Springer Optimization and Its Applications 84

Lily Elefteriadou

An Introduction to Traffic Flow Theory

Springer

Springer Optimization and Its Applications

VOLUME 84

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown evenmore profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics, and other sciences.

The series Springer Optimization and Its Applications publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository work that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multiobjective programming, description of software packages, approximation techniques and heuristic approaches.

For further volumes:
http://www.springer.com/series/7393

Lily Elefteriadou

An Introduction to Traffic Flow Theory

Springer

Lily Elefteriadou
Department of Civil
and Coastal Engineering
University of Florida
Gainesville, FL, USA

ISSN 1931-6828
ISBN 978-1-4614-8434-9 ISBN 978-1-4614-8435-6 (eBook)
DOI 10.1007/978-1-4614-8435-6
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013949149

Mathematics Subject Classifications: 90B06, 90B20

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Dennis and Jason

Preface

This book is intended for use in a one-semester, first-year graduate course for students focusing in transportation engineering. It can also be used in a senior-level undergraduate class for civil engineering students focusing in transportation engineering. The book can be used as a reference by transportation professionals interested in a refresher on traffic flow theory and applications. Students using the book should have a basic knowledge of transportation engineering and highway design (having attended an introductory course in transportation) as well as a basic knowledge of algebra and statistics. It provides the fundamental principles of traffic flow theory as well as discussion of the application of those principles in the context of specific facility types (freeways, signalized, intersections, etc.). The book does not contain any significant amount of material that cannot be found elsewhere; rather it assembles and presents in a concise manner what the author considers are the most important principles and tools that today's transportation professionals specializing in transportation operations should be well versed in. The book considers advanced technologies to the degree that they are relevant to the principles of traffic flow. When deemed appropriate, the book provides references for obtaining additional information on specific technologies and topics relevant today. The text is supplemented by illustrative examples and applications, and it provides references and resources for further reading.

This book focuses on the traffic operational quality of a facility assuming that the demand for using a particular facility is known; travel demand forecasting or traffic assignment is not within the scope of the book. The emphasis of the book is on highway transportation, primarily because the vast majority of the research and applications of traffic flow theory to date have been automobile focused. Some discussion and examples related to pedestrian flow models are provided throughout the text. Given the recent emphasis on multimodal transportation, it is very likely that the next few years will continue to bring increasing emphasis on the operations of alternative modes as well as the interaction among various modes. Many of the mathematical principles of traffic flow as well as the tools described in this book can be modified and applied to other modes, considering the characteristics of the respective units of traffic (pedestrians, bicycles, buses, etc.).

I have been teaching courses in traffic flow theory for nearly 20 years. This book was developed based on the outline and course notes developed for these courses, which evolved over time to consider new research findings and implementation priorities. Most of the work to develop the book material was undertaken during a sabbatical from the University of Florida during the academic year 2010–2011.

I am indebted to the numerous students who attended my classes and significantly contributed in many different ways to the development of this book. I am particularly indebted to Dr. Cuie Lu, Dr. Alexandra Kondyli, and Ms. Barbara Martin who made significant contributions through their detailed reviews and comments. I am also thankful to the four anonymous reviewers who provided invaluable feedback on an earlier version of the manuscript.

Gainesville, FL, USA Lily Elefteriadou

Notation

α	Acceleration or deceleration
AR	All-red signal duration
AWSC	All-way stop-controlled intersection
C	Cycle length
c	Capacity
D	Density
d	Distance
DCD	Double crossover interchange
DDI	Diverging diamond interchange
D_j	Jam density
DL	Delay (total for all vehicles delayed during a cycle), seconds, or hours
DL_{avg}	Average delay per vehicle (s/veh)
D_{opt}	Optimum density
F	Flow
FFS	Free-flow speed
G	Actual green
g	Effective green
GPSS	General-purpose simulation system
h	Headway
HCM	Highway Capacity Manual
HOT	High-occupancy toll lanes
HOV	High-occupancy vehicle lanes
L_C	Total lost time in a cycle
LOS	Level of service
L_P	Total lost time in a phase
L_s	Start-up lost time
MOE	Measures of effectiveness
mph	Miles per hour
O	Offset
pcphpl	Passenger cars per hour per lane
PHF	Peak hour factor

