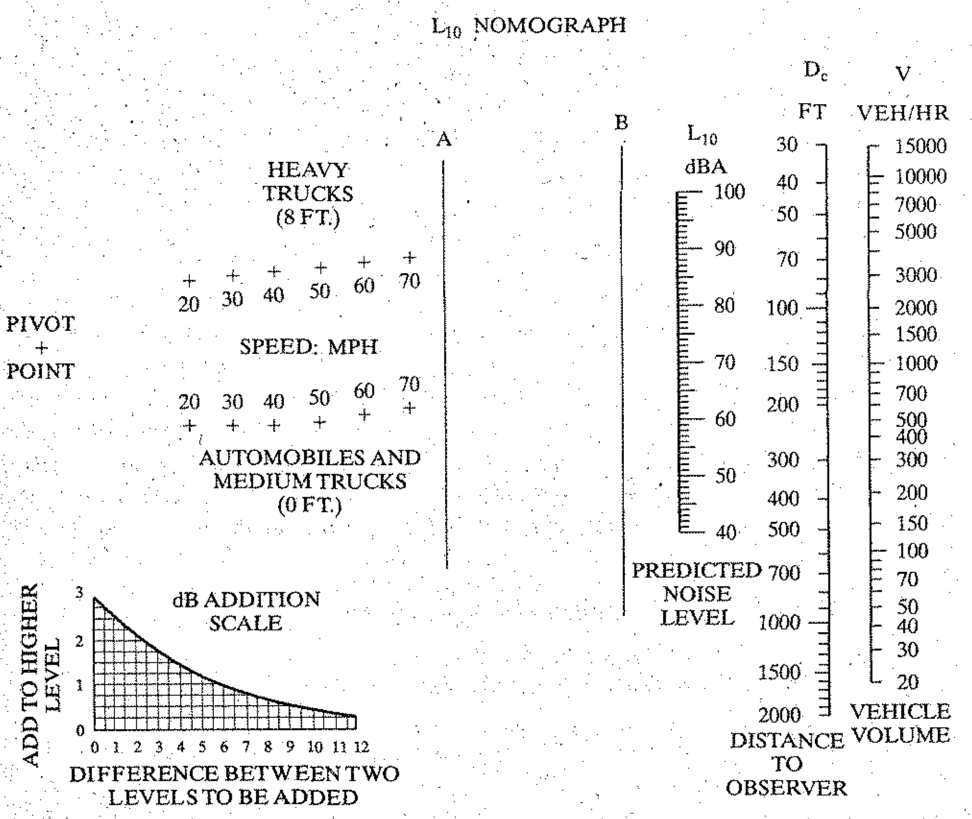


Figure 10.3.4 L_{10} nomograph. (From Kugler et al. [10.14].)

described next and added logarithmically to arrive at the total highway noise level. The medium truck volume is converted to automobile equivalents by a factor of 10. If the speeds of medium trucks and automobiles are equal, the two can be combined into one group. Heavy trucks are always analyzed separately. Automobiles and medium trucks differ from heavy trucks in that the major part of the noise emitted by the former is at the pavement level due to the interaction between the tires and the pavement. Heavy truck noise on the other hand is emitted from exhaust systems, which are located about 8 ft above the pavement level. This difference is denoted on the noise nomograph but is more important to the design, and analysis of barriers than to the simplified estimation of unshielded noise addressed here. To estimate the noise level produced by highway traffic, the first three steps of the following procedure are applied to each vehicular volume component and the results are combined, as explained in step 4.

Step 1. A straight line joining the pivot point at the extreme left-hand side of the nomograph to the point corresponding to the mean speed is extended until it intersects line *A*. Note that two sets of speed-related points are included: one for automobiles and medium trucks and the other for heavy trucks.

Step 2: A second straight line is drawn from the point obtained in step 1 to the traffic volume V on the scale located at the extreme right-hand side of the chart. The point of intersection of the second line and line B is noted.

Step 3. A third line is drawn from the point of intersection of line B to the distance D_c for which the noise level is calculated. The intersection of the third line and the L_{10} scale next to line B represents the required L_{10} estimate. The distance to the observer D_c may be either taken approximately from the middle of the highway or, if desired, from the middle of individual lanes. In the latter case the volume and speed inputs must be known by lane.

Step 4. The calculated L_{10} levels for the various highway flow components are combined. Because of the logarithmic nature of the dBA scale, the L_{10} levels cannot be added arithmetically. Instead, if two sounds are to be added, an incremental amount, depending on the difference in the two noise levels, is added to the higher of the two. The insert located at the lower left-hand side of the chart gives the magnitude of the incremental amount. Thus two sounds of equal intensity (i.e., zero difference) combine into a level that is only 3 dBA higher. The proof of this fact is given in Example 10.3.

Example 10.3

Prove that two sounds of equal intensity produce a decibel level that is only 3 dBA higher.

Solution The energy contained in the two sounds combined is equal to two times the sound energy I of either of the two alone. By Eq. 10.3.3, the combined level L is

$$\begin{aligned} L &= 10 \log_{10}\left(\frac{2I}{I_0}\right) \\ &= 10 \log_{10}\left(\frac{I}{I_0}\right) + 10 \log_{10}2 \\ &= 10 \log_{10}\left(\frac{I}{I_0}\right) + 3 \end{aligned}$$

Discussion The first term on the righthand side is the decibel level corresponding to the sound intensity I to which 3 dB is added when the two sounds are combined. In conjunction with the sound addition insert to Figure 10.3.4, when more than two noise sources are to be added, their magnitudes are first listed in decreasing order. The increment contributed by the lowest level to the next lowest is read from the graph and added to the latter. The result is then combined with the next list entry, and the procedure continues until all components have been included.

Example 10.4

A straight at-grade highway accommodates 3600 passenger cars and 40 medium trucks per hour. The average speed of the two vehicle types is the same and equals 40 mi/h. Plot the relationship between the noise level, L_{10}, and the distance from the highway.

Solution The procedure relating to Fig. 10.3.4 is applied, using a combined volume of

$$3600 + 40 \times 10 = 4000 \text{ automobile equivalents per hour}$$

and the results obtained in relation to various distances D_c are plotted in Fig. 10.3.5. The figure illustrates the attenuation of noise over distance from about 78 dBA at a distance of 30 ft to about 51 dBA at a distance of 1500 ft.

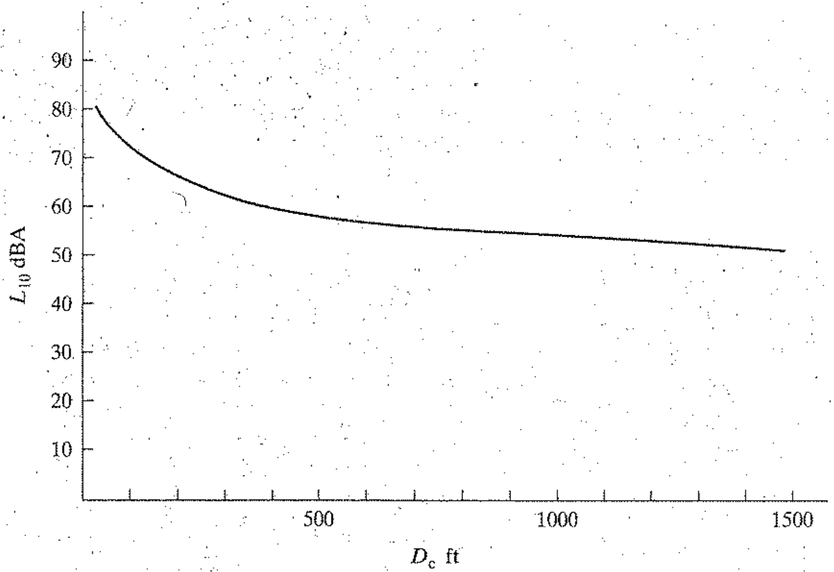


Figure 10.3.5 Noise level and distance from the highway.

In 1998, after several years of extensive research, measurement, calibration, and validation, FHWA issued a new computerized noise model that was simply called the FHWA *Traffic Noise Model* (TNM) and declared that no other noise model would be acceptable for federally supported highway projects [10.15, 10.16]. TNM uses a complex sound "ray tracing" routine to compute the noise level at user-specified receptor locations. It captures the contribution of five vehicle types (i.e., automobiles, medium trucks, heavy trucks, buses, and motorcycles), four pavement types, and the effects of traffic controls (i.e., stop signs, tollbooths, traffic signals, and on-ramp start points). Sound propagation accounts for atmospheric absorption, intervening ground acoustical and topographic characteristics, barriers (i.e., walls and berms), rows of buildings, and areas of heavy vegetation. The software is capable of presenting the results in the form of noise contours and contour maps showing the *differences* in noise levels between two noise barrier designs. In early 1999 FHWA issued a set of look-up tables to provide designers with a quick screening tool of potential mitigation designs [10.17].

10.4 ENERGY CONSUMPTION

10.4.1 Background

The enormous strides in industrial and economic growth that occurred in the United States during the twentieth century have been closely related to an ample supply of inexpensive energy, particularly energy derived from fossil fuels. Around 1970 the population of the United States constituted about 6% of the world's population but used approximately 30% of the global petroleum consumption. A little more than half of the petroleum used in this country is expended for transportation-related purposes, and of this amount the private automobile accounts for close to two-thirds, or about one-third of the total petroleum consumed in the United States. The potential impact that energy shortages can have was experienced during World War II, when strict rationing and allocation of energy and other resources had to be imposed. After the war energy consumption resumed its upward spiral and the problem came to the forefront in 1973 when the Organization of Petroleum Exporting Countries (OPEC) imposed an oil embargo and subsequently raised the price of crude oil. The economic effects of this action reverberated around the globe.

10.4.2 National Response to the Energy Embargo

The immediate response of the nation to the 1973–1974 energy embargo was to deal first with the emergency situation at hand. Among the earliest actions of the U.S. Congress was the passage of the 1973 Emergency Petroleum Allocation Act, which empowered the executive branch to establish an allocation plan for various sectors of the economy and geographical regions. Related actions included extensions of the daylight saving time, the establishment of a national highway speed limit of 55 mi/h, which became effective in 1974, and the creation of a Federal Energy Office to deal with the problem. The 1975 Energy Policy and Conservation Act provided for the development of a national energy contingency plan, which was issued by the Federal Energy Administration in 1976. The same act mandated a schedule for improving the fuel economy of new automobiles sold in the United States. At that time energy-related responsibilities were scattered among several agencies and programs. Pursuant to Executive Order 12009, the U.S. Congress established in 1977 the Department of Energy (DOE). This new federal department combined the Energy Research and Development Administration (ERDA) and the Atomic Energy Commission (AEC) and was charged with the lead role in coordinating the national response to the problem of energy and to seek short-, medium-, and long-term solutions under a multifaceted program, which came to be known as "Project Independence" [10.18].

A second fuel shortfall occurred in 1979 and caused significant disruptions despite the implementation in the meantime of various gasoline rationing schemes. In the same year the U.S. Congress passed the 1979 Emergency Energy Conservation Act, which directed the executive branch to establish energy conservation targets for the federal government and the states and required the states to submit their plans, including a transportation element, within 45 days of the issuance of the targets. In 1980 the newly elected administration ushered in a different perspective toward the problem by shifting the emphasis from central management to a reliance on a free-market approach. Proposals directed toward the abolishment of the DOE were sent forth, and in 1981 Executive Order 12287 was issued, which eliminated the then existing allocation and price controls on crude oil and petroleum products. Both approaches to the problem of energy have their strong proponents, and the nation's response to the problem continued to be the subject of national debate. By 1990 energy concerns became less of a national priority.

10.4.3 Transportation-User Reactions

The 1973–1974 oil embargo found the nation's transportation system ill prepared, and long queues at gasoline stations became commonplace. According to subsequent reviews of the major events that occurred during the period of low fuel supplies, the general reaction of highway users was to curtail automobile usage by about 20%, mainly by reducing recreational and nonessential trips [10.19]. Localized differences notwithstanding, modal shifts to transit on a national scale were minimal during the emergency. This has been attributed to a lack of adequate transit capacity and to uncertainties about the expected duration of the emergency situation. However, a trend toward the use of more efficient motor vehicles and other transportation equipment became evident. In this connection the 1975 Energy Policy and Conservation Act prescribed a time schedule for fuel economy improvements and required the sales-weighted average fuel economy of each domestic manufacturer to adhere to this schedule.

Other transportation-intensive sectors of the economy responded similarly to the fact that an increasing share of their operating costs were attributed to fuel costs. For example, Johnson and Saricks [10.20] report that most intercity freight carriers began to convert to more fuel-efficient equipment and devices through replacement and retrofit programs and to modify their maintenance and scheduling practices. Similarly, Horn [10.21] reports that the airline industry also moved toward the purchase of fuel-efficient aircraft, implemented new operational and maintenance practices, and reduced cruise speeds. In 1974 the airlines dropped a few thousand daily flights in order to increase passenger-load factors and thus to minimize their consumption of fuel. Highway- and transit-operating agencies also took measures to improve their own consumption rates. Among the actions taken by highway agencies was a conversion to fuel-efficient highway-lighting systems.

10.4.4. Energy-Related Transportation Actions

The predominant view of the energy problem among transportation planning agencies at the local, state, and federal levels was driven by the possibility of petroleum supply interruptions. The secret to the solution of the energy problem was understood to lie in emergency preparedness and in conservation. This view is reflected in the requirement of the 1979 Emergency Energy Conservation Act of Transportation Energy Contingency Plans [10.22, 10.23] and in proposed rules issued by the DOT in 1980, which required that energy conservation be considered in transportation planning programs receiving federal support. Possible energy-conservation strategies may be classified into those that are aimed to cause:

1. Technological innovations
2. Improvements in traffic flow
3. Reductions in the total vehicle miles of travel (VMT)

Technological innovations include improvements in the fuel efficiency of in-use technology. By converting to more fuel-efficient vehicles, highway users were able to sustain their trip-making levels while expending less, although more expensive fuel. An interesting side effect of this development has been its impact on the revenues of agencies that are responsible for the construction, operation, and maintenance of highway facilities because the source of funding for these activities had been primarily in the form of user charges, mainly gasoline taxes levied on a per gallon basis. Another development in relation to technological innovation was an accelerated level of often federally sponsored research and development in the areas of new engines and toward the utilization of alternate transportation fuels. Examples of new engine types include various external combustion engines such as the Stirling and the Rankine engines, continuous-combustion turbine devices, and various configurations of electric and hybrid electric vehicles. Fuels that have been proposed as replacements for conventionally derived petroleum products in existing and new engine designs include gasoline and distillates (diesel fuel), which can be derived from coal and shale, alcohol fuels such as methanol and fuel blends, hydrogen, and electricity, which can be derived from various sources.

Highway level of service and the search for traffic-flow improvements have always been matters of direct concern to transportation engineers. In addition, the effect of highway design and maintenance (e.g., grades, curvature, and pavement condition) on fuel-consumption rates had traditionally been included in the calculation of motor vehicle operating costs. It was, therefore, natural that the inclusion of fuel-consumption considerations in the evaluation of highway designs and congestion-reducing schemes would be given new emphasis in view of the evolution of the global energy situation. A sampling of related investigations includes the use of freeway shoulders as low-cost traffic lanes to improve traffic flow, allowing right turns on red to reduce idling times, providing left-turn lanes at signalized and unsignalized intersections, improving arterial access, and implementing signal systems. The energy effects of other strategies such as the spreading of the peak-period demand for highway travel through the implementation of staggered work schedules and other transportation system management (TSM) actions have also been addressed.

Strategies aimed at reducing the vehicle-miles traveled represent a departure from the other two categories of actions in that they require significant changes in the public's travel habits. Ways to reduce the VMT range from policies that encourage high vehicle occupancies to urban planning processes that emphasize the joint development of transportation and land use to minimize the need for travel without adversely affecting the accessibility of the population to activities.

10.4.5 Vehicle-Propulsion Energy

The propulsion energy expended by individual vehicles is typically reported in terms of either *energy economy rates* (i.e., distance traveled per unit energy) or its reciprocal, that is, the *energy consumption rate.* The energy measure is usually specified in terms of either the amount of a particular fuel or, when applicable, electrical energy. Thus the energy economy of passenger cars is specified as gallons of gasoline per vehicle mile and for electrically propelled transit vehicles, as kilowatt-hours per vehicle mile. In order to be able to compare the energy efficiency of vehicles using different types of fuel, several analysts resort to the conversion of energy requirements to a common unit such as the British thermal unit (Btu) or the joule. Comparisons based on such conversions, of course, are not sensitive to the particular source of the energy used (i.e., crude oil, coal, or nuclear energy), which have certain important policy implications.

A considerable body of research exists relative to the propulsion efficiency of highway vehicles and the factors that influence it. Among these factors are vehicular characteristics (e.g., vehicle type, weight, age, and engine displacement), highway geometrics and condition (e.g., grades, curvature, and pavement maintenance), and traffic-flow conditions (i.e., free-flow to jammed). Figure 10.4.1 illustrates the general shape of the relationship between sustained uniform speed and fuel consumption for highway vehicles. This figure shows the minimum fuel consumption for highway vehicles. It indicates that the minimum fuel-consumption rate corresponds to a uniform speed of about 35 mi/h, depending on the vehicle type and other factors just mentioned. Fuel-consumption curves similar to Fig. 10.4.1 are available in the technical literature for different types of vehicles, including passenger cars, light and heavy trucks, buses, and composite vehicles reflecting various vehicle combinations.

Regarding the propulsion efficiency of nonhighway transit vehicles and systems, it suffices to state that a great variability is found, depending on the type of system, its propulsion technology, and geometric characteristics, including station spacing and gradients.

To calculate the propulsion-energy requirements for a transportation network, estimates of the vehicle mixes and traffic-flow conditions are required, which in a planning context may be provided by transportation demand forecasting model systems such as those described in Chapters 8 and 9. Observed before and after traffic-flow conditions may also be

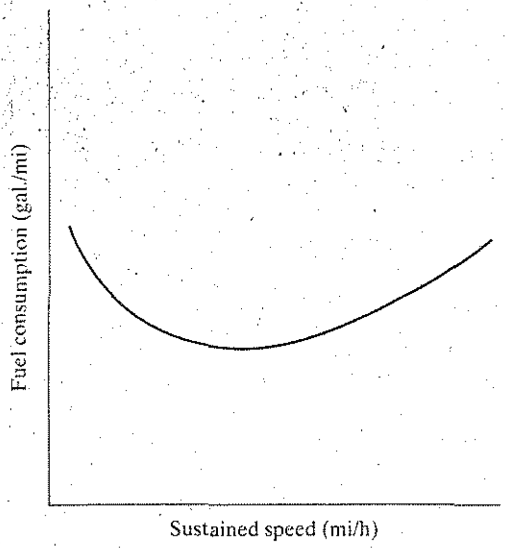


Figure 10.4.1 General relationship between sustained speed and fuel consumption.

used to assess the energy effects of various short-term policies. Vehicular volumes can then be translated to fuel consumption by the use of appropriately calibrated fuel relationships.

In the case of travel on urban arterials the traffic-flow characteristics involve interruptions by the control system and flow variations due to factors that are internal to traffic streams. The General Motors Research Laboratory [10.24] has calibrated a model relating consumption to travel time for various types of vehicles and vehicle mixes in urban arterial driving conditions. The following linear relationship was found to apply for average arterial system speeds of up to 35 mi/h.

$$f = k_1 + k_2 t \tag{10.4.1}$$

where

f = fuel-consumption rate, in gal/mi

t = travel time, in h/mi

k_1 = calibration constant, in gal/veh-mi

k_2 = calibration constant, in gal/h

Equation 10.4.1 may be rewritten in terms of average speed as

$$f = k_1 + \frac{k_2}{u} \quad u < 35 \text{ mi/h} \tag{10.4.2}$$

which has a shape similar to Fig. 10.4.1 but is actually calibrated for the average speed over an urban trip cycle rather than for sustained uniform speeds.

To calculate the fuel-consumption F for a single vehicular trip in urban traffic, Eq. 10.4.1 is multiplied by the length of the trip D to yield

$$F = k_1 D + k_2 T \tag{10.4.3}$$

where T is the travel time for the entire trip.

Figure 10.4.2 represents General Motors' (GM) calibrated relationship for a typical passenger-car mix consisting of 1973 to 1976 car models [10.3] with calibration constants $k_1 = 0.0362$ gal/veh-mi and $k_2 = 0.746$ gal/h. A similar relationship relating to diesel engine tractor-trailers with gross vehicle weights (GVW) of 33,000 lb and over is illustrated in Fig. 10.4.3. Finally, Fig. 10.4.4 shows the fuel-consumption rates of city buses in refer-

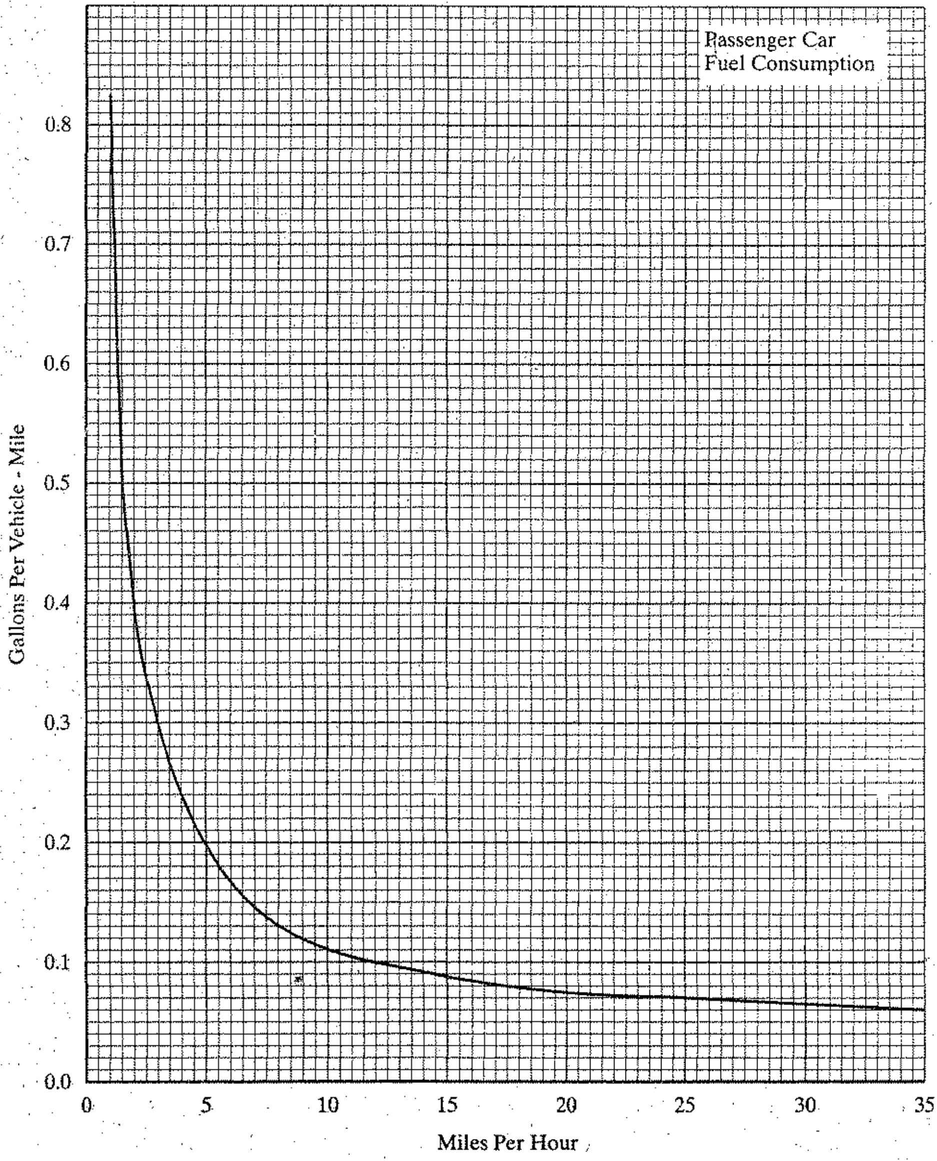


Figure 10.4.2 Passenger-car fuel consumption. (From Raus [10.3].)

ence to the number of stops per mile rather than to the average trip speed. The number of stops per mile used to enter the graph includes scheduled stops, stops that are caused by the traffic-flow conditions, and interruptions due to the control system.

Studies of the energy and air quality impacts of transportation systems continued into the twenty-first century [10.25].

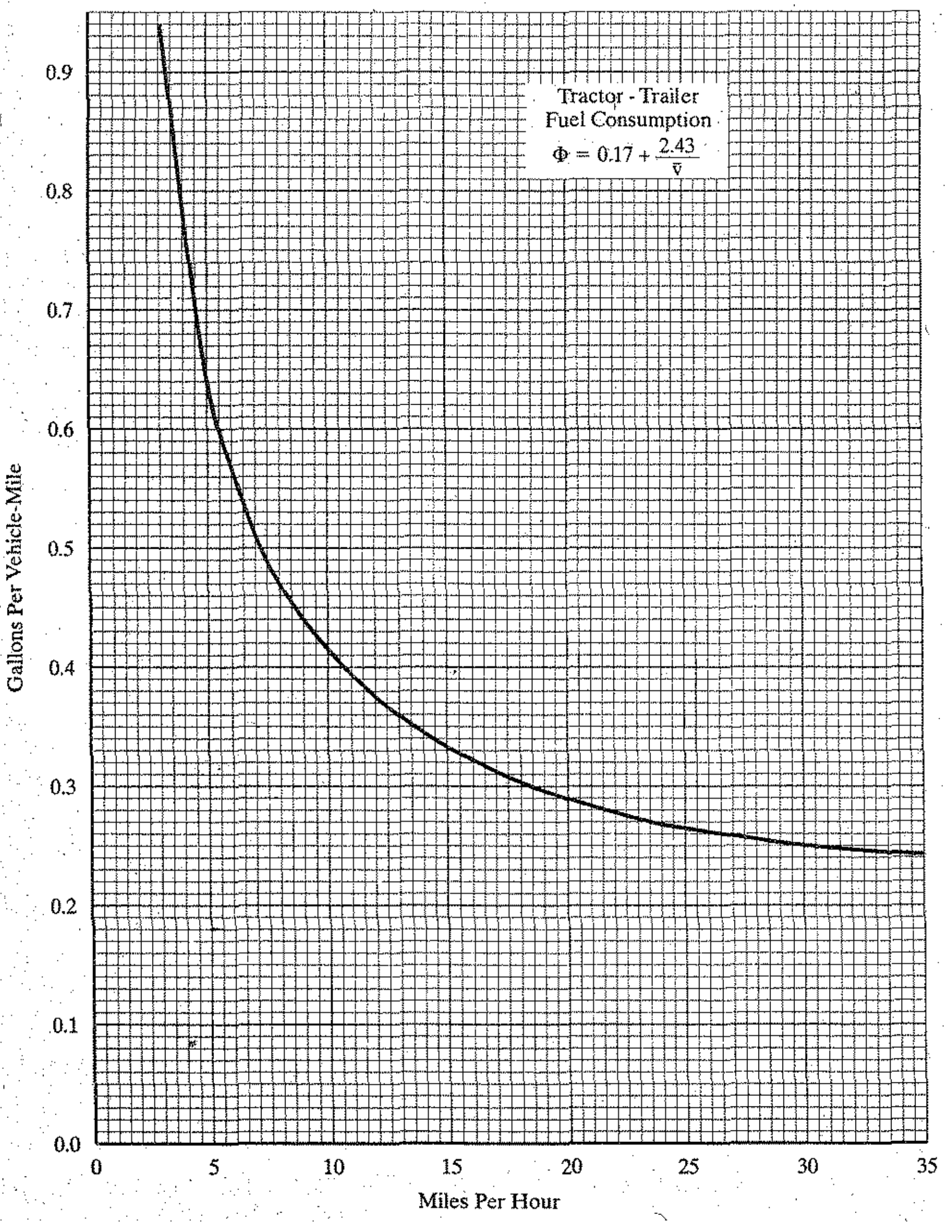


Figure 10.4.3 Tractor-trailer fuel consumption. (From Raus [10.3].)

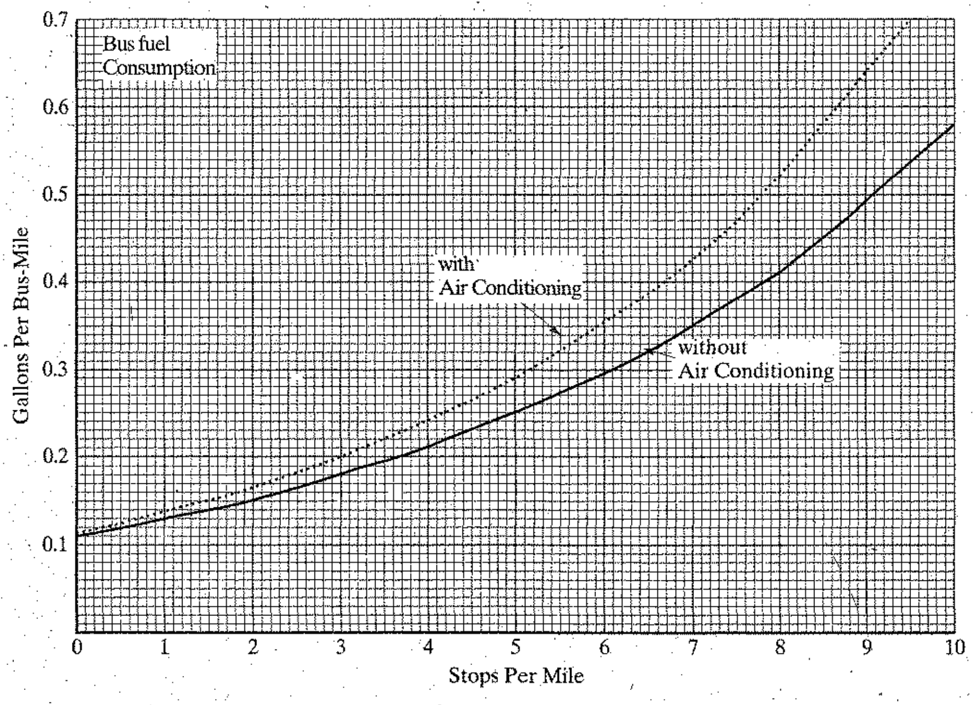


Figure 10.4.4 Bus fuel consumption. (From Raus [10.3].)

Example 10.5

During the typical 1976 weekday peak period, 4000 passenger cars traveled from a suburb to the CBD along a 6-mi arterial route at an average speed of 18 mi/h. One of the lanes on the route was subsequently reserved for car pools. This action resulted in a mild reduction in the peak-period vehicle trips. A postimplementation count showed that the special lane was used by 1000 vehicles (which included previous and new carpoolers) at an improved speed of 24 mi/h. However, 2800 vehicles used the regular lanes, and this caused a speed reduction to 12 mi/h for this component. Calculate the fuel consumed during the peak period (a) prior to and (b) subsequent to the opening of the car-pool lane.

Solution (a) Prior to the project the fuel consumption of the average vehicle over the 6-mi route was, according to Eq. 10.4.3 and the GM calibration constants,

$$F = (0.0362)(6) + \frac{0.746}{3} = 0.466 \text{ gal/veh}$$

A total of 4000 vehicles traversed the route during the peak period. Hence the total fuel consumption was

$$(4000)(0.466) = 1864 \text{ gal per peak period}$$

(b) Following the opening of the car-pool lane, the flow was segregated into regular lane traffic and car-pool lane traffic. Applying Eq. 10.4.3 twice and summing the results, the total consumption became

$$(2800)(0.590) + (1000)(0.0404) = 2066 \text{ gal/peak}$$

Discussion The peak-period propulsive fuel consumption on an arterial route was calculated for two operational strategies. In this particular case the fuel consumption following the opening of a car-pool lane increased even though the vehicle miles traveled during the peak actually decreased from 24,000 (i.e., 4000 veh × 6 mi) to 22,800. This was due to the resulting traffic-flow conditions given in the problem and should not be considered as the inevitable result of all carpool lane situations. The route fuel consumption for each case may be reported in terms of the aggregate economy rate (AER) by dividing the total vehicle miles by the total fuel consumption. Thus the AER for the pre- and the post-car-pool-lane situations was, respectively, 12.9 and 11.0 veh-mi/gal. The conclusion reached by comparing these two rates is identical to that drawn on the basis of the fuel consumption alone. However, neither of these two vehicle-mile-based measures provides definite information about the number of passenger miles accommodated by the two alternatives, which may be an important policy question.

10.4.6 Indirect Energy Consumption

The foregoing discussion has concentrated on the propulsive, or direct, energy consumption of transportation systems. A complete accounting of the energy requirements of transportation systems on the other hand also includes indirect energy expenditures, consisting of construction, maintenance, and operational energy expenditures. Several analysts have attempted to estimate the total (i.e., direct *and* indirect) energy needs of various modes and systems. It suffices to state that these estimates depend on the components of indirect energy that each analyst chose for inclusion in the calculation. Any attempt to trace the full energy implications of transportation systems is ultimately difficult, as it may include the energy expended for the extraction, refinement, conversion, and transportation of energy resources and fuels, and even items such as the energy embedded in the manufacturer of the vehicles. Consequently a detailed review of total transportation energy studies and their energy policy and economic implications is beyond the scope of this book.

10.5 SUMMARY

This chapter discussed the air quality, noise, and energy impacts of transportation, described and illustrated several models that can be used to estimate these impacts, and presented strategies that have the potential of addressing these issues.

The major contribution of transportation to air pollution is in the form of carbon monoxide, hydrocarbons, nitrogen oxides, and particulate matter, and photochemical smog. The degree of this contribution depends on emission levels, which are related to vehicle technology, traffic-flow levels, and traffic characteristics, and the subsequent processes of mixing, diffusion, and chemical oxidation. A method developed by Raus of the FHWA for calculating the emission rates on highway facilities and a simple mathematical pollutant diffusion model (the box model) were described.

Noise was defined as undesirable or unwanted sound and was related to physical and mental health problems. It is typically measured in terms of A-weighted decibel levels on a logarithmic scale that simulates human responses. The intensity of noise decreases with distance from the source because of spreading and absorption energy losses and is also intercepted and reflected by solid objects. These attributes of noise suggest mitigation strategies that include the placement of noise barriers and buffer zones between the source and the receiver in addition to vehicle-related and other actions. The simplest of several noise-estimation models was included. This model applies to long, straight segments of highways in the absence of noise barriers.

A significant portion of the national consumption of energy, particularly petroleum-based, is expended for transportation purposes. A recognition of the ultimate depletion of crude oil and international developments involving oil producing countries have brought this issue into sharp focus. The problem elicited differing reactions from several perspectives, including the users of transportation fuels, the highway- and transit-operating agencies, the transportation planning organizations, and the regional and national energy policy analysts. A method for estimating the propulsive energy requirements of transportation systems was included.

Suggested actions that have important implications with respect to all three impacts covered in this chapter were classified into those that aim at the technological performance of vehicles; are concerned with geometric design and traffic-flow operations; encourage significant changes in travel behavior, particularly modal choice; and propose alternative urban structural forms.

EXERCISES

1. The one-directional speed-concentration relationship for a 10-mi-long segment of a two-lane rural highway is

$$u = 45.0 - 0.3k$$

Apply the Raus model to estimate the total emissions of carbon monoxide assuming that the highway operates at capacity for an entire hour. The ambient temperature at the low altitude highway is 40°F.

2. A proposed increase of parking fees in a downtown area is expected to cause a reduction in the one-directional peak-hour flow of a radial six-lane highway from 5100 to 4200 veh/h. Given that the flow-concentration relationship for each highway lane is

$$q = 42.0k - 0.25k^2 \text{ mi/h per lane}$$

estimate the effect that the parking policy would have on the emission of carbon monoxide during the peak hour. Assume that the flow is distributed equally among the three lanes of traffic, the ambient temperature is 0°F, and the highway was originally operating at level-of-service F.

3. The pollutant emission rate E has been estimated to change with respect to time as

$$E = Ae^{-Bt}$$

where A and B are constants. Assuming a constant airflow F and a box volume V, apply the box model to express the pollutant concentration as a function of time. The initial concentration in the box at time $t = 0$ is K.

4. Repeat Exercise 3 assuming that the emission rate is given by

$$E = A(1 - e^{-Bt})$$

5. The air pollution emission rate E in a parking lot may be approximated by the step function shown in Fig. E10.5. Assuming a constant airflow F and zero initial pollutant concentration, plot the C as a function of time.

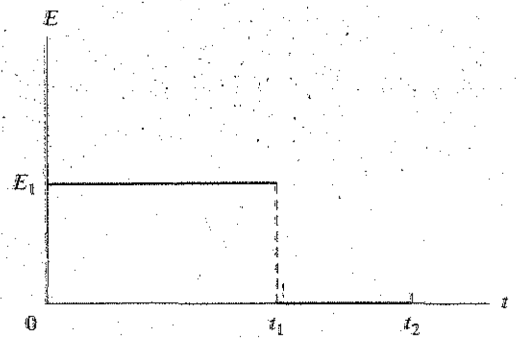


Figure E10.5

6. At a given location the measured noise level during 15 consecutive time intervals was

80, 75, 76, 75, 71, 72, 72, 73, 74, 76, 75, 72, 74, 73, 72

dB. Use this limited set of data to (a) approximate the cumulative distribution of noise level, (b) estimate the L_{10}, L_{50}, and L_{90} levels, and (c) calculate the L_{eq} noise level.

7. Bicyclists A and B rode along the bikeways shown at 15 and 20 mi/h, respectively. Bicyclist A encountered 180 vehicles during his 1.2-mi ride against traffic, and bicyclist B was overtaken by 30 more vehicles than she overtook during her 1.2-mi ride with traffic (see Fig. E10.7). Assuming that the traffic stream consisted of passenger cars only, calculate the L_{10} noise level to which each of the bicyclists was exposed.

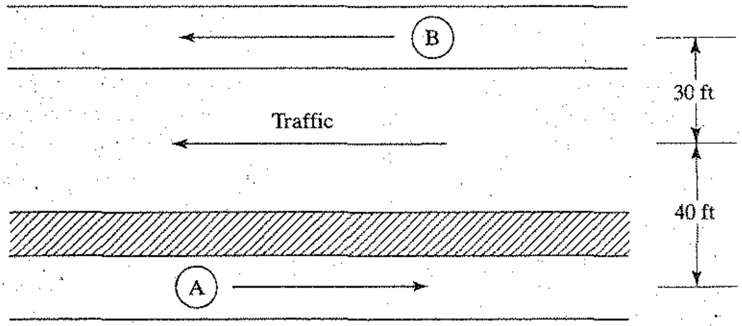


Figure E10.7

8. The traffic flows on each of the two lanes of a highway are shown in Fig. E10.8. Calculate the L_{10} noise level at point P.

9. Apply Eq. 10.3.2 to calculate the combined effect of the following four decibel levels:

71, 77, 72, 73

10. Combine the four noise levels given in Exercise 9 by means of the insert to Fig. 10.3.5.

11. The combined noise level from two sources is 68.5 dBA. The noisier of the two sources produces a noise level of 68.0 dBA. Estimate mathematically the level produced by the other source.

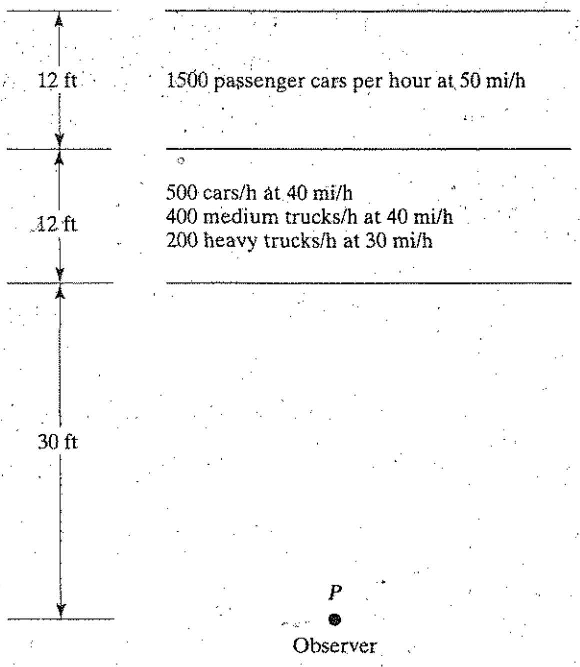

12 ft
1500 passenger cars per hour at 50 mi/h
12 ft
500 cars/h at 40 mi/h
400 medium trucks/h at 40 mi/h
200 heavy trucks/h at 30 mi/h
30 ft
P
Observer

Figure E10.8

12. The flow on a highway consists of 100 heavy trucks per hour traveling at 50 mi/h, 30 medium trucks per hour traveling at 40 mi/h, and 600 passenger cars per hour traveling at 50 mi/h. Assuming a highway width of 50 ft, specify the width of a buffer zone that ensures that the noise level in an adjacent park will not exceed the 1973 FHWA standard.

13. How close to the highway of Exercise 12 can a single-glazed masonry school building be located and still meet the FHWA noise standard?

14. The flow-concentration relationship for a roadway is

$$q = 60.0k - 4.0k^{1.5} \text{ veh/h}$$

Using the GM model, derive and plot the fuel-consumption rate f as a function of traffic stream concentration (veh/mi). Assume that the traffic stream consists of passenger cars only.

15. Find the fuel-consumption rate corresponding to $q_{\max}$ for the roadway of Exercise 14.

16. A bus line operates in mixed traffic and carries 4000 passengers per peak hour at an average speed of 10 mi/h. Typically each bus makes two scheduled stops per mile and is interrupted by the traffic control system and other vehicles six times per mile. Given an average occupancy of 50 persons per bus and a 5-mi trip, calculate the number of buses needed to serve the passenger demand and the total amount of fuel consumed during a typical peak hour.

17. Calculate the effect on the fleet size and the fuel consumption of the bus system of Exercise 16 assuming that an exclusive bus lane were to be implemented. The resulting conditions include a reduction of nonscheduled stops from six to two, and an average speed of 15 mi/h.

REFERENCES

10.1 Perkins, H. C., *Air Pollution,* McGraw-Hill, New York, 1974.

10.2 Environmental Protection Agency, *User's Guide to MOBILE I: Mobile Source Emissions Model,* Office of Air, Noise, and Radiation, EPA, Washington, DC, 1978.

10.3 Raus, J., *A Method for Estimating Fuel Consumption and Vehicle Emissions on Urban Arterials and Networks,* Report FHWA-TS-81-210, Office of Research and Development, Federal Highway Administration, Washington, DC, 1981.

10.4 Zimmerman, J. R., and R. S. Thompson, *User's Guide for HIWAY, A Highway Air Pollution Model,* Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1975.

10.5 National Cooperative Highway Research Program, *Development of an Improved Framework for the Analysis of Air Quality and Other Benefits and Costs of Transportation Control Measures,* Research Results Digest No. 223, Transportation Research Board, National Research Council, Washington, DC, 1998.

10.6 ———, *Relationship between Pavement Surface Texture and Highway Traffic Noise,* NCHRP Synthesis 268, Transportation Research Board, National Research Council, Washington, DC, 1998.

10.7 Seiff, H. E., "Enforcement of Control of Interstate Motor Carrier Noise: A Federal Perspective," *Motor Vehicle Noise Control,* Special Report 152, Transportation Research Board, National Research Council, Washington, DC, 1975, pp. 66–72.

10.8 Federal Highway Administration, *Noise Standards and Procedures,* Policy and Procedure Memorandum, Transmittal 279, 90-2, FHWA, Washington, DC, February 8, 1973.

10.9 Environmental Protection Agency, *Noise and Its Measurement,* Office of Public Affairs, EPA, Washington, DC, February 1977.

10.10 U.S. Department of Transportation, *Organization and Content of Environmental Assessment Materials,* Notebook 5, U.S. Government Printing Office, Stock No. 050-000-00109-1, Washington, DC, 1975.

10.11 Hall, F. L., and B. L. Allen, *Toward a Community Impact Measure for Assessment of Transportation Noise,* Transportation Research Record 580, Transportation Research Board, National Research Council, Washington, DC, 1976, pp. 22–35.

10.12 Wesler, J. E., *Introduction and History of Highway Noise Prediction Methods,* Transportation Research Circular 174, Transportation Research Board, National Research Council, Washington, DC, 1975, pp. 9–13.

10.13 Wright Air Development Center, *Handbook of Acoustic Noise Control,* WADC Technical Report 52-204, 1952.

10.14 Kugler, B. A., D. E. Commins, and W. J. Galloway, *Highway Noise: A Design Guide for Prediction and Control,* National Cooperative Highway Research Program Report 174, Transportation Research Board, National Research Council, Washington, DC, 1976.

10.15 Federal Highway Administration, *FHWA Traffic Noise Model: Technical Manual,* Final Report, DOT-VNTSC-FHWA-98-1, U.S. Department of Transportation, Washington, DC, 1998.

10.16 ———, *FHWA Traffic Noise Model: User's Guide,* Final Report, DOT-VNTSC-FHWA-98-1, U.S. Department of Transportation, Washington, DC, 1998.

10.17 ———, *FHWA Traffic Noise Model: Look-Up Tables,* Final Report, DOT-VNTSC-FHWA-98-5, U.S. Department of Transportation, Washington, DC, 1998.

10.18 U.S. Federal Energy Administration, *Project Independence Report,* U.S. Government Printing Office, Stock No. 4118-000019, Washington, DC, November 1974.

10.19 Cheslow, M. D., "Potential Use of Carpooling during Periods of Energy Shortages," *Considerations in Transportation Energy Contingency Planning,* Special Report 191, Transportation Research Board, National Research Council, Washington, DC, 1980, pp. 38–43.

10.20 Johnson, L. R., and C. L. Saricks, *An Evaluation of Options for Freight Carriers during a Fuel Crisis,* Transportation Research Record 935, National Research Council, Washington, DC, 1983, pp. 5–11.

10.21 Horn, K. W., "Energy and the Airline Industry," *Considerations in Transportation Energy Contingency Planning,* Special Report 191, Transportation Research Board, National Research Council, Washington, DC, 1980, pp. 69–70.

10.22 Transportation Research Board, *Considerations in Transportation Energy Contingency Planning,* Special Report 191, National Research Council, Washington, DC, 1980.

10.23 ———, *Proceedings of the Conference on Energy Contingency Planning in Urban Areas,* Special Report 203, National Research Council, Washington, DC, 1983.

10.24 Chang, M. F. et al., *The Influence of Vehicle Characteristics, Driver Behavior, and Ambient Temperature on Gasoline Consumption in Urban Areas,* General Motors Corporation, Warren, MI, 1976.

10.25 Transportation Research Board, "Effects of Transportation on Energy and Air Quality," *Transportation Research Record 1587,* National Research Council, Washington, DC, 1997.

11

Evaluation and Choice

11.1 INTRODUCTION

Even when presented a single proposal, decision-makers have a choice between it and doing nothing. Therefore every decision involves at least two options. *Evaluation* facilitates decision making by appraising the merits (*positive impacts*) and demerits (*negative impacts*) of alternative options in terms of either a single or multiple decision criteria. Determining which impacts are relevant to a particular decision and specifying the appropriate decision criteria are related to the value system within which the choice is to be made. In the case of transportation decisions in the public sector the operating value system is not that of any single individual or subgroup but that of the community as a whole. In Chapters 1 and 7 we recognized the existence of conflicting value systems within society. Consequently transportation decision making also entails the resolution of conflicts.

Two types of evaluation studies are commonly undertaken: preimplementation studies, which facilitate the choice of the best course of action from among several alternative proposals, and postimplementation studies, which assess the performance of already implemented actions. Postimplementation studies are important for two reasons: (1) They help to discover whether or not the implemented alternative performs well and (2) they help to determine whether or not it continues to perform properly over time. This is especially important in the case of transportation systems, which are subject to changing conditions and also to evolving goals and objectives. Continuous monitoring and periodic performance evaluation can help to identify emerging problems and also to provide guidance to the design of possible improvements.

To be selected for implementation, an alternativ: must be both feasible *and* superior to all other alternatives. The prerequisites to the admission of an alternative to the list of acceptable options include the conditions of technological feasibility, economic efficiency, and cost-effectiveness, and availability of the needed resources. In this chapter we present the fundamental elements of efficiency and effectiveness evaluation techniques, along with a brief description of their conceptual foundations, major strengths, and weaknesses.

11.2 FEASIBILITY AND IMPACT ENUMERATION

11.2.1 Measures of Feasibility

Technological feasibility refers to the ability of a system to function according to the laws of nature and not to its desirability: A perpetual motion machine may be highly desirable but technologically impossible. Engineers and other technologists are qualified to deal with questions relating to technology. Research and development are ongoing activities that occasionally lead to technological breakthroughs. The vast majority of practical applications, however, involve the use of existing technology. Even then innovative and creative ways of combining off-the-shelf technology are common. Consequently the question of technological feasibility is an aspect of evaluation that cannot be ignored.

Efficiency is defined as the ratio of the quantity produced (*output*) to the resources required for its production (*input*). *Physical* or *machine efficiency* is the ratio of the energy delivered by a machine or a process to the energy supplied to it. Although expressed in the same unit of measurement, the input and the output energy differ in form, for example, energy in the form of electricity vis-à-vis energy in the form of work done by the system. Machine efficiency is always less than unity because of the unavoidable energy losses that are incurred in the process. This waste can be justified only when the usefulness, or utility, of the output exceeds that of the input. When both the numerator and the denominator are converted to the same measure of economic value, their ratio is referred to as the *economic efficiency* of the machine, which must be greater than unity if the machine is to be economically feasible. The idea of economic efficiency has been extended to the evaluation of systems to contrast the economic value of the advantages (or benefits) that are derived from the system to its disadvantages (or costs).

Effectiveness is defined as the degree to which an action accomplishes its stated objectives. It differs from efficiency in that it does not need to express explicitly all impacts in the same scale of measurement. For example, the effectiveness of a regional transportation system for elderly and handicapped persons may be expressed as the proportion of eligible users that live within the service area of the system or as the total number of persons served, whereas its operating costs may be expressed in terms of dollars. *Cost-effectiveness* evaluation is the attempt to determine the efficacy of alternatives by comparing their cost to their effectiveness. Of course, if an objective method for collapsing all impacts to the same dimension were available, efficiency and effectiveness would lead to identical results, but no such method exists. Consequently both evaluative methods are used, sometimes separately and sometimes in cômbination [11.1, 11.2].

An alternative may be technologically feasible, economically efficient, and cost-effective and yet not be a prudent choice for implementation because of the unavailability of the financial and other resources that are needed for its implementation. Problems of affordability or resource availability are not uncommon. Consider, for example, the case of financial resources. Usually there exists a lag between the time when financial resources are expended and the time when the returns of the investment are realized. Lack of access to financial resources during this critical time lag would render the investment infeasible. Another common problem of financial affordability that is especially true in the case of public projects is related to the fact that the benefits derived by a public investment do not usually return in the form of money to the agency that expends the financial costs for the project. Unless the agency is in a position to afford these expenditures it would not be able to produce the benefits for whomever they would otherwise accrue.

11.2.2 Impact Trade-Offs

Determining the feasibility of each alternative is only half of the evaluation process. The other half involves the comparison of all proposals (including the do-nothing alternative) in order to select the best one among them. Based on the assignment of relative weights to the impacts of each alternative, this step involves *impact trade-offs*. Consider, for example, a choice between two transportation alternatives requiring equal and available financial expenditures. Further, assume that one of the two would provide a higher level of mobility than the other but would also discharge higher quantities of atmospheric pollutants. This statement implies that three impacts have been identified as relevant to the choice, appropriate measures of performance have been established to express them, and the likely levels of these impacts have been predicted for each alternative, perhaps using the methods of Chapters 8 through 10. When comparing the two alternatives, a trade-off between mobility on one hand and environmental quality on the other becomes apparent. In the final analysis the evaluation method used to aid this decision must incorporate the assignment of relative weights to the impacts.

11.2.3 Generalized Impact Matrices

The foregoing example of evaluation raises a problem that is inherent in situations where the decision-makers are faced with multiple decision criteria. On the side of costs, the direct cost of operating the system and a *negative externality* (i.e., the unintended undesirable impact of air pollution) were identified. Direct benefits (e.g., mobility) and potential *positive externalities* are typically included in the evaluative calculus. All recognizable impacts, whether intended or concomitant, can be classed into positive impacts (i.e., advantages or benefits) and negative impacts (i.e., disadvantages or costs) and the results of the impact estimation process that precedes the evaluation phase (see Chapters 8, 9, and 10) may be summarized in an *impact matrix*, as illustrated in Fig. 11.2.1.

This array lists the estimated impacts associated with each alternative expressed in terms of the applicable measures of performance, which differ with regard to their units of measurement. Moreover, some are expressed in terms of quantitative measures (i.e., carbon monoxide concentration), and others are qualitative.

Table 11.2.1 lists the impacts that were considered in the environmental impact statement (EIS) for a proposed Honolulu Area Rapid Transit (HART) System [11.3]. The first column summarizes the goals and objectives set forth in the regional general plan for the island of Oahu, where the city of Honolulu is located. The second column presents the specific goals identified by an earlier Oahu transportation study. Following are the objectives established by two previous Preliminary Engineering Evaluation Program studies of transit alternatives (PEEP I and II). The fourth column lists the criteria that were selected to measure the performance of alternative systems. Also noted is the potential applicability of these criteria to three characteristics of alternative proposals, that is, route location, transit system type, and system length. The rapid-transit alternatives are augmented by feeder bus services.

Impact category	$ costs		Mobility		Environmental quality			Social
Measures of performance	Capital	O & M	Travel time	Travel cost	Air	Noise	. . .	. . .
Do nothing								
Alternative A								
Alternative B								
⋮								

Figure 11.2.1 Generalized impact matrix.

Table 11.2.2 summarizes the analytical results obtained by applying the sequential transportation demand-forecasting process described in Chapter 8. This table includes only direct impacts.

Table 11.2.3 is the generalized impact matrix developed in connection with the HART EIS. It includes the direct and indirect impacts of each alternative either in quantitative terms or qualitatively.

11.3 ENGINEERING ECONOMIC ANALYSIS

11.3.1 Background

Traditional engineering economic analysis is based on the principle that the quantified impacts of alternatives should and can be converted to their monetary equivalents and treated just as if they were money. With this conversion, the calculation of economic efficiency and the comparison of alternatives on the basis of their costs and benefits can be conducted. The basic unit of measurement employed (i.e., money) has certain attributes that must be retained in the calculation of benefits and costs. A fundamental characteristic of money is its time value. Simply stated, this says that "a dollar today is not the same as a dollar tomorrow." To illustrate this point, consider the situation where an amount of $100 is deposited in a bank at an interest rate of 10%. One year from the day of deposit, $110 may be withdrawn from the bank. In this case $100 today is equivalent to $110 dollars a year from today. The *interest* or *discount* rate affects this equivalency.

11.3.2 Project Evaluation

Based on the axiom that the consequences that are relevant to the impending decision can be equated with money, each alternative may be considered to consist of two cash flows: a cash flow of benefits and a cash flow of costs, both shown as money equivalents at the times when they are expected to occur (Fig. 11.3.1). A proposed alternative is considered to be economically feasible when the benefits to be derived from it exceed its costs. This comparison between benefits and costs is legitimate only when the two cash flows are placed

TABLE 11.2.1 Example of Goals, Objectives, and Criteria

Transportation goals Oahu general plan	Transportation goals Oahu transportation study	Transit development objectives PEEP I and II	Transit development criteria for alternatives analysis	Applicable criteria for specific alternative analysis		
				Route location	System type	System length
1. Provide transportation facilities to enable travel from any point in the region to any other point within reasonable travel time by one or more modes	1. Provide transportation facilities for ease of movement throughout Oahu and provide a variety of modes of travel which will best serve the different requirements of the community	1. *Improve accessibility* by service and interconnecting existing and future urbanized areas of Oahu	a. Availability & coverage	—	✓	✓
			b. Average trip time	✓	✓	✓
			c. Service reliability	—	✓	✓
			d. Rider convenience	✓	✓	✓
			e. Rider comfort	✓	✓	✓
2. A transportation system which will provide the greatest efficiency and service to the community with the least overall expenditure of resources	2. Provide a balanced transportation system which will result in optimum service with the least public expenditure	2. *Provide a balanced transportation system* of transit and highways	a. System patronage	✓	✓	✓
			b. System capacity	✓	✓	✓
		3. *Minimize expenditure of resources* and disruption to community	a. Consumption of land	—	✓	✓
			b. Displacement of residents	✓	✓	✓
			c. Displacement of businesses	✓	✓	✓
			d. Reduction of community amenities	✓	✓	✓
			e. Disruption to future dvlopmt.	✓	✓	✓
			f. Disruption to local circulation	✓	✓	✓
			g. Disruption: constr. activity	—	✓	✓
			h. Savings in energy	—	✓	✓
			i. Technical risk	—	✓	✓

(continued)

TABLE 11.2.1 Example of Goals, Objectives, and Criteria—*(continued)*

Transportation goals Oahu general plan	Transportation goals Oahu transportation study	Transit development objectives PEEP I and II	Transit development criteria for alternatives analysis	Applicable criteria for specific alternative analysis		
				Route location	System type	System length
3. A transportation system to be designed as an integral part of and complementary to land-use policies	3. Integration of the transportation system with land use	4. *Support land-use* and development policies	a. Support regional development b. Support comm. development	✓ ✓	✓ ✓	✓ ✓
4. Preserve and maintain significant historic sites, scenery, and natural assets of Oahu[a]	4. Preserve Oahu's beauty and amenities	5. *Preserve environment*	a. Reduction air pollution b. Noise level c. Visual intrusion d. Vistas e. Historic sites	✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓	✓ ✓ ✓ ✓ ✓
5. Safety	5. Safety	6. *Safety*	a. Reduce accident exposure b. Security	— —	✓ ✓	✓ ✓
6. Provide a transportation system which will provide the greatest efficiency and service to the community with the least overall cost[b]	6. Provide a balanced transportation system which will result in optimum service at the least cost to the public[b]	7. *Provide the most economical system* which best meets all other objectives	a. Total annual cost b. Cost per trip c. Benefit-cost ratio	✓ — —	✓ ✓ ✓	✓ ✓ —
[a]Stated as one of the general goals [b]Stated separately from 2 to differentiate between expenditure of resources and least cost	[a]Combines goals 1 & 3 of OTS [b]Stated separately from 2 to differentiate between expenditure of resources and least cost	[a]Goal 2 stated as two separate objectives				

Source: Urban Mass Transportation Administration [11.3].

TABLE 11.2.2 Example Summary of Analytical Results

	Travel characteristics						Operating characteristics				Cost[a]	
	Total daily transit trips	Mode split (total) (%)	Daily trips on gwy.	% Peak-hour work trips by transit	Avg. trip length (mi)	Avg. trip time (min)	Vehicles required (with spares)		Vehicle miles operated daily		Total capital cost ($1,000)	Annual operating cost ($1,000)
System							Gwy.	Feeder	Gwy.	Feeder		
7-mi busway	456,250	13.8	288,200	42.4	7.31	36.3	179	752	30,308	121,408	414,411	49,510
LRT 28 mi	474,520	14.4	358,750	44.6	7.22	32.4	410	443	79,515	71,645	712,289	47,660
23 mi	474,520	14.4	353,700	44.6	7.21	32.3	325	477	75,530	75,910	646,537	46,320
14 mi	473,300	14.3	306,900	44.2	7.26	33.7	198	580	45,840	94,260	529,321	44,310
7 mi	459,300	13.9	277,300	42.8	7.19	35.2	109	774	23,580	124,605	406,808	50,170
Fixed guideway 23 mi	490,000	14.8	332,600	46.0	7.50	31.6	421	493	111,495	78,390	647,900	46,940
14 mi	473,300	14.3	306,900	44.2	7.26	33.7	264	580	64,225	94,260	517,318	43,890
7 mi	459,300	13.9	277,300	42.8	7.19	35.2	161	774	34,335	124,605	398,676	50,070

[a] All costs shown are in 1975 dollars.

Source: Urban Mass Transportation Administration [11.3].

TABLE 11.2.3 Example of Comparison Matrix for Alternative Systems

	Short 7-mi length			Medium 14-mi length		Long 23- and 28-mi lengths		
	Busway	LRT	Fixed guideway	LRT	Fixed guide	23-mi LRT	28-mi LRT	23-mi F.G.
Objective 1								
a. Availability & coverage	Same	Same	Same	Same	Same	Same	Same	Same
b. Avg. trip time (min)	36.3	35.2	35.2	33.7	33.7	32.3	32.4	31.6
c. Service reliability	(2)[a]	(1)	(1)	Same	Same	(2)	(2)	(1)
d. Rider convenience	(1)	(2)	(2)	Same	Same	(2)	(1)	(2)
e. Rider comfort	(2)	(1)	(1)	Same	Same	Same	Same	Same
Objective 2								
a. System patronage (million)	137.8	138.7	138.7	142.9	142.9	143.3	143.3	148.0
b. System capacity	[b]	Same	Same	Same	Same	Same	Same	Same
Objective 3								
a. Consumption of land (acres)	42	21	20	23	22	36	36	32
b. Displacement of residents (units)	233	152	148	166	162	179	179	167
c. Displacement of businesses (units)	257	168	164	187	183	194	194	184
d. Reduction of community amenities	Same	Same	Same	Same	Same	Same	Same	Same
e. Disruption to future development	Same	Same	Same	Same	Same	Same	Same	Same
f. Disruption to local circulation	Same	Same	Same	Same	Same	(2)	(3)	(1)
g. Disruption from constr. activities	Same	Same	Same	Same	Same	(1)	(2)	(1)
h. Savings in energy (million gal/yr.)	10.0	8.5	8.9	8.5	9.4	5.0	4.8	8.5
i. Technical risk	(3)	(1)	(2)	(1)	(2)	(1)	(1)	(2)

(continued)

TABLE 11.2.3 Example of Comparison Matrix for Alternative Systems—*(continued)*

	Short 7-mi length						Medium 14-mi length				Long 23- and 28-mi lengths					
	Busway		LRT		Fixed guideway		LRT		Fixed guide		23-mi LRT		28-mi LRT		23-mi F.G.	
Objective 4																
a. Support regional dvlpmt.	Same		Same		Same		Same		Same		Same		Same		Same	
b. Support comm. dvlpmt.	Same		Same		Same		Same		Same		Same		Same		Same	
Objective 5																
a. Reduction air pollution (ton/yr.)	2,970		3,240		3,260		4,110		4,150		4,120		4,140		4,930	
b. Noise level (dBA)	86–88		77–81		77		77–81		77		77–81		77–81		77	
c. Visual intrusion	(3)		(2)		(1)		(2)		(1)		(2)		(2)		(1)	
d. Vistas	(2)		(3)		(1)		(2)		(1)		(1)		(1)		(2)	
e. Historic sites	Same		Same		Same		Same		Same		Same		Same		Same	
Objective 6																
a. Reduce accident exposure	Same		Same		Same		Same		Same		(2)		(2)		(1)	
b. Security	Same		Same		Same		Same		Same		Same		Same		Same	
Objective 7[c]																
Interest rates	4%	10%	4%	10%	4%	10%	4%	10%	4%	10%	4%	10%	4%	10%	4%	10%
a. Total annual cost ($ million)	77.43	96.90	76.98	96.17	76.41	95.21	77.38	102.60	76.26	101.01	85.73	116.66	90.73	124.84	86.51	117.48
b. Cost per trip	56.2¢	70.3¢	55.5¢	69.3¢	55.1¢	68.6¢	54.1¢	71.8¢	53.4¢	70.7¢	59.8¢	81.4¢	63.3¢	87.1¢	58.5¢	78.4¢
c. Benefit-cost ratio	2.24	1.50	2.28	1.55	2.31	1.58	2.40	1.47	2.47	1.51	2.06	1.24	1.87	1.13	2.19	1.32

[a] Numbers in parentheses show ranking of alternatives based on how well they met the objective.

[b] Practical capacity for busways is unknown and assumed to be less than that for guided systems.

[c] All costs shown are in 1975 dollars.

Source: Urban Mass Transportation Administration [11.3].

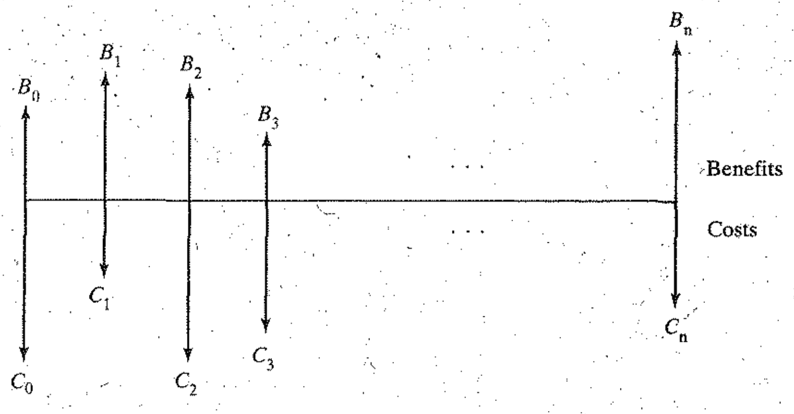


Figure 11.3.1 Streams of benefits and costs.

on the same time basis. Given an appropriate interest rate, the *present worth of benefits* (PWB) and the *present worth of costs* (PWC), or their equal series cash-flow equivalents, may be calculated. Chapter 12 develops the appropriate formulas that can be used for this task, which the reader may wish to review before continuing with the rest of this chapter.

The *net present worth* (NPW) of an alternative is defined as the present worth of its benefits minus the present worth of its costs. Hence a positive NPW implies economic feasibility. Another way of contrasting benefits and costs is the use of the *benefit-cost* (B/C) *ratio,* in which case the economic feasibility criterion requires a B/C ratio that is greater than unity. A third method of assessing the economic feasibility of an alternative is one that calculates the *internal rate of return* (IRR), which is defined as the interest rate that just equates benefits and costs, that is, the rate at which the NPW equals zero and the B/C ratio equals unity. This rate is then compared with a predetermined *minimum attractive rate of return* (MARR) reflecting managerial policy and profit expectations to assess whether or not the project is attractive.

11.3.3 Independent and Mutually Exclusive Alternatives

Before discussing the mechanics of economic evaluation of alternatives, it is appropriate to explain several principles that are explicitly or implicitly encompassed by the final choice. First, the set of alternatives being considered should include the do-nothing, or baseline, alternative. Second, pairs of alternatives can be either *independent* or *mutually exclusive.* Two alternatives are *independent* when the selection of one does not necessarily prohibit the selection of the other. For example, a state department of transportation may be contemplating the provision of subsidies to the bus systems of two different cities. Assuming that the necessary resources are available to the department, a decision to subsidize one city does not necessarily eliminate a favorable outcome for the other. By contrast, a pair of alternatives are said to be *mutually exclusive* if the choice of one renders the other impossible. A metropolitan transit authority engaged in the comparative evaluation of two technologically incompatible transit systems on a single alignment is faced with mutually exclusive alternatives. Third, the do-nothing alternative and each of the do-something alternatives are mutually exclusive. Fourth, the list of options under consideration includes all possible

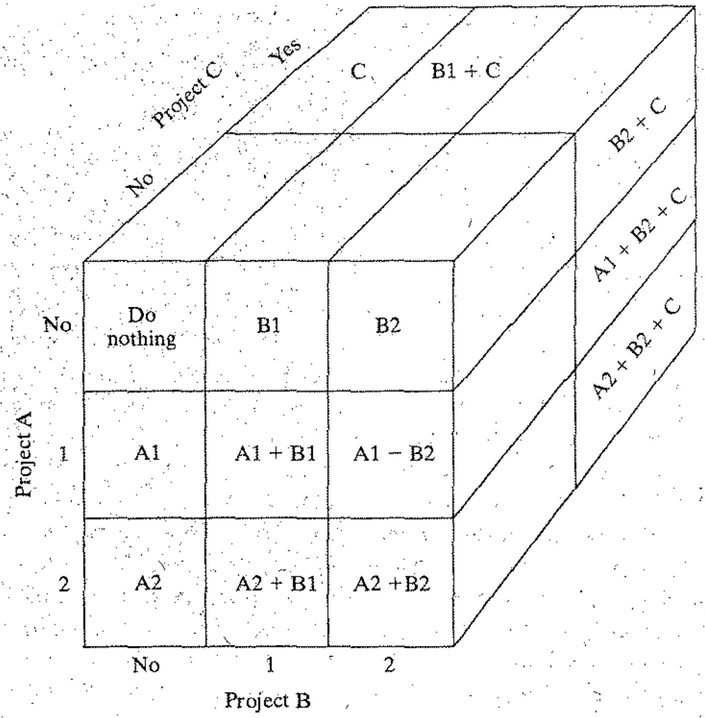


Figure 11.3.2 Alternative combinations of options.

combinations of independent alternatives. For example, when two independent projects are being considered, the list of available options contains four entries: the do-nothing alternative, each of the two projects alone, and the two in combination. When viewed in this manner, the four options are actually mutually exclusive, as it is not possible to implement one project alone and both projects together at the same time. The problem of economic evaluation and project selection becomes one of discovering the alternative combination of feasible projects that maximizes the benefits to be derived from the expenditure of available resources.

Example 11.1

A regional planning organization is considering the following proposals: two mutually exclusive alignments for a highway in county A (projects A1 and A2), two mutually exclusive alignments for a highway in county B (projects B1 and B2), and a special transportation system for handicapped persons in city C. What is the number of available options?

Solution Considering that with regard to the first and second highways, three possibilities exist (i.e., not building, selecting alternative 1, and selecting alternative 2) and that two choices are possible with regard to alternative C, the total number of proper combinations is 3 × 3 × 2 = 18, as illustrated in Fig. 11.3.2. If any one of the projects is infeasible, the total number of options is reduced accordingly. If, for example, alternative A1 were judged to be infeasible, the total number of options would become 3 × 2 × 2 = 12. Similarly, if project C were found to be infeasible, the remaining options would number 3 × 3 × 1 = 9.

11.3.4 Evaluation of Mutually Exclusive Alternatives

Consider two mutually exclusive do-something alternatives with the following discounted benefits and costs expressed in millions of dollars.

Alternative	PWB	PWC	NPW	B/C
A	1.8	1.2	0.6	1.50
B	2.9	2.2	0.7	1.32

According to the NPW criterion, alternative B is superior to alternative A, but according to the B/C criterion, alternative A is better than alternative B. This inconsistency between the two methods can be rectified by augmenting the B/C evaluation with an *incremental analysis*. To understand the rationale of incremental analysis, consider the simplified situation where the total \$2.2 million is in hand and no other investment option is possible; that is, the available amount of money could be either expended in one of the two projects or placed in a safe deposit box, where it would earn no interest. Under these assumptions the *overall* investment strategies associated with each of the two alternatives are (1) to invest \$1.2 of the \$2.2 million in the less costly alternative A, which will return \$1.8 million in benefits, and place the remaining \$1.0 million in the safe deposit box, which will return no additional benefits, and (2) to invest the entire \$2.2 million in the more costly alternative, which will derive total benefits of \$2.9 million. The present worth of the benefits resulting from the first strategy would equal \$1.8 million plus \$1.0 million, or \$2.8 million, as compared to the \$2.9 million associated with the more costly alternative. Hence investing in the second option is the more prudent choice. Another way of stating the above is that the *incremental benefits* (\$2.9 − \$1.8 = \$1.1 million) derived from the costlier alternative outweigh the *incremental costs* (\$2.2 − \$1.2 = \$1.0 million) it entails, or that the *incremental B/C* ratio between the two options is greater than unity. Thus when both alternatives are feasible in themselves, the incremental B/C ratio and the NPW criteria lead to identical conclusions. The incremental ratio analysis of feasible options is conducted as follows: The feasible alternatives are listed according to increasing cost, with the least costly alternative at the top of the list. If the incremental ratio between the first two entries is greater than unity, the more costly alternative is selected; otherwise the less costly alternative is retained. The chosen alternative is then compared with the next list entry and the procedure continues until all alternatives have been considered and all but the best alternative have been eliminated.

Example 11.2

The benefits and costs associated with the following five mutually exclusive alternatives have been discounted to their present worth and the alternatives have been listed according to increasing cost. Apply the B/C ratio method to select the best option.

Alternative	PWC	PWB	B/C
A	100	150	1.50
B	150	190	1.27
C	200	270	1.35
D	300	290	0.97
E	320	350	1.09

Solution After forming the B/C ratio, alternative D is found to be infeasible and therefore is dropped from further consideration. The incremental B/C between A and B is (190 − 150)/(150 − 100) = 0.8, and the costlier alternative B is dropped. The incremental ratio between A and the next feasible alternative in the list (i.e., C) is equal to(270 − 150)/(200 − 100) = 1.2 and C is favored over A. Finally, the comparison between C and E yields an incremental ratio of 0.67. Since this is less than unity, the less costly alternative C is retained as the best option.

Discussion Alternative C has been selected even though alternative A has a larger individual B/C ratio. It can easily be shown that the NPW criterion leads to the selection of the same alternative. The incremental ratio analysis must be preceded by an individual ratio analysis to eliminate all infeasible alternatives. If the incremental ratio analysis were to be applied directly to a list that happened to contain only infeasible alternatives, it would result in the selection of the least infeasible without any indication that the selected option is in fact infeasible. Serious problems related to this point arise in situations where, for practical reasons, the benefits and costs of do-something alternatives are calculated relative to the do-nothing alternative. For example, travel time or fuel savings are often considered to be benefits associated with proposed highways as compared to the do-nothing alternative. In that case the B/C analysis is an incremental analysis from the start. To illustrate this point, consider the following simple example: The benefits and costs associated with an existing highway (do-nothing) and a proposed highway are:

Alternative	PWB	PWC
Existing	1.2	1.8
Proposed	1.9	2.4

Clearly neither of the two is feasible. However, if the benefits and costs of the proposed highway were to be reported only in relation to the do-nothing alternative, the proposed project would have an appearance of feasibility and an incremental B/C ratio of 1.17. Moreover, the NPW of the relative benefits and costs would also be misleading. To avoid problems of this nature, ways of measuring benefits and costs in absolute rather than relative terms have been proposed. One such method is based on the *theory of consumer surplus* [11.4], but although conceptually attractive, these attempts are not without practical difficulties.

11.3.5 Identification and Valuation of Benefits and Costs

The conduct of economic evaluation procedures for the selection of the best alternative requires the conversion or valuation of quantified impacts to monetary terms. Impact valuation presents varying degrees of difficulty. Some impacts, such as construction and maintenance costs, are already expressed in monetary terms. The rest must be translated into monetary equivalents. As an illustration, Fig. 11.3.3 presents a family of curves suggested by a 1977 AASHTO manual [11.5] for the conversion of travel-time savings to dollars. These curves are based on extensive economic explorations into the matter, and unlike earlier versions, which assumed a linear relationship between time saved and dollar value irrespective of trip purpose, the 1977 version provides for a nonlinear relationship and a sensitivity to trip purpose. A linear relationship at, say, \$1.50 or \$2.80 per hour saved would be highly inappropriate if millions of daily trips, each saving a few minutes, were to be simply added together. According to Fig. 11.3.3, such small time savings are insignificant individually.

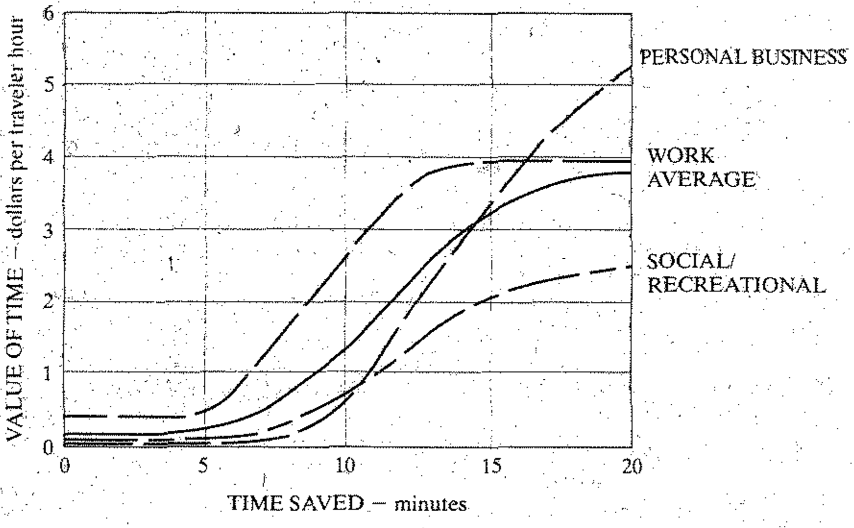


Figure 11.3.3 Value of time saved by trip purpose. (From *A Manual on User Benefit Analysis and Bus Transit Improvements*, Copyright 1977, by American Association of State Highway and Transportation Officials, Washington, D.C. Used by permission.)

Other impacts of transportation projects (e.g., effect on rural lifestyles or aesthetics) are much more difficult to quantify, let alone express in dollar equivalents. However, in order to be included in a B/C economic evaluation, they must be quantified and valuated.

11.3.6 Limitations of Economic Evaluation

The foregoing commentary brings to light the fact that economic efficiency analysis is not as objective as it may seem at first glance. Its strongest advantage is that it provides a useful quantitative but partial picture of the subject matter. Its major limitations may be classed into problems of impact enumeration, valuation, and distribution. The selection of an appropriate interest rate and the treatment of price inflation and deflation are also problematic.

The question of *impact enumeration* refers to the fact that not all impacts considered to be important can be included in the analysis. Even though no evaluation technique can possibly include all ramifications of major transportation projects, economic efficiency analysis further restricts the admissible set.

The problem of *impact distribution* refers to the fact that the benefits and costs are distributed unevenly between individuals and groups. For example, some persons may have to relocate their residences or businesses to permit the construction of a highway that could result in travel time and fuel savings for another group, the users of the new highway. Similarly, it may be argued that subsidizing a public transportation system entails the taking of tax dollars from everyone in order to enhance the mobility of the few that ride the system. In this connection the first piece of federal legislation to require a B/C analysis for public projects explicitly stated that public projects are justified:

> . . . if the benefits *to whomsoever they may accrue* are in excess of the estimated costs ([11.6], emphasis added).

Of course, counterarguments are possible in both examples just cited, but this is not the proper place to address them. It is clear, however, that *those who perceive that they will be adversely affected by a proposed project are not obliged to acquiesce on the grounds that the calculated overall B/C ratio is greater than unity.*

11.4 EFFECTIVENESS ANALYSIS

11.4.1 Background

The preceding discussion has pointed out that even when they can be quantified in terms of specific measures of performance the various impacts associated with proposed alternatives are often difficult to express in monetary terms. *Effectiveness,* which has been defined as the degree to which the performance of an alternative attains its stated objectives, seeks to rectify this problem by explicitly accounting for such impacts and providing a framework within which these impacts can be clearly defined and traded off via the choice of alternative. The effectiveness approach to evaluation and decision making is founded on the axiom that more informed, and hence better, decisions would result if the decision-makers were presented with the maximum amount of available information about the subject. Within this framework the basic role of the analyst becomes more concerned with facilitating the decision-making process by devising well-organized ways to summarize and transmit to the decision-makers the data required for the decision and less concerned with applying a specific technique that presumes to determine unambiguously the "best" alternative. At the same time the role of the decision-makers becomes more demanding as they are given the added responsibility of ultimately assigning relative values to the merits and demerits of the alternatives being considered. The vast technical literature on specific techniques that may be used to measure effectiveness (ranging from purely subjective to highly quantitative) as well as the decision-making processes and the institutional structures for which these techniques are best suited spans several disciplines. Only the basic elements of effectiveness analysis are discussed here.

11.4.2 Cost-Effectiveness

The application of economic efficiency methods to public projects had its origins in the civilian sector and the provision flood protection. Cost-effectiveness on the other hand was first applied in connection with the evaluation of military systems. In its strictest sense cost-effectiveness was an extension of the principles of economic efficiency, as it was concerned with maximizing the returns (in terms of effectiveness) of public expenditures described in terms of the monetary costs associated with the life cycles of proposed systems. The following simple example illustrates the essence of the method.

Suppose that the administration team of a university (i.e., the decision-making body) is faced with the task of selecting a new computer system for the college of engineering. Because the system is to be used primarily for undergraduate instruction, it has been agreed that the system should maximize the number of users it can accommodate simultaneously and that this number must be at least equal to that supported by the existing system. On the other hand, the administration has established a maximum cost constraint as well. Several computer manufacturers responded to a request for bids with the six mutually exclusive

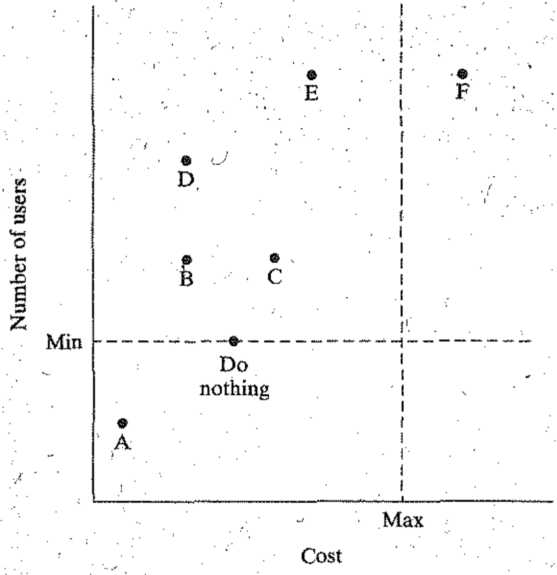


Figure 11.4.1 Example of cost versus effectiveness.

proposals shown in Fig. 11.4.1. In accordance with the agreed-upon rules, alternative A would be dropped because it fails to meet the minimum effectiveness level and alternative F would be eliminated as its costs exceed the maximum available resources. Furthermore, alternatives B and C and the do-nothing alternative would also be eliminated, as they cost at least as much but offer no better level of effectiveness than alternative D. The final choice would rest between alternatives D and E and would involve a trade-off between dollars and the number of potential users. This choice is an incremental consideration, but unlike the incremental analysis applied to the B/C ratio, the relative worth of the two impacts being traded off would be implicit in the final choice. The choice of alternative E would imply that the extra benefits are *at least* equivalent to the extra costs required. Conversely, the selection of alternative D would carry the implication that the worth of the incremental effectiveness associated with alternative E is less than the worth of the incremental dollar costs it would entail.

The problem of selecting the best computer system would be further complicated if the system's effectiveness were multidimensional, for example, if the availability of engineering software (however measured) were also considered to be important. Thus, as the dimensions of effectiveness increases, so does the complexity of determining the relative worth of the alternatives (i.e., evaluating them). Consequently an individual decision-maker soon becomes overwhelmed with vast amounts of often-conflicting information. The matter becomes worse as the number of individuals constituting the decision-making group increases. Hence a need arises to organize the available information and to establish procedures that aid the attainment of consensus.

Several ways by which the relative assessment of alternatives may be accomplished are as follows:

1. The decision-makers select the best based on their unexpressed subjective judgments.

2. Aided by the analyst, decision-makers *rank* alternative options in an ordinal sense (i.e., A is better than B) and make selection by one of several *rank-ordering* procedures.
3. Aided by the analyst and other sources, the decision-makers assign a score (usually based on the relative weights of impacts) to each alternative and select the one with the highest score.

11.4.3 Rank-Ordering Techniques

A rank-ordering technique with obvious application to the topic is one that Sage [11.7] treats with mathematical formalism and which has been applied in several variations by others. In simplified form the method works as follows.

The decision-maker, faced with n alternatives, is asked to compare them in pairs according to a contextual relationship, such as "alternative i is superior to alternative j." After the decision-maker has completed the consideration of all pairs, the following rules are examined to ensure consistency: (1) An alternative cannot be superior to itself. (2) If i is superior to j, then j cannot be superior to i. (3) If i is superior to j and j is superior to k, then i is superior to k. If the decision-maker violates any of these rules, an inconsistency is detected that should be resolved.

Example 11.3

Considering four options, a decision-maker has completed the following array by placing a 1 in cell (i, j) if the answer to the question "option i is superior to option j" was affirmative and a 0 otherwise. Check for any inconsistencies in the decision-maker's logic, and if none are found, identify the rank order of the four options.

i j	A	B	C	D
A	0	0	1	1
B	1	0	0	1
C	1	1	0	1
D	0	0	0	0

Solution The diagonal elements are all 0, as expected. However, A has been designated to be superior to C at the same time that C was considered to be superior to A. Moreover, a circularity exists between A, B, and C: B was considered to be superior to A, C superior to A, and C superior to B. Therefore a defect in the assessment is revealed. The same conclusion may be reached by drawing the directed graph shown in Fig. 11.4.2, where each arrow is directed from the inferior to the superior option.

Example 11.4

Assume that the inconsistency of Example 11.3 was pointed out to the decision-maker and that after considerable thought the decision-maker revised the original assessment by rating option A inferior to option C. Revise the solution of Example 11.3.

Solution The revised directed graph is shown in Fig. 11.4.3. Furthermore, by eliminating redundant arrows (e.g., from D to C), a clear rank order emerges.

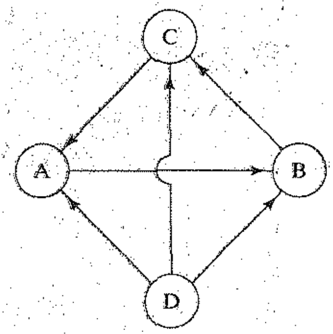


Figure 11.4.2 Graphic identification of rank-ordering deficiencies.

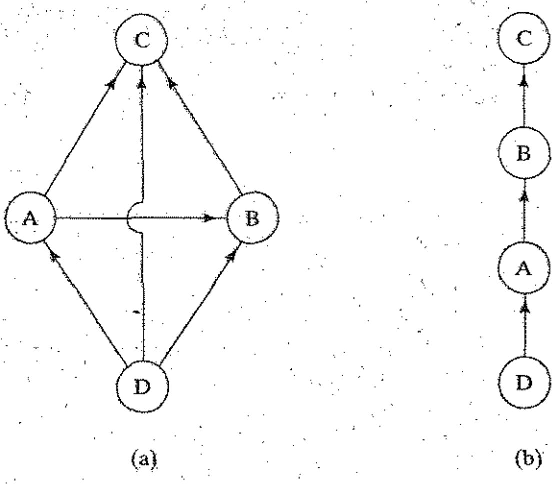


Figure 11.4.3 Clear rank order.

Discussion This method produces a consistent ranking of the options based on the subjective judgment of the decision-makers that can provide guidance to the discovery of inconsistencies that need to be clarified and resolved. When the decision-making body consists of many individuals, an overall compromise must be made. One way of accomplishing this is the method by which the "number 1 college football team" is selected in the United States: A panel of experts (i.e., football coaches and sports reporters) are asked to rank the top teams and the team that receives the most first-place votes is ranked as being the best. Alternatively, each first-place, second-place, and so on, vote is weighted and combined to derive an overall score.

Table 11.4.1 illustrates the application of the ranking techniques to eight alternatives considered in the 1979 Honolulu study, which is described in subsection 11.2.3. Note that the number of firsts, seconds, and thirds have been tabulated as well. Table 11.4.2 is a similar summary of rankings for the baseline alternative, an alternative consisting of a combination of TSM strategies and a 14-mi-long fixed-guideway rapid-transit alternative. Table 11.4.3 presents some details relating to the basis on which several of the rankings of Table 11.4.2 were derived.

TABLE 11.4.1 Summary of Rankings for Eight Alternatives

		7		14		23	28	23
		mi		mi		mi	mi	mi
	Bus	LRT	FG[a]	LRT	FG[a]	LRT	LRT	FG[a]
Objective 1								
a. Availability & coverage	—	—	—	—	—	—	—	—
b. Avg. trip time (min.)	2	1	1	—	—	2	3	1
c. Service reliability	2	1	1	—	—	2	2	1
d. Rider convenience (transfers per trip)	1	2	2	—	—	2	1	2
e. Rider comfort	2	1	1	—	—	—	—	—
Objective 2								
a. System patronage	2	1	1	—	—	2	2	1
b. System capacity	2	1	1	—	—	—	—	—
Objective 3								
a. Consumption of land (acres)	3	2	1	2	1	2	2	1
b. Displacement of residents (units)	3	2	1	2	1	2	2	1
c. Displacement of businesses (units)	3	2	1	2	1	2	2	1
d. Reduction of community amenities	—	—	—	—	—	—	—	—
e. Disruption to future development	—	—	—	—	—	—	—	—
f. Disruption to local circulation	—	—	—	—	—	2	3	1
g. Disruption from constr. activities	—	—	—	—	—	1	2	1
h. Savings in energy (million gal/yr.)	1	3	2	2	1	2	3	1
i. Technical risk	3	1	2	1	2	1	1	2
Objective 4								
a. Support regional development	—	—	—	—	—	—	—	—
b. Support comm. development	—	—	—	—	—	—	—	—
Objective 5								
a. Reduction air pollution (ton/yr.)	3	2	1	2	1	3	2	1
b. Noise level (dBA)	3	2	1	2	1	2	2	1
c. Visual intrusion	3	2	1	2	1	2	2	1
d. Vistas	2	3	1	2	1	1	1	2
e. Historic sites	—	—	—	—	—	—	—	—
Objective 6								
a. Reduce accident exposure	—	—	—	—	—	2	2	1
b. Security	—	—	—	—	—	—	—	—
Objective 7								
a. Total annual cost	3	2	1	2	1	1	3	2
b. Cost per trip	3	2	1	2	1	2	3	1
c. Benefit-cost ratio	3	2	1	2	1	2	3	1
No. of firsts	2	6	15	1	11	4	3	15
No. of seconds	6	10	3	11	1	14	10	4
No. of thirds	10	2	0	0	0	1	6	0

[a] FG = fixed guideway.

Source: Urban Mass Transportation Administration [11.3].

TABLE 11.4.2 Summary of Rankings: Baseline, TSM, and Fixed Guideway

	Approach A		
Evaluation factors	Baseline	TSM	14-mi. fixed gwy.
Objective 1			
a. Availability & coverage	2	1	1
b. Avg. trip time	3	2	1
c. Service reliability	3	2	1
d. Rider convenience	2	2	1
e. Rider comfort	2	2	1
Objective 2			
a. System patronage	3	2	1
b. System capacity	3	2	1
Objective 3			
a. Consumption of land	1	2	3
b. Displacement of residents	1	1	2
c. Displacement of businesses	1	2	3
d. Reduction of community amenities	1	1	2
e. Disruption of future development	1	1	2
f. Disruption of local circulation	3	2	1
g. Disruption from constr. activities	1	1	2
h. Savings in energy	3	2	1
i. Technical risk	1	1	2
Objective 4			
a. Support regional development	2	2	1
b. Support comm. development	2	2	1
Objective 5			
a. Reduction air pollution	3	2	1
b. Noise level	2	2	1
c. Visual intrusion	1	1	2
d. Vistas	1	1	2
e. Historic sites	1	1	2
Objective 6			
a. Reduce accident exposure	3	2	1
b. Security	2	2	1
Objective 7			
a. Total annual cost	1	2	3
b. Total annual cost per trip	1	2	3
c. Benefit-cost ratio	—	2	1
No. of firsts	12	9	16
No. of seconds	7	19	8
No. of thirds	8	0	4

Source: Urban Mass Transportation Administration [11.3].

TABLE 11.4.3 Comparative Evaluation Matrix: Baseline, TSM, and Fixed Guideway

Evaluation factors	Approach A		
	Baseline	TSM	14-mi. fixed gwy.
Objective 1			
a. Availability & coverage[a]	(2)	(1)	(1)
b. Avg. trip time (min)	40.7	40.1	33.7
c. Service reliability[a]	(3)	(2)	(1)
d. Rider convenience[a]	(2)	(2)	(1)
e. Rider comfort	(2)	(2)	(1)
Objective 2			
a. System patronage—1985 (million/yr.)	64.7	83.6	102.4
b. System capacity[a]	(3)	(2)	(1)
Objective 3			
a. Consumption of land (acres)	—	3	22
b. Displacement of residents (units)	—	—	162
c. Displacement of businesses (units)[a]	—	2	183
d. Reduction of community amenities[a]	(1)	(1)	(2)
e. Disruption of future development[a]	(1)	(1)	(2)
f. Disruption of local circulation[a]	(3)	(2)	(1)
g. Disruption from constr. activities[a]	(1)	(1)	(2)
h. Savings in energy (million gal/yr.)	—	0.9	4.5
i. Technical risk[a]	(1)	(1)	(2)
Objective 4			
a. Support regional development[a]	(2)	(2)	(1)
b. Support comm. development[a]	(2)	(2)	(1)
Objective 5			
a. Reduction air pollution (ton/yr.)	—	220	2260
b. Noise level (dBA)	86–88	86–88	77
c. Visual intrusion[a]	(1)	(1)	(2)
d. Vistas[a]	(1)	(1)	(2)
e. Historic sites[a]	(1)	(1)	(2)
Objective 6			
a. Reduce accident exposure[a]	(3)	(2)	(1)
b. Security[a]	(2)	(2)	(1)
Objective 7			
a. Total annual cost[b]—1985 ($ million)	32.9	45.0	66.2
b. Total annual cost per trip ($)	0.508	0.538	0.647
c. Benefit-cost ratio[c]	—	1.12	1.13

[a] For comparative measures, alternatives are ranked in the order of how well they met the objective.

[b] All costs based on constant 1975 dollars and an interest rate of 7%.

[c] Based on constant 1975 dollars.

Source: Urban Mass Transportation Administration [11.1].

11.4.4 Scoring Techniques

The objective of *scoring techniques* is the assignment of meaningful grades to the alternatives in a manner that reflects the degree to which they differ from each other. Numerous scoring techniques and procedures are reported in the technical literature. The following discussion is an amalgamation of these methods, emphasizing their rationale rather than an in-depth examination of any one in particular.

Figure 11.4.4 is an expanded version of the generalized impact matrix of Fig. 11.4.1 as it relates to one of the alternatives being evaluated. The impacts that are considered to be relevant to the evaluation are listed in the first row of Fig. 11.4.4. Related *impacts* are combined into a smaller number of evaluation *criteria,* which are themselves combined to yield the alternative's *overall score.* Conceptually the combination of a set of impacts into a criterion is identical to the derivation of the overall score from a set of quantified criteria. Furthermore, the evaluation of very complex systems may require more than the three levels of aggregation illustrated. At the other extreme the simplest case involves a scoring scheme that is based on a single criterion, which is identical to a single impact. A slightly more complex case entails a single criterion that is composed of several impacts. The composite grade of any criterion involves the following steps:

1. The impacts that constitute the criterion are identified and quantified, usually on different scales of measurement.
2. The quantified impacts are placed on the same scale of measurement.
3. The scaled impacts are assigned relative weights and combined.

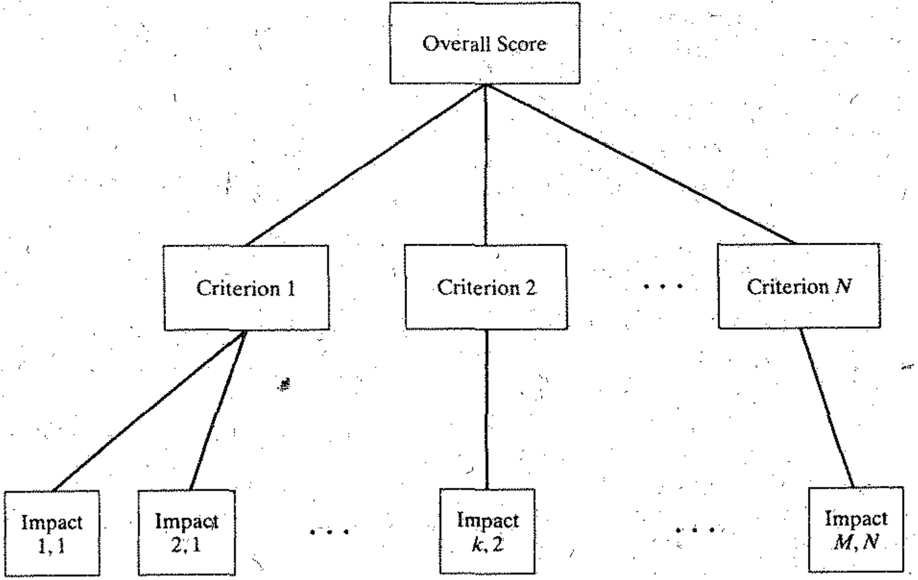


Figure 11.4.4 Impacts, criteria, and overall score.

For instance, the NPW, which incorporates the net effect of many impacts, may serve as one of several evaluation criteria. As explained in Section 11.3, the NPW of an alternative is derived by first predicting the likely impacts of the alternative (step 1), translating the disparately quantified impacts to dollar equivalents (step 2), and weighting them equally as if they were in fact dollars (step 3). The general scoring methods discussed here allow for (1) the use of a common scale other than a monetary scale and (2) the assignment of unequal weights to the impacts. This is the essential difference between measures of economic efficiency on one hand and measures of effectiveness on the other.