R	Red signal indication duration
r	Effective red
S	Saturation flow
s	Spacing
S_h	Saturation headway
t	Time
t_c	Critical gap (s)
TT	Travel time
TWSC	Two-way stop-controlled intersection
V	Volume
v	Speed
v_{FF}	Free-flow speed
VHT	Vehicle hours traveled
VMT	Vehicle miles traveled
v_{opt}	Optimum speed
vph	Vehicles per hour
vphpl	Vehicles per hour per lane
vps	Vehicles per second
VSL	Variable speed limits
Y	Yellow

Contents

Introduction

The Role of Traffic Flow Theory in Transportation

Transportation is generally concerned with the efficient, safe, and sustainable movement of people and goods. Transportation engineers work on various aspects of the five stages essential in the life cycle of a transportation facility: planning, designing, building, operating, and maintaining. In the planning stage, we typically forecast traffic demands for a future year/analysis period, perform a preliminary evaluation of alternative solutions, or identify priorities for system improvements. We are typically interested in answering questions such as how many travelers are expected to use our study facility daily, and what are the anticipated changes in the land uses in our study area. In the design stage, we are interested in the specific geometric elements of the selected alternative, such as horizontal and vertical alignment for the proposed facility. After building our facility (which is the purview of construction engineers), the focus shifts to operations and maintenance. In the operations stage, we are concerned with control algorithms (such as ramp metering), traffic management, and other aspects of operations such as incident removal. The maintenance stage involves regular upkeep and repairs such as resurfacing and restripping as well as traffic signal control maintenance.

Traffic flow theory relates primarily to the operations stage, but its tools and methods are used throughout the spectrum of transportation analysis. Traffic flow theory is that part of transportation that is concerned with the capacity and traffic operational quality of transportation facilities. In other words, the objective of traffic flow theory is to evaluate the operational quality of a traffic stream given a set of prevailing conditions, such as the highway design of the facility and the percent of trucks in the traffic stream. Questions of particular interest to traffic engineers are:

- How many lanes should the highway facility have in order to be operating below its capacity?
- How much traffic will the facility handle if it is designed with three lanes?

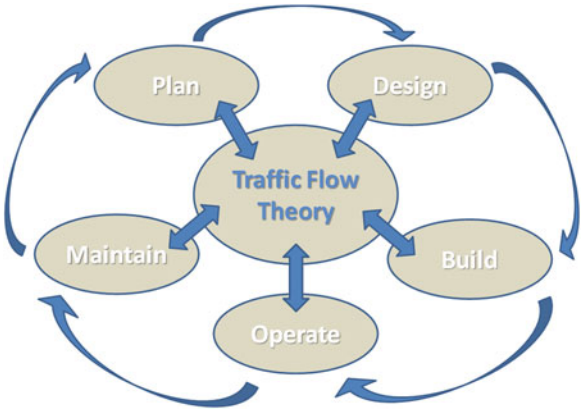


Fig. 1 Traffic flow theory and its relationship to transportation infrastructure

- What would be the operating speed of a given facility if traffic increases by 15 % in the next 3 years?
- Can the signalized intersection operations improve with a change in signal timings, or are geometric design changes needed?

Figure 1 illustrates the interrelationship between traffic flow theory and the five stages of transportation discussed above. Principles of traffic flow theory affect almost every stage of the transportation engineering process. During the planning process, traffic flow theory principles are used to evaluate operations under alternative scenarios. During the design process, specific geometric characteristics are selected to achieve a desired traffic operational quality. During the building process, traffic operations may need to be considered around work zones, and any anticipated congestion in the surrounding network needs to be managed using traffic flow theory principles. The operations stage relies most significantly on traffic flow theory principles, as it focuses primarily on optimizing the efficiency of the network given traffic flow principles, relating traffic control to the built environment and traffic stream characteristics. In the maintenance stage, changes in the traffic patterns and their impact on traffic operations are addressed. Eventually, maintenance activities lead to a new cycle of planning, designing, etc., as land use and traffic patterns change or infrastructure deteriorates.