Example 11.5

An elected official wishes to evaluate three transportation proposals on the basis of three criteria: economic worth, aesthetic quality, and electorate reaction. The economic worth of the alternatives is measured by their NPW, which has been calculated by a consulting firm according to accepted practice. The aesthetic attributes of the alternatives have been assessed by a survey conducted by a marketing research company and is measured by the percent of respondents that are pleased with each alternative. The probable electorate reaction has been reported by the official's staff, who maintain contacts with the voters in the official's district. The following table summarizes the available information:

Alternative	NPW (millions of dollars)	Aesthetics (%)	Electorate
A	6	70	Neutral
B	13	40	Favorable
C	14	90	Unfavorable

Solution Three possible ways of scoring are presented:

1. *Combined rankings.* The alternatives may first be ranked according to each criterion from the worst (i.e., lowest ranking) to the best (i.e., highest ranking), and a composite score may be derived by summing the rankings of each alternative. Thus

$$S_i = \sum_j R_{ij} \tag{11.4.1}$$

where

S_i = score of alternative i

R_{ij} = rank of alternative i with respect to criterion j

For the current example the scores of the three alternatives become:

Alternative	NPW	Aesthetics	Electorate	Score
A	1	2	2	5
B	2	1	3	6
C	3	3	1	7

This method applies equal weights to the criteria. Moreover, it is oblivious to the degree to which the alternatives differ from each other with respect to each criterion.

2. *Weighted rankings.* The criteria may be assigned relative weights, which will affect the contribution of each criterion to the overall scores:

$$S_i = \sum_j w_j R_{ij} \tag{11.4.2}$$

where w_j is the relative weight of criterion j. Assuming that the official considers satisfying the electorate to be four times as important as aesthetics and twice as important as economic worth; that is,

$$w(\text{NPW}) = 2 \qquad w(\text{aesthetics}) = 1 \qquad w(\text{elect.}) = 4$$

the scores of three alternatives become:

Alternative	Score
A	$2 \times 1 + 1 \times 2 + 4 \times 2 = 12$
B	$2 \times 2 + 1 \times 1 + 4 \times 3 = 18$
C	$2 \times 3 + 1 \times 3 + 4 \times 1 = 13$

The weights assigned to each criterion are reflected in the overall score of the alternatives. However, the problem of scaling the magnitudes associated with the alternatives with respect to each criterion still remains unsolved. Thus the differences in NPW between alternatives A versus B on one hand and B versus C on the other are not captured by this method.

3. *Scaled criteria.* The three criteria used in this example are measured on different scales: The NPW is a quantitative measure that is unbounded at either end. The scale that has been selected to measure aesthetic quality ranges from 0 to 100. Finally, elector reaction has been reported on an ordinal scale. If the three criteria are to be combined into a single score, they must be placed on a common scale. For the sake of illustration, consider a common ordinal scale ranging from 0 to 100. The criterion relating to aesthetics is already measured on this scale. The NPW of each alternative may be mapped onto the common scale by assigning a grade, say 90, to alternative C and proportioning accordingly the grades of the other two. With respect to the third criterion, neutral reaction may be used as an anchor midway on the scale.

Alternative	NPW	Aesthetics	Electorate	Score I	Score II
A	40	70	50	160	350
B	85	40	80	205	580
C	90	90	40	200	450

Two scores are shown in this table. Score I was derived using a function similar to Eq. 11.4.1 with the criteria levels associated with each alternative replacing the raw rankings. Score II is based on a weighting scheme, as in Eq. 11.4.2.

Discussion This example illustrates the mechanics of only four out of a very large number of possible scoring techniques. Theoretically inclined individuals may even be tempted to apply one or more of the scoring techniques described here to combine the various scores derived above into a super score, but such a process has no bounds. The potential for an infinite number of scoring variations, each leading to a different decision, may give to the process the appearance of capriciousness or arbitrariness. But no evaluation technique can be an end in itself. Thus the usefulness of any technique lies in its ability to help organize the decision-making process in an explicit and systemic way and not in its ability automatically to yield an unequivocable result. This presupposes a predisposition on the part of the decision-maker to participate actively in all stages of the process, including the identification of impacts and criteria, their scaling and weighing, and ultimately the final decision. As to the choice of technique, it is largely situational, depending on the quantity and quality of the available information.

11.4.5 Group Consensus

Perhaps the major source of difficulty associated with effectiveness analysis is the dependence on the subjective judgment of the decision-maker. This dependence is often moderated by the reliance on decision-making bodies that consist of many individuals. But precisely because of the differences that exist between individuals, group decision making requires the attainment of group consensus. Traditional means for reaching consensus include group discussion, debate, argumentation, and brainstorming. The advantages of these methods include the exposure of the group to differing points of view. A major drawback is that certain individuals tend to dominate the process because of rank, strength of conviction, or persuasive ability. Several methods that attempt to eliminate this difficulty have been devised. Theoretically the group's consensus may be revealed by statistically summarizing the responses of the members of a panel to the questions required by the ranking and scoring techniques discussed previously. The *delphi method,* originally proposed by the Rand Corporation [11.8], encompasses several procedures that attempt to facilitate collective decisions via a series of questionnaires administered to all members of a panel and accompanied by summaries of the panel's earlier responses. The final decision is enhanced by anonymity, equal treatments of all points of view, and the fact that the participants are free to revise their positions.

11.5 Summary

In this chapter we introduced the concepts of project evaluation and described the elements of some commonly used methods that can aid the evaluation of alternative courses of action and can facilitate the selection of an alternative for implementation. The complex nature of transportation-related decisions was conveyed only by implication because a detailed examination of the political, legislative, and judicial reverberations of transportation decisions is beyond the scope of this introductory book.

Evaluation methods were classified into economic efficiency methods and effectiveness methods. The former require that the quantified impacts that are relevant to evaluation should be translated into money equivalents and treated as such. The traditional economic efficiency evaluation measures of net present worth and B/C analysis were then described and illustrated. Finally, the case was made for expanding the evaluation framework to incorporate impacts that are either impossible or difficult to quantify in terms of dollars. Within this expanded framework various measures of effectiveness as well as measures of economic efficiency can serve as evaluative criteria for the ranking and scoring of alternatives.

EXERCISES

1. Drawing on the store of knowledge you have amassed so far, discuss the contents of Table 11.2.1.
2. Discuss several possible ways by which each of the travel, operating characteristics, and costs listed in Table 11.2.2 could have been estimated. Be as specific as you can.
3. Referring to Table 11.2.2, explain why the projected total daily transit patronage is different for each transportation alternative studied. Which part of the sequential travel-demand-forecasting methodology do you think produced these results? Explain specifically the most likely model variables that capture this effect.
4. In reference to Table 11.2.2, why do you think the estimated daily trips on the 7-mi fixed-guideway alternative are the same as those corresponding to the 7-mi LRT?
5. Perform an incremental B/C analysis of the alternatives listed in Table 11.2.3, assuming an interest rate of 4%. Why is the interest rate important in the B/C ratio method of evaluation?
6. Repeat Exercise 5 assuming an interest rate of 10%.
7. With reference to objective 7 of Table 11.4.2, why was not the B/C ratio of the baseline alternative reported? What implication can this fact have on the feasibility of the other two alternatives?
8. Perform an incremental analysis of the three alternatives listed in Table 11.4.3 and discuss your results.
9. Determine the preferred alternative of a decision-maker who has completed the following array on the basis of the contextual relationship alternative *i* is better than alternative *j*.

j *i*	A	B	C	D
A	0	0	0	0
B	1	0	0	1
C	1	1	0	1
D	0	0	0	0

10. Using the data given in Table 11.4.2
    (a) Calculate the simple combined rankings score of each of the three alternatives with respect to each of the seven objectives.
    (b) Rank the alternatives for each objective according to the scores derived in part a.
    (c) Apply the simple combined rankings technique to the results of part b.
    (d) Use the following weights to calculate the weighted-ranking scores derived in part (b).

i	1	2	3	4	5	6	7
w	2	4	1	4	5	2	7

Explain any assumptions that you felt were necessary to complete this exercise and explain why a universally applicable effectiveness analysis method is not possible.

11. Three alternative plans (A, B, and C) have been ranked with respect to four criteria (I to IV) as follows:

	I	II	III	IV
A	2	1	3	2
B	1	2	2	1
C	3	3	1	3

where 1 means the best. Apply the single combined-rankings technique to derive an overall score for each alternative.

12. Given the following weights for the four criteria of Exercise 11, compute the weighted-ranking scores for the three alternatives.

i	I	II	III	IV
w	2	4	1	2

13. Discuss the steps you would follow in order to apply the B/C ratio method to evaluate alternative highway-safety proposals consisting of all possible combinations of 12 accident-reducing actions, such as signalization, curve widening, street lighting, and so on.

14. From the government documents section of your school's library, obtain a planning study for a major transportation action and report on the evaluation method employed.

15. Discuss the possible differences in the perspectives of an environmentalist, a construction firm president, and a federal judge regarding a proposal to build a multilane highway through a conservation district. Use Appendix A as a guide to your answer.

REFERENCES

11.1 Thomas, E. N., and J. L. Schofer, *Strategies for the Evaluation of Alternative Transportation Plans,* National Cooperative Highway Research Program Report 96, Highway Research Board, National Research Council, Washington, DC, 1970.

11.2 Manheim, M. L. et al., *Transportation Decision-Making: A Guide to Social and Environmental Considerations,* National Cooperative Highway Research Program Report 156, Transportation Research Board, National Research Council, Washington, DC, 1975.

11.3 Urban Mass Transportation Administration, *Draft Environmental Impact Statement: Honolulu Area Rapid Transit Project,* UMTA Project Report HI-03-0005, U.S. Department of Transportation, Washington, DC, July 1979.

11.4 Wohl, M., *Transportation Investment Planning: An Introduction for Engineers and Planners,* Lexington Books, Lexington, MA, 1972.

11.5 American Association of State Highway and Transportation Officials, *A Manual on User Benefit Analysis of Highway and Bus-Transit Improvements 1977,* AASHTO, Washington, DC, 1978.

11.6 *United States Code,* U.S. Government Printing Office, Washington, DC, 1940, p. 2964.

11.7 Sage, A. P., *Methodology for Large-Scale Systems,* McGraw-Hill, New York, 1977.

11.8 Dalkey, N., and O. Helmer, "An Experimental Application of the Delphi Method to the Use of Experts," *Management Science,* 9, 3 (April 1963): 458–467.

PART 4

Supporting Elements

12

Elements of Engineering Economy

12.1 MONEY AND ITS TIME VALUE

Money's *raison d'être* is its acceptability as a medium of exchange. It can be used in exchange for goods and services much more efficiently than the direct trading of goods and services (i.e., barter). Because money can be used for the purchase of many items, it can serve as a standard of value for them, at least in a relative sense. Because it can retain the ability to be exchanged for other commodities at various times, money is also a store of value.

In one view the term *value* of a commodity is synonymous with the number of monetary units (or the *price*) that it commands. Others use the term value to refer to the degree to which a particular good or service satisfies the needs of individuals and employ the term *utility* to clarify this difference. The fact is that such a difference between price and utility exists and that the term value is often used for both. The context in which it is used commonly clarifies the connotation intended.

Often the *value of money* is defined as the reciprocal of the prices of the goods and services for which it can be exchanged. Thus, if for various reasons the number of monetary units (e.g., dollars or yen) required to obtain a given item were to increase, the situation could be described as either an increase in the price of that item or, conversely, a decrease in the value of the monetary unit. A major difficulty associated with this definition for the value of money lies in the fact that the prices of the myriad of goods and services that are daily exchanged in markets do not all behave in the same way. The prices of some may be on the decrease, whereas the prices of others are either stable or increasing. These price changes are caused by many conditions, including changes in the quantities demanded, technological breakthroughs that result in more efficient or less costly production methods, and changes in the availability of resources (or factors of production) through depletion, new discoveries, or even catastrophic events such as wars. To complicate matters, the supply of money in the form of currency and credit also affects prices. Even though the prices of individual goods and services vary differentially, the general behavior of prices may, nevertheless, be described by several indicators. The most well known of these indicators is the *consumer price index* (CPI), which is compiled by the U.S. Bureau of Labor Statistics to capture the price changes of a combination of goods that the typical family considers essential. When the general price level is on an upswing, the economy is said to experience price *inflation.* When prices are falling, the economy is in a *deflationary period.* The value of money as defined in this way decreases with price inflation and increases with price deflation.

People and firms exchange goods and services in order to maximize the satisfaction (i.e., utility) they derive from them. For example, firms give up money (and, indirectly, other goods and services) to purchase the services of employees and needed factors of production for the purpose of deriving profits from the sale of the goods and services they produce. Exchange is possible because the utility that individuals and other economic entities attach to the items being exchanged is not identical. Both parties to the exchange give up something they consider less desirable for something they consider more desirable. Since, as a store of value, money stands for the opportunities it represents to consumers and producers, it is imbued with the characteristics shared by all commodities, including the ability to satisfy human wants, that is, utility. As a major factor of production, it carries the ability to earn profits, and this earning power of money is reflected in the *time value of money.* Simply stated, a dollar in hand at present is not the same as a dollar in hand at some future date because in the future the present dollar would be incremented by the return it would earn in the meantime. Economic entities are willing to borrow and lend money at a premium because of the various profit-making opportunities they are able to pursue. Typically the lender is satisfied with the "rental" to be received from a borrower for the use of the lender's money, whereas the borrower is looking for an opportunity to put the borrowed money to some use that would gain for the borrower a satisfactory profit above the cost of "renting" the money.

12.2 INTEREST AND DISCOUNT

The premium paid or received for the use of money is known as *interest.* The rate at which interest accumulates (i.e., the *interest rate*) is quoted as the percentage gained over a specified time period, known as the *interest period.* Thus the interest rate relates a sum of money presently in hand to its equivalent sum at some future date. The rate that relates a sum of money at some future date to its equivalent at present is known as the *discount rate.*

The value of money is affected by price instabilities. Consequently the interest rate that lenders seek and borrowers are willing to pay is affected by (1) their expectations relating to potential movements of the price level (i.e., the *purchasing power* of money) and (2) their desired return or profit (i.e., the *earning power* of money). Assuming constant dollars (i.e., ignoring inflation) facilitates the understanding of the basic concepts covered by this section. Incidentally, the term *current dollar* refers to the reciprocal of the general price level at any given time, and therefore includes the effect of inflation. Current or real rates account for inflation. The real interest rate results after subtracting inflation from the prime rate (i.e., 7% − 2% = 5%).*

*More precisely, the real interest rate is derived from (1 + prime rate) ÷ (1 + inflation rate) − 1 = real interest. The aforementioned example becomes (1.07 ÷ 1.02) − 1 = 0.04902 or prime rate = 4.902%.

Example 12.1

A business firm borrows $10,000 and agrees to pay back $10,200 at the end of one month. Calculate the interest rate involved in the transaction.

Solution The total interest paid at the end of the month is $200. The interest period is 1 month and the interest rate is

$$\frac{\$200}{\$10{,}000} = 0.02 \text{ or } 2\% \textit{ per month}$$

Discussion The interest rate translates a present sum to a future sum. The terms *present* and *future* are defined in relation to each other and not to specific calendar times. In other words the previous calculation would be the same as long as the two times are separated by 1 month. The discount involved in this transaction is also $200. It relates the future sum ($10,200) to its present equivalent ($10,000). By definition, the discount rate is 2% per month, that is, the difference between the future and present sums divided by the *present sum* (see Example 12.2).

Example 12.2

An investor purchases a zero coupon bond for $867.98. The bond has a face value of $1000 and matures in 1 year. This means that the bond can be cashed after 1 year for $1000. Calculate the discount rate involved in the transaction.

Solution The future sum ($1000) was discounted by $132.02. The discount rate was 132.02/867.98 = 0.1521, or 15.21% per year.

Discussion The same transaction may be viewed as follows: The investor's principal of $867.98 earned an interest of $132.02 in 1 year at an interest rate of 15.21% per year.

12.3 SIMPLE AND COMPOUND INTEREST

In the preceding two examples the time over which the interest was earned (i.e., 1 month and 1 year, respectively) coincided with the interest period for which the interest and the discount rates were either calculated or quoted. This does not always need to be the case. The interval of time between the present and the future can be longer than a single interest period, in which case the present sum continues to earn interest at the quoted rate. *Simple interest* refers to the case where the percentage of the *original* sum of money is added at the end of each interest period. In the case of *compound interest* both the original sum (principal) and the interest earned are allowed to earn interest during subsequent periods. The difference between the two is illustrated next.

Example 12.3

A person who has a sum of $10,000 to invest is faced with the options of (a) earning simple interest at an annual rate of 9% per year or (b) earning compound interest at an annual rate of 8% per year. In both cases the principal and interest are to be withdrawn at the end of a 5-year period. Compare the two investment options.

Solution The consequences of each of the two options are tabulated as follows:

Year	Principal and interest at the start of the year	Interest added at the end of the year	Principal and interest at the end of the year
		Option (a): Simple interest at 9% per year	
1	$10,000.00	$900.00	$10,900.00
2	10,900.00	900.00	11,800.00
3	11,800.00	900.00	12,700.00
4	12,700.00	900.00	13,600.00
5	13,600.00	900.00	14,500.00
		Option (b): Compound interest at 8% per year	
1	$10,000.00	$ 800.00	$10,800.00
2	10,800.00	864.00	11,664.00
3	11,664.00	933.12	12,597.12
4	12,597.12	1007.77	13,604.89
5	13,604.89	1088.39	14,693.28

Discussion From the borrower's perspective, the annual 9% simple interest rate is superior to the 8% annually compounded rate up to the end of the third year. Beyond that time, the latter becomes the better option. From the lender's point of view, the reverse is true. Thus both the magnitude of the interest rate and the number of interest periods affect the relative consequences of the two types of interest. The future worth F of a present sum P can be calculated by

$$F = P(1 + in) \tag{12.3.1}$$

for the case of simple interest, and by

$$F = P(1 + i)^n \tag{12.3.2}$$

for the case of compound interest, where

i = interest rate (percent per period divided by 100)

n = number of interest periods separating P and F

The term multiplying the single sum P in Eq. 12.3.2 is one of several useful factors and is known as the *single-sum* (or single-payment) *compound-amount factor* for an interest rate i per period and n periods separating P and F (CAF′, i, n). Solving Eq. 12.3.2 for P yields

$$P = F\frac{1}{(1 + i)^n} \tag{12.3.3}$$

and the factor that discounts a future sum F to a present sum P is known as the *single-sum present-worth factor* (PWF′, i, n).

Example 12.4

A sum of $100,000 is invested at an annually compounded interest rate of 8% per year. Calculate its equivalent at the end of 20 years.

Solution For P = $100,000, $i = 0.08$, and $n = 20$, Eq. 12.3.2. yields

$$F = \$100{,}000(1 + 0.08)^{20}$$

$$F = \$100{,}000(4.66096) = \$466{,}096$$

The single-sum compound-amount factor is 4.66096.

12.4 NOMINAL AND EFFECTIVE INTEREST RATES

Frequently interest rates are specified on the basis of a period (usually a year) when compounding occurs more frequently than the specified period. By convention, the magnitude of the quoted *nominal interest rate* is equal to the product of the interest rate per interest period times the number of interest periods in the specified period. For example, a nominal annual rate of 12% compounded semiannually means that the interest rate is 6% per 6-month interest period. Similarly, a nominal interest rate of 12% compounded monthly implies an interest rate of 1% per month. The following example shows that the *effective interest rate* is larger than the nominal interest rate because the interest earned at the end of each interest period is subsequently allowed to earn interest as well.

✓ **Example 12.5**

Compute the equivalent of $1,000,000 at the end of 5 years if the annual interest rate is (a) 8% per year compounded quarterly and (b) 8% per year compounded semiannually. For each case, calculate the effective *annual* interest rate.

Solution For part (a) the effective quarterly rate is 2% per quarter. A 5-year period contains 20 quarters. Therefore substituting $i = 0.02$ and $n = 20$ in Eq. 12.3.2 results in

$$F = \$1{,}000{,}000(1 + 0.02)^{20} = \$1{,}485{,}947$$

To find the effective annual rate, consider the compound-amount factor for 1 year expressed in terms of the annual effective rate i and the annual nominal rate r compounded m times a year. The two must yield the same relationship between P and F 1 year apart. Therefore

$$(1 + i) = \left(1 + \frac{r}{m}\right)^m$$

Consequently the effective annual rate i is

$$i = \left(1 + \frac{r}{m}\right)^m - 1 \tag{12.4.1}$$

Therefore the effective annual rate for part (a) of this example is

$$i = (1 + 0.02)^4 - 1 = 0.082432 \quad \text{or} \quad 8.2432\% \text{ per year}$$

To illustrate its use, Eq. 12.4.1 is applied to the solution of part (b) of this example. With $m = 2$ interest periods per year, the effective annual rate is

$$i = (1 + 0.04)^2 - 1 = 0.0816 \quad \text{or} \quad 8.16\% \text{ per year}$$

Using this effective annual rate, at the end of 5 years the initial $P = \$1{,}000{,}000$ becomes

$$F = \$1{,}000{,}000(1 + 0.0816)^5 = \$1{,}480{,}244$$

The same result would have been obtained by using an effective semiannual interest rate of 4% per 6-month period and ten interest periods (i.e., 5 years times two 6-month periods in a year).

12.5 DISCRETE AND CONTINUOUS COMPOUNDING

In the foregoing examples compounding was considered to occur with finite interest periods, such as a year, 6-months, a quarter of a year, a month, and so forth. This is called *discrete compounding* because the interest is paid at the end of each discrete interest period. *Continuous compounding* represents the limiting case when the interest period approaches zero. Given a nominal interest rate r under continuous compounding, the effective interest rate would be the limit of i in Eq. 12.4.1 as the number of interest periods m approaches infinity. In that case the *effective interest rate for continuous compounding* is equal to $(e^r - 1)$, where e is the base of natural logarithms.

According to Eq. 12.3.2, when interest is compounded continuously at a nominal interest rate r per a specified period, the relationship between single sums separated by n periods is

$$F = Pe^{rn} \tag{12.5.1}$$

$$P = Fe^{-rn} \tag{12.5.2}$$

The multiplier of P in Eq. 12.5.1 is known as the *single-sum compound-amount factor* (CAF′, r, n) and the multiplier of F in Eq. 12.5.2. is called the *single-sum present-worth factor* (PWF′, r, n), both with continuous compounding. The subscript r refers to the *nominal* interest rate per period.

Example 12.6

Solve Example 12.5 for the case of continuous compounding.

Solution The nominal annual rate is still 8%. Hence for a 5-year period between P and F

$$F = Pe^{(0.08)(5)} = \$1,491,825$$

The effective annual interest rate is approximately equal to 8.33% per year.

Discussion For the same nominal interest rate the magnitude of F increases as the number of compounding interest periods m increases. The maximum occurs when m approaches infinity, that is, when the interest is compounded continuously.

12.6 CASH FLOWS

Up to this point the discussion of money equivalencies was restricted to the case of two single sums (or payments) separated by a time interval. The general case of a *cash flow* such as the one illustrated in Fig. 12.6.1 is frequently encountered, and the equivalent single sum E_k of the cash flow at a time k may be desired. This task can be accomplished by summing the equivalent sums at time k of each single payment in the cash flow using the appropriate single-sum present-worth or compound-amount factors as follows:

$$E_k = \sum_{j=0}^{k} S_j\,(\text{CAF}',*, k-j) + \sum_{j=k+1}^{n} S_j(\text{PWF}',*, j-k) \tag{12.6.1}$$

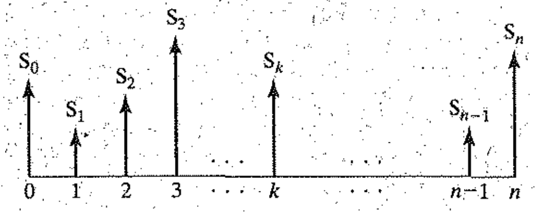


Figure 12.6.1 Cash flow.

The first subscript of the factors, which is designated by the asterisk, is either the effective interest rate per period, i, for discrete compounding or the nominal interest rate, r, per period for continuous compounding. Note that the single payment S_k has been included as the last term of the first summation but the fact that when $j = k$, $(k - j) = 0$ leaves that single payment intact. The terms S_j can be positive, zero, or negative. When they are specified by the difference between receipts and payments at each time j, the cash flow is referred to as the *net cash flow*.

Example 12.7

Given the following cash flow, calculate its equivalent single sum at (a) the end of the sixth period, (b) the end of the fourth period, and (c) time zero. Assume discrete compounding at 10% per period.

Solution This problem entails the application of Eq. 12.6.1. The solution to part (a) is tabulated next. It involves only the first summation shown in the equation because $k = n$.

j	0	1	2	3	4	5	6
S_j	100.00	200.00	50.00	0.00	150.00	50.00	200.00
$(k - j)$	6	5	4	3	2	1	0
S_j(CAF′)	177.16	322.10	73.21	0.00	181.50	55.00	200.00
							Sum = 1,008.97

The solution to part (b) may be obtained by applying Eq. 12.6.1 with $k = 4$. However, since the single sum obtained in part (a) is equivalent to the original cash flow, all that is needed is to find its equivalent at the end of the fourth period, or two periods earlier than the sixth. The single-sum present-worth factor for discrete compounding with an interest rate $i = 0.10$ and $n = 2$ yields

$$E_4 = (1008.96)(0.82645) = 833.86$$

Similarly, the equivalent single payment at time zero can be computed by Eq. 12.6.1 with $k = 0$. Alternatively, it can be calculated by discounting E_6 to its equivalent six periods earlier, or by discounting E_4 to its equivalent four periods earlier. Either one of the last two methods is simpler than the direct application of the long equation.
Hence

$$E_0 = (1008.96)(0.56447) = 569.54$$

or

$$E_0 = (833.86)(0.68301) = 569.54$$

Discussion This example illustrates the fact that individual components of a cash flow can be treated separately and the final result can be obtained by superposition. Care must be taken to move the individual components to the desired point in time using (1) the appropriate time separation by noting that the terms *present* and *future* refer to a *relative* time difference and (2) the correct present-worth and compound-amount factors. Equation 12.6.1 automatically takes care of these requirements. It also accounts properly for the situation where no payment exists at the end of one or more periods (e.g., the third period in this problem) because the corresponding term in the summation reduces to zero. The same equation may be used for discrete cash flows under continuous compounding. In that event the present-worth and compound-amount factors corresponding to continuous compounding must be used.

12.7 EQUAL SERIES OF PAYMENTS

A special cash-flow profile is a series of n equal payments as shown in Fig. 12.7.1(a). The first and last payments occur *at the end* of the first and last periods, respectively. The single-sum equivalent of the entire series at either time zero [i.e, the beginning of the first period; see Fig. 12.7.1(b)] or the end of the last period may be desired [Fig. 12.7.1(c)]. Conversely, a single sum at the beginning of the first period or the end of the last period may be converted to their equivalent series of equal payments [Fig. 12.7.1(a)]. Rather than having to apply Eq. 12.6.1 each time such a conversion is needed, four *equal-payment factors* have been developed. These are summarized in Table 12.7.1 along with the two single-sum factors that were discussed earlier. Note that the acronyms for the single-sum factors

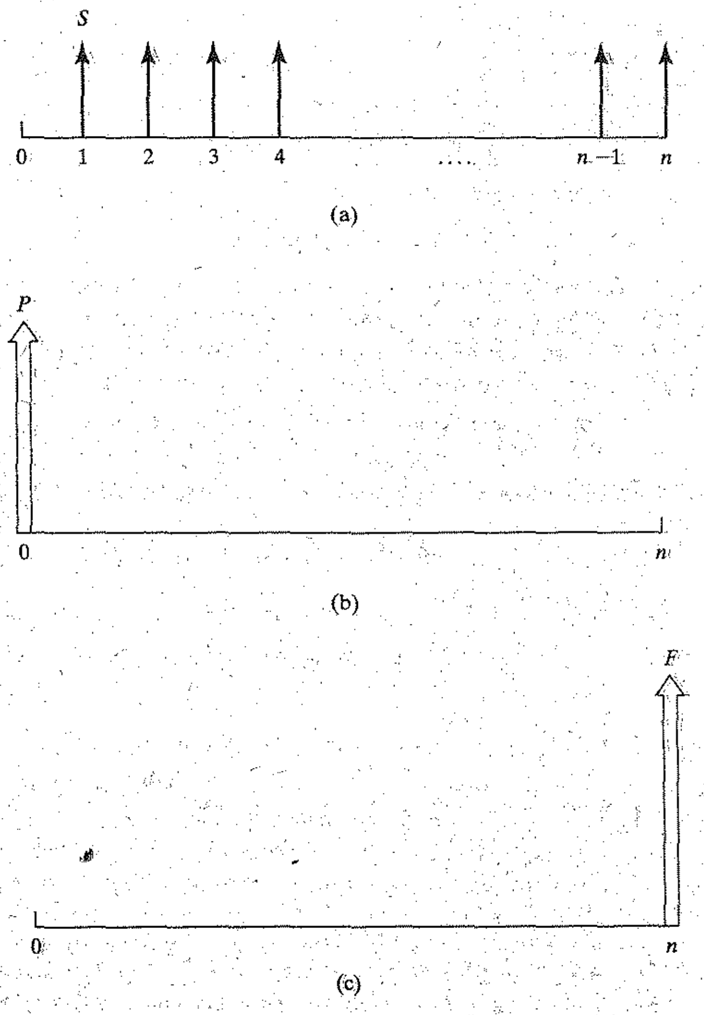


Figure 12.7.1 Equal series, present, and future single-payment equivalents.

TABLE 12.7.1 Discrete Compounding Factors[a]

Factor	Notation	Formula	Given	Find
1. Single-sum factors				
a. Compound-amount factor	(CAF′, i, n)	$(1+i)^n$	P	F
b. Present-worth factor	(PWF′, i, n)	$\dfrac{1}{(1+i)^n}$	F	P
2. Equal-series factors				
a. Compound-amount factor	(CAF, i, n)	$\dfrac{(1+i)^n-1}{i}$	S	F
b. Sinking fund factor	(SFF, i, n)	$\dfrac{i}{(1+i)^n-1}$	F	S
c. Present-worth factor	(PWF, i, n)	$\dfrac{(1+i)^n-1}{i(1+i)^n}$	S	P
d. Capital recovery factor	(CRF, i, n)	$\dfrac{i(1+i)^n}{(1+i)^n-1}$	P	S

[a] P, "present" single sum; F, "future" single sum; S, single sum in a series; n, number of periods.

have primes attached to distinguish them from the equal-payment factors of the same name.

As a mathematical illustration, the formulas for the equal-payment compound amount and sinking factors under discrete compounding are derived as follows:

Applying Eq. 12.6.1 to the series of equal payments of Fig. 12.7.1(a) with $h = 0$, $k = n$, $S_0 = 0$, and all other $S_j = S$, we obtain

$$F = S[(1+i)^{n-1} + \cdots + (1+i)^2 + (1+i) + 1]$$

Multiplying this equation by $(1 + i)$ and subtracting it from the result gives

$$F(1+i) - F = S[(1+i)^n - 1]$$

Solving for F yields

$$F = S\left[\frac{(1+i)^n - 1}{i}\right] \tag{12.7.1}$$

The bracketed term is the *equal-series compound-amount factor for discrete compounding* (CAF, i, n). Solving Eq. 12.7.1 for S, the corresponding *sinking-fund factor* (SFF, i, n) that converts a single sum F to an equal-payment series is seen to be the reciprocal of the (CAF, i, n). The rest of the equal-payment factors for discrete compounding as well as those corresponding to continuous compounding (see Table 12.7.1) can be derived in a similar manner. Also, useful relationships between the six factors for each compounding method may be reasoned out. For example, to find S given P, the latter is multiplied by the CRF. If P were to be multiplied by CAF′ to find F, and if this result were to be multiplied by SFF, the same equivalent equal-payment series would result. Therefore

$$\text{CRF} = (\text{CAF}')(\text{SFF}) \tag{12.7.2}$$

In fact all six factors can be expressed in terms of one single sum and one of the equal-payment factors. Moreover, the continuous-compounding factors can be derived by substituting the effective rate per period $i = e^r - 1$ into the discrete compounding factors.

Example 12.8

An automobile salesperson has offered the following terms to a customer who is interested in purchasing a $10,000 car: no down payment and 48 equal monthly payments, the first to be paid at the end of the first month after the purchase. Calculate the monthly payments and the equivalent single sum at the end of the 48-month period if the interest rate were 12% per year compounded monthly.

Solution The effective interest rate per month is 1%, $i = 0.01$, and $n = 48$. Hence the monthly payment S is

$$S = P(\text{CRF}, i, 48) = \$(10{,}000)(0.026338) = \$263.34 \text{ per month}$$

To find the equivalent single sum after 48 months, this equal-payment series can be converted using the CAF:

$$F = S(\text{CAF}, i, 48) = (263.38)(61.22258) = \$16{,}122.2\ 6$$

or the original single sum can be converted using the single-sum compound-amount factor:

$$F = P(\text{CAF}', i, 48) = (10{,}000)(1.612226) = \$16{,}122.\ 26$$

Moreover, the same result can be obtained by first determining the effective annual interest rate via Eq. 12.4.1 to be equal to 0.126825 and then applying the discretely compounded single-sum compound-amount factor with this rate and two periods (i.e., 48 months = 2 years).

Discussion In addition to the use of equal-payment factors, this example reviews several important principles. First, the interest period was matched with the payment period by converting the given monthly compounded annual nominal rate of 12% to a monthly effective rate of 1%. Second, three possible alternative methods of obtaining the single future sum were illustrated. The first two imply the following relationship between factors:

$$(\text{CRF})(\text{CAF}) = (\text{CAF}')$$

The third method by which F was obtained shows that the effective annual interest rate is larger than the quoted nominal annual rate (i.e., 12.6825% versus 12.00%).

12.8 SUPERPOSITION OF CASH FLOWS

A cash flow may be described as the superposition of its individual single-payment components. The same principle of superposition applies to the case of complex cash flows that can be decomposed into several simpler cash flows. In each case several alternative ways of decomposing and superposing the cash flow's parts are possible. It is advisable to contemplate these alternatives in order to discern the simplest way of solving the problem prior to undertaking any calculations. This principle is illustrated by the following example.

Example 12.9

Find the *present worth* (i.e., the equivalent single sum at time zero) of the cash flow shown in Fig. 12.8.1. The effective interest rate is 8% per period.

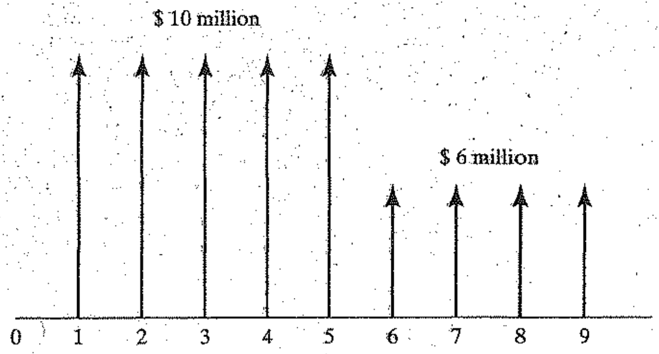


Figure 12.8.1 Example cash flow.

Solution The series shown may be decomposed into a simpler series in several ways. Three possibilities are:

1. A series of nine $6 million payments plus a series of five $4 million payments, both series beginning at the end of the first period.
2. A series of five $10 million payments beginning at the end of the first period plus a series of four $6 million payments beginning at the end of the fifth period.
3. A series of nine $10 million payments beginning at the end of the first period minus a series of four $4 million payments beginning at the end of the fifth period.

In this particular case the first method of decomposition seems to be the simpler of the three. However, for the purpose of illustration several solutions are attempted.

Using the first method of decomposition, the present worth of the original series is equal to

$$P = 6(\text{PWF}, i, 9) + 4(\text{PWF}, i, 5)$$
$$= 6(6.24689) + 4(3.99271) = \$53.45 \text{ million}$$

Using the second method of decomposition, the following two solutions are equivalent:

(a) $P = 10(\text{PWF}, i, 5) + 6(\text{PWF}, i, 4)(\text{PWF}', i, 5)$

$$= 10(3.99271) + 6(3.31213)(0.68058) = \$53.45 \text{ million}$$

or

(b) $P = 10(\text{PWF}, i, 5) + 6(\text{CAF}, i, 4)(\text{PWF}', i, 9)$

$$= 10(3.99271) + 6(4.50611)(0.50025) = \$53.45 \text{ million}$$

Using the third method of decomposition, the following two solutions are also equivalent:

(a) $P = 10(\text{PWF}, i, 9) - 4(\text{PWF}, i, 4)(\text{PWF}', i, 5)$

$$= 10(6.24689) - 4(3.31213)(0.68058) = \$53.45 \text{ million}$$

or

(b) $P = 10(\text{PWF}, i, 9) - 4(\text{CAF}, i, 4)(\text{PWF}', i, 9)$

$$= 10(6.24689) - 4(4.50611)(0.50025) = \$53.45 \text{ million}$$

EXERCISES

1. A school transportation company has purchased a new bus on the following terms: $20,000 down and a monthly payment of $2704 for 5 years at a nominal annual interest rate of 9%. What was the cash price?
2. Find the equivalent of the cash price of the bus described in Exercise 1 at the end of the 5-year period.
3. Using an annual interest rate $i = 8\%$, find the present worth of the cash flows given in Fig. E12.3.

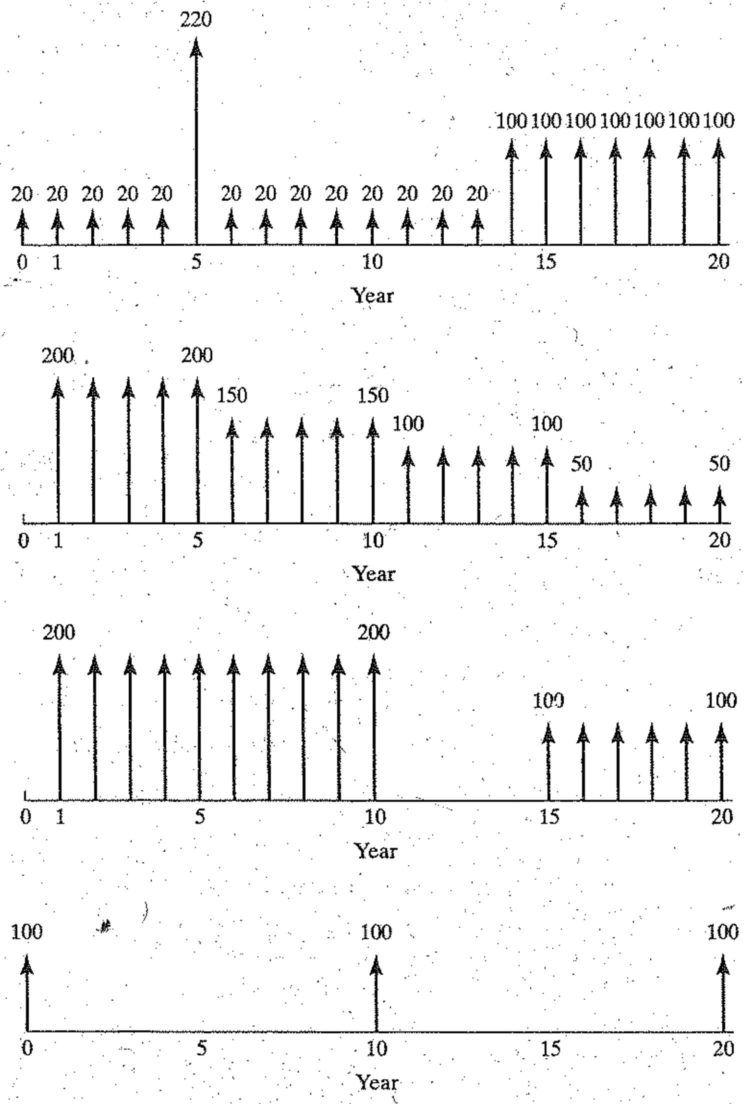


Figure E12.3

4. Find the worth of the cash flows of Exercise 3 at the end of the tenth year.

5. Referring to the cash flow shown in Fig. E12.5, which of the following equations are true?

$$P = 20 + [4(\text{CAF}, i, 5) + 4(\text{PWF}, i, 10)](\text{PWF}, i, 5)$$

$$P = 20 + 8(\text{PWF}, i, 15) - 4(\text{PWF}, i, 5)$$

$$P = 20 + [4(\text{CAF}, i, 15) + 4(\text{CAF}, i, 10)](\text{SFF}, i, 15)$$

$$P = 8(\text{CAF}', i, 15) + 4(\text{PWF}, i, 15)(\text{CAF}', i, 15) + 4(\text{CAF}, i, 10)$$

$$P = 20 + 4(\text{PWF}, i, 15) + 4(\text{CAF}, i, 10)(\text{PWF}', i, 15)$$

$$P = [20(\text{CAF}', i, 15) + 4(\text{CAF}, i, 15) + 4(\text{CAF}, i, 10)](\text{PWF}', i, 15)$$

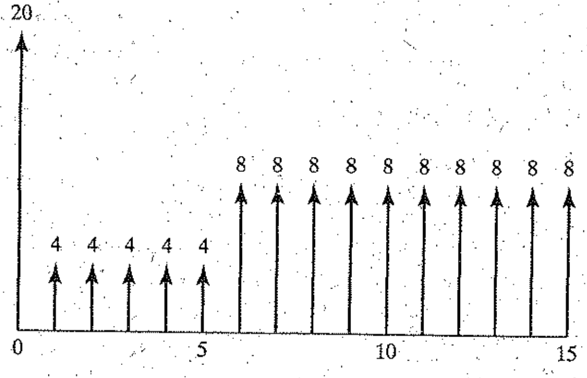


Figure E12.5

13

Probability and Statistics

13.1 INTRODUCTION

When the outcome of a situation or process can be known in advance with absolute certainty, the situation or process is said to be *deterministic.* It may be argued that in a cause-and-effect universe the outcome of every situation can be anticipated, assuming that all the factors affecting the situation are clearly understood. Knowing everything about something, however, requires knowing everything about everything else, and this is not a pragmatic claim. Engineering decisions are almost always based on limited information. Hence all situations entail some degree of uncertainty. When the degree of uncertainty is very low, the situation may be treated as if it were deterministic. Otherwise some way of incorporating uncertainty is needed. The *theory of probability* is the branch of mathematics that addresses this question. The theory had its origins in an attempt by Pascal to predict the likely outcome of interrupted gambling games to aid a group of friends in settling their bets and has since been applied to innumerable situations including the study of traffic systems.

The practical methods of analysis discussed in Chapter 4 are based on knowledge that has been obtained from observing the operation of many facilities, and they approximate the effects of the various factors that influence capacity by the use of empirically derived charts and tables. The percentage of the total approach volume that wishes to execute turning maneuvers, for example, is clearly a major factor affecting the operation of an intersection. However, it is reasonable to expect that not only the percentage, but also the pattern with which the left-turning vehicles arrive at the intersection would affect its operation. For example, turning vehicles arriving one after another during a short interval of time within the hour would result in different conditions from that which would result if they were spread out uniformly throughout the hour. Thus having the knowledge that 20%, for example, of the approach volume consists of left-turning vehicles and that 10% of the approaching vehicles plan to turn right does not specify unequivocably either the sequence in which turning vehicles appear or the movement desire of any one vehicle in the approaching stream; it simply quantifies the likelihood, or probability, with which each vehicle is expected to execute each maneuver. When a more detailed investigation of a system characterized by a high degree of variability is desired, methods that are explicitly based on the theory of probability are often employed.

13.2 ELEMENTS OF PROBABILITY THEORY

13.2.1 Background

The theory of probability is concerned with situations (conventionally referred to as *experiments*) that have many possible *outcomes*. For example, a vehicle approaching an intersection (experiment) may choose one of three possible movements (outcomes): It may go straight through, turn right, or turn left. Similarly, casting a single die has six possible outcomes: the numbers 1 through 6. Thus for each experiment there exists a set of possible outcomes collectively known as the *sample, outcome,* or *event space,* usually denoted by the uppercase Greek capital letter omega (Ω). An *event* is a subset of the sample space. It is *simple* if it consists of a single outcome and *compound* if it contains a combination of single outcomes. For example, when casting a die, "rolling a 2" is a simple event A, whereas "rolling an even number" is a compound event B such that

$$A = \{2\} \quad \text{and} \quad B = \{2,4,6\}$$

The following definitions from set theory are also relevant to the discussion of probability. The *complement* of an event A is denoted by $\bar{A}$ and is defined as the subset of Ω that contains all of the elements not belonging to A. Thus the complements of the events A and B are

$$\bar{A} = \{1,3,4,5,6\} \quad \text{and} \quad \bar{B} = \{1,3,5\}$$

The *empty* (or *null*) *set,* denoted by $\varnothing$, is a set that contains no elements. The *union* of two events A and B, denoted by $A \cup B$, is the set that contains the elements belonging to A or B or both. For example,

$$A \cup \mathrm{B} = \{2,4,6\} \quad \text{and} \quad A \cup \bar{\mathrm{B}} = \{1,2,3,5\}$$

The *intersection* of two events A and B, denoted by $A \cap B$, is the set that contains only the elements that the two events share in common. Thus

$$A \cap B = \{2\} \quad \text{and} \quad \bar{A} \cap B = \{4,6\}$$

When two events have no elements in common, they are *mutually exclusive.* Clearly an event and its complements are mutually exclusive. As just defined, pairs of simple events (i.e., single outcomes) are also mutually exclusive.

13.2.2 Definition of Probability

Probability is a measure of the likelihood with which events are expected to occur. Suppose we toss a coin 10 times and record the outcome of each trial as follows:

$$H\,H\,T\,T\,T\,H\,H\,T\,H\,H$$

The observed frequencies of heads (H) and tails (T) were 6 and 4 times, respectively. The *relative frequencies* of the two outcomes can be obtained by dividing the observed frequencies by the total number of trials. In this example

$$f(H) = 0.6 \quad \text{and} \quad f(T) = 0.4$$

Intuitively one would expect the two outcomes of tossing a fair coin to occur with equal frequencies. This could be the case when the number of trials is very large. Note that after only one trial this long-term expectation cannot possibly be satisfied as the relative frequencies would be 1 for one outcome (H in this case) and 0 for the other. The *limiting value* of the relative frequency of an event as the number of trials approaches infinity is defined as the *probability* of occurrence of that event on any one trial. The probability of each outcome may be derived intuitively, or it may be estimated by repeating the experiment a large number of times and computing the relative frequencies of the outcomes. One must keep in mind that when estimating the probabilities of events by experimentation, the chance always exists that the computed frequencies may deviate significantly from the theoretical probabilities.

It follows from this definition that the probability, $P[A]$, of an event A cannot be negative and cannot exceed 1. Also, the sum of the probabilities of all simple events contained in the sample space of an experiment always equals 1.

Other useful axioms of probability include the following. The probability of the event defined as the union of two events A and B is

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \tag{13.2.1}$$

The probability of the event defined by the intersection of A and B is subtracted from the sum of the probabilities of the two events to avoid double counting. Since two mutually exclusive events share no elements in common, the probability of the event defined by their union is equal to the sum of the probabilities of the two events. Also,

$$P[\Omega] = 1 \tag{13.2.2}$$

and

$$P[\varnothing] = 0 \tag{13.2.3}$$

Equation 13.2.2 states that the probability of the event defined as the union of all (mutually exclusive) simple events associated with an experiment is equal to 1. Thus on any trial *one* of the possible outcomes is certain to occur. Equation 13.2.3 is an alternative way of stating the same concept: It is impossible to obtain none of the single outcomes of an experiment on a given trial. Since the intersection of the two mutually exclusive events E and $\bar{E}$ is equal to the null set, it follows from Eqs. 13.2.1 through 13.2.3 that

$$P[E \cup \bar{E}] = P[\Omega] = P[E] + P[\bar{E}] - P[\varnothing] = 1$$

and

$$P[\bar{E}] = 1 - P[E] \tag{13.2.4}$$

Example 13.1

Consider the experiment of casting a single die. Identify all simple events (outcomes) of the experiment, and calculate the probability of each.

Solution The outcome space for this experiment contains six possible outcomes:

$$\{1,2,3,4,5,6\}$$

Assuming that the die is not loaded, the six simple events are equiprobable, each having a probability of occurrence on any trial of $\frac{1}{6}$. Note that all probabilities are nonnegative, and their sum equals unity. Moreover, the probability associated with any outcome that is not included in the sample space of the experiment (e.g., 9) is 0; that is, such an event is *impossible* given this experiment.

Example 13.2

For the experiment of Example 13.1, define two events A and B as

A: the outcome is odd and B: the outcome is less than 5

Find (a) the probability of each of the two compound events, (b) the probability of their union, and (c) the probability of their intersection.

Solution The two events are defined as

$$A = \{1,3,5\} \quad \text{and} \quad B = \{1,2,3,4\}$$

Events A and B contain three and four equiprobable (mutually exclusive) simple events, respectively, each having a probability of $\frac{1}{6}$. Therefore

$$P[A] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

and

$$P[B] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$$

The union of the two events, $\{1,2,3,4,5\}$, contains five equiprobable events. Hence $P[A \cup B] = \frac{5}{6}$. Their intersection contains only the elements 1 and 3. So $P[A \cap B] = \frac{2}{6} = \frac{1}{3}$.

Discussion According to Eq. 13.2.1, the probability of the event defined as the union of A and B is less than the sum of their individual probabilities by an amount that equals the probability of their intersection. Thus

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[A \cap B] \\ &= \frac{3}{6} + \frac{4}{6} - \frac{2}{6} = \frac{5}{6} \end{aligned}$$

which is the same result obtained by enumerating the simple events that belong to the union of A and B.

13.2.3 Conditional Probability and Independence

The estimation of the probability of event A is often affected by prior knowledge that event B has occurred because the outcome is restricted to the subset of the sample space B. The *conditional probability* of A given B is

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} \tag{13.2.5}$$

If events A and B are mutually exclusive, their intersection is the null set, which has a probability of 0. Hence in the case of mutually exclusive events the probability that A has occurred given that B has taken place is 0. In other words the occurrence of B excludes the occurrence of A.

When the occurrence of B does not alter the probability associated with A, the two events are known as *independent events.* For independent events

$$P[A \mid B] = P[A] \tag{13.2.6a}$$

Substituting Eq. 13.2.6a into Eq. 13.2.5 and solving for the probability of the intersection of A and B, we obtain

$$P[A \cap B] = P[A]P[B] \tag{13.2.6b}$$

Thus for *independent events* the probability of joint occurrence is equal to the product of the probabilities of the individual events.

Example 13.3

For the two events defined in Example 13.2, calculate the conditional probability $P[A \mid B]$.

Solution Substitution of the probabilities obtained in the earlier example into Eq. 13.2.5 yields

$$P[A \mid B] = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

Discussion The calculated conditional probability for event A is equal to the unconditional probability of the same event. Therefore the knowledge that event B has occurred does not affect the probability of event A. Hence the two events are independent.

Example 13.4

If the probability that a vehicle approaching an intersection will turn left is $P[L] = 0.25$ and the probability that it will turn right is $P[R] = 0.15$, calculate the probability that the vehicle will turn right given that it does not turn left.

Solution Event B (i.e., "the vehicle does not turn left") is the complement of event L. According to Eq. 13.2.4, $P[B] = 1 - P[L] = 0.75$. Compound event B contains the simple events of "turning right" (R) and "going through" (T), which are mutually exclusive. Hence the probability of the intersection of events B and R is simply equal to the probability of R, or 0.15. The probability that the vehicle will turn right given that it does not turn left is

$$P[R \mid B] = \frac{P[R \cap B]}{P[B]} = \frac{(0.15)}{(0.75)} = 0.20$$

Discussion Since $P[R \mid B]$ does not equal $P[R]$, events R and B are not independent. The results just obtained could have been reached by intuitive reasoning. The likelihood that the vehicle will turn right increases when it is known that it will not turn left. The solution, however, applies the formal relationships to illustrate the concept of conditional probability.

13.2.4 Discrete Distributions

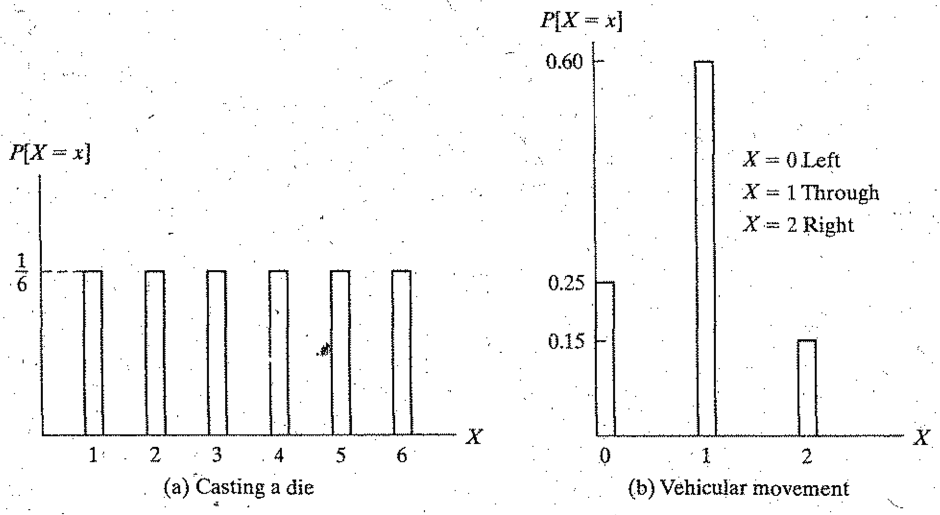
A *random variable* is a variable that takes on the values of the outcomes of an experiment. When the number of possible outcomes is finitely or infinitely countable, the random variable is said to be *discrete*. Otherwise the random variable is *continuous*. Examples of discrete variables include household vehicle ownership that takes the values 0, 1, 2 and so on, and binary (0, 1 or yes, no) or switch variables (e.g., 1 if a trip is oriented toward the CBD and 0 otherwise). Examples of continuous random variables include the time between failures of a machine and the distance from a reference point on a highway where the next accident will occur. In this textbook random variables are denoted by uppercase letters (e.g., X, Y, or Z) and the particular values that they assume are denoted by lowercase letters (e.g., x, y, z). Thus $X = x$ means the random variables X takes on a particular value x.

A function $p(x) = P[X = x]$, which associates each value of a discrete random variable to its probability, is known as the *probability (mass) function* (pmf), or the *discrete probability distribution*. The histogram of Fig. 13.2.1(a) illustrates graphically the probability function associated with casting a die, and Fig. 13.2.1(b) illustrates the probability function of the vehicle described in Example 13.4. In the case of the former the values assumed by the random variable are identical to the numerical values of the outcomes of the experiment and can be given that interpretation. In the case of the latter the numerical values assigned to the random variable are arbitrary and selected merely for convenience. Since the values of $p(x)$ represent the probabilities of simple events, they must satisfy the following conditions:

$$0 \leq p(x) \leq 1 \quad \text{for all } x \tag{13.2.7a}$$

$$\Sigma\, p(x) = 1 \tag{13.2.7b}$$

The probability of the union of several (mutually exclusive) outcomes is equal to the sum of the probabilities of the individual outcomes. The condition of Eq. 13.2.7b simply states that the sum of the probabilities of all single outcomes in the sample space of the experiment is equal to unity. This is merely a restatement of Eq. 13.2.2.



(a) Casting a die

(b) Vehicular movement

Figure 13.2.1 Discrete probability functions.

Another useful function of a discrete random variable is its *cumulative distribution function* (cdf), which is defined as

$$F(x) = P[X \leqslant x] \tag{13.2.8}$$

In words, the cdf is a function that assumes the values of the probability that a random variable X is less than or equal to a particular value x. Because $p(x)$ is nonnegative, a plot of the cdf against increasing values of x must necessarily be a nondecreasing step function, as illustrated in Fig. 13.2.2. Moreover, the lower and upper limits of this function are 0 and 1, respectively. The upper limit corresponds to the condition described by Eq. 13.2.7b.

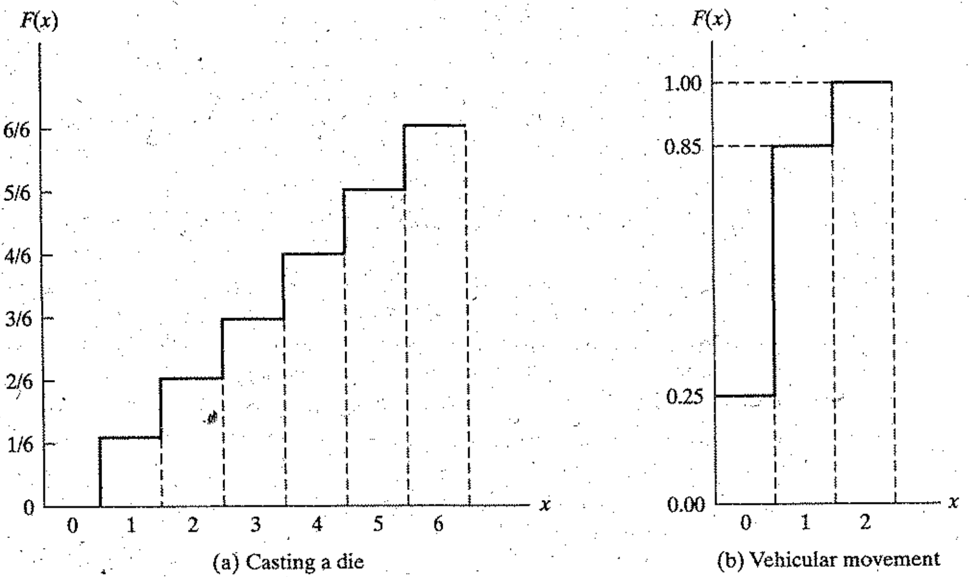
Two characteristics of a probability function are its *mean* (or *expected value*) and its *variance*. The mean is a measure of the central tendency or the average value of the distribution, and the variance is a measure of dispersion or the degree to which the distribution is spread out around the mean.

The mean is usually denoted by the Greek lowercase letter mu (μ) or by $E[X]$, the latter read "expected value." The mean is calculated by

$$E[X] = \sum_x x\,p(x) \tag{13.2.9}$$

The variance of a discrete distribution, denoted by σ^2 or by $V[X]$, is defined as the second moment about the mean, or

$$V[X] = \sum_x (X - E[X])^2\,p(x) \tag{13.2.10}$$



(a) Casting a die

(b) Vehicular movement

Figure 13.2.2 Discrete cumulative functions.

The square root of the variance is known as the *standard deviation* of the distribution, that is,

$$\sigma = S[X] = V[X]^{1/2} \tag{13.2.11}$$

Example 13.5

Calculate the mean, variance, and standard deviation of the discrete distributions of Fig. 13.2.1.

Solution (a) By Eq. 13.2.9, the mean of the distribution shown on Fig. 13.2.1(a) is

$$E[X] = (1)\left(\frac{1}{6}\right) + (2)\left(\frac{1}{6}\right) + \cdots + (6)\left(\frac{1}{6}\right) = \frac{21}{6} = 3.5$$

The variance is given by Eq. 13.2.10:

$$V[X] = (1 - 3.5)^2\left(\frac{1}{6}\right) + (2 - 3.5)^2\left(\frac{1}{6}\right) + \cdots + (6 - 3.5)^2\left(\frac{1}{6}\right) = 2.92$$

The standard deviation is the square root of the variance, or

$$S[X] = 1.71$$

(b) Similarly, for the distribution of Fig. 13.2.1(b)

$$\begin{aligned} E[X] &= (0)(0.20) + (1)(0.60) + (2)(0.15) = 0.9 \\ V[X] &= (0 - 0.9)^2(0.20) + (1 - 0.9)^2(0.60) \\ &\quad + (2 - 0.9)^2(0.15) = 0.35 \\ S[X] &= 0.6 \end{aligned}$$

Discussion The mean of the distribution does not necessarily coincide with one of the outcomes of the experiment. It simply represents the average of a series of trials as the number of trials approaches infinity. In part (a) of this problem the average represents a value that is meaningful in terms of the magnitudes of the outputs of the experiment. By contrast, the mean value of part (b) depends on the numerical codes selected for the three outcomes. The variance and standard deviation measure the dispersion of each distribution about its mean value.

13.2.5 Some Common Discrete Distributions

Any discrete function that satisfies the conditions of Eqs. 13.2.7 can conceivably be the distribution for some experiment. This subsection presents a number of discrete distributions that are frequently encountered in practice, the characteristics of which are well known. The work of the analyst can be simplified when the problem at hand fits the specifications of one of these common distributions. Formal mathematical derivations for these functions may be found in the technical literature (e.g., Ref. [13.1]).

The *uniform distribution* describes experiments that have a finite number of N equiprobable outcomes. The casting of a fair die and the tossing of a fair coin are but two examples of this distribution. In general,

$$p(x) = P[X = x] = \frac{1}{N} \quad \text{for all } x \text{ in } \Omega \tag{13.2.12}$$

The mean and variance of this distribution can be calculated by Eqs. 13.2.9 and 13.2.10. In the special case when X takes on the values $x = 1, 2, \ldots, N$ the following apply:

$$E[X] = \frac{N+1}{2} \quad \text{and} \quad V[X] = \frac{N^2-1}{12} \tag{13.2.13}$$

The *Bernoulli distribution* applies to experiments that have only two outcomes, often referred to as a "success" S and a "failure" F. Tossing a coin, fair or unfair, is a Bernoulli trial. If on any trial the probability of success is p and the probability of failure is q, by Eq. 13.2.7

$$P[F] = q = 1 - P[S] = 1 - p \tag{13.2.14}$$

in the special case when a success is coded as 1 and a failure as 0, the mean and variance of the Bernoulli distribution become

$$E[X] = 1p + 0q = p \tag{13.2.15a}$$

and

$$V[X] = pq = p(1-p) \tag{13.2.15b}$$

The *binomial distribution* expresses the probability of x successes in a sequence of n independent Bernoulli trials. The random variable takes on the values $x = 0, 1, 2, \ldots, n$. The binomial probability function is

$$p(x) = P[X = x] = \frac{n!}{x!(n-x)!}p^x q^{n-x} \tag{13.2.16}$$

The mean and variance of the binomial distribution are

$$E[X] = np \tag{13.2.17a}$$

and

$$V[X] = npq \tag{13.2.17b}$$

The Bernoulli distribution is a particular case of the binomial when $n = 1$.

The *geometric distribution* is also based on a sequence of independent Bernoulli trials. It represents the probability that the first success will occur on the xth trial. This means that A: the first $(x - 1)$ independent trials result in a failure and B: the last trial is a success. Considering that events A and B are, by definition, independent, Eq. 13.2.6b yields the following probability that the first success will occur on the xth trial:

$$p(x) = q^{x-1}p \qquad x = 1, 2, \ldots \tag{13.2.18}$$

This is an example of a discrete random variable X that takes on an *infinitely countable* number of values. The mean and variance of the geometric distribution are

$$E[X] = p^{-1} \tag{13.2.19a}$$

and

$$V[X] = qp^{-2} \tag{13.2.19b}$$

The *negative binomial* (or *Pascal*) *distribution* measures the probability that the kth success will occur on the xth trial of a Bernoulli process. For this to be true two events must occur, A: there must be $(k-1)$ successes in the first $(x-1)$ trials *and* B: the last trial must be a success. The probability of event A is given by the binomial (Eq. 13.2.16) and the probability of event B is p. Since the two events are independent, the probability that the kth success will occur on the xth trial becomes

$$p(x) = \frac{(x-1)!}{(k-1)!(x-k)!} p^k q^{x-k} \qquad \text{for } x = k, k+1, \ldots \tag{13.2.20}$$

The mean and variance of the negative binomial are

$$E[X] = kp^{-1} \tag{13.2.21a}$$

and

$$V[X] = kqp^{-2} \tag{13.2.21b}$$

The geometric distribution is a special case of the Pascal distribution when $k = 1$.

Finally, a discrete distribution that has found wide application in traffic situations is the *Poisson distribution.* It describes the probability of x occurrences of an event (successes) within a given interval of time (or space) t and applies to experiments that satisfy the following conditions:

1. There exists a small interval dt within which the probability of one occurrence is $\lambda\, dt$, whereas the probability of additional occurrences is negligible.
2. The occurrences (or nonoccurrences) of the event in nonoverlapping intervals are mutually *independent.*

The total interval t may be thought to consist of a sequence of small intervals dt, each representing an independent Bernoulli trial with $p = \lambda\, dt$. In other words the probability of x successes within the total interval t is given by the binomial. In the limit, that is, when the number of small intervals is very large and the probability of success p is small, the Poisson distribution is obtained:

$$p(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \qquad \text{for } x = 0, 1, \ldots \tag{13.2.22}$$

The mean and variance of the Poisson distribution are

$$E[X] = V[X] = \lambda t \tag{13.2.23}$$

Example 13.6

The probability that a vehicle will turn left at an intersection is known to be 0.15. Assuming independence, calculate the probabilities of the following events:

(a) The tenth vehicle is not turning left.
(b) Exactly three out of ten vehicles will turn left.
(c) At least three out of ten vehicles will turn left.
(d) No more than three out of ten vehicles will turn left.

(e) The first left-turning vehicle will be the fourth vehicle.
(f) The eighth vehicle will be the third to turn left.

Solution The movement of each vehicle is a Bernoulli trial with $p = 0.15$ and $q = 0.85$.

(a) The probability that the tenth vehicle, and any other vehicle, will not turn left is $q = 0.85$ (Bernoulli distribution).

(b) According to the binomial distribution, the probability that three out of ten vehicles will turn left is

$$p(3) = 0.130$$

(c) The complement of event A: at least three out of ten is event B: zero or one or two out of ten and

$$P[A] = P[X > 2] = 1 - P[B] \qquad \text{for } n = 10$$

But B is the union of three mutually exclusive simple events, the probability of which is given by the binomial. Hence

$$\begin{aligned} P[A] &= 1 - \{p(0) + p(1) + p(2)\} \\ &= 1 - (0.197 + 0.347 + 0.276) = 0.180 \end{aligned}$$

(d) The binomial still applies. The probability of the compound event A: no more than three in ten is

$$\begin{aligned} P(A) &= P[X < 4] = p(0) + p(1) + p(2) + p(3) \\ &= 0.197 + 0.347 + 0.276 + 0.130 = 0.950 \end{aligned}$$

(e) This question may be answered by using either the geometric distribution or the negative binomial (Pascal) distribution with $k = 1$. Thus

$$p(4) = (0.85)^3(0.15) = 0.092$$

(f) The Pascal distribution provides the answer to the question of the probability that the kth (i.e., third) left turner is the eighth vehicle, x, is the sequence

$$p(8) = 0.031 \qquad \text{for } k = 3$$

Discussion The event of which the probability is being sought must be clearly understood. For example, the difference between "exactly three," "at least three," and "no more than three" was illustrated in parts (a), (b), and (c). Also, the appropriate distribution must be selected. It should be remembered that the notation $p(x)$ has a different meaning depending on the particular distribution used. Thus in the case of the binomial $p(4)$ represents "the probability of four occurrences in n trials," and in the case of the geometric, it represents "the probability that the first occurrence is on the fourth trial."

Example 13.7

It is known that 8% of the drivers in a resort town drive under the influence of alcohol. Assuming independence, calculate the probability of the following events:

(a) No driver in five stopped is under the influence.
(b) Exactly 10 out of 100 stopped are drunk.
(c) Fifty out of 500 are under the influence.

Solution Considering the discovery of a drunken driver to be a "success," each act of stopping a driver is a Bernoulli trial with $p = 0.08$.

(a) The probability of finding no driver to be under the influence of alcohol in five independent Bernoulli trials is the product $q^5 = 0.659$ (see Eq. 13.2.6b). The same result can be obtained by applying the binomial distribution with zero successes in five trials.

(b) In this case the binomial yields a probability of discovering exactly ten drunken drivers:

$$p(10) = 0.102 \qquad \text{when } n = 100$$

Since p is small and n is large, the Poisson with a mean $np = 8$ (Eq. 13.2.17a) may be applied to approximate the binomial distribution; that is,

$$p(10) = P[X = 10] = 0.099 \qquad \text{when } n = 100$$

(c) Repeating the procedure of part (b), the binomial distribution yields $p(50) = 0.0167$, and the Poisson approximation (with mean 40) results in $p(50) = 0.0177$.

Discussion The Poisson approximation to a binomial improves as the number of trials increases. In this particular case the same conclusion is reached by comparing the mean and variance of the binomial distribution. For part (b) the mean is 8 and the variance is $npq = 7.36$. For part (c) the mean and variance are 40 and 36.8, respectively.

Example 13.8

Cars arrive at a parking garage at a rate of 90 veh/h according to the Poisson distribution. Compute the cumulative distribution for the random variable X that represents "the number of arrivals per minute."

Solution The mean arrival rate is 1.5 veh/min. Hence Eq. 13.2.22 becomes

$$p(x) = P[X = x] = \frac{(1.5)^x}{x!} e^{-1.5} \qquad x = 0, 1, 2, \ldots$$

The cumulative distribution $F(x) = P[X \leq x]$ is obtained by summing the probabilities of the simple events $0, 1, \ldots, x$. The results are:

x	$p(x) = P[X = x]$	$F(x) = P[X \leq x]$
0	0.223	0.223
1	0.335	0.558
2	0.251	0.809
3	0.126	0.935
4	0.047	0.982
5	0.014	0.996
—	—	—
—	—	—
—	—	—

Discussion Properly, the cumulative distribution approaches 1 as x approaches infinity. According to the Poisson distribution, the random variable X is defined up to this limit. In practical situations, however, X is usually bounded at some lower value. In the situation examined here, for example, the number of cars arriving at the garage within a minute cannot be very large. Thus, at best, the Poisson distribution can serve only as an approximation to this situation. Noting, however, that the calculated probability of more than five arrivals per minute is only 0.004, it is not unreasonable to expect that this approximation would be satisfactory.

13.2.6 Continuous Random Variables

The random variables discussed in the preceding section were allowed to assume a countable number of values. On the other hand, many situations are characterized by an uncountable number of possible outcomes that can be described only by continuous random variables. For example, consider the measurement of the headways between persons as they enter a hall. Conceivably the headway between any two persons can vary from zero (when they enter simultaneously) to infinity (following the last person *ever* to enter the hall). Assuming that the headways can be measured with absolute precision, the probability of obtaining *exactly* any particular value is zero. In the continuous case probability is associated with *ranges of the outcome* rather than with single values. The following definitions clarify this statement.

The range over which a continuous random variable is defined (e.g., zero to infinity in the example relating to headways) can be divided into infinitesimal intervals, dx. If the probability that the outcome of an experiment will fall within dx is equal to the area $f(x)\,dx$, then the function $f(x)$ is defined as the *probability density function* (pdf) or the *continuous probability distribution* for that experiment.

Since the probability of the occurrence of any event is nonnegative, a pdf is also nonnegative. Moreover, according to Eq. 13.2.2, the sum of the probabilities of the (mutually exclusive) events defined by each dx over the range of the random variable X must necessarily equal unity. In mathematical terms

$$f(x) \geqslant 0 \quad \text{for all } x \tag{13.2.24a}$$

and

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 \tag{13.2.24b}$$

These conditions correspond to the requirements that must be satisfied by discrete probability distributions expressed by Eqs. 13.2.7. The area under the pdf between points a and b is equal to the probability that the outcome will fall in the interval $\{a,b\}$, or

$$\int_{a}^{b} f(x)\,dx = P\,[a \leqslant X \leqslant b] \tag{13.2.25}$$

When $a = b$, the area under the curve is zero. Hence the probability associated with any single value of the random variable is zero.

The mean, the variance, and the standard deviation of a pdf are defined as

$$E[X] = \int_{-\infty}^{+\infty} xf(x)\,dx \tag{13.2.26}$$

$$V[X] = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x)\,dx \tag{13.2.27}$$

and

$$S[X] = \{V[X]\}^{1/2} \tag{13.2.28}$$

Equation 13.2.26 describes the first moment of the area under the curve about the y-axis. Considering that this area equals unity, the expected value is the x-coordinate of its centroid. Equation 13.2.27 describes the second moment of the area about the centroidal y-axis.

The *cumulative distribution function* (cdf) of a continuous random variable is defined as

$$P[X \leqslant x] = \int_{-\infty}^{x} f(x)\,dx \tag{13.2.29}$$

The cdf is a nondecreasing function with a lower limit of 0 and an upper limit of 1. In this respect it is identical to the cumulative distribution function of a discrete random variable.

13.2.7 Some Common Continuous Distributions

The *uniform distribution* is defined as

$$f(x)\begin{cases} \dfrac{1}{b-a} & \text{for } a \leqslant x \leqslant b \\ 0 & \text{otherwise} \end{cases} \tag{13.2.30}$$

In words, it consists of a horizontal line over the range $\{a,b\}$ in such a way that the area under the curve is equal to 1. The mean and variance of the uniform pdf are

$$E[X] = \frac{a+b}{2} \quad \text{and} \quad [X] = \frac{(b-a)^2}{12} \tag{13.2.31}$$

The *(negative) exponential distribution* bears a special relationship to the discrete Poisson distribution. When the occurrence of an event follows the Poisson distribution, the interval between occurrences is distributed according to the negative exponential. For example, if vehicles arrive at an intersection according to the Poisson distribution, the interarrival times (i.e., the headways between successive vehicles) are exponentially distributed. The negative exponential pdf is

$$f(x) = ae^{-ax} \tag{13.2.32}$$

where e is the base of natural logarithms. The mean and variance of the negative exponential are

$$E[X] = \frac{1}{a} \quad \text{and} \quad V[X] = \frac{1}{a^2} \tag{13.2.33}$$

The mean of this distribution is equal to its standard deviation. Moreover, it is the reciprocal of the mean of the Poisson distribution.

Numerous distributions have been used to describe vehicular headways and other traffic phenomena (e.g., Ref. [13.2]). Those just discussed will be sufficient for the purposes of this book.

Example 13.9

Given the arrival pattern of Example 13.8, calculate (a) the mean headway, (b) the probability that a headway is less than or equal to 45 s, and (c) the probability of headways longer than 2 min.

Solution (a) The average, or mean, headway is the reciprocal of the average number of arrivals per unit time. The latter was calculated in Example 13.8 to be 1.5 veh/min. Therefore

$$E[X] = \frac{1}{a} = \frac{1}{1.5} \text{ min (per vehicle)}$$

(b) This probability is given by the integral of the pdf from 0 to 0.75 min, or the value of the cdf at $x = 0.75$ min. Hence

$$P[X \leq 0.75] = F(0.75) = 1 - e^{-(1.5)(0.75)} = 0.675$$

(c) $P[X > 2] = 1 - P[X \leq 2] = e^{-3} = 0.050$

Example 13.10

Given that the instantaneous location of vehicles along a highway is Poisson distributed and that the average concentration is 100 veh/mi, calculate (a) the probability that 30 vehicles will be found on any quarter of a mile and (b) the probability that the spacing between any two vehicles is less than or equal to 0.02 mi.

Solution (a) The average number of vehicles per quarter mile is 25 and the probability that exactly 30 vehicles occupy a quarter mile, according to the Poisson distribution, is

$$p(30) = P[X = 30] = 0.045$$

(b) If the location of the vehicles is Poisson distributed, their spacing is exponentially distributed. With an average spacing of 0.01 mi (i.e., $a = 100$),

$$F(0.02) = P[X \leq 0.02] = 1 - e^{-(100)(0.02)} = 0.865$$

Discussion The relationship between the negative exponential and the Poisson distributions is clearly illustrated by the last two examples. Both are "memoryless." This property implies that the occurrences of events described by the Poisson are mutually independent. Similarly, the intervals between events described by the negative exponential are mutually independent. Specifically, when the negative exponential is used to describe the time interval between successive occurrences, the memoryless property means that occurrences in the future are not influenced by what has happened in the past. In the case of vehicular headways the negative exponential is most appropriate in the case of low concentration conditions, when the interactions between vehicles are at a minimum and vehicular events occur at random. When used to describe the headways between vehicular arrivals at an intersection, it seems reasonable that the intersection should be isolated, or removed, from other intersections, which, because of the signal control regularity, may impart a definite pattern to the arrivals at the intersection under study.

Perhaps the best-known probability density function is the *normal* (or *Gaussian*) *distribution.* Many natural phenomena tend to be approximated by this distribution. Moreover, during the eighteenth century it was discovered that measurement errors tended to follow the "bell-shaped" symmetrical pattern of the normal distribution. Consequently this distribution played a very important role in the development of statistical theory.

The equation of the normal distribution is given in terms of its mean μ and standard deviation σ as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \text{for } -\infty < x < +\infty \qquad (13.2.34)$$

Figure 13.2.3 shows that the normal distribution is centered about the mean, whereas the spread of the distribution depends on the standard deviation. The notation $N[\mu, \sigma]$ is used often to designate a normal distribution.

The cumulative normal distribution is given by Eq. 13.2.29. However, in the case of the normal distribution this equation cannot be integrated in closed form. Integration is accomplished through the use of *standard normal distribution* tables and a change of variables, as follows.

It is well known that if a variable x is distributed as $N(\mu, \sigma)$, the variable

$$z = \frac{x - \mu}{\sigma} \qquad (13.2.35)$$

is distributed as $N[0, 1]$. The cumulative $N[0, 1]$ distribution is given in Table 13.2.1. The value of z is read to two significant figures (first column followed by first row). Thus the cumulative distribution for $z = 2.05$ is 0.9798.

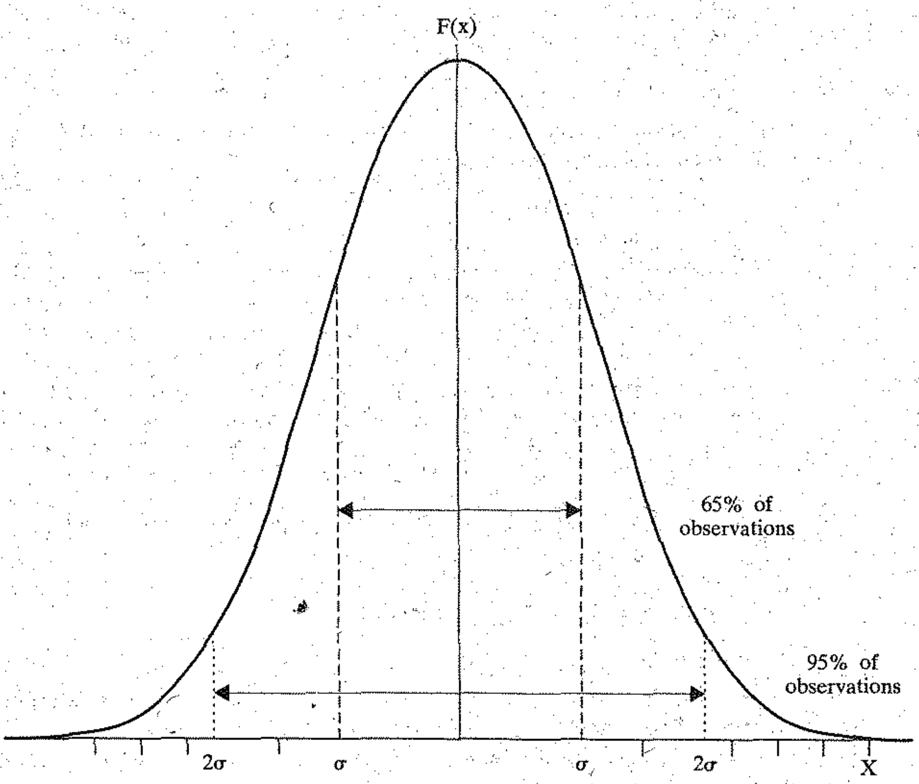


Figure 13.2.3 Normal distribution.

TABLE 13.2.1 Standard Normal Cumulative Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

From Irwin Miller and John E. Freund, *Probability and Statistics for Engineers*, 2nd ed., © 1977, p. 487. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, NJ.

Example 13.11

The scores of students taking a national examination follow the normal distribution with mean 500 and standard deviation 100.

(a) Calculate the proportion of scores above 643.

(b) Calculate the minimum score that places a student in the top 5%.

(c) Calculate the proportion of students scoring between 400 and 700.

Solution. According to Eq. 13.2.35, $z = (x - 500)/100$ is distributed as $N[1, 0]$.

(a) For $x = 643$, $z = 1.43$, consulting Table 13.2.1, the probability that z is less than or equal to 1.43 (another way of stating "x is less than or equal to 643") is 0.9222. Therefore the probability that z is greater than 1.43 is given as $1 - 0.9222 = 0.0764$; that is, 7.64% of the examinees are expected to score above 643.

(b) $F(z) = 0.95$ lies between $z = 1.64$ and $z = 1.65$. The corresponding values of x are 664 and 665, respectively. Conservatively, a grade above 665 ensures that the student is in the top 5%.

(c) For $x = 400$, $z = -1.00$ and for $x = 700$, $z = 2.00$. Note that Table 13.2.1 contains only positive values of z. However, because of the symmetry of the normal distribution, the area $F(-z) = 1 - F(z)$. As a result, $F(-1.00) = 1 - F(1.00) = 1 - 0.8413 = 0.1587$. The area $F(2.00) = 0.9772$. Consequently the area between $z = -1.00$ (corresponding to $x = 400$) and $z = 2.00$ (corresponding to $x = 700$) is $F(2.00) - F(-1.00) = 0.8185$. This means that 81.85% of the examinees are expected to score between 400 and 700.

13.3 EXPERIMENTAL DATA AND MODEL PARAMETERS

In many engineering and scientific applications relationships between variables are established by conducting experimental studies in either the laboratory or the field. The data collected in this manner may be plotted and the relationship between them discerned. Figure 13.3.1 represents a plot (known as a *scatter diagram* or *scatterplot*) of such observations, each described in terms of a pair of values X and Y that resulted from an experiment. The two variables may represent the stress and strain of steel samples, the speed and concentration of a traffic stream, or city population and volume of long-distance telephone calls. Because of experimental and other errors of measurement, the points shown on the scatter diagram will not fall precisely on a smooth curve. For this reason the task of the analyst becomes threefold: to hypothesize the mathematical form of the relationship between the two variables (*model postulation*), to estimate the parameters of the model based on the experimental data (*model calibration*), and to determine how well the calibrated relationship explains the observed data (*goodness of fit*).

One method of deriving the relationship between the two variables X and Y plotted in Fig. 13.3.1 is freehand approximation. However, the resulting relationship between the variables as well as the assessment of the goodness of fit will be highly subjective. For this reason a well-defined and rigorous technique of curve fitting is usually preferred. The *method of least squares* is a technique that yields the best-fitting line of a postulated form to a set of data. For example, the following are two possible mathematical forms that may be postulated in the case of a relationship involving two variables Y and X:

$$Y = a + bX \tag{13.3.1a}$$

$$Y = c + dX + eX^2 \tag{13.3.1b}$$

where X is the independent, or explanatory, variable, Y is the dependent, or explained, variable, and the constant coefficients are the model parameters. Because a set of paired values of (X_i, Y_i) are the known results of an experiment, *calibrating the model means determining the unknown values of the parameters that fix the postulated equation to the one that best fits the data.*

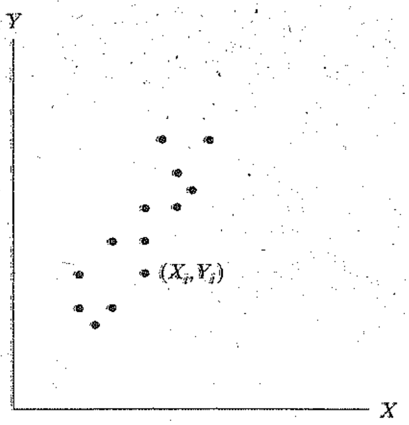


Figure 13.3.1 Scatter diagram.

13.4 LINEAR AND NONLINEAR REGRESSION

The method of least squares determines the numerical values of the coefficients that minimize the sum of square deviations between the observed values of the dependent variable Y_i and the estimated values $\hat{Y}_i$ that would be obtained by applying the calibrated relationship.

13.4.1 Simple Linear Regression

Consider the scatter diagram of Fig. 13.4.1. If it can be assumed that the relationship between X and Y is linear, then the method of least squares linear regression can be used to find the one straight line that best fits the data shown. An infinite number of straight lines can be drawn through the scatter diagram, each having its unique pair of parameters, that is, the Y-intercept a and the slope b. Hence the problem reduces to finding those values of a and b that define the best-fitting straight line. This line will then be used to describe the relationship between X and Y as

$$Y = a + bX \tag{13.4.1}$$

Considering the ith observation shown in Fig. 13.4.1, a difference exists between the observed value of Y_i corresponding to X_i and the estimated value of Y that would be obtained by substituting X_i in Eq. 13.4.1. The estimated value of Y is denoted by $\hat{Y}_i$ to distinguish it from the observed value Y_i. The difference between the two is known as the *error, deviation*, or *residual*. The straight line that minimizes some measure of the sum of all such deviations would appear to be the best-fitting straight line. In order to weigh equally the positive and negative deviations (in other words in order to ensure that the straight line passes *through* the scatter diagram), the deviations are squared and their sum is minimized; that is, the specific values of a and b are selected in such a way as to minimize the sum of square deviations.

In mathematical terms, find the values of a and b that minimize the sum:

$$S = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \tag{13.4.2}$$

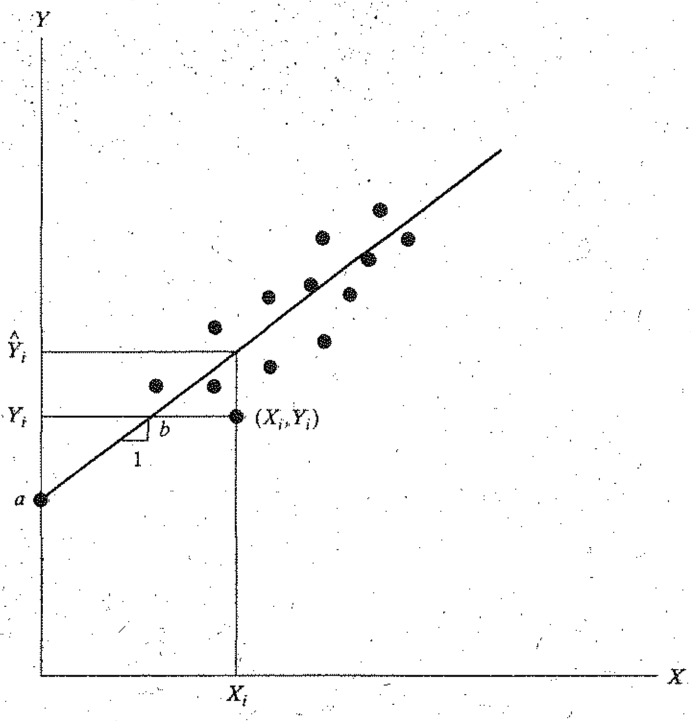


Figure 13.4.1 Linear regression.

To minimize S with respect to a and b, Eq. 13.4.2 must be expressed in terms of these two parameters. This is accomplished by substituting Eq. 13.4.1 in Eq. 13.4.2:

$$S = \sum_{i=1}^{N} (Y_i - a - bX_i)^2 \tag{13.4.3}$$

Setting the partial derivatives of S with respect to a and b equal to zero, we have

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{N} [2(Y_i - a - bX_i)(-1)] = 0 \tag{13.4.4a}$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{N} [2(Y_i - a - bX_i)(-X_i)] = 0 \tag{13.4.4b}$$

Dividing by 2 and rearranging terms gives

$$Na + \left(\sum_{i=1}^{N} X_i\right)b = \sum_{i=1}^{N} Y_i \tag{13.4.5a}$$

$$\left(\sum_{i=1}^{N} X_i\right)a + \left(\sum_{i=1}^{N} X_i^2\right)b = \left(\sum_{i=1}^{N} X_iY_i\right) \tag{13.4.5b}$$

These two equations, known as the *characteristic equations*, are linear with two unknowns a and b because the coefficients of the two unknowns and the constant terms on the right-hand side can be computed using the known data of the original experiment.

Application of *Cramer's rule* leads to

$$b = \frac{\begin{vmatrix} N & \Sigma Y_i \\ \Sigma X_i & \Sigma X_iY_i \end{vmatrix}}{\begin{vmatrix} N & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{vmatrix}} = \frac{N(\Sigma X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{N(\Sigma X_i^2) - (\Sigma X_i)^2} \tag{13.4.6}$$

Substituting the mean values of the observations $\overline{X}$ and $\overline{Y}$ defined as

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \quad \text{and} \quad \overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \tag{13.4.7}$$

where N is the total number of experimental data, Eq. 13.4.6 can be rewritten as

$$b = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2} \tag{13.4.8}$$

Dividing Eq. 13.4.5a by the number of observations N and substituting Eq. 13.4.7 we obtain

$$\overline{Y} = a + b\overline{X} \tag{13.4.9}$$

In other words the point $(\overline{X}, \overline{Y})$ satisfies the equation of the best-fitting line. This means that *the best-fitting straight line always passes through the mean of the observations.*

By substituting the value of b obtained from either Eq. 13.4.6 or Eq. 13.4.8 into Eq. 13.2.9, the Y-intercept a becomes

$$a = \overline{Y} - b\overline{X} \tag{13.4.10}$$

Thus, given a set of N observations (X_i, Y_i), the parameters of the best-fitting straight line are given by Eqs. 13.4.8 and 13.4.10.

Example 13.12

Given the following measurements of traffic speed u and concentration k, apply the method of least squares to find the best-fitting straight line $u = a + bk$.

u	50	45	40	30	25
k	10	20	36	39	70

Solution Speed is the dependent variable, concentration is the independent variable, and a and b are the desired parameters of the postulated linear model. These parameters are calibrated by using Eqs. 13.4.8 and 13.4.10 arranged in tabular form here:

u	k	$u - \bar{u}$	$k - \bar{k}$	$(u - \bar{u})(k - \bar{k})$	$(k - \bar{k})^2$
50	10	12	−25	−300	625
45	20	7	−15	−105	225
40	36	2	1	2	1
30	39	−8	4	−32	16
25	70	−13	35	−455	1225
190	175	0	0	−890	2092

where

$$\bar{u} = 190/5 = 38$$

$$\bar{k} = 175/5 = 35$$

Hence

$$b = \frac{-890}{2092} = -0.43$$

$$a = \bar{u} - b\bar{k} = 38 + (0.43)(35) = 53.05$$

and

$$u = 53.05 - 0.43k$$

The scatter diagram and the best-fitting straight line are shown in Fig. 13.4.2.

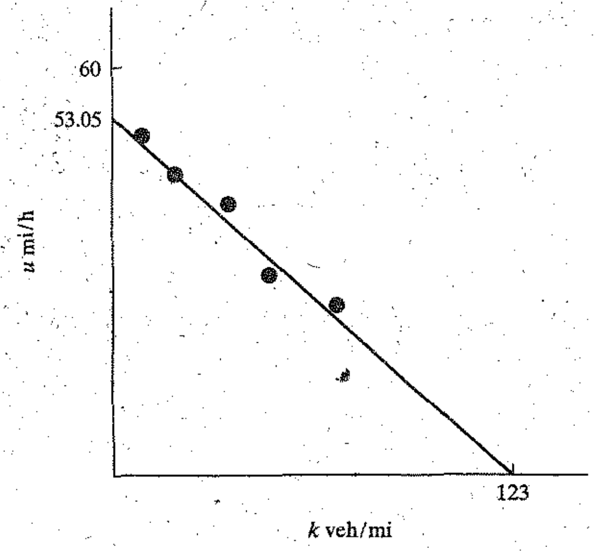


Figure 13.4.2 Scatter diagrams and best-fitting line (small spread of data).

Discussion The coefficient b turned out to be negative, as one would expect from the discussion of the general speed-concentration relationship. Also, it should be noted that the technique did not determine the mathematical form of the relationship: A linear form was postulated and the best-fitting straight line according to the regression criterion was the result.

Example 13.13

Repeat the procedure of Example 13.12 using the following set of observations:

u	70	20	15	50	35
k	10	23	39	38	65

Solution Proceeding as before, we obtain

u	k	$u - \bar{u}$	$k - \bar{k}$	$(u - \bar{u})(k - \bar{k})$	$(k - \bar{k})^2$
70	10	32	−25	−800	625
20	23	−18	−12	216	144
15	39	−23	4	−92	16
50	38	12	3	36	16
35	65	−3	30	−90	1225
190	175	0	0	−730	1694

where

$$\bar{u} = 190/5 = 38$$

$$\bar{k} = 175/5 = 35$$

Hence

$$b = \frac{-730}{1694} = -0.43$$

$$a = \bar{u} - b\bar{k} = 38 + (0.43)(35) = 53.05$$

and

$$u = 53.05 - 0.43k$$

The best-fitting straight line through the given data is shown in the accompanying Fig. 13.4.3.

Discussion As far as the calculation of the model's parameters is concerned, this problem is identical to the preceding problem. In both cases the best-fitting straight lines passing through the respective scatter diagrams were found. Note, however, that both problems led to identical regression equations. But a comparison of the two diagrams reveals that the first represents a tighter fit to the data than the second. Although the two sets of data led to the same relationship, the analyst would have more confidence in using the equation in the case represented by the first than the second set of experimental observations. A quantitative way of quantifying the "goodness of fit" is needed. The next section addresses this question.

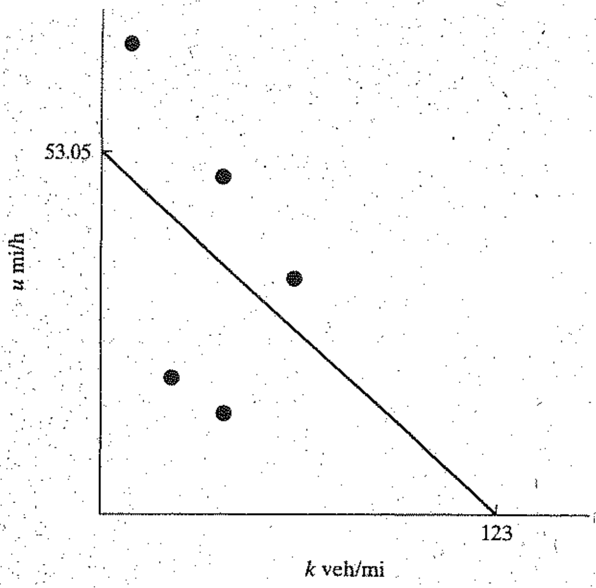


Figure 13.4.3 Scatter diagrams and best-fitting line (large spread of data).

13.4.2 Correlation

The sum

$$\mathrm{TSS} = \Sigma(Y_i - \bar{Y})^2 \tag{13.4.11}$$

called the *total sum of square deviations from the mean,* or the *total variation,* is a measure of the degree to which the Y observations are spread around their average value. It can be shown (see the exercises) that

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2 \tag{13.4.12}$$

The first term on the right-hand side of Eq. 13.4.12 is the *error sum of squares* (ESS), which the regression technique minimizes. It is also known as the *unexplained variation.* The second term represents the sum of squares of the difference between the estimated values of $\hat{Y}_i$ that lie on the regression line and the average value of $\bar{Y}$, which, as proven in the previous section, also lies on the regression line. Thus these differences are explained by the presence of the line. The sum of the squares of these quantities is known as the *explained variation.* The goodness of fit of a regression line increases with the proportion of the total variation that is explained by the line. The *coefficient of determination*

$$r^2 = \frac{\mathrm{TSS} - \mathrm{ESS}}{\mathrm{TSS}} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} \tag{13.4.13}$$

quantifies this fact. It ranges from zero when none of the total variation is explained by the regression line to unity when all of the variation is explained by the line. It is denoted as a squared quantity to capture the fact that it is always nonnegative. The square root of the coefficient of determination is called the *coefficient of correlation.* Its value can range from −1 to +1. In the case of linear regression the sign of r is the same as the sign of the slope b

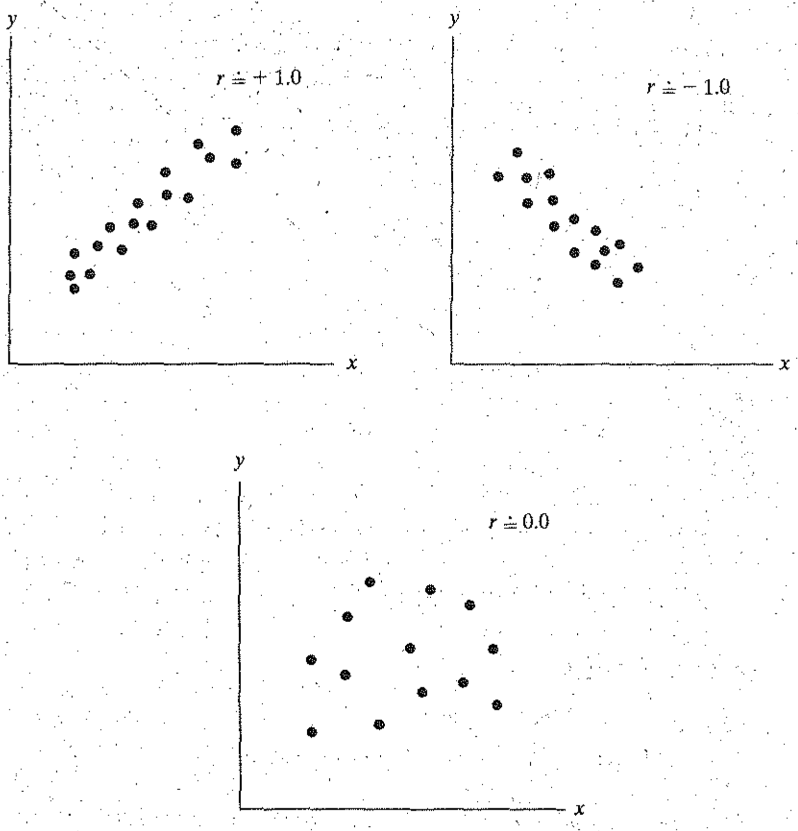


Figure 13.4.4 Correlation.

of the regression line. Figure 13.4.4 illustrates that if r is near $+1$, there exists a high positive correlation; if it is near -1, there exists a high negative correlation; and if it is around zero, there exists no correlation between X and Y. The following formula gives the proper magnitude *and* sign of r:

$$r = \frac{N(\sum X_iY_i) - (\sum X_i)(\sum Y_i)}{\{[N(\sum X_i^2) - (\sum X_i)^2][N(\sum Y_i^2) - (\sum Y_i)^2]\}^{1/2}} \tag{13.4.14}$$

Example 13.14

Compute the coefficient of correlation between X and Y using the data of Examples 13.12 and 13.13.