Key Elements and Interactions in Traffic Flow Theory

As indicated above, the objective of traffic flow theory is to evaluate the operational quality of a traffic stream given a set of prevailing conditions. Operational quality refers to specific performance measures such as speed, delay, or travel time. The traffic stream consists of vehicles with specific size and performance characteristics (e.g., length, weight-to-horsepower, etc.) as well as drivers with specific characteristics and behaviors (e.g., aggressive, unfamiliar with the facility, etc.).

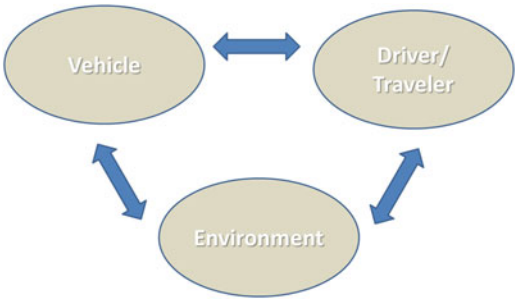


Fig. 2 Key elements affecting traffic flow: environment, vehicle, driver/traveler

The traffic stream may also consist of pedestrians, bicycles, or mass transit vehicles. Traffic flow theory is also concerned with the operations and interactions between these different modes. Prevailing conditions refer to the design (e.g., horizontal alignment, grade, etc.) and control elements (e.g., presence of a signal or a STOP sign) of a facility as well as other external conditions such as rain or incidents. In summary, traffic flow theory seeks to understand the interactions between three fundamental elements: the vehicle, the driver/traveler, and the environment (Fig. 2). The characteristics of each of those elements as well as their interactions are of primary importance to traffic engineers.

As an example, let us consider the movement of a truck along a level-terrain highway. The truck characteristics (such as its weight-to-horsepower) would affect the acceleration and maximum speed it can reach. If the highway has a steep upgrade, that acceleration and maximum speed would be significantly reduced. Thus, the interaction of the vehicle characteristics with the highway characteristics has an effect on the performance of the vehicle. Now consider that the truck is traveling on a two-lane highway along a no-passing zone. Its speed would significantly affect the speed of the following vehicles and ultimately the operations of the facility (average travel time, capacity, etc.).

As a second example, let us consider a driver in a passenger vehicle waiting to make a left turn into a major highway. The driver needs to select a suitable gap in the mainline traffic and complete the maneuver. If the driver is aggressive, he or she would select a shorter gap. If the majority of the drivers making this maneuver tend to be aggressive (say, making a left turn out of a college campus), the total number of vehicles turning into the mainline would tend to be higher and the capacity of the approach would be higher. In this case, the driver characteristics are quite significant. Now let us consider a driver who is distracted talking on the cell phone and does not immediately notice the availability of a suitable gap. The capacity of the approach could be significantly reduced by the particular driver action which would occur rather randomly.

Traffic flow theory draws from many different disciplines for developing appropriate analysis tools. It is affected by vehicle dynamics and vehicle performance issues (mechanical engineering) as well as by human factors, psychology, and sociology. It also draws on statistics and operations research principles and physics. The highway environment is generally defined by policy decisions (land use,

economic, etc.) which also impact traffic operational quality. Thus, it is important to understand each of the elements shown in Fig. 2 as well as their interactions so that we can better develop and operate our transportation networks.

Book Organization

This book is organized into four parts. The first part focuses on individual vehicle movement and presents equations of motion for modeling the acceleration and deceleration of single vehicles as well as interactions at the vehicle level (car-following, lane changing). The second part focuses on operations at the traffic stream level. It presents the three fundamental characteristics of traffic (speed, flow, density) as well as their interrelationships in the form of traffic stream models. It also discusses in detail capacity and other performance measures used to describe a traffic stream. The third part formulates and uses several different traffic analysis techniques for bottlenecks and queuing systems, including mathematical modeling, empirical modeling, and simulation. The last part discusses the characteristics and operations of specific highway facilities focusing on freeways, signalized and unsignalized intersections, and two-lane highways.