Solution Using Eq. 13.4.14, we obtain

$$r = -0.95 \qquad \text{for the data of Example 13.12}$$

and

$$r = -0.39 \qquad \text{for the data of Example 13.13}$$

Discussion As expected, both correlations are negative and the first case represents a better fit than the second.

13.4.3 Multiple Linear Regression

Simple linear regression involves only *two* variables. But often it is appropriate to postulate relationships that include two or more independent variables, each of which partially explains the value of the dependent variable Y. A relationship between the dependent and the independent variables of the form

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_pX_p \tag{13.4.15}$$

calibrated by the method of least squares is known as a *multiple linear regression* model.

Although a detailed treatment of the method of multiple regression is beyond the scope of this book, it is of some value to point out certain characteristics of the model. First, it should be understood that every experimental observation used in the calibration process must consist of $(p + 1)$ observations $(Y_i, X_{1i}, X_{2i}, \ldots, X_{pi})$. Calibration of the relationship means, as before, the estimation of the numerical values of the parameters of the model (i.e., the constant a_0 and the coefficients $a_1, \ldots, a_p$ in order to minimize the sum of squared deviations).

Second, the independent variables to be included in the relationship must be chosen so that they are not highly correlated among themselves. The simple correlation coefficient between pairs of potential independent variables may be computed via Eq. 13.4.14. When this criterion is satisfied, each of the terms on the right-hand side of Eq. 13.4.15 would be independent of the rest, capturing the effect of that specific variable X on the value of the dependent variable Y. If on the other hand two X's included in the equation were highly correlated, then it would be very difficult to examine the effect of each on the dependent variable because varying one of the two X's necessarily involves a change in the other.

Third, each of the selected independent variables must be highly correlated with the dependent variable Y; otherwise it would have no explanatory power.

Several calibration procedures are available for multiple regression relationships. One technique adds the independent variables one at a time and assesses the degree to which the addition of the last variable improves the relationship. On this basis a final relationship emerges, which includes the set of independent variables that provide the best fit. A *coefficient of multiple correlation* and various statistical tests can aid in assessing the goodness of fit.

13.4.4 Direct Nonlinear Regression

Linear regression (whether simple or multiple) assumes that the relationship between dependent and independent variables is in fact linear. Thus when linear regression is applied to the observations illustrated by the scatter diagram in Fig. 13.4.5, the result will be the best-fitting straight line (see the dashed line), even though the underlying relationship is clearly not linear.

To calibrate nonlinear relationships, one of two methods is frequently used. The first method involves specifying a nonlinear model and proceeding through the minimization of the sum of squared deviations, as in the case of simple linear regression, except that the postulated nonlinear form is substituted in Eq. 13.4.2 prior to the minimization step. This is illustrated next for the best-fitting parabola to a set of experimental data.

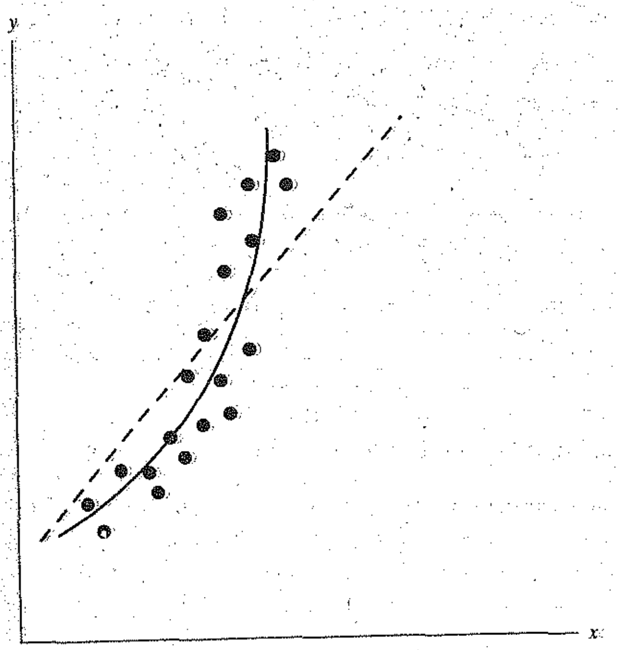


Figure 13.4.5 Nonlinearity.

Example 13.15: Least Squares Parabola

Fit an equation of the form $Y = a + bX + cX^2$ to the following data:

Y	30	40	65	85
X	2	3	4	5

Solution The unknown parameters a, b, and c are computed by minimizing the sum of squared deviations, or

$$\begin{aligned}\min S &= \Sigma(Y_i - \hat{Y}_i)^2\\ &= \Sigma(Y_i - a - bX_i - cX_i^2)^2\end{aligned}$$

The following characteristic equations result from setting the partial derivatives of S with respect to a, b, and c equal to zero and separating variables:

$$\begin{aligned}\Sigma\, Y_i &= aN = b(\Sigma\, X_i) + c(\Sigma\, X_i^2)\\ \Sigma(X_iY_i) &= a(\Sigma\, X_i) + b(\Sigma\, X_i^2) + c(\Sigma\, X_i^3)\\ \Sigma(X_i^2Y_i) &= a(\Sigma\, X_i^2) + b(\Sigma\, X_i^3) + c(\Sigma\, X_i^4)\end{aligned}$$

Substituting the given data in the characteristic equations, we obtain

$$\begin{aligned}220 &= 4a + 14b + 54c\\ 865 &= 14a + 54b + 224c\\ 3645 &= 54a + 224b + 978c\end{aligned}$$

which, when solved simultaneously, yield the following values:

$$a = 16, \quad b = 1.5, \quad \text{and} \quad c = 2.5$$

Thus the best-fitting parabola of the postulated type becomes

$$Y = 16 + 1.5X + 2.5X^2$$

Discussion Again, it must be emphasized that the technique did not select the functional form but merely determined the best line of the form supplied by the analyst. Care must be taken to express the sum of squared deviations to be minimized in terms of the postulated form. The reader is encouraged to find the least squares equations of the forms $Y = aX + bX^2$, $Y = a + bX^2$, and $Y = bX^2$ using the same data and to compare the results of the three regression lines. The first and last of these equations "force" the line to pass through the origin.

13.4.5 Linear Regression with Transformed Variables

The second method of calibrating nonlinear relationships applies when a nonlinear relationship can be transformed to a linear relationship, as Fig. 13.4.6 illustrates. In this case *linear regression is applied to the transformed relationship* to determine the values of *its* parameters, which are then transformed back to the parameters of the original model.

Example 13.16

The following speed and concentration measurements were taken on a highway:

u	50	35	35	25	20
k	10	40	50	80	100

It is desired to calibrate a speed-concentration equation of the form proposed by Underwood; that is, $u = u_f \exp(-k/k_m)$. Determine the parameters of this model using simple linear regression.

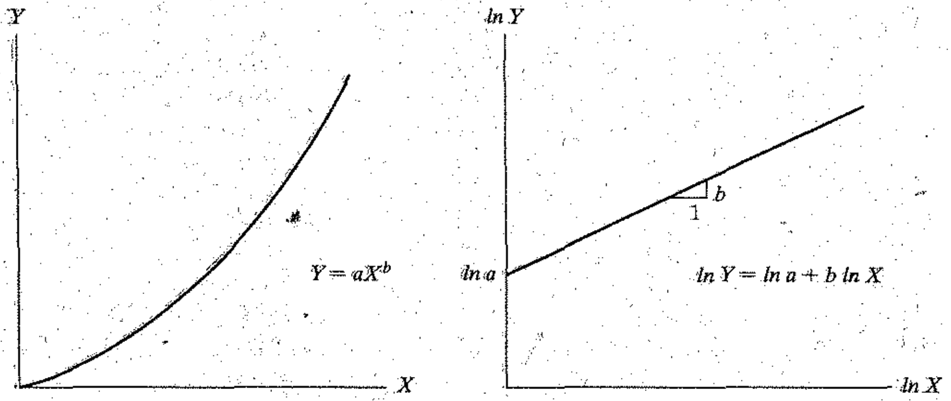


Figure 13.4.6 Variable transformation.

Solution The postulated relationship is not linear. However, taking the natural logarithm of both sides, we have

$$\ln u = \ln u_f - \frac{k}{k_m}$$

The following substitutions render this equation in the proper linear regression form:

$$Y = \ln u \qquad X = k \qquad a = \ln u_f \qquad b = \frac{-1}{k_m}$$

Performing simple linear regression of X on Y as before leads to

$$a = 4.01 \qquad b = -0.01 \qquad Y = 4.01 - 0.01X$$

To find the values of the parameters of the original model, make the inverse transformation:

$$u_f = e^a = 55 \text{ mi/h} \quad \text{and} \quad k_m = \frac{-1}{b} = 100 \text{ veh/mi}$$

and

$$u = 55\, e^{-k/100}$$

Discussion The free-flow speed of 55 mi/h and the concentration at capacity of 100 veh/mi represent the best-fitting Underwood relationship to the given data. This equation may be applied as in Section 3.4 to find the implied $q - k$ and $u - k$ curves and to estimate the capacity of the roadway, which (the reader could verify) happens to be 2023 veh/h.

13.4.6 Selection of Explanatory Variables

When estimating a multiple regression model, the analyst is faced with the questions of how many and which independent variables to include in the equation. The following rules of thumb provide some guidance in this respect. Regarding the number of independent variables to be included in a model, practical experience has shown that the law of diminishing returns holds true with respect to accuracy resulting from increasing the number of independent variables. Figure 13.4.7 illustrates the point by plotting the accuracy obtained by increasingly complex models versus the number of variables employed [13.3]. The figure shows that a point is reached beyond which the extra cost and complexity associated with adding another variable (which includes the need to forecast this variable toward the target year) may not be warranted by the increasingly smaller improvements in accuracy obtained.

The following four guidelines are helpful in deciding which explanatory variables to include in a linear regression model. The selected explanatory variables:

1. Must be linearly related to the dependent variable
2. Must be highly correlated with the dependent variable
3. Must not be highly correlated between themselves
4. Must lend themselves to relatively easy projection

The first rule states that the relationship between a selected explanatory variable and the dependent variable must be linear, as required by the mathematical specification

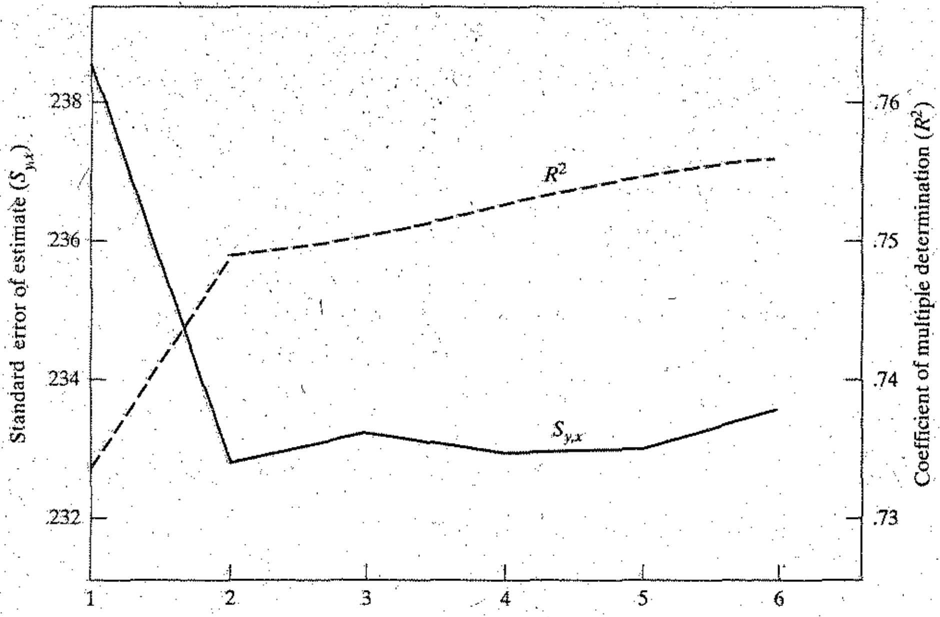


Figure 13.4.7 Graphical representation of the stepwise change in standard error of estimate and coefficient of multiple determination.
(From Federal Highway Administration [13.3].)

of the model. If this is not true, an appropriate transformation of the explanatory variable may be performed as explained in Section 13.4.5. The second rule states that the explanatory variable must be highly associated with the dependent variable; otherwise it would have no explanatory power. The third rule states that variables that are highly correlated among themselves must not be included in the same equation. If two potential explanatory variables are highly correlated, they essentially measure the same effect or in other words they are not independent. If both were to be included in the same equation, double counting would result. Moreover, the resulting equation would not be easy to interpret, as the sensitivity of the dependent variable to a single explanatory variable could not be captured by that variable's coefficient alone. The fourth rule states that the selected explanatory variables must be such that they can be forecast toward the target year with relative ease. The reason models are used in the first place is that it is extremely difficult to project the dependent variable directly. Consequently the model is calibrated in terms of a set of factors (or independent variables) that explain the dependent variable. Unless these factors are relatively easy to project into the future, the entire effort will prove to be of little value.

By applying these rules, the number of potentially useful alternative specifications of a model can be considerably reduced. What remains is to select the best model for this reduced set on the basis of statistical tests for goodness of fit and, not to be underestimated, the application of professional judgment to ensure the reasonableness of results.

Example 13.17

The following correlation matrix contains the simple correlation coefficients between pairs of variables computed by Eq. 13.4.14 using base-year data. Discuss the question of which explanatory variables X should be included in a linear multiple regression model.

	Y	X_1	X_2	X_3	X_4
Y	1.00	0.32	0.92	0.95	0.62
X_1		1.00	0.25	0.19	0.03
X_2			1.00	0.99	0.29
X_3				1.00	0.33
X_4					1.00

Solution Variable X_1 is not highly correlated with the dependent variable Y. Hence it may be eliminated from further consideration. Although highly correlated with Y, variables X_2 and X_3 are also highly correlated with each other. Therefore they should not appear together in the same equation. Variable X_4 is not highly correlated with either X_1 or X_2. Hence it can appear in the same equation with either of the two. Based on this discussion, the following alternative linear multiple regression models may be considered:

1. $Y = a_0 + a_2X_2$
2. $Y = b_0 + b_3X_3$
3. $Y = c_0 + c_4X_4$
4. $Y = d_0 + d_2X_2 + d_4X_4$
5. $Y = e_0 + e_3X_3 + e_4X_4$

Discussion The simple correlation matrix is symmetric, and only the upper or lower half should be specified. The diagonal elements are equal to unity because the correlation between any variable with itself is perfect. The five potential models specified in the solution meet the rules of selection given earlier. Further analysis is required to discover the best among them.

13.5 HYPOTHESIS TESTING AND MODEL EVALUATION

A number of hypotheses are formulated and tested statistically to evaluate the goodness of fit of a linear model estimated with empirical data. The basic steps of hypothesis testing are the following:

Step 1: Formulation of a null hypothesis (H_0); for example, parameter value is equal to the estimated value.

Step 2: Formulation of an alternative hypothesis (H_1); for example, parameter value is equal to zero or a value different from the estimated one.

Step 3: Identification of a test statistic distribution based on H_0; usually Student's t or F statistic is used, depending on the test.

Step 4: Performance of comparison and rejection of H_0 if test statistic has very low probability of occurrence when H_0 is true.

These basic steps are applied to three common tests: single-parameter test, test of a linear model as a whole, and test of equality of segmented linear models. These three tests are detailed next.

13.5.1 Single-Parameter Test

As the name of the test indicates, one parameter at a time is tested. It is compared with an alternative value (most commonly with zero). The test is formulated as follows:

$$H_0\text{: } \beta_1 = \beta_1^*$$

$$H_1\text{: } \beta_1 \neq \beta_1^*$$

$$\text{Statistic} = |\hat{\beta}_1 - \beta_1^*|/S_{\beta 1}$$

$$\text{Criterion: reject } H_0 \text{ if } |\hat{\beta}_1 - \beta_1^*|/S_{\beta 1}/\sqrt{N} > t_{N-J-1,\alpha/2}$$

where

$\hat{\beta}_1$ = average estimated value (mean)

$S_{\beta 1}$ = standard deviation of β_1

N = sample size

J = number of degrees of freedom (i.e., number of independent variables in the model)

$1 - \alpha$ = level of statistical confidence desired (e.g., for $\alpha = 0.05$ the level of statistical confidence is 95%)

The most common test of this type is setting H_0: $\beta_1 = 0$ and H_1: $\beta_1 \neq 0$.

Example 13.18

The following data were gathered for saturation flow (see Chapter 4 for definition):

1820 1700 1780 1620 1810 1850 1690 1750 1750 1900 1860 1830

Compare these measurements with the recommended value of 1800 for a 95% level of statistical confidence and decide which is the preferred value for the saturation flow.

Solution First the mean and standard deviation of the sample must be estimated. Treat saturation flow as parameter β.

$$\beta(\text{mean}) = 1780$$

$$S_\beta = 81.2$$

$$N = 12$$

Then the hypotheses are formulated:

$$H_0\text{: } \beta_1 = 1800$$

$$H_1\text{: } \beta_1 \neq 1800$$

$$\text{Statistic} = |1780 - 1800|/81.2/\sqrt{12}) = 0.85 < 2.19 = t_{12-1,\,0.025}$$

The criterion value of the t statistic (2.19) was taken from statistical tables attached to any standard book on probability and statistics ([13.1] and many others). Since this is a single-parameter estimate, the J factor is not applicable.

Discussion The test does not support the rejection of the null hypothesis that the saturation flow should be equal to the recommended value of 1800.

13.5.2 Test of a Linear Model

This procedure tests the overall validity of a linear model by comparing all of its parameter estimates to zero. If the test is not passed, this means that there is substantial probability that all parameters may be equal to zero, and thus the model is by and large worthless. The hypothesis testing for this is as follows:

$$H_0: \overline{\boldsymbol{\beta}} = \overline{\mathbf{0}}$$

$$H_1: \overline{\boldsymbol{\beta}} \neq \overline{\mathbf{0}}$$

$$\text{Statistic} = \frac{\text{RSS/J}}{[\text{ESS}/(N - J - 1)]}$$

$$\text{Criterion: reject } H_0 \text{ if } \frac{\text{RSS/J}}{[\text{ESS}/(N - J - 1)]} > F_{J,N-J-1,(1-\alpha)}$$

where

$\overline{\boldsymbol{\beta}}, \overline{\mathbf{0}}$ = vector matrices

RSS = TSS − ESS as defined by Eqs. 13.4.11 and 13.4.12

N = sample size

J = number of estimated parameters

F statistic = criterion number taken from statistical tables [13.1]

Example 13.19

A multiple linear regression model was estimated on a computer with the aid of a statistical package. The output includes the following: ESS = 45.0, TSS = 95.8, N = 1220, J = 8. Check whether the equality of all dependent variable parameter estimates to zero can be rejected with 95% statistical confidence.

Solution From tables [13.1] the F statistic results as $F_{8,1211,95\%} = 1.94$. The statistic and comparison is as follows:

$$\frac{(95.8 - 45.0)/8}{45.0/(1220 - 8 - 1)} = 171 > 1.94$$

Discussion The F test is generally not a powerful test; it helps to reject models with very weak associations between the dependent and independent variables. For $N - J - 1 > 10$ the F-statistic value is less than 3.0, and usually the estimated statistic can easily exceed the F-statistic value, unless if ESS approaches TSS. For instance, in the previous example, if ESS is 95.0, the statistic becomes 1.27 < 1.94.

13.5.3 Test of Equality of Segmented Linear Models

Data for the analysis of a specific factor (dependent variable) are gathered from a population. Often specific characteristics of the population suggest or necessitate the estimation of the same linear model for various subgroups of the population. Examples of this are (1) flow characteristics of two-, four-, or six-lane highways may be basically similar, but segmentation according to the size of the facility may provide better estimates for a specific flow characteristic; (2) trip characteristics of people may be assessed more accurately if segmentation by gender is applied; (3) automobile ownership may vary significantly between urban and rural communities; and so forth.

In the case of segmented estimation the corresponding parameters of each model should be checked for equality. If the hypothesis of equality between the corresponding parameters of the models for each segment cannot be rejected, the segmentation is largely worthless.

Assume a straightforward linear model with three independent variables, which was estimated as follows:

Pooled model (p):

$$\text{All data: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Segmented models (g):

$$\text{Segment 1: } Y = \beta_{01} + \beta_{11} X_{11} + \beta_{21} X_{21} + \beta_{31} X_{31}$$

$$\text{Segment 2: } Y = \beta_{02} + \beta_{12} X_{12} + \beta_{22} X_{22} + \beta_{32} X_{32}$$

The hypothesis testing is formulated as follows:

H_0: $\beta_{11} = \beta_{12}, \beta_{21} = \beta_{22}, \beta_{31} = \beta_{32}$

H_1: $\beta_{11} \neq \beta_{12}, \beta_{21} \neq \beta_{22}, \beta_{31} \neq \beta_{32}$

$$\text{Statistic} = \frac{[\text{ESS}(p) - (\text{ESS}_1 + \text{ESS}_2)]/(J+1)}{(\text{ESS}_1 + \text{ESS}_2)/[(N_1 - J_1 - 1) + (N_2 - J_2 - 1)]}$$

Criterion: reject H_0 if

$$\frac{[\text{ESS}(p) - (\text{ESS}_1 + \text{ESS}_2)]/(J+1)}{(\text{ESS}_1 + \text{ESS}_2)/[(N_1 - J_1 - 1) + (N_2 - J_2 - 1)]}$$

$$> F_{J+1,[(N_1 - J_1 - 1) + (N_2 - J_2 - 1)],(1-\alpha)}$$

In this illustration J is equal to 3 (β_1, β_2, β_3). ESS (p) is the ESS of the pooled model (an identical model estimated with all the data, without segmentation).

Example 13.20

The following summary statistics were obtained after the pooled and segmented estimation of a linear model. Test the equality of parameter estimates between the segmented models for $\alpha = 0.05$.

	Pooled	Segment 1	Segment 2
ESS	180	60	73
TSS	400	190	210
J	5	5	5
N	645	305	340

Solution The F statistic is estimated from tables [13.1]: $F_{6,633,95\%} = 2.10$. The test statistic is calculated as follows:

$$\frac{[180 - (60 + 73)]/6}{(60 + 73)/633} = 37.3 > 2.10$$

Thus the equality of parameter estimates between the segmented models can be rejected.

13.5.4 Comprehensive Judgment of a Linear Model

Five major checks should be conducted for judging the quality and acceptability of an empirically estimated linear model:

1. Overall model fit (coefficient of determination, r^2 or R^2)
2. Significance of parameter estimates (t test of all β_i)
3. Equality of all parameters to zero (overall model worthiness, F test)
4. Standard error of estimate of the dependent variable
5. Sign and size of parameter estimates (intuitiveness of parameters)

A model estimated from household survey data is utilized to illustrate the application of the whole set of tests:

$$\begin{array}{cccccccccccc} Y & = & 0.26 & + & 0.71X_1 & + & 0.71X_2 & + & 0.052X_3 & - & 0.28X_4 \\ (0.63) & & (0.07) & & (0.02) & & (0.03) & & (0.013) & & (0.05) \end{array}$$

The results are

$$R^2 = 0.49 \qquad F = 326 \qquad N = 1400$$

Dependent variable:

Y = number of household automobiles (automobile-ownership level: 0, 1, 2, . . .)

Independent variables:

X_1 = number of drivers in the household

X_2 = residential density variable:
1 if household resides in a low-density (suburban) location,
0 if household resides in a high-density (urban) location

X_3 = income per person calculated as the ratio of the total household income to the number of people in the household (in thousand 1989 U.S. dollars)

X_4 = number of household workers who commute to work by public transportation

The numbers in parentheses represent the standard error of estimate.

The overall model fit is deemed good since $R^2 = 0.49$, which means that 49% of the variance of the dependent variable can be explained by the model. R^2 scores ranging between 0.30 and 0.60 generally denote a good model fit. R^2 scores less than 0.20 represent a poor model fit, and such models should be used with particular caution for real-world applications. R^2 scores greater than 0.70 when derived from large samples (i.e., a few hundred observations or more) may be suspicious. Collinearity effects and other problems may be resulting in exaggerated goodness of fit.

The t test for each parameter shows that all of them are statistically significant at the 99% level ($\alpha = 0.01$). The t statistic (tables [13.1]) for more than 120 observations and $\alpha = 0.01$ is 2.6. The t value for each parameter is estimated by dividing the parameter estimate by the corresponding standard error of estimate. For example, for X_4, $t = 0.28/0.05 = 5.6$, which is greater than 2.6. Thus the hypothesis that this parameter may be equal to zero can be rejected with 99% certainty. The same applies to all of the other independent variables of the model.

The F score for the model is very high (i.e., $326 > 2.37$), which permits the rejection of the hypothesis that all parameter estimates are equal to zero, with 99% certainty.

The standard error of estimate for the dependent variable is 0.63. The standard error is equivalent to the standard deviation, and this means that 68% of the observations fall within one standard deviation and 95% fall within two standard deviations from the mean. The smaller the standard deviation is, the more confident one can be for the estimates of the dependent variable. For the data set utilized for the estimation of this model the mean automobile-ownership level (dependent variable) is 2.02, thus the magnitude of the standard deviation is acceptable. In the case where the standard deviation is larger than the mean the model is clearly unacceptable.

Finally, an intuitive test should be done with respect to the sign and size of parameters. The parameter estimates of the model examined are intuitive:

X_1 = (positive) larger number of drivers corresponds to higher automobile-ownership level because of the higher mobility needs and the ability to drive a car

X_2 = (positive) low density tends to increase automobile-ownership level (indeed, public transportation is not extensive in low-density locations and walking distances tend to be long, thus the use of the automobile may be necessary to fulfill mobility needs)

X_3 = (positive) higher income per person corresponds to more funds being available for automobile acquisition

X_4 = (negative) high number of household drivers commuting by public transportation translates into less need for automobiles for trips to work; thus automobile ownership tends to decrease

Regarding the size of variables, prior experience is necessary for judgment. Based on historical evidence, the contribution of income should be small compared with the contribution of other variables such as the number of drivers. This is because automobile ownership tends to be driven largely by the need for mobility. Income plays a secondary role:

Low-income households tend to purchase inexpensive automobiles, and high-income households tend to purchase more luxurious automobiles. The model examined fulfills this anticipation. Consider a household with three drivers and $15,000 income per person. The respective contributions to automobile ownership (refer to the model) are as follows:

$$\text{Drivers:} \quad 3(0.71) = 2.13$$

$$\text{Income per person: } 15(0.052) = 0.78$$

Overall the model passes all tests and it should be considered reliable for real-world use (in the context where the model was derived and estimated).

13.6 SUMMARY

In this chapter we introduced the basic concepts and definitions of the theory of probability. The probability distributions associated with several discrete and continuous random variables were then presented, and the applications of these distributions to traffic phenomena were illustrated by some simple examples. We also covered the method of least square regression that can be used to estimate the parameters of linear and nonlinear models and tests for evaluating these models.

EXERCISES

1. The outcome of an experiment is the sum obtained by casting two dice. Enumerate the simple outcomes of the experiment and calculate each probability.
   (a) Each simple outcome
   (b) An even outcome
   (c) An odd outcome greater than 6
   (d) An outcome that is either odd or greater than 4
   (e) An outcome that is less than 5 but greater than 2
2. For the experiment of Exercise 1, determine whether the following two events are (a) mutually exclusive and (b) independent:

A	the outcome is less than 8
B	the outcome is greater than 5

3. For the experiment of Exercise 1, determine if the following two events are (a) mutually exclusive and (b) independent:

A	{2,3,4,5,6,10}
B	{2,4,6,8,9,10,11,12}

4. Intersection *A* is located downstream of intersection *B*. A traffic engineer observed that during the morning peak period the probability that intersection *A* is congested is 0.30. There is a 0.50 probability of *A* being congested given the knowledge that *B* is congested. On the other hand the

probability of B being congested given the knowledge that A is congested is 0.90. Calculate the probability that (a) B is congested during the peak period and (b) at least one of the two intersections is congested.

5. There are only two traffic control signals in downtown Kahului, Maui. The *independent* probabilities that each of the signals will malfunction on a given day are 0.05 and 0.04, respectively. System failure is defined by the condition that at least one signal is out. Calculate the probability that the system will fail on any particular day.

6. A study has shown that in cars with a driver and a passenger, the probability that the driver wears a seat belt is 0.35 and the probability that the passenger buckles up is 0.50. If 80% of the passengers of drivers that buckle up do the same, calculate (a) the probability that both the driver *and* the passenger of a car are wearing seat belts and (b) the probability that a particular driver is wearing a seat belt given the knowledge that the passenger is buckled up.

7. The probability that downtown parking garage A is not full is 0.20 and the probability that garage B is full is 0.50. Knowing that at least one of the two garages is full 90% of the time, calculate the probability that a shopper will find parking in garage B given that garage A is full.

8. Prove Bayes' theorem, which states that

$$P[A \mid B] = \frac{P[B \mid A]\,P[A]}{P[B]}$$

9. Two alternative highway routes connect a suburb with a downtown area. Define event A: Highway 1 is jammed during the peak period, and event B: Highway 2 is jammed during the peak period. If $P[A]$ is 0.70, $P[B]$ is 0.60, and $P[B \mid A]$ is 0.50, calculate and interpret $P[A \cap B]$, $P[A \cup B]$, and $P[A \mid B]$.

10. Draw the probability mass function of the experiment described in Exercise 1 and compute its mean and variance. Also, draw the cumulative distribution.

11. Decide which of the following functions can possibly be discrete probability distributions for the specified sample space:

(a)

X	1	2	3	4	5	6
$p(x)$	0.2	0.3	0.2	0.2	0.1	0.1

(b)

X	1	2	3	4	5	6
$p(x)$	0.5	0.4	0.3	0.1	−0.4	0.1

(c)

X	1	2	3	4	5	6
$p(x)$	0.4	0.2	0.1	0.1	0.1	0.1

12. At a particular intersection approach, 30% of the left-turning vehicles fail to signal their intentions to turn. Assuming independence, calculate each probability:

(a) Three vehicles in a row will fail to signal.
(b) Three vehicles in a row will signal.
(c) The second vehicle will not signal.
(d) The second vehicle to signal will be the fifth vehicle observed.
(e) Two vehicles in five will fail to signal.
(f) The first vehicle to signal will be the fourth vehicle observed.
(g) No more than two vehicles in five will signal.

13. Draw the probability function and the cumulative distribution for the Poisson case described in Example 13.8.

14. Vehicles arrive at an isolated intersection according to the Poisson distribution. Given that the mean arrival rate is 500 veh/h, calculate (a) the probability that zero vehicles will arrive during a 10-s interval and (b) the probability that at least five vehicles will arrive during a 10-s interval.

15. Students arrive at a lecture room at the rate of 15 per minute according to the Poisson distribution. Calculate the probability of (a) *exactly* three arrivals in 20 s, (b) *no more* than three arrivals in 20 s, and (c) *at least* three arrivals in 20 s.

16. The average concentration of vehicles on a highway section is 70 veh/mi. Given that the concentration is Poisson distributed, calculate the probability of finding (a) exactly ten vehicles and (b) five or more vehicles on any particular tenth of a mile.

17. Which of the following continuous functions can serve as probability distributions for the specified range of outcomes?

    (a) $f(x) = -1.0 + 0.2x, 0 \leqslant x \leqslant 10.0$

    (b) $f(x) = x(1.0 - 2.0x), 0 \leqslant x \leqslant 0.5$

    (c) $f(x) = 24.0x(1.0 - 2.0x), 0 \leqslant x \leqslant 0.5$

18. Prove that Eq. 13.2.31 is correct.

19. Airplanes arrive at an airport area at an average rate of six per hour. Assuming that the arrival pattern is Poisson distributed, calculate the probability that the headway between two successive arrivals will be greater than 20 min.

20. The airport control tower (see Exercise 19) processes airplanes in their order of arrival. Assuming that the service time is negative exponential and that the service rate is ten landings per hour, calculate (a) the average number of airplanes in the system (i.e., being served and stacked), (b) the average number of airplanes awaiting clearance to land, (c) the average time spent in the system, and (d) the average time an airplane is in the queue.

21. A turnpike toll area contains four toll booths arranged in parallel. The arriving vehicles conform to the Poisson distribution, with an average headway of 12 s. Assuming that the average service time is 5 s, the service time is negative exponential, and the queue discipline is FIFO (first in, first out), find the average queue length and the expected time in the system if (a) two of the booths are in operation and (b) only one booth is open.

22. Speed (mi/h) measurements at a suburban location with a speed limit of 35 mi/h showed that speeds were distributed according to $N[27, 9]$. Calculate the following:

    (a) the 85th percentile speed

    (b) the 20th percentile speed

    (c) the percent of vehicles exceeding the speed limit.

23. A transportation engineer was hired by the city planning department to calibrate a multiple regression model for trip productions. The department has collected base-year data for the following variables:

$$P_I = \text{trip productions}$$

$$X_1 = \text{zone population}$$

$$X_2 = \text{median income}$$

$$X_3 = \text{median age}$$

$$X_4 = \text{car registrations}$$

$$X_5 = \text{number of dwelling units}$$

A preliminary analysis of the data resulted in the following simple correlation matrix:

	P_t	X_1	X_2	X_3	X_4	X_5
P_t	1.00	0.95	0.83	0.41	0.82	0.85
X_1		1.00	−0.21	0.22	−0.29	0.91
X_2			1.00	0.82	0.89	−0.43
X_3				1.00	−0.19	−0.15
X_4					1.00	−0.22

Specify at least five possible equations that may be tried and give the specific reasons for their selection.

24. The following data were obtained from an experiment:

X	1	3	4	6	8	9	11	14
Y	9	8	7	5	4	4	2	1

Fit a line of the form $Y = a + bX$, and calculate and interpret the coefficient of correlation.

25. Use the data of Example 13.5 to fit a curve of the form $Y = aX^2$ by direct nonlinear regression.

26. Repeat Exercise 24 assuming that $Y = aX^b$ by linear regression with transformed variables.

27. Plot and compare the results of Example 13.5 and Exercises 25 and 26.

28. Show that Eqs. 13.4.6 and 13.4.8 are equivalent.

REFERENCES

13.1 Benjamin, J. R., and C. A. Cornell, *Probability, Statistics and Decision for Civil Engineers,* McGraw-Hill, New York, 1970.

13.2 Transportation Research Board, *Traffic Flow Theory: A State of the Art Report,* draft, Committee on Theory of Traffic Flow (A3A11), Washington, DC, 1997.

13.3 Federal Highway Administration, *Guidelines to Trip Generation Analysis,* U.S. Department of Transportation, Washington, DC, 1967.

14

Queuing and Simulation

14.1 INTRODUCTION

A *queue* is simply a waiting line. Therefore systems that involve waiting lines are called queuing systems and mathematical descriptions of queuing systems are known as queuing models. Transportation systems often involve queues. For example, vehicles accumulating at an intersection approach during red can be thought of as forming a waiting line waiting to be served during the subsequent green display. Similarly, transit vehicles arriving at a station may form waiting lines as they load and unload passengers, who in turn form queues to enter or alight vehicles, purchase tickets, exit through turnstiles, and so forth. Other examples include vehicles waiting to be served at a gasoline station, airplanes awaiting clearance for takeoff or landing, patients scheduled for use of a hospital's operating room, component parts stockpiled at an assembly plant, computer jobs awaiting execution or printing, vehicles at an intersection, and so forth. Continuous processes may also be described as queuing systems: for example, drinking water in a reservoir "waiting" to be used by households.

This chapter describes the characteristics of simple queuing models. Section 14.2 shows that as the queuing patterns become complex, analytical solutions become intractable. In such cases an alternative approach is the use of *computer simulation*. Section 14.3 presents the fundamental elements of simulation models and the fundamental characteristics of the *Monte Carlo* technique. As explained in Chapter 8, the Monte Carlo technique is finding increasing application to advanced methods of simulating probabilistically regional travel patterns due to the computational requirements of alternate methods such as market segmentation.

14.2 QUEUING MODELS

14.2.1 Background

Queuing, or waiting-line, phenomena are everyday occurrences. No matter how complex, queuing systems are characterized by an *arrival pattern,* a *service facility,* and a *queue discipline.* When all three components are constant, the system can be analyzed by deterministic methods. Probabilistic systems, however, are more common.

The arrival pattern describes the way in which the items (or "customers") to be served enter the system. For instance, some examples in Chapter 13 considered the case of vehicles arriving according to the Poisson distribution, which meant that their interarrival times were exponentially distributed.

The service facility is characterized by the *number and arrangement of servers* and by a *service pattern.* A service facility can be a *single-server* or a *multiserver* facility. Service counters may be arranged in parallel, in series, or in any combination of the two. The service pattern usually measures either the rate at which customers are processed (i.e., vehicles per minute) or the time required to serve individual customers. These characteristics may also be described by appropriate probability functions.

The queue discipline refers to the rules by which the next customer to be served is chosen. Some of the most common rules include the following. A *first-in, first-out* (FIFO) scheme serves customers in the order they arrive: The customer at the front of the waiting line is always selected. Another rule is the *last-in, first-out* (LIFO) rule. For example, program instructions in a computer's stack memory are executed in the reverse order of placement into the stack. The service rule may allow either a *single queue* or *multiple queues.* The customers may be treated equally or they may be treated according to some *priority.* Examples of priority rules include car-pool lanes for use by vehicles carrying a specified minimum number of passengers, express lanes at a supermarket, business-transactions-only counters at a bank, the scheduling of patients for an operation according to the severity of their maladies, and computer processing of administrative jobs before the jobs submitted by faculty and students. The priority rule may be either *preemptive* or *nonpreemptive,* depending on whether or not a higher priority customer is permitted to interrupt the processing of a lower priority customer. For example, emergency vehicles on the roadways have a preemptive priority over other vehicles.

A signalized intersection may be modeled as a multichannel system with complex queue disciplines controlled by the traffic signal. The description of transit stations in Chapter 4 reveals the possibility of analyzing their operations by employing queuing models. Also, platoons of vehicles on the road may be viewed as moving queues. The previous discussion illustrates that queuing systems can range from the very simple to the very complex. Relatively simple systems may be examined by formulating the appropriate mathematical equations and solving them *analytically.* More complex systems become mathematically intractable and are usually solved by *numerical* methods.

The solution to a queuing problem entails the assessment of a system's performance, which in turn is described by a set of *measures of performance.* These may include the number of customers served per unit time, the average delay per customer, the average and maximum length of the waiting lines, the percent of time each service counter is idle, the cost of operating the system, and so forth.

14.2.2 Single-Server FIFO Systems

One of the simplest queuing problems that is amenable to analytical solution is the single-server FIFO system with Poisson arrivals and exponentially distributed customer service times. When in the system, customers are assumed to be patient; that is, they do not leave prematurely. The system is assumed to have an unlimited holding capacity: There is no upper limit on the number of customers that can be in the waiting line. The *state of the system* is described by the random variable X representing the number of customers in the system at any given time, including those that are being served. Any reference book on the queuing theory (e.g., Refs. [14.1–14.3]) may be consulted for a mathematical proof that the steady-state conditions of the above system when the mean arrival rate (λ arrivals per unit time) is less than the mean service rate (μ items served per unit time), X is distributed according to the following function:

$$f(x) = P[X = x] = r^x(1 - r), \qquad x = 0, 1, 2, \ldots \tag{14.2.1}$$

where

$$r = \frac{\lambda}{\mu}$$

The expected value, that is, the average number of customers in the system at any time, is

$$E[X] = \frac{r}{1 - r} \tag{14.2.2}$$

Other useful measures of performance include the following. The average number of customers in the waiting line (*queue length, L_q*) is

$$E[L_q] = \frac{r^2}{1 - r} \tag{14.2.3}$$

The expected time each customer spends in the system and in the queue are calculated, respectively, by dividing the last two equations by λ. Thus

$$E[T] = \frac{1}{\mu - \lambda} \tag{14.2.4}$$

and

$$E[T_q] = \frac{\lambda}{\mu(\mu - \lambda)} \tag{14.2.5}$$

Example 14.1

Bank customers arrive at a single drive-in window at an average rate of 15 veh/h. On the average the customers need 3 min each to transact their business. Given that the arrival pattern is described by the Poisson distribution and that the departure time is exponentially distributed, calculate the following:

(a) The percent of time that the bank teller will be idle

(b) The probabilitv that five customers will be in the system

(c) The average number of customers in the system

(d) The average queue length

(e) The average time each customer spends in the system

Solution The average arrival rate is 15 customers per hour, the average service rate is 60/3 = 20 customers per hour and $r = 0.75$. Since the service rate of this FIFO system is larger than the arrival rate, the equations just developed apply.

(a) The teller is idle when there are no customers in the system. Hence

$$p(0) = P[X = 0] = (0.75)^0(1 - 0.75) = 0.25$$

or 25% of the time.

(b) The probability that five vehicles will be in the system is given by Eq. 14.2.1:

$$p(5) = (0.75)^5(1 - 0.75) = 0.059$$

or 5.9% of the time.

(c) The average number of customers in the system is

$$E[X] = \frac{15}{20 - 15} = 3 \text{ customers}$$

(d) By Eq. 14.3.3, the average queue length is

$$E[L_q] = 2.25 \text{ customers}$$

(e) The average time in the system is

$$E[T] = (20 - 15)^{-1} = 0.2 \text{ h, or } 12 \text{ min}$$

Discussion The average number of customers in the queue is not equal to the average number of customers in the system less one because when the system is empty (25% of the time in this example), one customer cannot be meaningfully subtracted from zero customers in the system. The average of a discrete distribution does not need to coincide with an outcome. Thus an average queue length of 2.25 customers is meaningful as a long-term average.

14.2.3 Multiserver FIFO Systems

A more complex queuing system is a FIFO system with N identical service counters in parallel. The average service rate *per counter* is μ and the remaining variables are defined as before. In this case the distribution of X is as follows.

For $x = 0$:

$$p(0) = \left[\sum_{x=0}^{N-1} \frac{r^x}{x!} + \frac{r^N}{(N-1)!(N-r)}\right]^{-1} \tag{14.2.6a}$$

For $1 \le x \le N$:

$$p(x) = \frac{r^x}{N!} p(0) \tag{14.2.6b}$$

For $x > N$:

$$p(x) = \frac{r^x}{N! N^{x-N}} p(0) \tag{14.2.6c}$$

The average number of customers in the system is

$$E[X] = r + \left[\frac{r^{N+1}}{(N-1)!(N-r)^2}\right] p(0) \tag{14.2.7}$$

The average queue length is

$$E[L_q] = \left[\frac{r^{N+1}}{(N-1)!(N-r)^2}\right] p(0) \tag{14.2.8}$$

The expected time in the system is

$$E[T] = \frac{E[X]}{\lambda} \tag{14.2.9}$$

Finally, the expected time in the queue is

$$E[T_q] = \frac{E[L_q]}{\lambda} \tag{14.2.10}$$

Example 14.2

Solve Example 14.1 again, assuming that an identical service counter is added in parallel to the existing one.

Solution The new arrangement is also a FIFO system but one that provides two service counters (i.e., $N = 2$). This means that customers line up in single file and up to two customers can be served simultaneously. The ratio of the arrival to the service rate of a single counter is $r = 0.75$, as before.

(a) The percent of time that *both* tellers are idle is given by Eq. 14.2.6a:

$$p(0) = \left[\left(\frac{0.75^0}{0!} + \frac{0.75^1}{1!}\right) + \frac{0.75^2}{1!(2-0.75)}\right]^{-1}$$

$$= 0.455 \text{ or } 45.5\% \text{ of the time}$$

The percent of the time that only one of the two tellers is idle is given by the probability that only one customer is in the system. By Eq. 14.2.6b,

$$p(1) = \frac{0.75}{2!}(0.455) = 0.171 \quad \text{or} \quad 17.1\% \text{ of the time}$$

Thus 45.5 + 17.1 = 62.6% of the time either one or both of the tellers are idle.

(b) The probability that five customers are in the system is given by Eq. 14.2.6c:

$$p(5) = \frac{0.75^5}{2!\,2^{(5-2)}}(0.455) = 0.007 \quad \text{or} \quad 0.7\% \text{ of the time}$$

(c) Equation 14.2.7 gives the average number of customers in the system:

$$E[X] = 0.75 + \left[\frac{0.75^3}{1!(2-0.75)^2}\right](0.455) = 0.873 \text{ customers}$$

(d) According to Eqs. 14.2.7 and 14.2.8, the relationship between the average queue length and the average number of customers in the system is

$$E[L_q] = E[X] - r$$

Hence

$$E[L_q] = 0.873 - 0.75 = 0.123 \text{ customer}$$

(e) By Eq. 14.2.9, the average time each customer spends in the system is

$$E[T] = \frac{0.873}{15} = 0.058 \text{ h} \qquad \text{or} \qquad 3.49 \text{ min}$$

Discussion The new arrangement greatly reduces the queue length from 2.25 to 0.123 customers and the average waiting time from 12 to 3.49 min, but at the expense of having either one or both of the tellers idle 62.6% of the time as compared to 25% of the time for the single-server situation of Example 14.1. Thus there exists a trade-off between the customers' convenience and the cost of running the system. The equations governing the two simple queuing systems just examined clearly show that the mathematical complexity of queuing models increases rapidly. In fact satisfactory analytical models are unavailable for many problems, especially for those that involve priority scheduling. The next section discusses the elements of digital computer simulation, a numerical technique that can be applied to the investigation and design of complex systems, including queuing systems.

14.3 COMPUTER SIMULATION

14.3.1 Background

The study of complex systems that cannot be sufficiently simplified to be amenable to analytical solution requires alternative methods; the use of simulation models is one possibility. A successful simulation model is an abstraction of a real system that retains the system's essential aspects. The model can be used either to enhance the understanding of how the system works or to investigate the potential effects of proposed modifications to the system. Being an abstraction of a real system, a model cannot be identical to it in all respects. Consequently a model is employed when direct experimentation with a real system is impossible, too costly, or unsafe. To be useful, any model of a system must realistically represent the system. Although more involved, the steps of model calibration and validation are very important, but a detailed discussion of these topics is beyond the scope of this book.

Simulation models can be either deterministic or probabilistic. In addition, simulation models can be either *physical* or *mathematical*. Vehicle crash tests using anthropomorphic dummies are examples of physical models. A model of the Mississippi River system used by the U.S. Corps of Engineers at Vicksburg, MS, is another example. This section is concerned with mathematical models.

14.3.2 Monte Carlo Simulation

Monte Carlo simulation employs an artificial probabilistic experiment (model), the repeated application of which leads to an approximation of the outcome of a system or process. The basic idea of this method is illustrated next.

Consider the now familiar process of a sequence of independent Bernoulli trials, specifically the tossing of a fair coin. The authors repeated a Bernoulli sequence of ten tosses four times and obtained the following results:

1. $T\quad T\quad T\quad H\quad T\quad H\quad H\quad H\quad H\quad T$
2. $T\quad H\quad H\quad T\quad H\quad H\quad H\quad H\quad H\quad H$
3. $T\quad H\quad T\quad T\quad T\quad T\quad H\quad H\quad H\quad H$
4. $T\quad H\quad H\quad T\quad T\quad T\quad T\quad H\quad T\quad T$

Without any knowledge of the underlying process, these four *realizations* of the process may seem to be totally random. Yet it is known in this case that each single outcome listed was the result of a Bernoulli trial with $p = q = 0.5$. In fact the frequencies of heads and tails in all 40 trials taken together are nearly equal, as would be expected in the long run.

Chapter 13 discussed various distributions that may be used to calculate the probability of compound events, such as the probability of getting x successes in n trials and the probability of obtaining the first success on the xth trial. By contrast, a Monte Carlo model of the previous experiment is intended to produce sequences of outcomes (i.e., realizations) that are consistent with the underlying process, in this case a series of independent Bernoulli trials. If desired, the long-run probabilities of various events can be approximated by analyzing the results of the model for a large number of repetitions. To understand how the technique works, consider the cumulative probability distribution of the coin toss experiment shown in Fig. 14.3.1. The horizontal axis shows the two possible outcomes of each trial, labeled $X = H$ and $X = T$ for head and tail, respectively. The vertical axis represents the cumulative probability $P[X \le x]$. The difference in the values of the cumulative function for two adjacent outcomes is the probability of the second outcome. Consequently the range from 0 to 1 on the vertical axis has been divided according to the probabilities of the outcomes of the experiment under discussion. For a sequence of independent and uniformly distributed numbers in the range between 0 and 1 the long-run frequencies with which these *random numbers* would fall within each segment of the vertical axis would be proportional to the probability of the corresponding outcome.

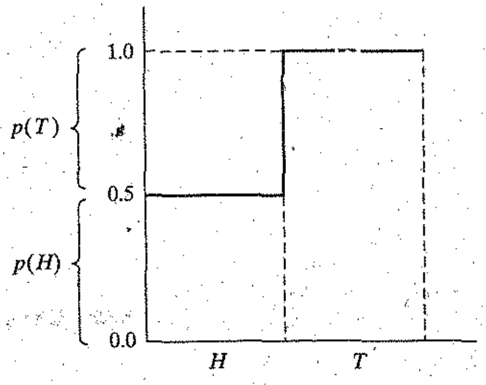


Figure 14.3.1 Generation of discrete outcomes.

Example 14.3

It is known that 15% of the vehicles approaching an intersection will turn left, 60% will go straight through, and the rest will turn right. Construct the corresponding cumulative distribution and translate the following ten *random numbers* to outcomes of this process: 0.5954, 0.4501, 0.2590, 0.7081, 0.1405, 0.9740, 0.8676, 0.2729, 0.4474, 0.0166.

Solution On the cumulative distribution for each trial (i.e., approaching vehicle) shown the ranges on the vertical axis corresponding to the left, right, and through movements are {0.00, 0.15}, {0.15, 0.40}, and {0.40, 1.00}, respectively (see Fig. 14.3.2). The first random number (0.5954) falls in the range corresponding to a through movement. Thus the first *simulated* vehicle is going straight ahead. Continuing with the remaining random numbers, the following results are obtained:

Vehicle number	Movement
1	Through
2	Through
3	Right
4	Through
5	Left
6	Through
7	Through
8	Right
9	Through
10	Left

Discussion The coding scheme selected to represent the three possible outcomes was arbitrary in this case because there exists no natural ordering of the acts of turning left, turning

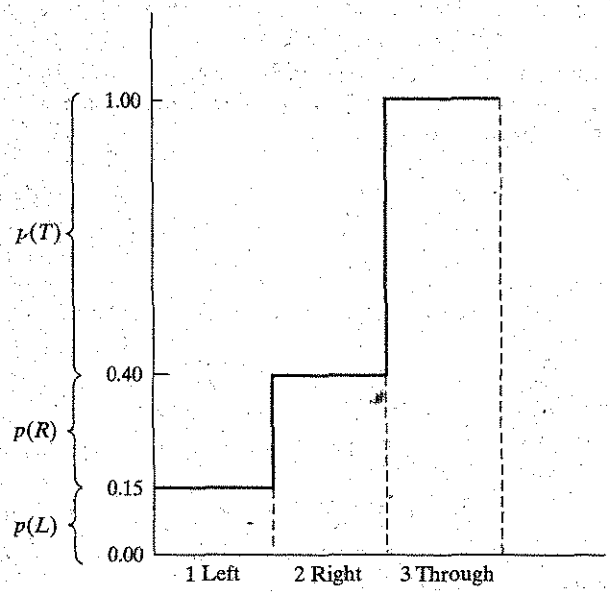


Figure 14.3.2 Turning movements.

right, and going straight through. Had a different coding scheme been chosen, the generated sequence of approaching vehicles would not have been identical to the one shown in this example. However, both would be consistent with the underlying distribution.

14.3.3 Simulation of Outcomes of a Continuous Random Variable

A similar method can be used to generate a sequence of outcomes for a continuous random variable. Figure 14.3.3 illustrates how a random number in the range of 0.0 to 1.0 can be transformed to a particular outcome x. The transformation entails equating the cumulative distribution to a random number R_N and solving the resulting equation for x. For example, consider the cumulative negative exponential distribution

$$F(x) = 1 - e^{-ax}$$

Equating $F(x)$ to a random number in the range {0.0, 1.0} and solving for x, we obtain

$$x = -\frac{1}{a}\ln(1 - R_N) \qquad (14.3.1)$$

Since R_N is a uniformly distributed random number, its complement $(1 - R_N)$ is also a uniformly distributed random number. For this reason it is simpler to use the following equation:

$$x = -\frac{1}{a}\ln R_N \qquad (14.3.2)$$

Example 14.4

Transform the first five of the random numbers given in Example 14.3 to a sequence of vehicular headways, assuming that the average headway is 6.0 s.

Solution In Chapter 13 the parameter a of the negative exponential was shown to be the reciprocal of its mean value, in this case $a = \frac{1}{6}$. For the purpose of illustration both Eqs. 14.3.1 and 14.3.2 are applied to this problem as follows:

		Headway (s)	
Vehicle	R_N	(Eq. 14.3.1)	(Eq. 14.3.2)
1	0.5954	5.43	3.11
2	0.4501	3.59	4.79
3	0.2590	1.80	8.11
4	0.7081	7.39	2.07
5	0.1405	0.91	11.78

Discussion These two sequences of headways conform to the same underlying distribution. Hence they are two realizations of the same process. If needed, more or longer realizations may be produced by using different lists of random numbers.

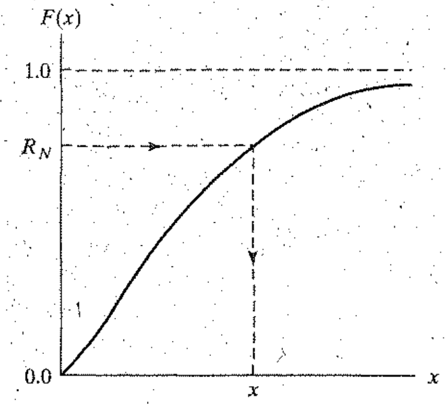


Figure 14.3.3 Generation of continuous random variable outcomes.

14.3.4 Generation of Random Numbers

By definition, true random numbers are independent and uniformly distributed. The independence property means that a sequence of random numbers does not follow any systematic pattern. The fact that they are drawn from a uniform distribution implies that any number in the appropriate range is equally likely to be drawn. Theoretically the uniform distribution of random numbers is continuous. However, in most practical applications, the numbers are limited to the number of significant figures required by the problem at hand. Many sophisticated methods of random number generation have been devised. For example, the RAND Corporation has published, for use in scientific applications, a sequence of 1 million random digits that were generated by periodically sampling a random electronic noise [14.4].

A simple method for obtaining random digits is the *top hat method:* Ten cards, each bearing one of the ten numbers from 0 to 9, are thoroughly mixed in a receptacle. A card is drawn, its number is recorded, and the card is placed back in the receptacle; the cards are again mixed in preparation for the drawing of the next number. If random numbers within the range from 0 to 1 are desired with, say, five significant figures, the digits drawn from the hat are grouped in ordered sets of five and the decimal point is placed in front of each group.

Although the top hat method can produce a series of as many true random numbers as desired, it is a time-consuming and inefficient method, especially if the random numbers are required at the speed that a computer processes the instructions constituting a simulation program. For this reason it is necessary to devise rapid generation methods based on carefully developed computer algorithms [14.5–14.8]. It should be noted that the numbers generated in this way are not truly random since they obey the rules or pattern of the routine used in their generation. For this reason they are referred to as *pseudorandom* numbers. Nevertheless, the better of the pseudorandom generators are considered to be adequate for most engineering applications.

The *middle-square method* suggested by von Newmann is one of the simplest pseudorandom number generators available. An initial number (the *seed*) is squared and the middle digits of the result are taken to represent the first pseudorandom number, which becomes the seed for the generation of the next number. The sequence of random numbers given in Example 14.3 was generated by this method, using the number 3549 as the seed. Note that the square of the seed is equal to 12595401, of which the middle part constitutes the four digits given as the first random number. A problem with this method is that it eventually tends to degenerate to zeros.

14.3.5 The Simulation Model

The generation of random deviates by the Monte Carlo method is only a part of a larger simulation model. The heart of the model consists of a computer program that imitates the behavior of the system over time. For example, suppose that it is desired to simulate a signalized intersection. The intersection (system) involves (1) certain *entities* (e.g., arriving vehicles, the signal, or the traffic lanes provided), each having several essential *attributes* (e.g., the arrival time and movement desire of a vehicle, the cycle pattern of the signal, the permitted use of lanes); and (2) a set of rules that govern the *interactions* between entities (e.g., left-turning vehicles must use an exclusive lane, or left-turning vehicles must wait for adequate gaps in the opposing traffic). In this example the headways between vehicular arrivals may be simulated by the method of Example 14.4 for each intersection approach, and the turning desire of each arriving vehicle may be simulated by the method of Example 14.3. The simulation program would then process these arrivals through the intersection according to the rules of interaction and collect data on the *measures of performance* that are relevant to the problem at hand, such as vehicular delays or queue lengths. To simulate the passage of time, the model establishes a *simulation clock*, which is advanced periodically as the system changes from one *state* to another. The following two examples illustrate two ways by which the simulation clock may be advanced. The first, *interval-oriented simulation*, updates the clock by a constant time interval, and the second, *event-oriented simulation*, advances the clock to the next event that triggers a change in the system's state.

Example 14.5: Interval-Oriented Simulation

Vehicles arrive at a parking garage at an average rate of 30 veh/h. Assuming that the arrival pattern is described by the Poisson distribution and that the arriving vehicles are served by a single attendant at a *constant* rate of one vehicle per 2.5 min, use the random numbers of Examples 14.3 and 14.4 to simulate a sequence of ten 5-min intervals. Assume that at the start (i.e., clock = 0) the system is empty.

Solution The state of the system may be described as the number of vehicles in the queue awaiting to be served. The clock is advanced by 5-min intervals, and the model generates the number of arrivals during each interval according to the Poisson distribution. A maximum of two vehicles per interval are processed and the (nonnegative) difference between the number of vehicles that arrived and the number of vehicles processed represents the number of vehicles that join the queue at the end of each 5-min interval.

To generate vehicular arrivals, the cumulative Poisson distribution with a mean value of 2.5 vehicles per 5-min period is first calculated as follows:

x	$p(x) = P[X = x]$		$P(x) = P[X \le x]$
0	0.08		$0.08\}p(0)$
1	0.21		$0.29\}p(1)$
2	0.26		$0.55\}p(2)$
3	0.21		$0.76\}p(3)$
4	0.13		$0.89\}p(4)$
5	0.07		$0.96\}p(5)$
6	0.03	$0.99 \simeq$	$1.00\}p(6)$

An upper limit of six vehicles per period was placed on the distribution as a reasonable approximation, and the ranges in the values of the cumulative distribution corresponding to the seven possible outcomes are noted in the last column of the table.

The simulation entails the generation of the number of vehicular arrivals during each 5-min interval and the processing of up to two vehicles (if present) during the same interval to determine the queue length at the end of the interval. The first random number in the given sequence (i.e., 0.5954; Example 14.3) translates to three arrivals during the first 5-min interval because

$$0.55 \leq 0.5954 \leq 0.76$$

Since two of these arrivals can be served during the interval, a queue consisting of one vehicle will remain at the end of the interval. The following table summarizes the results obtained for the ten consecutive 5-min intervals:

Interval	Arrivals	Departures	Queue length
1	3	2	1
2	2	2	1
3	1	2	0
4	3	2	1
5	1	2	0
6	6	2	4
7	4	2	6
8	1	2	5
9	2	2	5
10	0	2	3

Discussion This admittedly simple simulation illustrates the interval scanning method of advancing the clock. A total of 23 vehicular arrivals were simulated for the first 50 min. Twenty of these vehicles were served, leaving three vehicles in the waiting line at the end of the tenth period. In this particular example the parking-lot attendant was busy all the time since there were always vehicles awaiting service. For simplicity the service time was assumed to be constant. A more realistic model would allow some variability in the number of vehicles that can be served during any 5-min interval. Moreover, the model can be further extended to allow for multiple service channels and service priorities. More complex models necessitate the preparation of a computer program describing the operation of the subject system. Many simulation models have been developed for use in specific contexts including traffic situations (e.g., Refs. 14.9–14.11). Returning to Example 14.5, it should be noted that a limitation of interval-oriented simulation models is the fact that they are oblivious to the detailed behavior of the system *within* each interval. For example, three vehicles were simulated to arrive during the first 5-min interval, but the precise arrival times of these vehicles within the interval are not known. Consequently certain system characteristics (e.g., the average delay per vehicle) can only be approximated by this model.

Example 14.6: Event-Oriented Simulation

Prepare an event-oriented model of the system described in the previous example and apply this model to simulate the first 11 vehicles that enter the system, assuming that the first vehicle arrives at time zero. The model should be able to calculate the time that each vehicle spends in the waiting line and the percent of time that the attendant is idle. If needed, use the same sequence of random numbers as before.

Solution The time that a vehicle spends in the queue is given by the time interval from the moment it arrives to the moment it begins to be served. The latter coincides with the moment when the servicing of the vehicle ahead has been completed. The arrival time of a vehicle can be computed by adding the headway between the previous arrival and the subject vehicle to the arrival time of the leader. Since the arrival pattern is described by the Poisson distribution, the interarrival times (i.e., headways) are described by the negative exponential. They can be generated by applying the procedure of Example 14.4 to transform a sequence of random numbers to a sequence of headways. Because in this case the arrival rate is 30 veh/h, the average headway is 2 min.

The first vehicle is assumed to arrive at time zero. It will receive the attention of the attendant immediately. Hence it will spend no time in the waiting line. Given a constant service time of 2.5 min, the first vehicle will be processed at clock time $0 + 2.5 = 2.5$ min and the attendant will be busy during this time. The headway between the first and the second vehicles can be generated by Eq. 14.3.2 using the first random number (0.5954). Thus

$$x = -2.0 \ln(0.5954) = 1.04 \text{ min}$$

The second vehicle will arrive 1.04 min after the first, which arrived at time zero. But the servicing of the first vehicle will not be finished until clock time 2.5 min. Hence the second vehicle must wait in line until then, for a total of $2.5 - 1.04 = 1.46$ min. The following table summarizes the results of the simulation for the first 11 vehicles to enter the system.

Vehicle number	Headway (min)	Arrival time	Service start	Delay (min)	Service finish	Idle time
1	—	0.00	0.00	0.00	2.50	0.00
2	1.04	1.04	2.50	1.46	5.00	0.00
3	1.60	2.64	5.00	2.36	7.50	0.00
4	2.70	5.34	7.50	2.16	10.00	0.00
5	0.69	6.03	10.00	3.97	12.50	0.00
6	3.93	9.96	12.50	2.54	15.00	0.00
7	0.05	10.01	15.00	4.99	17.50	0.00
8	0.28	10.29	17.50	7.21	20.00	0.00
9	2.60	12.89	20.00	7.11	22.50	0.00
10	1.61	14.50	22.50	8.00	25.00	0.00
11	8.20	22.70	25.00	2.30	27.50	0.00

Discussion This simulation model differs from that of the previous example in that it advanced the clock to the next significant occurrence, that is, the arrival time, the start of servicing, or the finish of servicing of each vehicle. The clock began at time zero when the first vehicle arrived. The next event was the arrival of the second vehicle 1.04 min later. This was followed by the finish of the servicing of the first vehicle at clock = 2.50 min. The next event was the arrival of the third vehicle at clock = 2.64 min, and so forth. A total of 11 vehicles were examined after ten iterations of the model. By contrast, the first ten iterations of the alternative model of Example 14.5 covered 23 vehicles. However, that model was not as detailed as the present model. This comparison stresses an important point: Often the analyst has a choice between alternative models of the same system. The choice of model involves a balance between the degree of detail required and the available resources.

Consistent with the results of the interval-oriented model of Example 14.5 are the results of the event-oriented model of the present example, which show that the attendant was continuously busy. The average time that each vehicle spent in the waiting line can be computed by dividing the sum of the delays shown in column 5 of the given table by the total number of vehicles to obtain $42.10/11 = 3.83$ min/veh. A total of 11 vehicles arrived during the 22.70 min of simulated time. This translates to 29.07 veh/h, which is close to the stipulated 30 veh/h. Considering the limited number of vehicles simulated, such close agreement is surprising. Normally larger deviations between the two values would be tolerated for such a small sample size.

14.4 SUMMARY

Queuing, or waiting-line, models were presented, and the analytical solutions to FIFO systems with Poisson arrivals and negative exponential service times were presented. The basic elements of the powerful numerical technique of computer simulation were presented. These included the generation of random and pseudorandom numbers, the transformation of these numbers to probable outcomes of underlying processes described by cumulative probability distributions, and the use of these realizations to follow the changes in the state of the simulated system to aid in assessing its likely behavior in terms of applicable measures of performance via simulation models.

EXERCISES

1. Airplanes arrive at an airport area at an average rate of six per hour according to the Poisson distribution. The airport control tower processes airplanes in their order of arrival. Assuming that the service time is negative exponential and that the service rate is ten landings per hour, calculate (a) the average number of airplanes in the system (i.e., being served and stacked), (b) the average number of airplanes awaiting clearance to land, (c) the average time spent in the system, and (d) the average time an airplane is in the queue.
2. A turnpike toll area contains four toll booths arranged in parallel. The arriving vehicles conform to the Poisson distribution, with an average headway of 12 s. Assuming that the average service time is 5 s, the service time is negative exponential, and the queue discipline is FIFO, find the average queue length and the expected time in the system if (a) two of the booths are in operation and (b) only one booth is open.
3. Use the random numbers provided in Example 14.3 to simulate, in two different ways, five tosses of two dice. Discuss your results.
4. A wheel of fortune is divided into ten equal sectors numbered from 1 to 10. Devise a Monte Carlo simulation of this roulette and produce the result of (a) five spins for which each outcome is a digit from 1 to 10 and (b) ten spins, assuming that the outcome is given by the following three events:

$$A: x < 3$$
$$B: 3 \leq x \leq 7$$
$$C: x > 7$$

5. Extend the simulation model of Example 14.5 to allow for exponentially distributed service time with an *average* of 2.5 min.
6. Expand Example 14.6 to allow for two service channels and a FIFO queue discipline.

7. Extend Example 14.6 to allow for two service lines and a queue discipline stating that the next arrival chooses the shorter of the two lines 80% of the time. When the two queues are equal, the choice of line is made on a 50/50 basis.
8. Construct a simulation model of Exercise 1 incorporating the following modifications: (a) commercial flights constitute 30% of the arrivals and are given priority over general aviation flights, (b) the average service times for commercial and general aviation operations are 8 and 6 min, respectively, and (c) your model should include as many measures of performance as practicable.
9. Computerize any of Exercises 3 through 8.

REFERENCES

14.1 WOLLF, R., *Stochastic Modeling and the Theory of Queues,* Prentice-Hall, Englewood Cliffs, NJ, 1989.

14.2 MORSE, P. M., *Queues, Inventory and Maintenance,* John Wiley, New York, 1958.

14.3 GERLOUGH, D. L., and M. J. HUBER, *Traffic Flow Theory: A Monograph,* Special Report 165, Transportation Research Board, National Research Council, Washington, DC, 1975.

14.4 RAND CORPORATION, *A Million Random Digits with 100,000 Normal Deviates,* Free Press, New York, 1955.

14.5 GALLER, B. A., *The Language of Computers,* McGraw-Hill, New York, 1962.

14.6 GORDON, G., *System Simulation,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1978.

14.7 GRAYBEAL, W., and U. W. POOCH, *Simulation: Principles and Methods,* Winthrop Publishers, Cambridge, MA, 1980.

14.8 JANNSON, B., *Random Number Generators,* Almqvist & Wiksell, Stockholm, Sweden, 1966.

14.9 PAPACOSTAS, C. S., "Capacity Characteristics of Downtown Bus Streets," *Transportation Quarterly,* 36, 4 (1982): 617–630.

14.10 TRANSPORTATION RESEARCH BOARD, *The Application of Traffic Simulation Models,* Special Report 194, National Research Council, Washington, DC, 1981.

14.11 WANG, Y., and P. PREVEDOUROS, "Comparison of CORSIM, INTEGRATION and WATSIM in Replicating Volumes and Speeds on Three Small Networks," *Transportation Research Record* 1644 (1998): 80–92.

15

Transportation Software

15.1 INTRODUCTION

Computers and software are involved in many aspects of transportation. Examples include highway design with computer-aided design (CAD), taxi, metro bus, and handicapped van dispatch with computer-aided dispatch (also CAD), bus, train, boat, and airplane scheduling, traffic signal analysis (isolated, arterial, and grid systems), and so forth.

In broad terms software can be classified based on its intended purpose as planning, engineering design, and operations. The latter can be on-line (for the real-time control of processes) or off-line for pre- or post-implementation analyses. Transportation-specific software is often complemented or interfaced with office, statistical, and mathematical software for record keeping, summary development, statistical analysis, and the estimation of models (e.g., estimation of a choice model for use in a planning model).

This chapter presents the following four categories of software*:

- Geographic information systems
- Traffic simulation software
- Traffic capacity software
- Planning software

These four categories were chosen because they are directly related with transportation engineering and are consistent with the subjects covered in this textbook. The software presented in this chapter is summarized in Table 15.1.1. Most of these software are used widely in traffic engineering and transportation planning applications and lend themselves to class-

*Land-use models were presented in Section 7.5.3. Air pollution models (MOBILE, EMFAC, and CALINE) and the FHWA traffic noise model (TNM) were presented in Chapter 10.

TABLE 15.1.1 Sample Transportation Planning and Traffic Engineering Software

- Geographic Information Systems in Transportation
  - ARC/INFO, INTERGRAPH, GisPlus
- Traffic simulation software
  - Urban Street Networks
    - Microscopic: SimTraffic, TRAF-NETSIM
    - Macroscopic: EVIPAS, NETFLO, PASSER, SYNCHRO, TRANSYT
    - Mesoscopic: CONTRAM, SATURN
  - Freeways and Freeway Corridors
    - Microscopic: INTRAS, FRESIM
    - Macroscopic: CORQ, FREQ, FREFLO, KRONOS
  - Mixed Networks
    - Microscopic: AIMSUN, CORSIM, INTEGRATION, PARAMICS SCOT, WATSim
    - Macroscopic: CORFLO
- Capacity analysis software
  - HCS, SIDRA, EZ-SIGNALS, HCM/Cinema, SIGNAL94
- Planning software
  - EMME/2, QRS II, TRANPLAN, MINUTP, TP+, TRANSCAD, TRANSIMS

room demonstration or incorporation into undergraduate and graduate courses: In each category popular and/or historically significant software are presented. The reader should contact the transportation software clearinghouses at the universities of Florida (McTrans) and Kansas (PC-Trans) for a complete list of software. Other sources include the extensive collection of traffic models compiled by researchers at the University of Leeds [15.1].

15.2 GEOGRAPHIC INFORMATION SYSTEMS

15.2.1 GIS Fundamentals

Regional transportation planning is a data-intensive activity: It requires a vast amount of information about the type, intensity, and geographical distribution of land uses and population characteristics within a region. In addition, it requires the specification of the existing and proposed multimodal transportation networks. The collection, management, and use of such data are expensive, labor-intensive, and time-consuming tasks that need the participation of many organizations and individuals. Computerized methods in general and geographic information system (GIS) technology in particular are well suited to the management and sharing of transportation-related data.

GIS has been defined as "an information technology composed of hardware, software, and data used to gather, store, edit, display, and analyze geographic information [15.2]. GIS is suitable to transportation systems because transportation systems have a strong geographical (spatial) aspect. In general a GIS retains data in several "layers." For example, a GIS layer may be allotted to land-use information, another to soil conditions, a third to environmentally sensitive areas, a fourth to the transportation network, and so forth. All of these layers use the same geographic coordinate system and make possible the simultaneous consideration of data from multiple layers. For example, sensitive environmental areas from one layer, soil conditions from a second layer, and the layer containing the existing highway network may be superimposed to help in the identification of possible network extensions

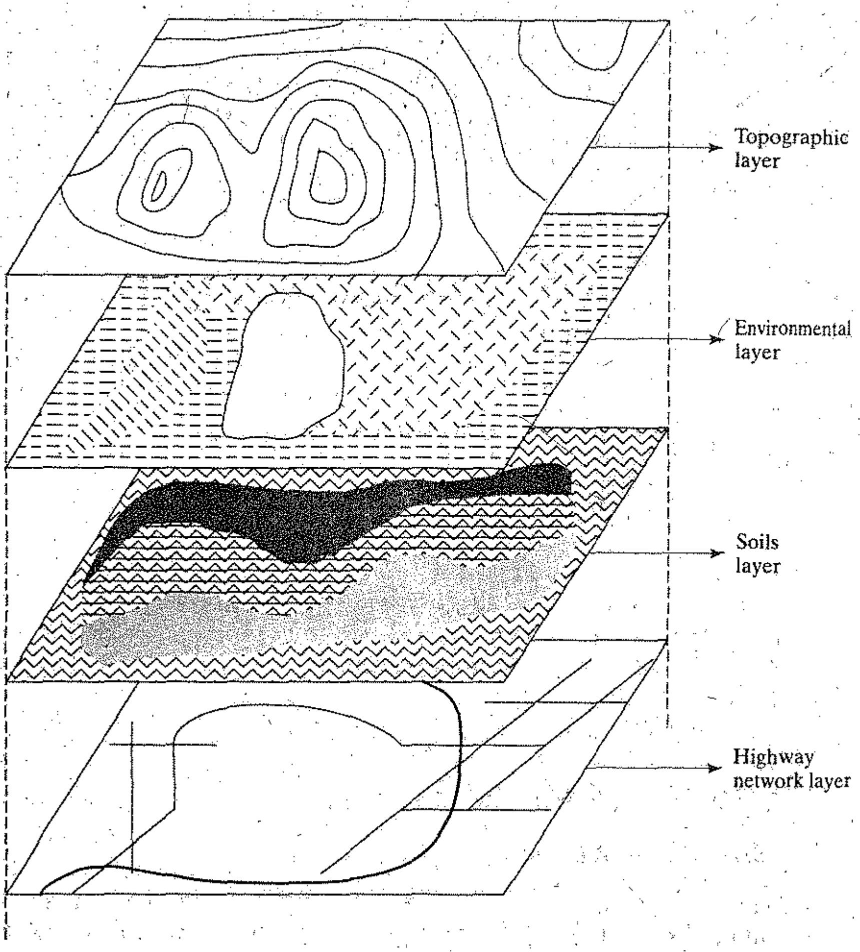


Figure 15.2.1 Typical multilayer structure of GPS.

that meet environmental and geotechnical requirements as in Fig. 15.2.1. GIS uses computer graphics and mapping that enhance the visualization of complex spatial data.

GIS technology combines the principles of topology and database management. Topology refers to the representation of features in terms of their location, shape, and spatial relationship. The features (or entities) may be described as points, lines, and polygons. In a particular application points may be used to represent the locations of traffic accidents, in another application they may represent zonal centroids, and so on. Lines are used to describe elements, such as roadway segments, utility lines, and the like. Polygons represent real features, such as traffic analysis zones, political jurisdictions, land subdivisions, and rainfall intensity zones. The principles of topology are applied to define both the location of entities and their spatial relationships (e.g., line *A* lies within polygon *P*, or polygons *C* and *D* are adjacent to each other). Other powerful spatial operations include topological *overlay* and *buffering*. Topological overlays can be used to develop new information based on existing data. For example, a polygon layer showing the land-use classifications (e.g., residential, agricultural, industrial) may be combined with a polygon layer containing zoning designations of land parcels to create a new layer showing a set of new polygons with attributes that are derived from the two constituent layers. Buffering is the spatial operation where a zone of a specified width is automatically generated around existing features. For example, a buffer of 10 mi may be generated around the location of a retail shop to show its likely market area, a buffer of 200 ft from the centerline of a stream may be drawn to identify areas within which development may be prohibited, and so forth.

The topological specification of transportation networks involves some special requirements [15.3]. One such requirement is the need to employ a *linear referencing system* (LRS) in order to identify locations that lie precisely on transportation network elements. The LRS has been in use traditionally on highway facilities by the use of *mile markers* or *mileposts* designating the distance along a route from an established point of origin. In turn these mileposts are used to reference other locations. For example, a crash site may be identified as located "0.55 m north of milepost number 25." It is of interest to note that, over time, physical mileposts may become "historical" in the sense that any realignment of the roadway between the point of origin and the milepost would alter the actual distance along the route between these two points. At any rate it should be clear that when using a GIS-based or any other digital transportation network, a need arises to reference locations (either points or roadway segments) along the network. The location of a crash site or the georeferencing of a street address are examples of locating points along the network, whereas the specification of a stretch of roadway that may require pavement rehabilitation is an example of referencing a roadway segment. A special technique that is used to identify such locations by reference without modifying the underlying digitized network is known as *dynamic segmentation*. In this connection it is important to point out that the resolution and accuracy with which network links are digitized has an effect on the precision with which map locations represent their actual counterparts. Typically roadway *links* or *arcs* are identified by their beginning and ending nodes, whereas the link shape between these nodes is captured by a series of *shape points* or *vertices* that are kept internally by the computer system for this purpose. The accuracy of representation is a function of the number of vertices (known as *densification*) per unit length. In other words the trajectory of a digitized link is only an approximation of its real-world counterpart. Thus points that lie on the real-world network that are specified by their planar (x, y) coordinates may appear to be off the network in its digitized version. Linear referencing on the other hand obviates this problem.

The nonspatial (or *thematic*) attributes of the features included in a GIS are usually described in a relational database. Examples of thematic attributes include the population of a traffic analysis zone; the length, capacity, and free-flow speed of a highway link; or the severity and property damage associated with a traffic accident. Thus a GIS system may be thought of as an "intelligent" map because, in addition to its ability to represent objects graphically, it "knows" how the objects are related and also associates thematic information with these objects. For this reason GIS technology is finding increasing use in many subject areas, including transportation engineering, planning, and decision making. The latter are usually referred to as GIS-T.

The nonspatial description of objects is commonly implemented through relational database management systems (RDBMS), that is, computer programs that allow the creation, maintenance, and administration of relational databases. Most of these employ the *standard query language* (SQL), an interactive programming language for getting information from and sending information to the database either directly or through an application such as a GIS program. SQL is both an ANSI and an ISO standard. Microsoft's Access and the most elaborate client/server ORACLE system are examples of RDBMS.

Briefly, a relational database consists of a number of interconnected tables (also known as *relations*). Each table consists of a number of records (table rows) and each record contains a number of *attributes* or *fields* (table columns). In the GIS context each record represents a geographical object and the attributes describe that object. For example, if a table's records represent the links found on a transportation network, the attributes of each record would correspond to relevant link characteristics such as the number of lanes, facility type, capacity, and so on. The level of sophistication and complexity of a relational database depends on the specific application. In all cases, however, careful database design and normalization at the start of a project are highly recommended. Database *normalization* (e.g., Ref. 15.4) is the process by which anomalous relationships are eliminated to ensure efficient, nonredundant, and accurate database operations such as inserting, deleting, and updating information. As an example, consider a case where some information about an entity appears on two different tables of a relational database. If updating this information becomes necessary, it must be modified at both locations. Otherwise an inconsistency will be created in the system. One aspect of normalization is to ensure that modifications to the database are effectuated in one step and propagated correctly throughout the database structure.

15.2.2 GIS Products

Since the mid-1980s, there has been an explosion in the use of GIS applications. The question of whether to use a GIS, particularly in spatially related fields such as engineering and planning, is no longer an issue. The difficulty lies in selecting the appropriate GIS product given the plethora of choices available. As with all decisions related to software acquisition, the answer lies in matching an organization's needs with the functionality and cost associated with competing products.

At the top of full GIS functionality (and cost) is the Arc/Info system developed by the Environmental Systems Research Institute (ESRI) of Redlands, CA. This system was developed from the bottom up as a spatial engine employing an efficient method of considering topology based on arcs (hence the name), which can be used to build either a line layer or a polygon layer. The "Info" part of the name refers to the name of the RDBMS originally used by the system. Many municipal and large GIS users have adopted Arc/Info for large-scale system development and application. Another full-functionality product that found favor with state departments of transportation was the MGE system developed by the Integraph Corporation of Huntsville, AL. Its advantage was derived from the fact that it ran on top of the company's computer-aided design (CAD) engine (Microstation), which along with special design modules had been adopted by these large organizations.

Offering varying levels of functionality is a variety of products (some developed or distributed by the two main GIS vendors mentioned earlier). These may be considered as "business class" GIS products, the best of which would be sufficient for all but the most demanding GIS applications in the area of transportation engineering and planning. Many non-GIS transportation software (e.g., planning packages, see Section 15.5) provide linkages to these types of GIS software for the purposes of sharing data displaying the results of their analyses. Among these "business class" GIS products are ArcView and AtlasGIS (both offered by ESRI), MapInfo (by MapInfo Corporation), Maptitude (by Caliber Corporation), and the GeoMedia suite of products offered by Integraph Corporation. The latter represents a leap on the part of the developer to produce a full-functionality GIS package using the Microsoft Windows/NT platforms.

15.2.3 GIS and GPS

Although its applicability ranges far beyond the subject of this section, the Global Positioning System (GPS) merits a brief mention in relation to GIS applications [15.5]. Fully named as the NAVigation System using Time And Ranging (NAVSTAR) Global Positioning System, this system was initially developed in 1973 by the U.S. Department of Defense. The basic idea was to use *trilateration* to determine the location of a GPS receiver antenna using its distance from orbiting satellite vehicles (SVs) at a known time (or *epoch*). The first prototype SV was launched in February 1978. Another ten were placed into orbit between 1978 and 1985. Together these 11 SVs are known as Block I. Their purpose was to help prove the concept and potential applications. By 1982 at least one commercial surveying company was offering GPS services not for navigation but for surveying applications. A year later the National Geodetic Survey (NGS) and the Texas Department of Highways and Public Transportation (SDHPT) purchased several receivers to support geodetic surveys. Production SVs (known as Block II) were placed in orbit between 1989 and 1993.

In 1994 the GPS system was declared to be fully operational. It consists of three major elements: the space segment, the control segment, and the user segments. The space segment is made up of a constellation of 24 SVs arranged in six groups of four. Each group occupies one of six orbital planes inclined to the equatorial plane by 55° and is spaced equally (i.e., at 60° apart) around the equator. At least four (and up to eight) satellites are visible at any given time from almost everywhere on the globe. The control segment includes a master control station located at the Falcon Air Base in Colorado and four tracking stations around the world. Information from the tracking stations is used to compute and upload the precise orbital data of each satellite (known as the "ephemeris"), and clock corrections and other data. The user segment requires GPS receivers and software that use signals transmitted periodically by each satellite to perform navigation, surveying, and other positioning tasks.

The SV signals are composed of two carrier frequencies ($L1$ and $L2$) modulated by two pseudorandom (PRN) binary codes generated by known and published mathematical equations. The two codes are called low accuracy *coarse acquisition* (C/A) code and high accuracy *precise* (P) code. When "antispoofing" is enabled, the P-code is replaced by a classified high accuracy Y-code known only to authorized users. The accuracy of GPS also depends on *selective availability* (SA), the deliberate degrading of the signals.

As of the late 1990s a new network of continuously operating reference stations (CORS) of known locations is being developed to support high accuracy positioning. The introduction of GPS necessitated the definition of a reference ellipsoid with a center at the mass center of the earth. Known as the *World Geodetic System 1984* (WGS 84), it is almost identical to the *Geodetic Reference System 1980* (GRS 80), which was adopted by the

International Union of Geodesy and Geophysics in 1979 and is used by the *North American Datum 1983* (NAD 83). However, the WGS 84 differs significantly from other ellipsoids, such as the Clarke 1866 used by the NAD 27.

The quick and accurate determination of geographical positioning offered by the GPS can enhance the development of GIS applications. It can greatly enhance the task of data collection relating to real-world objects (e.g., traffic signs for a sign inventory GIS and pavement conditions for a pavement management system), as well as more advanced applications and automatic vehicle location, emergency vehicle dispatching, transit system schedule adherence analyses, travel time surveys, and so forth.

15.3 TRAFFIC SIMULATION SOFTWARE*

15.3.1 Traffic Simulation Model Characteristics

Computer simulation is important for the analysis of freeway and urban street systems. Through simulation, transportation specialists can study the formation and dissipation of congestion on roadways, assess the impacts of control strategies, and compare alternative geometric configurations. Over the past three decades a considerable variety of sophisticated computer models that are capable of simulating various traffic operations have been developed. Simulation models have different characteristics: static or dynamic, deterministic or stochastic, microscopic or macroscopic. Each simulation model has its own logic and use limitations, and is applicable to specific components of a transportation system.

Widely used and newly developed traffic simulation models are included in this section. It is emphasized that most traffic simulation models are under continuous improvement. As a result, the text herein is a presentation of features and main attributes of each software and not a critique.† Important issues, such as model selection, data needs, variability and reliability of results, and output analysis as well as simulation limitations are discussed.

15.3.2 Classification

A variety of traffic simulation models have been developed since the 1960s. The simplest model classification may be based on the classification of facilities that the model can analyze. Gibson [15.7] classified simulation models as those for intersections, arterials, urban networks, freeways, and freeway corridors. The need for integrated control strategies has resulted in recent developments of simulation models for integrated freeway/signalized intersection networks. Each of these traffic subsystems, isolated, coordinated, or integrated, has unique problems and objectives.

A common classification method for simulation models is based on the *uncertainty content* that represents the deterministic or stochastic nature of simulation and the *time*

*A special thanks is due to Mr. Yuhao Wang for the assistance he provided in identifying and reviewing traffic simulation software during his Masters study at the University of Hawaii.

†Although software obsolescence occurs rather rapidly, several features of early versions are maintained in the newer versions. For example, a review of FHWA's *Traffic Models Overview Handbook*, dated June 1993, reveals that of the 11 models presented in it, most remain largely unchanged, except for the relaxation of some limitations and the great improvements to their input/output interfaces with Windows, shells, and so on [15.6].

horizon that represents the static or dynamic properties of simulation. A simulation model could be dynamic and stochastic or dynamic and deterministic in nature. Given the traffic characteristics in the real world, simulation models that fall into the static classification do not exist, although simulation based on time-slice static traffic flows is not rare. Traffic simulation models can also be classified into types of *interval scanning* (or time stepping) and *event scanning* based on how often the status of the traffic network is updated and the statistics on traffic performance are collected. When time scanning is used, the state of the traffic system is examined and performance statistics are collected at regular intervals of time. In the event-based models the traffic situation is updated when events of importance to traffic operations occur (e.g., signal turns red).

Much like demand forecasting models, the classification of traffic simulation models is based on the level of aggregation. Microscopic models consider the characteristics of each individual vehicle and its interactions with other vehicles in the traffic stream. Therefore they can simulate traffic operations in detail but usually require extensive inputs and long execution times. Macroscopic models are characterized by continuum fluid representations of traffic flow in terms of aggregate measures, such as flow rate, speed, and density. These models lose detail but gain the ability to deal with large problems within short execution times. Analytical procedures are incorporated into both microscopic and macroscopic models to evaluate existing conditions and to predict performance under different design and control scenarios.

Typical microscopic simulation modeling methods are based on car-following and lane-changing theories that can represent the traffic operations and vehicle/driver behaviors in detail. The car-following theory describes the longitudinal movement of vehicles. The classical car-following approach is quite straightforward, that is, each vehicle attempts to advance at its desired speed while maintaining a safe following distance from the vehicle ahead. The lane-changing theory describes the lateral traffic behavior. This may be considered in terms of a number of *perception thresholds* governing the consideration of the risk of accepting a gap in a neighboring lane. A set of decision rules is used to calculate whether a speed advantage may be obtained if a vehicle were to change lane. Microscopic simulation modeling incorporates queuing analysis, shock-wave analysis, and other analytical techniques. In addition, most microscopic simulation models are stochastic in nature, employing a Monte Carlo process to generate random numbers for representing the driver/vehicle behavior in real traffic conditions.

Macroscopic models model traffic as an aggregate fluid flow. Continuum models, simple or high-order, are usually employed in macroscopic simulation modeling [15.8]. The simple continuum model consists of a continuity equation representing the relationship among the speed, density, and flow-generation rate. The simple continuum model does not consider acceleration and inertia effects and cannot describe nonequilibrium traffic flow dynamics with precision [15.9]. A high-order continuum model takes into account acceleration and inertia effects by using a momentum equation in addition to the continuity equation. This momentum equation accounts for the dynamic speed-density relationships observed in real traffic flow. A well known momentum equation is Payne's equation [15.10], which is employed in FREFLO.

A limited number of simulation models fall into the third category of mesoscopic models. For example, macroscopic models usually do not simulate lane-changing, merging, and diverging behaviors. However, KRONOS, often classified as a macroscopic model, does simulate these behaviors and therefore it could also be a mesoscopic model. On the other hand, INTEGRATION, a microscopic model in the sense that individual vehicle movements are traced through the network, does not explicitly consider the details of vehicle lane-changing and car-following behavior, which is a core attribute of most microscopic simulation models. Instead, it considers the aggregate speed-volume interaction of traffic, which is a typical attribute of macroscopic models.

15.3.3 Traffic Simulation Models

The earliest computer simulation work in highway transportation was the intersection simulation undertaken by the Transport Road Research Laboratory in the United Kingdom in 1951, and the first simulation work in the United States was on the intersection and freeway models developed at UCLA in 1953. The development of simulation models has grown rapidly since then. Gibson [15.7], Van Aerde et al. [15.11], May [15.12], and Sabra and Stockfisch [15.13] reviewed simulation models for intersections, arterial networks, freeways, and freeway corridors up to mid-1995. One year later the "pool" of models was enriched with TSIS/CORSIM, WATSim, and INTEGRATION 2.

15.3.3.1 Urban Street Networks

Urban street traffic systems comprise intersections, grid-based networks, and a variety of complex traffic activities and control strategies, such as parking adjacent to traffic streams, bus blockage, one-way streets, reversible lane operations, and so forth. These systems exclude freeways, expressways, and all types of limited access facilities.

15.3.3.1.1 Microscopic: SimTraffic, TRAF-NETSIM

SimTraffic. This software is an accompaniment to SYNCHRO (see below). It is a microscopic simulation and animation software that produces NETSIM-like animation requiring a lesser user-effort. Trafficware, the developer of SimTraffic reports that most parameters in CORSIM, as derived through several FHWA research projects, are included in SimTraffic [15.14]. SimTraffic animation and simulation can be executed simultaneously, whereas CORSIM requires NETSIM simulation first, followed by TRAFVU animation. SimTraffic does not explicitly model bus routes, bus stops, car parking, and HOV lanes. On the other hand it can handle much larger networks (e.g., 300 versus 100 intersections) and three times as much total volume on a single run (30,000 versus 10,000 vehicles). SimTraffic's capabilities have been upgraded to include simple freeway modeling, which can be combined with an arterial network. The interaction between the two networks is limited because each freeway ramp is modeled as an unsignalized intersection requiring turning movement inputs.

NETSIM. The NETwork SIMulation model was originally called UTCS-1 because its development was supported by the Office of Research of the U.S. Federal Highway Administration (FHWA) as part of the Urban Traffic Control System (UTCS) program. Two earlier models, DYNET and TRANS, were incorporated in the development of UTCS-1 [15.15].

NETSIM is a microscopic, interval-scanning simulation model that is capable of representing complex urban networks, traffic control systems, and vehicle performance characteristics. It is microscopic because each vehicle is treated as an independent entity; it is interval-scanning because the state of the system is computed at regular time intervals, specifically every second. The model traces the trajectory of each vehicle as it progresses through the network. The motion of each vehicle is governed by car-following rules: lane-changing and overtaking behavior, turning movements, and response to the traffic control system. Several of the characteristics of each vehicle are assigned probabilistically using the Monte Carlo method described in Chapter 14. Therefore individual vehicle/driver combinations, vehicle turning movements on new links, and many other behavioral and operational decisions are all represented as random processes.

Being a microscopic simulation model, NETSIM requires a considerable amount of inputs, such as:

1. *Topology of the roadway network.* The network is described by nodes and one-way links. The nodes represent intersections and points where the roadway geometric characteristics change, such as lane drop locations. The links represent one-directional roadway segments between nodes.
2. *Characteristics of each roadway link.* These include the link's length; the link's free-flow speed; and the number and channelization of lanes including full lanes, turning lanes, and turning bays. In addition, the mean values of start-up delays, lost time, and discharge headways at the downstream end of each link are specified. Pedestrian interference at the upstream end of the link is also input. Related links that receive traffic from the subject link are described as illustrated in Fig. 15.3.1.
3. *Traffic control system.* The input stream includes the characteristics of the traffic control system. NETSIM is capable of simulating the operation of stop and yield sign control, pretimed signals, and actuated signal detector/controller combinations for semiactuated and fully actuated operations as described in Chapter 4.
4. *Traffic demand.* The traffic volumes entering and exiting the network and the distribution of turning volumes at each intersection are specified.
5. *Traffic composition.* The composition of the simulated vehicles is given in terms of four fleet components, that is, automobiles, trucks, car pools, and buses. Several types of vehicles may be specified for each fleet component. Each type is described in terms of its operational characteristics, including maximum acceleration, maximum speed,

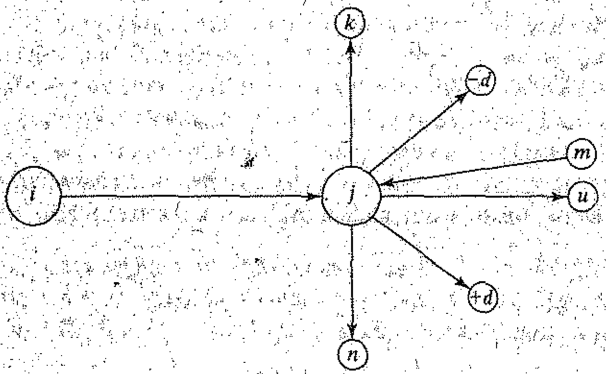


Figure 15.3.1 Possible configurations of links emanating from link (i, j). i, beginning of link; j, end of link; k, left turns; m, through traffic; n, right turns; $\pm d$, diagonal movements; u, movement opposing left turns.

speed-acceleration relationship, headway characteristics, fuel consumption rates, and pollutant emission rates.

6. *Bus operations.* Optionally, the user may specify the operating characteristics of a bus transit system in terms of bus routes, stations, and frequencies of service.

NETSIM provides a wealth of output on a link-specific basis that is aggregated over the entire network. Output MOEs include travel times, total and stopped delays, timing data, queue lengths, signal phase failures, vehicle occupancies, fuel consumption, pollutant emissions, and so on.

Recent microcomputer versions of NETSIM are accompanied with editing programs to facilitate the preparation of the input stream, enhanced graphics to display inputs and outputs, and animation software. The latest version of TRAF-NETSIM uses an identical seed number technique to represent identical traffic streams and to reduce output variability [15.16].

15.3.3.1.2 Macroscopic: EVIPAS, NETFLO, PASSER, SYNCHRO, TRANSYT

EVIPAS. The Enhanced Value Iteration Process Actuated Signals (EVIPAS) software optimizes actuated controller settings for isolated signalized intersections operating with a NEMA or Type 170 controller. Delay, fuel consumption, and several other MOEs can be used as a base for the optimization that can accommodate any combination of user-selected signal settings such as minimum and maximum green, vehicle extension, time before reduction, and so forth. [15.17]

NETFLO. The NETwork traffic FLOw simulation model can simulate the traffic flows at two levels. NETFLO I is a stochastic, event-based model. It moves each vehicle intermittently according to events and moves each vehicle as far downstream as possible in a single move. Although NETFLO I treats each vehicle on the network as an identifiable entity, car-following and lane-changing behaviors are not modeled explicitly. Therefore NETFLO I models traffic at a lower level of detail than NETSIM. NETFLO II is a deterministic, interval-based model. It is essentially a modified TRANSYT without optimization capability. In NETFLO II the traffic stream is represented in the form of movement-specific statistical histograms. NETFLO and FREFLO, a macroscopic freeway model, are combined into the integrated simulation system CORFLO.

PASSER. The Progression Analysis and Signal System Evaluation Routine, PASSER II-90, performs traffic signal optimization on a single arterial street based on bandwidth maximization. PASSER II is a part of the Arterial Analysis Package, which also provides input files for use with TRANSYT-7F. PASSER IV-96 is applicable to networks of arterial streets. PASSER III-90 and III-98 are separate software designed to perform signal optimization on dual-signal diamond interchanges. Up to 15 alternative phase sequences and several interchanges in tandem can be evaluated in one run. All PASSER software are macroscopic. They were developed by the Texas Transportation Institute for the Texas DOT [15.6].

SYNCHRO. SYNCHRO is a traffic signal timing software designed to generate optimal signal timings (cycles, splits, and offsets). A secondary product of the analysis is capacity and performance estimations similar to those in the HCM. SYNCHRO's unique feature is the choice between Webster's delay formula (and its derivative, as it appears in the Highway Capacity Manual) or the percentile delay formula. According to the latter, the simulation models the 90th, 70th, 50th, 30th, and 10th volume percentiles to better account for variations in traffic and capture saturated conditions. Different optimization criteria are used for under- and oversaturated conditions. Network partitioning for better subnetwork cycle length estimation is available. The computed results can be simulated by SimTraffic. Unlike TRANSYT and HCM-based models, SYNCHRO permits detailed input of actuated signals and progression settings [15.18]. An augmented version of SYNCHRO produces input files for PASSER II, TRANSYT-7F, and CORSIM.

Cycle length optimization is based on a simple formula that accounts for negative factors, the sum of which is attempted to be minimized:

$$PI = (D_{\%} + 10S + 20Q + 300U) / 3600 \tag{15.3.1}$$

where

$$PI = \text{performance index}$$
$$D_{\%} = \text{percentile delay}$$
$$S = \text{number of stops}$$
$$Q = \text{number of vehicles in queue}$$
$$U = \text{number of unserved vehicles}$$

The coefficients of 10, 20, and 300 are seconds of penalty.

The coordinability factor (CF) is generated by SYNCHRO to assess whether signals in a network should be coordinated. CF is estimated by Eq. 15.3.2. CF ranges between 0 and 100. Values above 80 indicate a necessity for coordination, whereas values below 20 suggest that coordination is useless or counterproductive.

$$CF = \max(CF_1, CF_2) + AP + AV + AC \tag{15.3.2}$$

where

$CF_1 = 100 - 1.3\,(T - 4)$,
T is the link travel time with $4 < T < 80$ s

$CF_2 = 100\,AT/SS$,
AT is average traffic and
SS is storage space; they are calculated as follows:

$AT = 3600^{-1}\,VC$,
V is the lane group volume and
C is the cycle length

$SS = NDL^{-1}$,
N is the number of lanes in the lane group,
D is the link distance, and
L is the vehicle length

$AP = 10 - 0.55(100 - 0.5(V_{30} + V_{60})V^{-1})$, is the platoon adjustment;
V_{30} are volume arrivals during the busiest 30% of the cycle and
V_{60} are volume arrivals during the busiest 60% of the cycle

$AV = 0.02(V_{\text{Two-Way}} - 700)$ if $V_{\text{Two-Way}} < 1200$

$AV = 0.01(V_{\text{Two-Way}} - 200)$ if $1200 \leq V_{\text{Two-Way}} < 2200$

$AV = 20$ if $V_{\text{Two-Way}} \geq 2200$;
AV is the volume adjustment and
$V_{\text{Two-Way}}$ is the two-way volume of the link

$AC = CI/2$
CI is the cycle increase in seconds; it represents the increase that needs to be imposed on smaller cycles to match the larger cycle in the network to achieve coordination

The inputs and outputs of SYNCHRO have many similarities to those of TRANSYT. SYNCHRO outputs can be microsimulated with SimTraffic for more refined results.

TRANSYT. TRAffic Network StudY Tool's original version was developed by Dennis Robertson at the Transport Road Research Laboratories (UK) in 1967. There is no representation of individual vehicles in TRANSYT, and all calculations are made on the basis of the average flow rates, turning movements, and queues. TRANSYT processes platoons of vehicles, as shown in Fig. 15.3.2. The smooth curves represent the actual process of vehicle dispersion as they move downstream from the intersection after they were released at the beginning of green. The dashed blocks represent the way TRANSYT handles this process. In the actual simulation process the blocks are substantially smaller, and the user can control the size of the simulation steps that essentially control the size of the platoons. In Fig. 15.3.2 the queue gradually dissipates, the speed of vehicles increases and thus, the average headway between successive vehicles elongates (i.e., car-following distance increases as speed increases), thereby causing a substantial decrease of the saturation flow.

TRANSYT-7F can perform plain simulation, which results in the performance of the existing network without any alterations. This output often serves as the base on which improvements are evaluated. In optimization mode TRANSYT-7F utilizes a disutility index (DI) that it attempts to minimize by manipulating signal settings, such as cycle length, green splits, and offsets [15.19]. DI is estimated as follows (it is similar to the performance index used in earlier versions):

$$DI = \sum_{i=1}^{n}\{(w_{di}d_i + kw_{Si}S_i) + U_i(w_{di-1}d_{i-1} + kw_{Si-1}S_{i-1}) + QP\} \qquad (15.3.3)$$

where

d_i = total delay on link i (and $i - 1$) out of n links; it consists of uniform and overflow delay as in HCM 2000 (see Chapter 4)

S_i = the total number of stops on link i (and $i - 1$)

w_{xi} = link-specific weight factors for delay and stops

k = a user-defined parameter determining the importance of stops relative to delays (i.e., greater emphasis on stops corresponds to a greater emphasis on arterial progression)

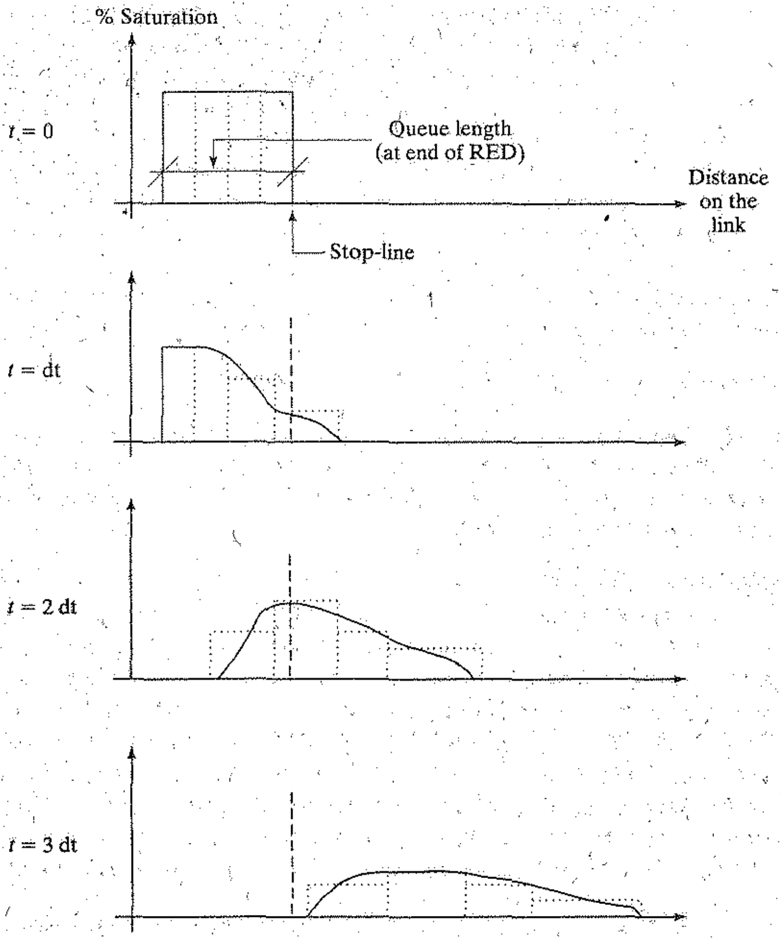


Figure 15.3.2 Example of platoon dispersion as modeled in TRANSYT.

U_i = 1 if link-to-link weights have been established, 0 otherwise

QP = queue penalty estimated as follows:

$$QP = Q\,W_F\,F_i \qquad (15.3.4)$$

where

Q = 1 if the maximum back of queue penalty has been selected by the user, 0 otherwise

W_F = networkwide penalty applied when the link is full

F_i = number of steps during which link i is full

An important variable that emphasizes the importance of arterial progression is progression opportunities (or PROS), which represents the number of successive signalized intersections that can be crossed at the design speed without stopping. Optimization is based on the performance index (*PI*), which can be equal to *DI*, PROS, or various combinations of the two.

TRANSYT uses volumes, saturation flows, average link lengths and speeds, signal phasing, minimum intervals (i.e., minimum duration of green to satisfy pedestrian crossing requirements), and cycle lengths as inputs. Then in an optimization run it calibrates green splits and the progression offsets to minimize delays and to improve progression. At the end of each run it produces output that includes average delays per approach and intersection, intersection and networkwide *DI*s, as well as fuel consumption and emissions estimates.

15.3.3.1.3 Mesoscopic; CONTRAM, SATURN

CONTRAM and SATURN were developed primarily for traffic assignment purposes. They can be used for simulating vehicle routing in a complex traffic system, and their modeling mechanisms have been modified and incorporated into integrated network simulation models such as INTEGRATION.

CONTRAM. The CONtinuous TRaffic Assignment Model is a traffic assignment and simulation model that treats a group of vehicles (called a packet) as a single entity. Thus vehicles that belong to a packet travel along the same minimum cost route and arrive at the same time. CONTRAM determines time-varying link flows and route costs, in terms of given time-varying route inflows, in a dynamic setting. As such, it is entirely different from TRANSYT and NETSIM [15.20]. In CONTRAM traffic demands are expressed as O-D (origin-destination) rates for each given time interval. These O-D rates are converted into an equivalent number of packets, which are assigned to the network at a uniform rate for each time interval. A traffic assignment equilibrium is achieved through iterations in which each packet is removed from the network and reassigned to a new minimum path.

SATURN. The Simulation and Assignment of Traffic in Urban Road Networks model also is a traffic assignment model based on the incorporation of two phases: a detailed simulation phase of intersection delays coupled with an assignment phase that determines the routes taken by O-D trips [15.21]. The complete model is based on an iterative loop between the assignment and simulation phases. Thus the simulation determines flow-delay curves based on a given set of turning movements and feeds them to the assignment. The assignment uses these curves to determine route choice and updated turning movements. These iterations continue until the turning movements reach reasonably stable values.

15.3.3.2 Freeways and Freeway Corridors

A corridor is a roadway system consisting of a few primary longitudinal roadways (freeways or major arterials) carrying a major traffic movement with interconnecting roads that offer the drivers alternative paths to their destinations. Freeway models usually simulate traffic flow on the integrated system of a mainline freeway and its ramps, whereas freeway corridor models can simulate the traffic on a mainline freeway and its ramps as well as on neighboring arterials.

15.3.3.2.1 Microscopic: INTRAS, FRESIM, and Others

INTRAS. The INtegrated TRAffic Simulation model is a stochastic model developed by KLD Associates in the late 1970s and was enhanced continuously through the 1980s [15.12]. It uses a vehicle-specific, time-stepping, detailed lane-changing and car-following logic to represent traffic flow and control of a freeway corridor including the surrounding surface street network (if desired). INTRAS requires detailed geometric and traffic information, including link lengths, lane numbers, location, free-flow speeds, vehicle composition, traffic volumes, O-D data, and so on.

FRESIM. The INTRAS model was reprogrammed by JFT and Associates according to structure design techniques and made more user-friendly. The revised model was called FRESIM and has been incorporated into the TRAF family [15.11]. FRESIM can simulate complex freeway geometrics, such as lane add/drop, inclusion of auxiliary lanes, and variation in slopes, superelevation, and radius of curvature. The model can handle freeway operational features, such as lane-changing, on-ramp metering, and representation of a variety of traffic behaviors in freeway facilities.

CARSIM, WEAVSIM, and FREESIM. CARSIM and WEAVSIM emphasize specific application purposes; both were based on INTRAS. INTRAS, as a general purpose analysis model, had certain limitations; that is, its car-following logic was not capable of realistically simulating the behavior of traffic under stop-and-go conditions on freeways. Therefore a new car-following model named CARSIM (CAR-following SIMulation), was developed to offer additional realistic features and capabilities for simulation of car-following behavior on freeways.

Similarly, because the INTRAS lane-changing logic could not represent the intensive lane-changing maneuvers at the weaving sections of freeways adequately, WEAVSIM was developed specifically for the study of the dynamics of traffic flow at weaving sections. FREESIM is a stochastic model whose logic is based on a rational description of the behavior of the drivers in a freeway lane-closure situation. A set of algorithms were established to simulate driver car-following/lane-changing behavior in response to advance MUTCD warning signs.

15.3.3.2.2 Macroscopic: CORQ, FREQ, FREFLO, KRONOS

CORQ. The CORridor Queuing model developed by Yagar in the early 1970s is a freeway corridor network assignment and simulation model. The corridor consists of a directional freeway, its ramps, major cross streets, and any competing alternative surface streets. Traffic flows are approximated as fluids, and travel times are calculated as simple step functions for both free-flowing and congested conditions. A key element of CORQ is the dynamic assignment technique for allocating time-slice O-D demands to a time-dependent traffic network. However, the travel time relationship is expressed as a static step function of link flows and intersection delays, which is a drawback [15.11]. The time relationship is insensitive to changes in signal timings on parallel arterials. Because CORQ was perhaps the most detailed corridor-level model throughout the 1980s, parts of its modeling approach were modified and incorporated into the design of the integrated network simulation model INTEGRATION.

FREQ. FREQ is a deterministic simulation model for a directional freeway corridor, developed at UC-Berkeley. Since 1968 the FREQ model has been under continuous development and a new version, FREQ10, is presently available. The FREQ10 system contains an entry control model (FREQ10PE) for analyzing ramp metering and an on-freeway priority model (FREQ10PL) for analyzing HOV (high occupancy vehicle) facilities. The simulation model consists of two parts: one for freeways and another for arterials in the corridor. The parallel arterial routes are aggregated and modeled as one after several simplifying assumptions are incorporated into the analysis.

FREFLO. FREFLO, developed by Payne [15.10], simulates traffic flow on freeways using a formulation of aggregate variables based on suitably modified analogies of fluid flows. Initial work with the FREFLO revealed that the model was limited in its ability to simulate realistically congested flow conditions. Many efforts were made to address this problem, including the development of another freeway model, FRECON. FREFLO itself was modified to resolve the difficulties in representing congested conditions and incorporated into TRAF, which allows FREFLO to interface with other models that can simulate the neighboring surface street systems. TRAF's traffic assignment model can provide FREFLO with volume and routing information.

KRONOS. KRONOS, developed by Michalopoulos in the early 1980s, is a freeway simulation model that uses a simple continuum model to represent traffic flow. KRONOS has been continuously enhanced since the inception and several versions are described in the literature (e.g., Ref. [15.22]). Unlike other macroscopic simulation programs, KRONOS explicitly models interrupted flow behaviors such as lane-changing, merging, diverging, weaving, and spillback, which were not taken into account by other macroscopic freeway programs. KRONOS has been applied for evaluating the effectiveness of different freeway design/operational alternatives (e.g., Ref. [15.23]). Updates made by Kwon include routines to handle HOV lanes and traffic responsive ramp metering.

15.3.3.3 Mixed Networks

Earlier methods for simulating mixed freeway-arterial networks combined existing subnetwork models through a traffic assignment subroutine to simulate an integrated system. Several *composite*, *synthetic*, or *fully integrated* simulation models have been developed since the 1980s. Most of the models in this category are able to model complex networks in considerable detail. Consequently a common disadvantage is the extensive requirement for input data and calibration.

15.3.3.3.1 Microscopic: AIMSUN, CORSIM, INTEGRATION, PARAMICS SCOT, WATSim

AIMSUN. The Advanced Interactive Microscopic Simulator for Urban and Nonurban Networks (AIMSUN) is the analytical part of the Generic Environment for Traffic Analysis and Modeling (GETRAM) developed at the Polytechnic University of Cataluna, Spain [15.24]. GETRAM's input processor is a traffic network graphical editor (TEDI). The data from TEDI are analyzed by AIMSUN, which contains interfaces with assignment models such as EMME/2 (presented in Section 15.5). AIMSUN is a microscopic simulator that deals with all types of streets in an integrated fashion. It accommodates traffic flows in either turning movement or O-D matrix form. When turning movements are used, traffic is distributed over the network stochastically; when O-D data are used, vehicles are assigned to specific routes. Several vehicle types can be defined; these are used both for traffic processing purposes (e.g., car-following, lane-changing, gap acceptance, etc.) but also for lane use restrictions (e.g., HOV lanes, bus-only lanes, etc.). AIMSUN's developers have made extensive efforts to accommodate ITS features such as variable message signs, in-vehicle travel guidance, detectors, actuated signal control, ramp metering, and elements of incident management. This sophisticated software has seen little use in the United States.

CORSIM. CORSIM is a combination of NETSIM and FRESIM. The model is capable of simultaneously simulating traffic operations on surface streets as well as on freeways in an integrated fashion. However, within the earlier integrated traffic simulation system (TRAF), the total freeway/urban street systems simulated by the combination of NETSIM and FRESIM could only be called composite networks rather than integrated networks, in terms of the TRAF system characteristics of distinct separation of the assignment and simulation phases of the analysis, independent control strategies in each subnetwork, data transfers between models/modules, and the lack of rerouting capability [15.25]. A traffic assignment model can be run to enter O-D trip information, and two assignment options, system optimal or user equilibrium, can be selected. The assignment results then interface the components of the CORSIM model.

A Windows version of TSIS (Traffic Software Integrated System) [15.26] was released by the FHWA in 1997 to provide an integrated, user-friendly, graphical user interface and environment for running CORSIM. Early evaluations show that CORSIM provides superior animation and that it is a competent software with limitations similar to NETSIM's and FRESIM's [15.27].

INTEGRATION. INTEGRATION was developed in the late 1980s by Van Aerde and Associates. Individual vehicle movements through the network are traced to monitor and control the behavior of vehicles that belong to a certain subpopulation. The model differs from most other microscopic models in that only the aggregate speed-volume interactions of traffic and not the details of a vehicle's lane-changing and car-following behaviors are considered [15.27].

INTEGRATION is routing-based; that is, a vehicle's trip origin, destination, and departure times are specified external to the model. The actual trip path and the arrival times at each link along the path to be derived within the simulation are based on the modeled interactions with other vehicles. Another distinctive feature of INTEGRATION is that it may be the first model that considers the ITS route guidance information in the vehicle routing/rerouting mechanism. User-specified detector location for data collection as well as basic signal optimization at user-defined intervals are additional features. While it provides intuitive graphical capability for viewing vehicles as they move through the network, it provides no graphical user interface for viewing and editing network data.

PARAMICS. PARAMICS is a network-level microsimulator developed in the 1990s in the U.K. It has one base and three optional components. The Modeller is the core simulation component and Processor, Analyser, and Programmer are optional components. Paramics developers claim a unique modeling methodology based on the viewpoint of the driver. This unique approach yields results that have been validated by comparison to longstanding software such as TRANSYT. Paramics' modeling approach makes it suitable for the assessment of detailed effects such as: response to signing, temporary lane closures (e.g., for incidents and road construction), loop detectors linked to variable speed signs, and different speed limits for specific types of vehicles. Paramics requires a Unix or X-Windows operating environment. Only a small number of Paramics applications have occurred outside the U.K. [http://www.paramics-online.com/].

SCOT. The Simulation of COrridor Traffic may be the earliest model for integrated networks [15.28]. The model is the synthesis of UTCS-1 and DAFT. UTCS-1 is the precursor of NETSIM and DAFT is a mesoscopic simulation model for freeways, ramps, and arterials, in which vehicles are grouped into platoons. Therefore SCOT may also be classified as a mesoscopic model. The key design element of SCOT is the interface features between the mesoscopic and microscopic characteristics of the two submodels. Although SCOT appears suitable for simulating area-level traffic networks, the model is no longer supported.

WATSim. The Wide Area Traffic SIMulation model was developed by KLD Associates. It is a stochastic, integrated network simulation model that extends the functionality of TRAF-NETSIM to incorporate both freeway and ramp operations with surface street traffic. Basically the internal processing of TRAF-NETSIM has been modified to create WATSim. Its operational features include those in TRAF-NETSIM plus HOV configurations, light rail vehicles, toll plazas, path tracing, ramp metering, and real-time simulation and animation [15.27, 15.29]. WATSim also includes an interface with a traffic assignment model.

15.3.3.3.2 Macroscopic: CORFLO

CORFLO. CORFLO is a combination of NETFLO I, NETFLO II, and FREFLO models, integrated within the TRAF/TSIS operation environment. FREFLO is used to simulate the traffic on the freeway subnetwork and NETFLO is used for the surface street network. Two important enhancements to CORFLO are the addition of new logic for user-optimal traffic assignment based on simulated link travel time and the introduction of capacity for en-route diversion modeling. Therefore, within the TRAF/TSIS system, the equilibrium traffic assignment model may be used to provide volume and routing information to FREFLO and NETFLO.

15.3.4 Model Selection, Output Variability, and Other Limitations

The preceding sections illustrate the large variety of computer simulation models for analyzing traffic systems. These models have characteristics that may or may not fit a specific application because of their specific attributes, strengths, and weaknesses. Thus selecting a model is an important step toward traffic problem resolution.

Model evaluation and selection depend heavily on the establishment of a set of criteria. These criteria are usually based on the purposes of model application that are considered. Van Aerde et al. [15.11] proposed a general list of criteria for guiding model evaluation: (1) quality of model in terms of traffic engineering theory; (2) quality of program code; (3) user friendliness and documentation; (4) field validation and verification; and (5) availability, implementation, cost, and support. For each of these criteria detailed subcriteria and corresponding weights can be listed depending on the objectives. A list of 16 criteria with their associated priority levels were also identified by Marcus and Krechmer [15.30] as a basis for selecting a candidate set of simulation models. After defining the list of criteria, a literature review and limited testing could be conducted to assist in the model selection process.

Validation for a macroscopic simulation model is usually undertaken at the macroscopic level. For a microscopic model, however, validation is conducted at both a microscopic and a macroscopic level [15.31]. At the microscopic level the attributes of individual vehicles, such as location, time, headway, and speed computed from the simulation model, are compared with those obtained from field data. At the macroscopic level the aggregate parameters, such as the average speed, density, and volume of vehicles, are compared between simulated results and field data.

Most microscopic, stochastic models employ Monte Carlo procedures to generate random numbers for representing the stochastic behavior of individual driver-vehicle combinations. As a result, the simulation output of stochastic simulation models may contain a great deal of variability from one run to the next. The variability in model output can lead to concern about the model's reliability, and the user may have difficulties in analyzing the simulation results under different control strategies.

One method frequently used for reducing the variability of model outputs is the method of *independent replications* [15.32]. This method uses an adequate number of independent runs, based on the desired statistical level of confidence, to get the means and variances for model parameters. The replication method for reducing model variability, in practice, could be very inefficient when many runs with long stabilization periods are needed to generate adequate observations for analyses.

The *batch means* method is a quicker alternative because it works with a single simulation run [15.32]. The total observations generated from a single long simulation are divided into subsequences or batches, and the observations in a given batch are essentially similar to those generated by a short replication, that is, for a 45-min simulation run, three results are obtained with batch sizes of 15 min. This concept of batch means-based variance reduction is appealing, but batch size determination is neither unique nor straightforward.

Two other variance techniques, *common random numbers* and *antithetical varieties,* have been used with TRAF-NETSIM. Both techniques reduce the variability of stochastic models by controlling the random number seeds used to drive the simulation model. These variance reduction techniques can be applied by using the "identical traffic stream" feature of the TRAF-NETSIM, one of the recent enhancements of the model that is intended to assure that the variance in the performance measures is primarily due to the control variables and not to random variation.

Broad limitations of traffic simulation include (1) imperfect simulation of driver behavior, (2) approximate representation of the reality, (3) excessive hardware demands, and (4) inconsistent simulation results [15.1]. Specific limitations of many models include the lack of overtaking, ignorance of pedestrians and two-wheelers, weak transit and HOV-lane treatment, inability to model realistically large networks, incompatible outputs (e.g., no options to format output to fit popular office software), and absence of links to and from GIS, CAD, and planning software. In addition, there are concerns about the effort involved in model calibration and validation, and suspicions that some models are essentially research products that have not been validated over a wide range of conditions.

15.4 CAPACITY SOFTWARE: HCS, SIDRA, AND OTHERS

The capacity analysis software presented in this section are computational models, as compared to the simulation models in the previous section [15.33]. All the software described next can analyze three- and four-leg intersections. Only SIDRA can analyze intersections with five or more legs.

HCS. The Highway Capacity Software is a precise replication of the Highway Capacity Manual (HCM) on a personal computer platform. Most chapters of the manual, including freeway and intersection analyses, are included in the HCS. A full Windows version of HCS became available before the turn of the century.

SIDRA. SIDRA was developed by the Australian Road Research Board. Its signalized and unsignalized intersection analysis is based on the HCM. SIDRA is one of only a few models available that can analyze roundabouts as well as unsignalized intersections; it can do these analyses for both left- and right-hand driving. In addition to delay and level of service, its output includes queue lengths, stop rates, energy consumption, and emissions statistics.

EZ-SIGNALS, HCM/Cinema, and SIGNAL94. EZ-SIGNALS was developed by Viggen Corp. It is a Windows application of the signalized intersections chapter of the HCM. It became available in 1997 and later was incorporated into the HCS. HCM/Cinema was developed by KLD Associates, and it is also a replication of HCM's signalized intersections chapter, with the additional option of executing NETSIM at the single intersection level, which produces a wealth of measures of effectiveness as well as animation. SIGNAL94 was developed by Strong Concepts. It too replicates HCM's signalized intersections analysis along with signal timing optimization and extensive ability to analyze intersection geometry and control alternatives.

15.5 PLANNING SOFTWARE: EMME/2, QRS II, TRANPLAN, MINUTP, TP+, TRANSCAD, TRANSIMS

Planning software automates the four-step process of trip generation, trip distribution, mode choice, and trip assignment. TRANSIMS is a departure from the traditional four-step process as explained later. A large number of inputs is required, such as a full description of the network, the existing traffic and transit volumes, and origin-destination (O-D) tables by zones. Most planning software allows the user to insert and calibrate models for each of the four steps.

The major advantage of all of the planning software is that after inputting the data, complex analyses can be done, and a large number of alternatives can be evaluated in a short time. The software presented next greatly facilitates analyses for new transportation facilities (i.e., new, extended or widened roadway facilities, new public transportation service, expansion of airport or port facilities, etc). They also are useful in analyzing large developments, such as a residential subdivision, a new shopping center, an office park, and so forth. At a grand scale these software permit the analysis of whole metropolitan areas and the evaluation of proposed new installations or alterations.

During the 1970s the FHWA and the Urban Mass Transportation Administration (UMTA; now Federal Transit Administration, FTA) embarked on a major initiative to develop a set of manual and mainframe computer-based tools to help localities implement the demand-forecasting models described in Chapter 8. This package of models was known as the Urban Transportation Planning System (UTPS). With the advent of personal computers during the 1980s, several private vendors began offering their PC versions of the UTPS package. Typically these packages included a graphical user interface to aid in the specification of modal transportation networks and a set of functions and utilities to facilitate the implementation of trip generation, trip distribution, mode choice, and trip assignment modules. Each package included a scripting language permitting users to tailor forecasting procedures to their particular needs. Based on user requirements and advancements in the area of travel demand modeling, these vendors continually offer improvements and added features, such as the capability to analyze HOV facilities, and methods to capture transportation policies (e.g., congestion pricing and parking restrictions). By the end of the 1990s most of these products offered linkages to GIS software as well. The major features of several popular transportation software are presented next.

EMME/2. The major feature of this software package is the incorporation of *multimodal equilibrium:* In all applications both automobile and transit related characteristics can be incorporated simultaneously, which closely approximates real world conditions (i.e., car and transit modes are competing in an urban environment). This property does not only offer the ability to assess the impact of transit services on road networks, but also it aids in the identification of more efficient routes for transit services.

The inputs require a network representation that can be input by coordinates, or it can be digitized directly from maps. On each node and link the pertinent modes, transit lines, turns, and volumes are input. Different types of transit vehicles can be incorporated and a total of 30 modes can be handled. Zone characteristics, such as demand, socioeconomic variables, and travel impedance are inputs. Network or zone data, such as traffic surveys, accident statistics, pavement characteristics, and other custom information, can be incorporated with user-defined attributes. In addition, existing traffic characteristics of roadway or transit links such as volumes, travel times, and speeds can be input for comparisons (i.e., observed versus estimated). The user can specify unlimited expressions (models) representing demand, volume-delay relationships, turn penalties, and mode choice behavior. INRO Consultants, the developers of EMME/2 stress that the package has been designed with a glass-box rather than a black-box philosophy [15.34].

EMME/2 provides a framework for implementing a wide variety of travel demand forecasting: from simple road or transit assignments or the classical four-step model to the implementation of multimodal equilibration procedures that *integrate demand functions* into the assignment procedures (i.e., multimodal traffic assignment under constant or variable demand conditions). Both aggregate and disaggregate input models can be used, in either a *sequential or simultaneous* manner. Other detailed inputs include the explicit modeling of dedicated lanes (i.e., HOV lanes and transit-ways), the incorporation of walk connections between transit lines, and the modeling of people's different perceptions of the various travel time components. Validation and checking procedures are supplied for the continuous identification and control of counterintuitive inputs (i.e., unconnected links, dead-end links with entering traffic, etc.). The program makes automatic corrections if requested, given a set of user-specified criteria.

The main output of EMME/2 is the overall network equilibrium assignment and the presentation of comprehensive results (most in a graphic, interactive way). This output can be used in traffic simulation models for the establishment of signal settings and evaluation of network performance. Updated network characteristics can be fed back to EMME/2 for the derivation of optimum network assignment. Other optional outputs are economic evaluation and traffic impact analysis. Detailed outputs include performance estimates of HOV lanes and truck traffic, location analysis of existing and future transit and roadway facilities, and computation of least-cost paths according to a user-defined cost function.

QRS II. The Quick Response System was developed in the 1970s as a set of manual travel demand analysis techniques intended to provide a means for the quick analysis of policy issues, particularly at a small area level. Detailed descriptions of these techniques for trip generation, trip distribution, modal choice, auto occupancy, time-of-day travel demand distribution, traffic assignment, and capacity analysis are documented in Ref. [15.35]. A standalone gravity model for trip distribution and a simultaneous trip distribution/modal choice model similar to the share model (i.e., Eq. 8.7.2) were included. In 1981 the FHWA released a microcomputer version of the system as Quick Response System I (QRS I). This implementation was capable of handling larger problems than those that could be analyzed via the earlier manual methods, but it proved to be awkward to use. A more flexible system, known as QRS II, was subsequently developed and upgraded at the Center for Urban Transportation Studies of the University of Wisconsin at Milwaukee [15.36].

QRS II features a powerful interactive graphics general network editor (GNE), which can be used to draw and quickly modify highway and transit networks on the computer screen and to display and plot networks and the results of travel demand analyses. All data needed by the system are entered via the GNE. The system is capable of performing both the routine calculations required by the manual techniques described previously and more complex and detailed analyses using the forecasting model combinations as described in this textbook. Algorithms for trip generation, trip distribution, modal choice, highway and transit path finding, traffic assignment, and transit assignment are part of QRS II. Default equations and parameters are provided for the three trip purposes. These may be overridden or adapted to local conditions by experienced users.

The trip generation step estimates the trip productions and trip attractions of each zone as person-trips per day. The embedded default trip production model first calculates the total zonal productions based on average household trip rates. Either household income or household automobile ownership may be selected as the independent variable. The total zonal productions are then split into three purposes (home-based work, home-based nonwork, and non-home-based) according to embedded parameters. Trip attractions are estimated via a multiple linear regression equation.

The trip distribution model is accomplished by a gravity-type of model with options as to the choice of travel time factor. One option is the *power* function expressed by Eq. 8.3.9; the other is an *exponential* form; that is:

$$F = e^{-aW} \tag{15.5.1}$$

The basic modal choice model employed by QRS II splits interzonal demands (by purpose) between highways and transit based on the difference in the disutilities of the two modes, and on the degree of "captivity" associated with the trip-producing zone. The form of the mathematical model is similar to the binomial logit equation. Additional modes can be included by multiple application of the model.

Directional, time-of-day distributions, and vehicle occupancy factors are applied prior to assignment as described in Chapter 8. QRS II is capable of performing all-or-nothing, iterative capacity restrained, and through a feature that averages the results of successive iterations, a true equilibrium traffic assignment. Transit assignment is based on a variation of the probabilistic multipath algorithm developed by Dial [15.37]. QRS II requires the specification of a separate network for each of the modes involved in a multimodal system. Consequently special attention is called for to ensure consistency between modal networks that share network elements.

TRANPLAN, MINUTP and TP+. The TRANPLAN software is a toolbox of more than 40 "functions" that are grouped in the following categories: trip generation, distribution/modal choice models, networks, paths, loading, matrix utilities, reporting, and plotting. A graphics package, the Network Information System (NIS), is available for the development, display, and maintenance of highway and transit networks and related data. The NIS is available in a standard edition and an extended edition. The latter is a rudimentary GIS that provides the capability for defining, displaying, and updating up to 15 types of polygon boundaries that may be used to represent traffic analysis zones, census tracks, and so on. TRANPLAN, however, is a *batch* rather than an interactive system. This means that a control file containing the instructions associated with a particular run of the model must be prepared off-line. The control file specifies the combination of functions to be run and, for each function, the needed inputs, options, and desired outputs. The user may develop certain parts of an application (e.g., trip generation) by other programs and interface the results with TRANPLAN functions. This is necessary when the user wishes to apply a model that is not directly supported by TRANPLAN.

The TRANPLAN trip generation model for the estimation of trip productions and trip attractions is of the multiple-regression form. For trip distribution TRANPLAN supports the gravity model and the Fratar model. The modal choice model is of the diversion-curve type that "splits" interzonal trips between two modes (i.e., automobile versus public transit) based on either the difference of the ratio of the corresponding interzonal impedances. The network functions allow for the definition and updating of highway and transit networks. Subarea networks can be extracted. The highway minimum paths are produced based on a vine algorithm that accounts for turn prohibitions and turn penalties; the transit path algorithm was originally developed by Alan M. Voorhees for the U.S. Department of Housing and Urban Development [15.38]. Supported traffic assignment (or highway loading) models include the free/all-or-nothing method, the capacity restrained algorithm, incremental loading, and a user equilibrium algorithm initially developed in connection with the UTPS. The transit assignment model loads interzonal passenger trips on the minimum paths generated by the transit path functions. Transit trips may be split among competing transit lines in one of three ways: in proportion to the frequencies of the competing lines, equally among them, or only on selected lines.

MINUTP [15.39] is also a library of programs that provides similar capabilities. Its graphical user interface (NETVUE) can display a transportation network for editing, visual inspection, and display of the results. At the heart of MINUTP is a module called MATRIX, which allows the manipulation of trip tables and skim tables. Each cell of an array can be modified by replacement, addition, and weighting factors. Arrays can also be combined in accordance with full mathematical expressions to implement almost any model structure. MATRIX can also be used to estimate trip-length distributions and curves. A variety of modules allow for path building to estimate interzonal impedances by mode and various types of trip assignment are supported.

Both TRANPLAN and MINUTP were written as 16-bit DOS applications. During the mid-1990s the developers of these two packages joined forces to create a Windows-based 32-bit application known as TP+, which combined the best features from each package along with additional enhancements. To allow users of the previous software, TP+ is capable of reading and writing files in the TRANPLAN and MINUTP formats. An improved graphical user interface named Viper is included in the TP+ package.

TransCAD. This package is basically a GIS-T application with augmented abilities for transportation planning because it encompasses zone building and the four-step planning process that can be enriched with user-input choice and assignment models. The package includes a variety of routing and scheduling routines permitting transit routing, hazardous material transportation (e.g., routes with exposure to the least population), and so on [15.40]. TransCAD is a convenient GIS-T for higher-level (aggregate) planning analyses.

TRANSIMS. The primary goal of the Travel Model Improvement Program (TMIP) sponsored by the FHWA is the development of a system of travel forecasting models, TRANSIMS. This Los Alamos National Laboratory-led effort is a considerable departure from the traditional, four-step travel forecasting model. It may be seen as a super-integrated transportation analysis model that begins with a household and commercial activity disaggregation module, which feeds into the intermodal trip planner module. Once trips have been defined, the traffic microsimulation module performs traffic analyses in an integrated fashion with the previous modules and feeds into the environmental simulation module that is designed to accommodate the requirements for the CAAA. A limited test application of TRANSIMS was completed in 1998 using data from the Dallas-Fort Worth metropolitan area in Texas. A large application and validation effort in Portland, OR, commenced early in 1999. The market introduction of TRANSIMS is expected sometime after 2002. The TMIP and TRANSIMS web sites provide much additional information.

REFERENCES

15.1 ALGERS, S., E. BERNAUER, M. BOERO, L. BREHERET, C. DI TARANTO, M. DOUGHERTY, K. FOX, and J.-F. GABARD, *Review of Micro-Simulation Models,* SMARTEST Project Deliverable D3, Contract No.: RO-97-SC. 1059, European Commission Transport RTD Programme, August 1997.

15.2 ABKOWITZ, M., S. WALSH, E. HAUSER, and L. MINOR, "Adaptation of Geographic Information Systems to Highway Management," *Journal of Transportation Engineering,* Vol. 116, No. 3 (May/June 1990): 310–327.

15.3 TRANSPORTATION RESEARCH BOARD, *Implementation of Geographic Information Systems (GIS) in State DOTs,* Research Results Digest, No. 180, Washington, DC, August 1991.

15.4 DATE, C. J., *An Introduction to Database Systems,* 6th ed., Addison-Wesley Programming Series, Addison-Wesley Publishing Company, Inc., 1995.

15.5 HURN, J., *GPS: A Guide to the Next Utility,* Trimble Navigation Ltd., Sunnyvale, CA, 1989.

15.6 FEDERAL HIGHWAY ADMINISTRATION, *Traffic Models Overview Handbook,* Report No. FHWA-SA-93-050, U.S. Department of Transportation, Washington, DC, June 1993.

15.7 GIBSON, D., *Available Computer Models for Traffic Operations Analysis,* Special Report 194, TRB, Washington, DC, 1981.

15.8 LEO, C-J., and R. L. PRETTY, "Numerical Simulation of Macroscopic Continuum Traffic Models," *Transportation Research B,* Vol. 26B, No. 3 (1992): 207–220.

15.9 MICHALOPOULOS, P. G., and P. YI, "Continuum Modeling of Traffic Dynamics for Congested Freeways," *Transportation Research B,* Vol. 27B, No. 4 (1993): 315–332.

15.10 PAYNE, H. J., "FREFLO: A Macroscopic Simulation Model of Freeway Traffic," *Transportation Research Record, 772,* TRB, Washington, DC, (1979): 68–75.

15.11 VAN AERDE, M., S. YAGAR, A. UGGE, and E. R. CASE, "A Review of Candidate Freeway-Arterial Corridor Traffic Models," *Transportation Research Record, 1132,* TRB, Washington, DC (1987): 53–65.

15.12 MAY, A. D., "Freeway Simulation Models Revisited," *Transportation Research Record 1132,* TRB, Washington, DC (1987): 94–99.

15.13 SABRA, Z. A., and C. R. STOCKFISCH, "Advanced Traffic Models: State-of-the-Art," *ITE Journal* (1995): 31–42.

15.14 TRAFFICWARE, INC., *SimTraffic—User Guide,* Berkeley, CA, 1998.

15.15 RATHI, A. K., and A. J. SANTIAGO, "The New NETSIM Simulation Program," *Traffic Engineering & Control,* No. 5 (1990): 317–320.

15.16 ———, "Identical Traffic Streams in the TRAF-NETSIM Simulation Program," *Traffic Engineering & Control,* No. 6 (1990): 351–355.

15.17 VIGGEN, CORP., *EzVIPAS 1.0—User Guide,* 1994.

15.18 TRAFFICWARE, INC., *SYNCHRO 3.2—User Guide,* Berkeley, CA, 1998.

15.19 FEDERAL HIGHWAY ADMINISTRATION, *TRANSYT 7-F: User's Manual,* Version 8.1, U.S. Department of Transportation, Washington, DC, March 1998.

15.20 SMITH, M. J., "A New Dynamic Traffic Model and the Existence and Calculation of Dynamic User Equilibria on Congested Capacity-Constrained Road Networks," *Transportation Research B,* Vol. 27B, No. 1 (1993): 49–63.

15.21 HALL, M. D., D. VAN VLIET, and L. G. WILLUMSEN, "SATURN—A Simulation-Assignment Model for the Evaluation of Traffic Management Schemes," *Traffic Engineering & Control,* No. 4 (1980): 168–176.

15.22 KWON, E., and P. MICHALOPOULOS, "Macroscopic Simulation of Traffic Flows in Complex Freeway Segments on a Personal Computer," *IEEE 1995 Vehicle Navigation & Information Systems Conference Proceedings,* Seattle, WA, pp. 342–345.

15.23 PREVEDOUROS, P. D., *Westbound Lunalilo St. On-Ramp Closure: Justification, Design and Analysis of Research Experiment,* Report prepared for the Hawaii DOT, Honolulu, 1999.

15.24 BARCELO, J., and J. L. FERRER, *AIMSUN2: Advanced Interactive Microscopic Simulation for Urban Networks,* Departamento de Estadistica e Investigation Operative, Faculdad de Informatica, Universidad Politecnica de Cataluna, 1998.

15.25 Van Aerde, M., and S. Yagar, "Dynamic Integrated Freeway/Traffic Signal Networks: Problems and Proposed Solutions," *Transportation Research A*, Vol. 22A, No. 6 (1988): 435–443.

15.26 Kaman Sciences Corp., *TSIS User's Guide*, Version 4.2, 1997.

15.27 Wang, Y., and P. D. Prevedouros, "Comparison of CORSIM, INTEGRATION and WATSim in Replicating Volumes and Speeds on Three Small Networks," *Transportation Research Record 1644*, TRB, Washington, DC (1998): 80–92.

15.28 Leiberman, E. B., "Simulation of Corridor Traffic—The SCOT Model," *Highway Research Record 409*, HRB, Washington, DC (1972): 34–45.

15.29 KLD Associates, Inc., *WATSim Model: User Guide*, April 1996.

15.30 Marcus, C. T., and D. Krechmer, "The Use of Simulation Models on the Central Artery/Third Harbor Tunnel Project," *IEEE 1995 Vehicle Navigation & Information Systems Conference Proceedings*, Seattle, WA, pp. 280–285.

15.31 Benekohal, R. F., "Procedure for Validation of Microscopic Traffic Flow Simulation Models," *Transportation Research Record 1320*, TRB, Washington, DC (1991): 190–202.

15.32 Chang, G-L., and A. Kanaan, "Variability Assessment for TRAF-NETSIM," *ASCE Journal of Transportation Engineering*, Vol. 116, No. 5 (1990): 636–657.

15.33 Prevedouros, P. D., "Signalized Intersection Capacity Analysis Software," *Traffic Congestion and Traffic Safety in the 21st Century ASCE Conference Proceedings*, Chicago, June 1997, pp. 69–75.

15.34 INRO Consultants, *EMME/2: User's Manual*, Montreal, Quebec, Canada, 1989.

15.35 Sosslau, A. B. et al., *Quick-Response Urban Travel Estimation Techniques and Transferable Parameters*, User's Guide, NCHRP Report 187, TRB, Washington, DC, 1978.

15.36 Horowitz, A. J., *Quick Response System II Reference Manual*, Version 2.2, prepared for the Federal Highway Administration, U.S. Department of Transportation, Center for Urban Transportation Studies, University of Wisconsin at Milwaukee, 1988.

15.37 Dial, R. B., G. S. Rutherford, and L. Quillian, Transit Network Analysis: INET, Report UMTA-UPM-20-79-3, U.S. Department of Transportation, Washington, DC, July 1979.

15.38 The Urban Analysis Group, *TRANPLAN User's Manual*, Danville, CA, 1990.

15.39 COMSIS Corp., *MINUTP: Technical User's Manual*, Silver Springs, MD, 1995.

15.40 Caliper Corp., *TransCAD—User Manual*, Version 3/Windows, 1996.

A

1982 Guidelines for the Preparation of Environmental Documents*

Federal Highway Administration
U.S. Department of Transportation

INTRODUCTION

The purpose of this material is to provide guidance to FHWA field offices and project applicants on National Environmental Policy Act (NEPA) actions and to provide the public with a further explanation of FHWA internal operating procedures in the development of the reports and documentation required by NEPA. This material also provides the guidance required by 23 U.S.C. 109(h) to assure the full consideration of possible adverse economic, social, and environmental effects of proposed FHWA projects. While the material was developed primarily to provide guidance in the development of environmental impact statements (EISs), it is also applicable, to the extent appropriate, for environmental assessments and other environmental studies deemed necessary prior to the advancement of a project with a categorical exclusion determination or a finding of no significant impact. This material is not regulatory, but has been developed to provide uniform and consistent guidance for the development of environmental documents. Each project will need to be carefully evaluated and the appropriate environmental document developed based on each individual situation.

The FHWA fully subscribes to the Council on Environmental Quality (CEQ) philosophy that the goal of the NEPA process is better decisions and not more documentation. As noted in the CEQ regulations, EISs should normally be less than 150 pages for most projects and not more than 300 pages for the most complex projects.

*FHWA Technical Advisory T 6640.8, February 24, 1982.

The FHWA considers the early coordination process to be a valuable tool to assist in identifying and focusing on the significant environmental issues. On April 30, 1981, the CEQ issued a memorandum entitled "Scoping Guidance," which discusses various techniques that will ensure participation in the scoping process. The CEQ also issued, on March 6, 1981, a memorandum entitled "Questions and Answers about the NEPA Regulations." Both of the documents are nonregulatory; however, they do provide CEQ views on various issues and are available from the FHWA Office of Environmental Policy (HEV-10).

A.1 ENVIRONMENTAL ASSESSMENT (EA)

Title 23, Code of Federal Regulations, Part 771, Environmental Impact and Related Procedures, describes those circumstances where the preparation of an EA is appropriate. The CEQ regulations require that an EA is to include the information listed in 40 CFR Part 1508.9. The following format, which assures this coverage, is suggested:

a. *Cover sheet.* There is no *required* format for the EA. However, it is recommended the EIS cover sheet format, as shown on page 540, be followed *where appropriate.* Since the EA is not formally circulated, there is no need to include the "comments due" paragraph on page 541.

b. *Description of the proposed action.* Describe the locations, length, termini, proposed improvements, etc.

c. *Need.* Identify and describe the problem which the proposed action is designed to correct. Any of the items discussed under the "Need" section in Section 3 (EIS—Format and Content) may be appropriate.

d. *Alternatives considered.* Discuss all reasonable alternatives to the proposed action which were considered. The EA may either discuss (1) the preferred alternative and the alternatives considered or (2) if the applicant has not identified a preferred alternative, the alternatives under consideration.

e. *Impacts.* Discuss the social, economic and environmental impacts of the alternatives considered and describe why these impacts are considered not significant.

f. *Comments and coordination.* Describe coordination efforts and comments received from government agencies and the public. If the EA includes a Section 4(f) evaluation, the EA and the Section 4(f) evaluation may be circulated to the appropriate agencies for Section 4(f) coordination, or the Section 4(f) evaluation may be supplemented by any additional information necessary to properly explain the project and circulated as a separate document.

g. *Appendices (if any).* Include only analytical information that substantiates an analysis which is important to the document. Other information should be incorporated by reference only.

A.2 FINDING OF NO SIGNIFICANT IMPACT (FONSI)

771.121 of 23 CFR 771, entitled Environmental Impact and Related Procedures, describes the approval process for a FONSI. Section 1508.13 of the CEQ regulations describes the content of a FONSI. The EA should be modified to reflect all applicable significant environmental comments received as a result of the public hearings or other significant environmental comments received as a result of the public and clearinghouse notification process. The EA, revised as appropriate, including appropriate responses to any comments received, is then submitted to the FHWA Division Administrator along with the applicant's recommendation. The basis for the applicant's recommendation should be documented in the EA. After review of the EA and any other appropriate information, the FHWA Division Administrator may determine that the proposed action has no significant impacts. This is documented by attaching to the EA a separate statement (example follows) which clearly sets forth the FHWA analysis of the EA along with any other supporting documentation that has resulted in a FONSI. As appropriate, the FHWA Division Administrator may choose to expand on the discussion in the sample FONSI to identify the basis for the decision. The EA/FONSI should document compliance with the requirements of all applicable environmental laws, Executive Orders, and other related requirements. If full compliance is not possible by the time the FONSI is prepared, it should reflect consultation with the appropriate agencies and provide reasonable assurance that the requirements will be met.

FEDERAL HIGHWAY ADMINISTRATION
FINDING OF NO SIGNIFICANT IMPACT
FOR
(Title of Proposed Action)

The FHWA has determined that this project will not have any significant impact on the human environment. This finding of no significant impact is based on the attached environmental assessment (reference other environmental documents as appropriate), which has been independently evaluated by the FHWA and determined to adequately and accurately discuss the environmental issues and impacts of the proposed project. It provides sufficient evidence and analysis for determining that an environmental impact statement is not required. The FHWA takes full responsibility for the accuracy, scope, and content of the attached environmental assessment.

__________ ____________________ __________
Date Responsible Official Title

A.3 EIS—FORMAT AND CONTENT

Each EIS should have a cover sheet containing:

(EIS number)

(Route, Termini, City or County, and State)
Draft (Final)
Environmental Impact Statement
Submitted Pursuant to 42 U.S.C. 4332(2)(c) (and
where applicable, 49 U.S.C. 1653(f) by the
U.S. Department of Transportation
Federal Highway Administration
and
State highway agency (HA)
and
(As applicable, local highway agency (HA))
Cooperating Agencies
List Here

__________________ ____
Date of Approval For FHWA Title

The following persons may be contacted for additional information concerning this document:

(Name, address, and telephone number of FHWA division office contact)

(Name, address, and telephone number of HA contact)

A one-paragraph abstract of the statement.
Comments on this draft EIS are due by (*date*) and should be sent to (*name and address*).

The top left-hand corner of the cover sheet of all draft and final EISs contains a number parallel to that in the following example:

FHWA-AZ-EIS-81-01-D(F)(S)

FHWA—name of Federal agency
AZ—name of State (cannot exceed four characters)
EIS—environmental impact statement
81—year draft statement was prepared
01—sequential number of draft statement for each calendar year
D—designates the statement as the draft statement
F—designates the statement as the final statement
S—designates supplemental statement

The EISs should be printed on $8\frac{1}{2} \times 11$-inch paper with all graphics folded for insertion to that size. The wider sheets should open to the right with the title or identification on the right. The use of a standard size will facilitate administrative recordkeeping.

Summary

The summary should include:

a. A brief description of the proposed FHWA action indicating route, termini, type of improvement, number of lanes, length, county, city, state, etc., as appropriate.

b. A description of any significant actions proposed by other government agencies in the same geographic area as the proposed FHWA action.

c. A summary of major alternatives considered. (The final EIS should identify the preferred alternative).

d. A summary of *significant* environmental impacts, both beneficial and adverse.

e. Any areas of controversy (including issues raised by both agencies and the public).

f. Any significant unresolved issues.

g. A list of other federal actions required because of this proposed action (i.e., permit approvals, etc.).

Table of Contents

a. Cover sheet

b. Summary

c. Table of contents

d. Purpose of and need for action

e. Alternatives including proposed action

f. Affected environment

g. Environmental consequences

h. List of preparers

i. List of agencies, organizations, and persons to whom copies of the statement are sent

j. Comments and coordination

k. Index

l. Appendices (if any)

Purpose of and Need for Action

Identify and describe the transportation problem(s) which the proposed action is designed to address. This section should clearly demonstrate that a "need" exists and must define the "need" in terms understandable to the general public. This discussion will form the basis for the "no action", discussion in the "Alternatives" section. The following is a list of items which may assist in the explanation of the need for the proposed action. It is by no means all-inclusive or applicable in every situation and is intended only as a guide.

a. *System linkage.* Is the proposed project a "connecting link"? How does it fit in the system? Is it an "essential gap" in the Interstate System?

b. *Capacity.* Is the capacity of the present facility inadequate for the present traffic? Projected traffic? What capacity is needed? What is the level of service?

c. *Transportation demand.* Includes relationship to any statewide plan or adopted urban transportation plan.

d. *Federal, state, or local governmental authority (legislation) directing the action.*

e. *Social demands or economic development.* New employment, schools, land use plans, recreation, etc. What projected economic development/land use changes indicate the need to improve or add to the highway capacity?

f. *Modal interrelationships.* How will the proposed facility interfere with and serve to complement airports, rail and port facilities, mass transit services, etc.

g. Is the proposed project necessary to correct an existing or potential safety hazard? Is the existing accident rate excessively high? Why? How will the proposed facility improve it?

Alternatives Including Proposed Action

The "Alternatives" section of the draft EIS should begin with a concise discussion of how the "reasonable alternatives" were selected for detailed study. It should also describe those "other alternatives" that were eliminated early in project development and the basis for their elimination. The alternatives to be considered in this section will normally include the following:

a. The "no-action" alternative, which would include those usual short-term minor reconstruction types of activities (safety improvements, etc.) that are a part of an ongoing plan for continuing operation of the existing roadway system in the project area.

b. A Transportation System Management (TSM) alternative which would include those types of activities designed to maximize the utilization and energy efficiency of the present system. Possible subject areas to include in this alternative are options such as fringe parking, ridesharing, high-occupancy vehicle (HOV) lanes on existing roadways, and traffic signal timing optimization. This limited construction alternative should be given appropriate consideration when major urbanized area construction activities are proposed. On major new urbanized area highway projects, the option of including and/or designating HOV lanes should be a consideration. Consideration of this alternative may be accomplished by reference to the regional transportation plan, when that plan considers this option. In the case of regional transportation plans which do not reflect consideration of this option, it may be necessary to evaluate the feasibility of this alternative. The effects that reducing the scale of a link in the regional transportation plan will have on the remainder of the system will need to be discussed during the evaluation of this alternative. While this discussion relates primarily to major projects in urbanized areas, the concept of achieving maximum utilization of existing facilities is equally important in rural areas. Before major projects on new location are proposed, it is important to demonstrate that reconstruction and rehabilitation of the existing system will not adequately correct the identified deficiencies. Appendix A of 23 CFR 450 provides additional discussion on the goals and scope of the TSM concept.

c. All other proposed "construction" alternatives discussions should include, where relevant, those reasonable and feasible alternatives (i.e., transit options) which may not be within the existing funding authority of FHWA. Some urban projects may be multimodal, thus requiring close coordination with the Urban Mass Transportation Administration (UMTA). In these situations, UMTA should be consulted early in the project development process. Depending on the extent of UMTA involvement and the possible use of UMTA funds for portions of the proposal, the need to request UMTA to be either a "lead agency" or a "cooperating agency" should be considered at the earliest stages of project development. Where applicable, cost-effectiveness studies that have been performed should be summarized in the EIS.

The discussion of alternatives in this section can be best accomplished by a brief written description of each alternative, supplemented with maps and other appropriate visual aids such as photographs, drawings, or sketches which would assist the reader in better understanding the various alternatives, impacts, and mitigation measures. In some situations, design level details may be appropriate to evaluate impacts. However, final design details are not normally available at this stage in project development. The material should provide a clear understanding of each alternative's termini, location, costs, and major design features (number of lanes, right-of-way requirements, median width, etc.) which will contribute to a reader's better understanding of each alternative's effects on its surroundings or the community.

Generally, each alternative should be developed to a comparable level of detail in the draft EIS. Normally, the draft EIS should state that all alternatives are under consideration and that a decision will be made only after the public hearing transcript and comments on the draft EIS have been evaluated. However, in those situations where the HA has identified a "preferred" alternative based on its early coordination and environmental studies, the HA may so indicate in the draft EIS. However, the EIS should include a comment to the effect that the final selection will not be made until the results of the EIS circulation and the public involvement process have been fully evaluated. The final EIS must identify the preferred alternative and discuss the basis for the selection.

Affected Environment

This section should provide a *concise* description of the existing social, economic, and environmental setting for the area affected by all of the alternative proposals. The description should be a single general description for the area rather than a separate one for each alternative. All environmentally sensitive locations or features should be identified. However, it may be desirable to exclude from environmental documents certain specific location data on archeological sites to prevent vandalism.

To reduce paperwork and eliminate the presentation of extraneous background material, the discussion should focus on significant issues and values. Prudent use of photographs, illustrations, and other graphics within the text can be effective in giving the reviewer an understanding of the area. The statement should describe other related Federal activities in the area, their interrelationships, and any significant cumulative environmental impacts.

Data and analyses in the statement should be in proportion to the significance of the impacts which will be discussed later in the document. Less important material should be summarized or referenced. This section should also describe the scope and status of the planning process for the area. The inclusion of a map of any adopted land use and transportation plan for the area would be helpful in relating the proposed project to the areawide planning process.

Environmental Consequences

This section will discuss the probable social, economic, and environmental effects of the alternatives and the measures to mitigate adverse impacts.

There are several ways of preparing this section. Normally, it is preferable to discuss the impacts and mitigation measures separately for each of the alternatives. However, in some cases (such as where there are few alternatives), it may be advantageous to present this section with the impacts as the headings. Where possible, a subsection should be included which would discuss the general impacts and mitigation measures that are the same regardless of the alternative selected. This would reduce or eliminate repetition under each of the alternative discussions.

When the final EIS is prepared, the impacts and mitigation measures associated with the selected alternative may need to be discussed in more detail than in the draft EIS. In discussing the impacts, both beneficial and adverse, the following should be included in both the draft and final EIS:

a. A summary of studies undertaken and major assumptions made, with enough data or cross referencing to determine the validity of the methodology.

b. Sufficient information to establish the reasonableness of the conclusions concerning impacts.

c. A discussion of mitigation measures. Prior to completion of the final EIS, these measures normally should be investigated in appropriate detail so that a commitment can be included in the final EIS.

Charts, tables, maps, and other graphics illustrating comparisons between the alternatives (i.e., costs, residential displacements, noise impacts, etc.) are useful as a presentation technique.

In addition to normal FHWA program monitoring of design and construction activities, special instances may arise when a formal program for monitoring impacts or mitigation measures will be appropriate. In these instances, the final EIS should describe the monitoring program.

Listed below are examples of the potentially significant impacts of highway projects. These factors should be discussed *to the extent applicable* for each alternative. This list is by no means all-inclusive and on specific projects there may be other significant impacts that require study.

Social and Economic Impacts

The statement should discuss:

a. Changes in the neighborhoods or community cohesion for various groups as a result of the proposed action. These changes may be beneficial or adverse, and may include splitting neighborhoods, isolating a portion of an ethnic group, new development, changed property values, or separation of residences from community facilities, etc.

b. Changes in travel patterns and accessibility (e.g., vehicular, commuter, bicycle, or pedestrian). If any cross streets are terminated, the EIS should reflect the views of the involved city or county on such street closings.

c. Impacts on school districts, recreation areas, churches, businesses, police and fire protection, etc.

d. The impacts of alternatives on highway and traffic safety as well as on overall public safety.

e. Regional economic impacts, such as the effects of the project on development, tax revenues and public expenditures, employment opportunities, accessibility and retail sales. Any significant impacts on the economic viability of affected municipalities should also be discussed together with a summary of any efforts taken and agreements reached for using the transportation investment to support both public and private development plans. To the extent possible, this discussion should rely upon reviews by affected state, county, and city officials and upon studies performed under 23 U.S.C. 134.

f. For projects that might lead to or support large commercial development, the EIS should provide information on any significant effects the pending action would have on established business districts, and any opportunities for mitigation by the public and/or private sectors.

g. The general social groups specially benefitted or harmed by the proposed action should be identified. Particular effects of a proposal on the elderly, handicapped, nondrivers, transit-dependent, or minorities should be described to the extent these can be reasonably predicted. For example, where minority impacts may be a significant concern, EISs should contain, when applicable, the following information, broken down by race, color, and national origin: the population in the study area, the number of displaced residents, the type and number of displaced businesses, and the type and number of displaced employees. Secondary sources of information such as census data reports can be utilized for obtaining this type of background information. Changes in minority employment opportunities, the relationship of the proposed action to other Federal actions which may serve or affect the minority population, and proposed mitigation measures to reduce or avoid impacts on minority populations should also be discussed.

Relocation Impacts

The relocation information necessary for the draft EIS may be included in the draft statement, either in the form of a complete conceptual stage relocation plan, or summarized in sufficient detail to adequately explain the relocation situation along with a resolution of anticipated or known problems. When the relocation information is summarized, the conceptual stage relocation plan should be referenced in the draft EIS.

A discussion of the information listed below is to be included in the draft EIS *to the extent appropriate* for the project.

a. An estimate of households to be displaced, including the family characteristics (e.g., minorities, handicapped, income levels, the elderly, large families, length of occupancy, and owner/tenant status). Where the project is not complex from a relocation viewpoint and the impact on the community is slight, this information may be obtained by visual inspection and from available secondary sources. On complex relocation projects where the relocation will have a major impact on the community, a survey of affected occupants may be needed. This survey may be accomplished by a sampling process.

b. A discussion of available housing in the area and the ability to provide suitable relocation housing for each type of family to be displaced within the financial capabilities of the relocatees.

c. A description of any special advisory services that will be necessary for unique relocation problems.

d. A discussion of the actions proposed to remedy insufficient relocation housing, including a commitment to housing of last resort, if necessary.

e. An estimate of the number, type, and size of businesses to be displaced. The approximate number of employees for each business should be included along with the general impact on the business dislocation(s) on the economy of the community.

f. A discussion of the results of early consultation with the local government(s) and any early consultation with businesses potentially subject to displacement, including any discussions of potential sources of funding, financing, planning for incentive packaging (e.g., tax abatement, flexible zoning, and building requirements), and advisory assistance which has been or will be furnished along with other appropriate information.

g. Impact on the neighborhood and housing community services where relocation is likely to take place. If there will be extensive residential and/or business displacement, the affected community may want to investigate other sources of funding from local and state entities as well as HUD, the Economic Development Administration, and other federal agencies, to assist in revitalization of the community.

h. The results of discussions with local officials, social agencies, and such groups as the elderly, handicapped, nondriver, transit-dependent, and minorities regarding the relocation impacts.

i. A statement that the housing resources are available to all relocatees without discrimination.

The effects on each group should be described to the extent reasonably predictable. The analysis should discuss how the relocation caused by the proposed project will facilitate or inhibit access to jobs, educational facilities, religious institutions, health and welfare services, recreational facilities, social and cultural facilities, pedestrian facilities, shopping facilities, and public transit services.

Air Quality Impacts

The EIS should contain a brief discussion of air quality effects or a summary of the carbon monoxide (CO) analysis if such an analysis is performed. The following provides additional guidance:

a. A microscale CO analysis to determine air quality impacts is probably unnecessary where such impacts are judged to be minimal or insignificant. The judgment on the degree of CO impacts may be based on: (1) previous analyses for similar projects, (2) previous general analyses for various classes of projects, or (3) simplified graphical or "table look-up" analyses.

b. If the impacts of CO are judged to be minimal or insignificant, a brief statement to this effect is sufficient. The basis for the statement should be given in the EIS.

c. If the project CO contribution plus the background level are known to be well below the 1- and 8-hour National Ambient Air Quality Standard or other applicable standard, then the air quality CO impact is judged to be insignificant.

d. For those projects where a CO microscale analysis is performed, then the total CO concentration (project contribution, plus estimated background) at identified reasonable receptor sites for all alternatives should be reported and compared with applicable State and national standards.

e. If a CO analysis is performed, a brief summary of the methodologies and assumptions used should be given in the EIS.

f. In addition to the CO impact assessment, one of the two following statements should be included in the EIS:

(1) This project is in an area where the State implementation plan does not contain any transportation control measures. Therefore, the conformity procedures of 23 CFR 770 do not apply to this project.

(2) This project is in an air quality nonattainment (or attainment) area which has transportation control measures in the State implementation plan (SIP) which was (conditionally) approved by the Environmental Protection Agency on (date). The FHWA has determined that both the transportation plan and the transportation improvement program conform to the SIP. The Federal Highway Administration has determined that this project is included in the transportation improvement program for the (indicate 3C planning area). Therefore, pursuant to 23 CFR 770, this project conforms to the SIP.

Noise Impacts

The EIS should contain a summary of the noise analysis including the following:

a. A brief description of noise sensitive areas, including information on the numbers and types of activities which may be affected. If the project has significant noise impacts, noise contours of the proposed action and alternatives may be appropriate to assist in understanding those impacts.

b. The extent of the impact (in decibels). This should include a comparison of the predicted noise levels with both the FHWA design noise levels and the existing noise levels.

c. Noise-abatement measures which have been considered and those measures that would likcly be incorporated into the proposed project.

d. Noise problems for which no prudent solution is reasonably available and the reasons why.

Energy

Draft and final EISs should discuss in *general terms* the energy requirements and conservation potential of various alternatives under consideration. This general discussion might recognize that the energy requirements of various construction alternatives are similar and are generally greater than the energy requirements of the no-build alternative. Additionally, the discussion could point out that the post-construction, operational energy requirements of the facility should be less with the build alternative as opposed to the no-build alternative. In such a situation, one might then conclude that the savings in operational energy requirements would more than offset construction energy requirements and thus, in the long term, result in a net saving in energy usage. For most projects, a detailed energy analysis including computations of Btu requirements, etc., is not needed, but the discussion should be reasonable and supportable.

For major projects with potentially significant energy impacts (an example would be the Westway project in New York City), both the draft and final EIS should discuss any *significant* direct and/or indirect energy impacts of the proposed action. Direct energy impacts refer to the energy consumed by vehicles using the facility. Indirect impacts include construction energy and such items as the effects of any changes in automobile usage. The action's relationship and consistency with any State and/or regional energy plan should also be indicated.

The final EIS should identify any energy conservation measures that will be implemented as a part of the recommended alternative. Measures to conserve energy include the use of high-occupancy vehicle incentives, measures to improve traffic flow, and also pedestrian and bicycle facilities.

Wild and Scenic Rivers

If the proposed action could have an adverse effect on a river on the National Wild and Scenic Rivers System or a river listed in the Nationwide Inventory of rivers with potential for inclusion in the National Wild and Scenic Rivers System, there should be early coordination with the National Park Service (NPS) or the Department of Agriculture (USDA). The EIS should identify any potential significant adverse effects on the natural, cultural, and recreational values of the inventory river. Adverse effects include alteration of the free-flowing nature of the river, alteration of the setting, or deterioration of water quality. If it is determined that the proposed action could foreclose options to designate the river under the act, the EIS should reflect consultation with the NPS or USDA on avoiding or mitigating the impacts. The final EIS should indicate measures which will be included in the action to avoid or mitigate impacts. The October 3, 1980, memorandum from the Office of Environmental Policy provides additional information on this subject area.

Floodplain Impacts

The draft EIS should contain a summary of the "Location Hydraulic Studies" required by FHPM 6-7-3-2, Location and Hydraulic Design of Encroachments on Floodplains. Exhibits defining the floodplains or regulatory floodway, as appropriate, should be provided whenever possible. When there is no practicable alternative to an action which includes a significant encroachment, the final EIS should contain the finding required by FHPM 6-7-3-2, paragraph 8, in a separate subsection titled "Only Practicable Alternative Finding." When there is a regulatory floodway affected by the proposed action, the final EIS should contain a discussion of the consistency of the project with the regulatory floodway.

Coastal Zone Impacts

Where the proposed action is within, or may affect land or water uses within, the area covered by a State Coastal Zone Management Program (CZMP) approved by the Department of Commerce, the environmental document should briefly describe the CZMP plan, identify the potential impacts, and include evidence of coordination with the State Coastal Zone Management agency or appropriate local agency. For FHWA assisted activities, the EIS should include the State Coastal Zone Management agency's determination as to whether the project is consistent with the State CZMP plan. For direct Federal actions, the EIS should include the lead agency's consistency determination. If it is determined that the proposed action is inconsistent with the state's approved CZMP, FHWA will not approve the action except upon a finding by the Secretary of Commerce that the proposed action is consistent with the purposes or objectives of the Coastal Zone Management Act or is necessary in the interest of national security. The final environmental document for the proposed action will document all findings.

Wetlands Impacts

a. All draft EISs for projects involving new construction in wetlands should include sufficient information to: (1) identify the type of wetlands involved, (2) describe the impacts to the wetlands, (3) evaluate alternatives which would avoid these wetlands, and (4) identify practicable measures to minimize harm to the wetlands. Exhibits showing the wetlands in relation to the alternatives, including the alternatives to avoid construction in the wetlands, should be provided.

b. Executive Order 11990, Protection of Wetlands, requires federal agencies ". . . to avoid to the extent possible the long and short term adverse impacts associated with the destruction or modification of wetlands and to avoid direct or indirect support of new construction in wetlands wherever there is a practicable alternative. . . ." In evaluating the impact of the proposed project on wetlands, the following two questions should be addressed: (1) what is the importance of the impacted wetlands? and (2) what is the significance of this impact on the wetlands? Merely listing the number of acres taken by the various alternatives of a highway proposal does not provide sufficient information upon which to determine the degree of impact on the wetland's ecosystem. The wetlands analysis should be sufficiently detailed to allow a meaningful discussion of these two questions.

c. In evaluating the importance of the impacted wetlands, the analysis should consider such factors as: (1) the primary functions of the wetlands (e.g., flood control, wildlife habitat, erosion control, etc.), (2) the relative importance of these functions to the total wetlands resource of the area, and (3) other factors such as uniqueness that may contribute to the wetlands importance.

d. In determining the significance of the highway impact, the analysis should focus on how the project affects the stability and quality of the wetlands. This analysis should consider the short- and long-term effects on the wetlands and the significance of any loss such as: (1) flood control capacity, (2) erosion control potential, (3) water pollution abatement capacity, and (4) wildlife habitat value. Knowing the importance of the wetlands involved and the significance of the impact, the SHA and FHWA will be in a better position to determine what mitigation efforts are necessary to minimize harm to these wetlands.

e. For purposes of analyzing alternatives and the wetlands finding, "located in wetlands" means that the proposed right-of-way or easement limits of the highway are located wholly or partially in wetlands or that the highway is located in the vicinity of the wetlands and there is evidence that the new construction will directly cause long-term damage or destruction of the wetlands.

f. Mitigation measures which should be considered include enhancement of existing wetlands, creation of new wetlands, and erosion control. It should be noted that any mitigation measure should be related to the actual adverse impact caused by the project and that acquisition of privately owned wetlands for purposes of protection should only be considered as a last resort.

g. When there is no practicable alternative to an action which involves new construction located in wetlands, the final EIS should contain the finding required by Executive Order 11990 and by DOT Order 5660.1A, entitled Preservation of the Nation's Wetlands, August 24, 1978, in a separate section or exhibit titled "Wetlands Finding." Approval of the final EIS containing this finding will document compliance with the requirements of Executive Order 11990. The finding should contain in summary form and with reference to the detailed discussions contained elsewhere in the final EIS:

(1) a reference to executive Order 11990;

(2) a discussion of the basis for the determination that there are no practicable alternatives to the proposed action;

(3) a discussion of the basis for the determination that the proposed action includes all practicable measures to minimize harm to wetlands; and

(4) a concluding statement as follows: "Based upon the above considerations, it is determined that there is no practicable alternative to the proposed new construction in wetlands and that the proposed action includes all practicable measures to minimize harm to wetlands which may result from such use."

h. A formal wetlands finding is required for all projects processed with EIS's or FONSI's that involve new construction in wetlands. In the case of a project processed as a categorical exclusion, the division office's administrative record should document evaluations of alternatives and measures to minimize harm for these actions.

Land-Use Impacts

This discussion should begin with a description of current development trends and the state and/or local government plans and policies with regard to land use and growth in the area. These plans and policies will be reflected in the area's comprehensive development plan, including land use, transportation, public facilities, housing, community services, and other areas.

The land-use impact analysis should assess the consistency of the alternatives with the comprehensive development plans adopted for the area. The secondary social, economic, and environmental impacts of *significant* induced development should be presented.

The EIS should note any proposed alternatives which will stimulate low-density, energy-intensive development in outlying areas and will have a significant adverse effect on existing communities. Throughout this discussion, the distinction between planned and unplanned growth should be clearly identified.

Joint Development

When applicable, the EIS should discuss how the implementation of joint development projects will preserve or enhance the community's social, economic, environmental, and visual values. This discussion should be included as part of the land-use impact presentation.

Historic and Archeological Preservation

The draft EIS should contain a discussion demonstrating that a survey meeting the requirements of 36 CFR Part 800.4 has been performed for each alternative under consideration. The discussion should begin by describing the resources and summarizing the impacts that each alternative will have on these resources that might meet the criteria for inclusion on the National Register of Historic Places. There should be a record of coordination with the State Historic Preservation officer concerning the significance of the identified resources, the likelihood of eligibility for the National Register, and an evaluation of the effect of the project on the resources.

The draft EIS can serve as a preliminary case report for Section 106 requirements if the document indicates this and it contains the necessary information (36 CFR 800.13). The transmittal memorandum to the Advisory Council on Historic Preservation should specifically request consultation.

The final EIS should demonstrate that all the requirements of 36 CFR Part 800 have been met. If the selected alternative has an effect on a resource that is on or eligible for inclusion on the National Register, the final EIS should contain (a) a determination of no adverse effect concurred in by the Executive Director of the Advisory Council on Historic Preservation or (b) an executed memorandum of agreement or (c) in the case of a unique situation where FHWA is unable to conclude the memorandum of agreement (MOA), a copy of comments transmitted from the Advisory Council to the Secretary of Transportation. When necessary, the discussion should indicate that archeological recovery will be performed. The proposed use of land from a site on or eligible for inclusion on the National Register will normally require a determination pursuant to Section 4(f) of the DOT Act. The treatment of archeological sites is discussed in 23 CFR 771.135(f). Additional details regarding the type of information needed at the draft EIS and final EIS stages are contained in the May 14, 1980, memorandum from the Office of Environmental Policy to all regional offices.

Water Quality Impacts

This discussion should include summaries of analyses and consultations with the state and/or local agency responsible for water quality. Coordination with the Environmental Protection Agency (EPA) under the Federal Clean Water Act may provide assistance in this area. The EIS should discuss any locations where roadway runoff may have a significant effect on downstream water uses, including existing wells. A 1981 FHWA research report entitled “Constituents of Highway Runoff” contains procedures for estimating pollutant loading from highway runoff.

Section 1424(e) of the Safe Drinking Water Act requires that proposed actions which may impact those areas that have been designated as principal or sole-source aquifers be coordinated with EPA. The EPA will furnish information on whether any of the alternatives affect the aquifer. If none of the alternatives affect the aquifer, the requirements of the Safe Drinking Water Act are satisfied. If an alternative is selected which affects the aquifer, a design must be developed to assure, to the satisfaction of EPA, that it will not contaminate the aquifer.

If a rest area is involved, a Section 402 permit is required for point source discharge. Any potential Section 402 permits should be identified in the EIS. Also, for both the Section 402 and Section 404 permits, a water quality certification from the State agency responsible for water quality is necessary.

The MOA with the Corps of Engineers allows for application for permit as soon as the preferred alternative is identified (i.e., final EIS stage). Use of the procedures in the MOA is encouraged to minimize possible delays in the processing of Section 404 permits later in project development. The final EIS should indicate the general location of the fill or dredged activity, approximate quantities of fill or dredged material, general construction grades, and proposed mitigation measures, and should include evidence of coordination with the corps.

Threatened or Endangered Species

The HA shall request from the Departments of the Interior (DOI) and/or Commerce (DOC) information on whether any species listed or proposed as endangered or threatened may be present in the area of the proposed construction project. If those departments advise that there are no such species in the area, the requirements of the Endangered Species Act have been met. If those departments advise that such a species may be present, the FHWA/HA shall undertake a biological assessment to identify any threatened or endangered species which are likely to be affected by the proposed action. This biological assessment should include:

a. An on-site inspection of the area affected by the proposed project.

b. Interviews with recognized experts on the species at issue.

c. A literature review to determine the species distribution, habitat needs, and other biological requirements.

d. An analysis of possible impacts to the species.

e. An analysis of measures to minimize impacts. This biological assessment should be forwarded to DOI/DOC for a biological opinion. The Fish and Wildlife Service (F&WS) is responsible for the protection of terrestrial and fresh-water species and the National Marine Fisheries Service (NMFS) is responsible for the protection of marine species.

Upon completing their review of the biological assessment, the F&WS/NMFS may request additional information and/or a meeting to discuss the project or issue a biological opinion stating that the project : (a) is not likely to jeopardize, or (b) will promote the conservation of or (c) is likely to jeopardize the threatened or endangered species. In selecting a preferred alternative, jeopardy to an endangered or threatened species must be avoided. If either a finding of (a) or (b) is given, the requirements of the Endangered Species act are met. If a detrimental finding is presented, the proposed action may be modified so that the species is no longer jeopardized. In unique circumstances, an exemption may be requested. If an exemption is denied, the action must be halted or modified. The final EIS should document the results of the coordination of the biological assessment with the appropriate agencies.

Prime and Unique Agricultural Lands

Information on prime and unique agricultural lands should be solicited through early consultation with the Department of Agriculture (USDA), and the EIS should identify the direct and indirect impacts of the proposed action on these lands, including:

a. An estimate of the number of acres that might be directly affected by right-of-way acquisition.

b. Areas where agricultural operation might be disrupted.

c. Potential indirect effects such as those related to project-induced changes in land use.

The EIS should contain a map showing the location of prime and unique agricultural lands in relation to the project alternatives, summarize the results of consultations with the USDA, and include copies of correspondence with USDA regarding the project. Specific actions to avoid or, if that is not possible, to reduce direct and indirect effects on these lands should be identified.

Construction Impacts

The EIS should discuss significant impacts (particularly air, noise, water, detours, safety, visual, etc.) associated with construction of each of the alternatives. Also, where applicable, the impacts of disposal and borrow areas should be discussed along with any proposed measures to minimize these impacts.

Considerations Relating to Pedestrians and Bicyclists

Section 682 of the National Energy Policy Act of 1978 recognizes that bicycles are an efficient means of transportation, represent a viable commuting alternative to many people, and deserve consideration in a comprehensive national energy plan. The FHWA recognizes that bicyclists are legitimate highway users and that FHWA has a responsibility to provide for their transportation needs. Section 109(n) of 23 U.S.C. provides that "the Secretary shall not approve any project under this title that will result in the severance or destruction of an existing major route for nonmotorized transportation traffic and light motorcycles, unless such project provides a reasonable alternate route or such a route exists." The FHWA policy regarding Bicycle Program Activities is further defined in an August 20, 1981, memorandum from Administrator Barnhart to all regional administrators. Where appropriate, the EIS should consider pedestrian and bicycle use as an integral feature of the project and include a discussion of the relationship of the proposed project to local plans for bicycles and pedestrian facilities and evidence that the project is consistent with 23 U.S.C. 109(n).

Stream Modification and Wildlife Impacts

Title 16 U.S.C. 662(a) requires consultation with the Fish and Wildlife Service and the appropriate State agency regarding any federal action which involves impoundment (surface area of 10 acres or more), diversion, channel deepening, or other modification of a stream or body of water. Exhibits should be used to identify stream modifications. The use of the stream or body of water for recreation or other purposes should be identified. It should also discuss any significant impacts on fish and wildlife resources, including direct impact to fish and wildlife, loss or modification of habitat, and degradation of water quality.

Visual Impacts

This discussion should include an assessment of the visual impacts of the proposed action, including the "view from the road" and the "view of the road." Where relevant, the EIS should document the consideration given to design quality, art, and architecture in the project planning. These values may be particularly important for facilities located in sensitive urban settings. Where relevant, the draft EIS should be circulated to officially designated state and local arts councils and, as appropriate, other organizations with an interest in design, art, and architecture.

List of Preparers

This section will include lists of:

- **a.** State (and local agency) personnel, including consultants, who were primarily responsible for preparing the EIS or performing environmental studies, and their qualifications, including educational background or experience
- **b.** The FHWA personnel primarily responsible for preparation or review of the EIS, and their qualifications
- **c.** The areas of EIS responsibility for each preparer

List of Agencies, Organizations, and Persons to Whom Copies of the Statement are Sent

List all entities from which comments are being requested (draft EIS) and identify those that submitted comments (final EIS).

Comments and Coordination

- **a.** The draft EIS should summarize the early coordination process, including scoping, meetings with community groups and individuals, and the key issues and pertinent information received from the public and government agencies through these efforts.
- **b.** The final EIS should include a copy of all substantive comments received (or summaries thereof, where the response has been exceptionally voluminous), along with a response to each substantive comment. When the EIS is revised as a result of the comments received, a copy of the comments should contain marginal references indicating where revisions were made, or the discussion of the comments should contain such references. The FHWA comment(s) on the draft EIS should not be included in the final EIS. However, the document should include adequate information for the FHWA reviewer to ascertain the disposition of the comment(s). Formal comments by the Department of Transportation should be included in the final EIS along with an appropriate response to each comment.
- **c.** The final EIS should document compliance with requirements of all applicable environmental laws, Executive Orders, and other related requirements. To the extent possible, all environmental issues should be resolved prior to the submission of the final EIS. Where this is not possible, the final EIS should clearly identify any remaining unresolved issues, the change taken to resolve the issues, and the positions of the respective parties.
- **d.** The final EIS should contain a summary and disposition of substantive comments on social, economic, and environmental issues made at any public hearing or other public involvement activity or which were otherwise considered.

Index

The index should include major subjects and areas of significant impacts so that a reviewer need not read the entire EIS to obtain information on a specific subject or impact.

23 CFR 771 requires compliance to the extent possible with other applicable environmental laws, Executive Orders, and other related requirements. This includes the certifications and reports required by 23 U.S.C. 128 relating to public hearings, considerations of social, economic, and environmental (SEE) effects and consistency of the project with urban planning goals promulgated by the community. The certifications normally are made at the time the final EIS or FONSI is submitted to the FHWA Division Administrator. The report of SEE effects required by 23 U.S.C. 128 will normally be satisfied by the final EIS, FONSI, or identification of the project as a categorical exclusion.

Appendices

Material prepared as appendices to the EIS should:

- **a.** consist of material prepared in connection with the EIS (is distinct from material which is not so prepared and which is incorporated by reference)
- **b.** consist of material which substantiates an analysis which is fundamental to the EIS
- **c.** be analytic and relevant to the decision to be made and
- **d.** be circulated with the EIS or be readily available on request. Other reports and studies referred to in the EIS should be readily available for review or for copying at a convenient location.

Alternate Process for Final EISs

Paragraph 1503.4 of the CEQ regulation (40 CFR 1500, et seq.) provides the opportunity for expediting final EIS preparation in those instances when, after receipt of comments resulting from circulation of the draft EIS, it is apparent that the changes in the proposal or in the EIS in response to the comments received are minor and that:

- **a.** all reasonable alternatives were studied and discussed in the draft EIS, and
- **b.** the analyses in the draft EIS adequately identified and quantified the environmental impacts of all reasonable alternatives.

When these two points can be established, the final EIS can consist of the draft EIS and an attachment containing the following:

- **a.** Errata sheets making corrections to the draft EIS, if applicable.
- **b.** A section identifying the preferred alternative and a discussion of the reasons it was selected. The following should also be included in this section, if applicable:
  - (1) Final Section 4(f) evaluations containing the information described in Section 6 of these guidelines
  - (2) Wetlands finding(s)
  - (3) Floodplain finding(s)
  - (4) A list of commitments for mitigation measures for the preferred alternative
- **c.** Copies (or summaries) of comments received from circulation of the draft EIS and public hearing and responses thereto.

A.4 DISTRIBUTION OF EISs AND SECTION 4(f) EVALUATIONS

Environmental Impact Statements

a. Copies of all draft EISs should be circulated for comments to all public officials, private interest groups, and members of the public having or expressing an interest in the proposed action or the draft EIS, and to all government agencies expected to have jurisdiction, responsibility, interest, or expertise in the proposed action. Internal FHWA distribution of draft and final EISs is subject to change and is noted in memorandums to the Regional Administrators as requirements change. The FHWA transmittal letter to the Washington Headquarters should include a recommendation regarding the need for the prior concurrence of the Washington Headquarters in accordance with 23 CFR 771(e).

b. Copies of all approved final EISs should be distributed to all cooperating agencies, to all federal, state, and local agencies and private organizations, and members of the public who commented substantively on the draft EIS. A copy of all approved delegated EISs should be forwarded to the FHWA Washington Headquarters (HEV-10) for recordkeeping purposes.

Copies of all draft and final EISs in the categories listed in 23 CFR 771(e) should be provided to the Regional Representative of the Secretary of Transportation at the same time as they are forwarded to the FHWA Washington Headquarters.

c. Copies of all EISs should normally be distributed as follows, unless the agency has indicated to the FHWA offices the need for a different number of copies:

(1) The EPA Headquarters: five copies of the draft EIS and five copies of the final EIS (this is the "filing requirement" in Section 1506.9 of the CEQ regulation; the correct address is listed therein).

(2) The appropriate EPA regional office responsible for EPA's review pursuant to Section 309 of the Clean Air Act: five copies of the draft EIS and five copies of the final EIS.

(3) The DOI Headquarters:

(a) All States in FHWA Regions 1, 3, 4, and 5, plus Hawaii, Guam, American Samoa, Arkansas, Iowa, Louisiana, Missouri, and Puerto Rico: 12 copies of the draft EIS and 7 copies of the final EIS.

(b) Kansas, Nebraska, North Dakota, Oklahoma, South Dakota, and Texas: 13 copies of the draft EIS and 8 copies of the final EIS.

(c) New Mexico and all states in FHWA Regions 8, 9, and 10, except Hawaii, North Dakota, and South Dakota: 14 copies of the draft EIS and 9 copies of the final EIS.

Section 4(f) Evaluation

If the Section 4(f) evaluation is included in an EIS, DOI Headquarters should receive the same number of copies listed above for EISs for consultation in accordance with the requirements of 23 U.S.C. 138. If the Section 4(f) evaluation is processed as a separate document or as part of an EA, the DOI should receive seven copies of the draft Section 4(f) evaluation for coordination and seven copies of the final Section 4(f) statement for information.

In addition, draft Section 4(f) evaluations, whether in a draft EIS, an EA, or a separate document, are required to be coordinated where appropriate with HUD and USDA.

A.5 RECORD OF DECISION—FORMAT AND CONTENT

The record of decision (ROD) must set forth the reasons for the project decision, based on the material contained in the environmental documents. While cross referencing and incorporation by reference of other documents is appropriate, the ROD should explain the basis for the project decision as completely as possible.

a. *Decision.* Identify the selected alternative. Reference to the final EIS may be used to reduce detail and repetition.

b. *Alternatives considered.* This information can be most clearly organized by briefly describing each alternative (with reference to the final EIS, as above), then explaining and discussing the balancing of values underlying the decision. This discussion must identify the alternative or alternatives which were considered preferable from a strictly environmental point of view. If the selected alternative is other than the environmentally preferable alternative, the ROD should clearly state the reasons for that decision. In addition, if use of Section 4(f) land is involved, the required Section 4(f) approval should be summarized.

For each individual decision (final EIS), the values (economic, environmental, safety, traffic service, community planning, etc.) which are significant factors in the decision-making process may be different and may be given different levels of relative importance. Accordingly, it is essential that this discussion clearly identifies each significant value and the reasons some values were considered more important than others. While any decision represents a balancing of the values, the ROD should reflect the manner in which these values were considered in arriving at the decision.

It is also essential that legislation requirements in 23 U.S.C. be given appropriate weight in this decision-making process. The mission of FHWA is to implement the federal-aid highway program to provide safe and efficient transportation. While this mission must be accomplished within the context of all other federal requirements, the beneficial impacts of transportation improvements must be given proper consideration and documentation in this ROD.

c. *Measures to minimize harm.* Describe all measures to minimize environmental harm which have been adopted for the proposed action. State whether all practicable measures to minimize environmental harm have been incorporated into the decision and, if not, why.

d. *Monitoring or enforcement program.* Describe any monitoring or enforcement program which has been adopted for specific mitigation measures, as outlined in the final EIS.

A.6. SECTION 4(f) EVALUATIONS—FORMAT AND CONTENT

Draft Evaluation—Format

a. Describe proposed action (if separate document)

b. Describe Section 4(f) resource

c. Impacts on resource (by alternative)

d. Avoidance alternatives and their impacts

e. Measures to minimize harm

f. Coordination with appropriate agencies

g. Concluding statement (final document only)

In the case of a complex Section 4(f) involvement, it is desirable to include the analysis in a separate section of the draft EIS, EA, or for projects processed as categorical exclusions, in a separate document. A Section 4(f) evaluation should be prepared for each location within the project where the use of Section 4(f) land is being considered.

Draft Evaluation—Content

The following information should be included in the Section 4(f) evaluation, as appropriate:

a. A brief description of the project and the need for the project (when the Section 4(f) evaluation is circulated separately).

b. A detailed map or drawing of sufficient scale to identify essential elements of the highway/Section 4(f) land involvement.

c. Size (acres or square feet) and location (maps or other exhibits such as photographs, sketches, etc.) of involvement.

d. Type of property (recreation, historic, etc.).

e. Available activities at the property (fishing, swimming, golfing, etc.).

f. Description and location of all existing and planned facilities (ball diamonds, tennis courts, etc.).

g. Usage (approximate number of users/visitors, etc.).

h. Relationship to other similarly used lands in the vicinity.

i. Access (pedestrian and vehicular).

j. Ownership (city, county, state, etc.).

k. Applicable clauses affecting the title, such as covenants, restrictions, or conditions, including forfeiture.

l. Unusual characteristics of the Section 4(f) land (flooding problems, terrain conditions, or other features that either reduce or enhance the value of portions of the area).

m. The location (using maps or other exhibits such as photographs or sketches) and the amount of land (acres or square feet) to be used by the proposed project including permanent and temporary easements.

n. The probable increase or decrease in environmental impacts (noise, air pollution, visual, etc.) of the alternative locations and designs considered on the Section 4(f) land users.

o. A description of all reasonable and practicable measures which are available to minimize the impacts of the proposed action on the Section 4(f) property. Discussions of alternatives in the draft EIS or EA may be referenced rather than repeated.

p. Sufficient information to evaluate all alternatives which would avoid the Section 4(f) property. Discussions of alternatives in the draft EIS or EA may be referenced rather than repeated. However, this section should include discussions of design alternatives (to avoid Section 4(f) use) in the immediate area of the Section 4(f) property.

q. The determination that there are no feasible and prudent alternatives is not normally addressed at the draft EIS, EA, or preliminary document stage until the results of the formal coordination have been completed.

r. The results of preliminary coordination with the public official having jurisdiction over the Section 4(f) property and with regional (or local) offices of DOI and, as appropriate, the regional (or local) office of USDA and HUD.

Section 4(f) Discussion in Final Document

When the selected alternative involves the use of Section 4(f) land, a Section 4(f) evaluation may be included as a separate section in the final EIS or FONSI or for projects processed as categorical exclusions, in a separate final Section 4(f) evaluation. The final evaluation should contain:

a. All information required above for a draft evaluation.

b. A discussion of the basis for the determination that there are no feasible and prudent alternatives to the use of the Section 4(f) land. The supporting information must demonstrate that there are unique problems or unusual factors involved in the use of alternatives and that the cost, environmental impact, or community disruption resulting from such alternatives reaches extraordinary magnitudes.

c. A discussion of the basis for the determination that the proposed action includes all possible planning to minimize harm to the Section 4(f) property.

d. A summary of the appropriate formal coordination with the Headquarters Offices of DOI, and as appropriate, the Headquarters Offices of USDA and HUD.

e. Copies of all formal coordination comments received and an analysis and response to any questions raised.

f. Concluding statement as follows: "Based upon the above considerations, it is determined that there is no feasible and prudent alternative to the use of land from the (Section 4(f) property) and that the proposed action includes all possible planning to minimize harm to the (Section 4(f) property) resulting from such use."

A Section 4(f) approval is the written administrative record which documents the approval required by 23 U.S.C. 138. The Section 4(f) approval will be incorporated into either the final EIS or the ROD. When the Section 4(f) approval is contained in the ROD, the information noted in items (a) through (e) above may be incorporated by reference to the EIS. For a project processed as a categorical exclusion, any required Section 4(f) approval will normally be prepared as a *separate* document.

A.7 PREDECISION REFERRALS TO CEQ

a. Any FHWA office receiving a notice of intent of referral from another agency should provide a copy of that intent of referral to the FHWA Washington Headquarters, Office of Environmental Policy (HEV-10), and the involved Regional Office, Division Office, and HA. This notice of intent of referral would generally be received as part of an agency's comments on the draft EIS. The exception would be when an agency indicates that the draft EIS did not contain adequate information to permit an assessment of the proposal's environmental acceptability. Every reasonable effort should be made to reach agreement with the agency prior to filing of the final EIS. If agreement cannot be reached, the final EIS should document the attempts to resolve the issues and summarize the remaining differences. Prior concurrence of the Washington Headquarters is necessary in the case of government opposition on environmental grounds.

b. The response to the notice of referral will be prepared by the Washington Headquarters with input from the regional, division, and state offices. The FHWA Washington Headquarters will obtain the concurrence of the Department of Transportation prior to the response to CEQ.

c. Upon reviewing the draft EIS from another federal agency, if the FHWA Regional or Division Office believes a referral will be necessary, it should so advise HEV-1. The Office of Environmental Policy (HEV-1) will review the proposed referral and, if appropriate, will advise the Departmental Office of Environment and Safety (P-20), which will coordinate DOT comments on the draft EIS, including the notice of intended referral. Every reasonable effort should be made to resolve the issues after providing notice of intent to refer and prior to the lead agency's filing of the final EIS with EPA. In the event that the issues have not been resolved, the appropriate field office should prepare a referral to CEQ to be submitted through HEV to P-20 for a determination as to whether a referral to CEQ is appropriate.

A.8 OTHER AGENCY STATEMENTS

a. The FHWA review of statements prepared by other agencies will consider the environmental impact of the proposal on areas within FHWA's functional area of responsibility or special expertise.

b. Agencies requesting comments on highway impacts usually forward the draft EIS to the FHWA Washington Headquarters for comment. The FHWA Washington Headquarters will normally distribute these EISs to the appropriate regional office and will indicate where the comments should be sent. The regional office may elect to forward the draft statement to the division office for response.

c. When a field office has received a draft EIS directly from another agency, it may comment directly to that agency if the proposal does not fall within the types indicated in item (d) of this section. If more than one DOT Administration is commenting at the regional level, the comments should be coordinated by the DOT Regional Representative to the Secretary or designee. Copies of the FHWA comments should be distributed as follows:

(1) Requesting agency—original and one copy
(2) P-20—one copy
(3) DOT Secretarial Representative—one copy
(4) HEV-10—one copy

d. The following types of action contained in the draft EIS require FHWA Washington Headquarters review and such EISs should be forwarded to the Associate Administrator for Right-of-Way and Environment (HRE-01), along with regional comments, for processing:

(1) Actions with national implications

(2) Legislation or regulations having national impacts or national program proposals

A.9 PROPOSALS FOR LEGISLATION OR REGULATIONS

Proposals for regulations and legislation will be evaluated by the initiating Washington Headquarters office for compliance with the appropriate NEPA requirements. The proposal may require the development of an EA and FONSI, or an EIS which will be the responsibility of the initiating office in consultation with HEV-10. When a draft EIS for proposed legislation is appropriate, it will be submitted to OST for transmittal to the Office of Management and Budget for circulation in the normal legislative clearance process. Any comments received on the EIS will be transmitted to Congress. Except as provided in 40 CFR Part 1506(b)(2) there need not be a final EIS.

Index