Part I
Modeling Traffic: Vehicle–Driver–Driving Environment Interactions

This first part of this book focuses on the movement of the basic unit of traffic, i.e., the single vehicle. This movement is a function of the fundamental equations of motion as well as the interactions between vehicle, driver, and environment (discussed in the Introduction and presented in Fig. 2). The principles presented in this part form the basis for understanding the performance of traffic streams and the principles of traffic flow theory.

Chapter 1 discusses the mathematical formulations describing the motion of a vehicle. The first part of the chapter focuses on the equations of motion and provides a series of acceleration, speed, and distance functions. The vehicle, driver, and environment characteristics and their effects on vehicle movement are discussed in the last part of the chapter. Chapter 2 presents the principles of the motion of a group of vehicles, i.e., car-following, lane-changing, and gap acceptance models. The chapter discusses primarily interactions between vehicles and also the effects of vehicle characteristics, driver characteristics, and driving environmental factors on these interactions.

Chapter 1
Modeling the Motion of a Single Vehicle

The movement of a single vehicle can impact significantly the performance of a traffic stream. Consider a slow-moving vehicle at the front of a long line of vehicles traveling behind it, with no-passing opportunities. The performance of this lead vehicle significantly affects the speed and travel time of the following vehicles. Understanding the characteristics, performance, and movement of each vehicle type allows us to model groups of vehicles and to evaluate the performance of the traffic stream. Furthermore, the movement of individual vehicles allows us to develop better highway design and traffic control solutions. For example, understanding the acceleration and deceleration constraints of various vehicle types can help us design more effective passing zones and to allocate appropriate yellow and all-red intervals at signalized intersections.

This chapter first discusses the basic equations of motion for a single vehicle. The second part of the chapter discusses vehicle, driver, and environmental characteristics and their effects on vehicle motion.

Motion of a Single Vehicle

The motion of a single vehicle can be fully described mathematically using equations of motion. These are based on laws of physics and provide relationships between key parameters such as acceleration, speed, distance traveled, and travel time. This section derives and presents these equations of motion for various types of movement (constant speed, constant acceleration, and varying acceleration). The discussion in this section is based to a large degree on [1].

Figure 1.1 presents a diagram with the *trajectory* of a vehicle, i.e., a plot providing the vehicle's position as a function of time. The horizontal axis of the figure indicates the time, while the vertical axis indicates the respective location of the vehicle. This type of diagram is called a *time–space diagram*. The trajectory of the vehicle is the line connecting the location–time points along the vehicle's path.

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_1,

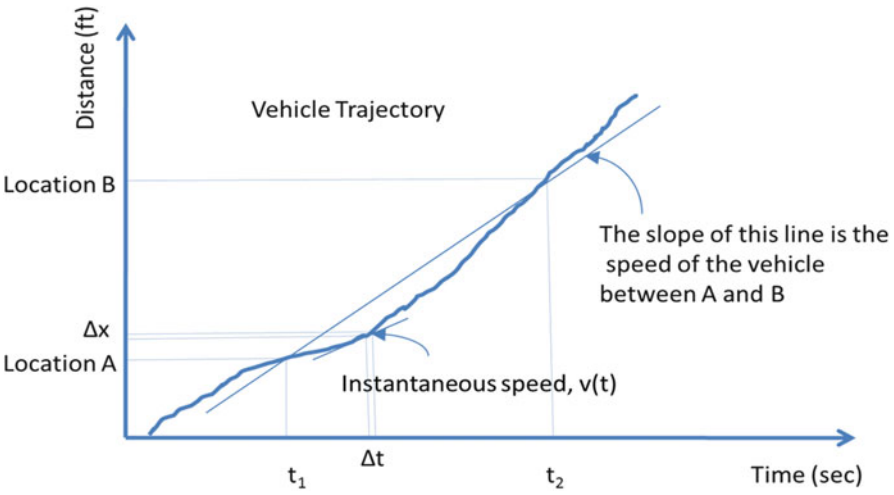


Fig. 1.1 Trajectory of a vehicle

The average speed of the vehicle between locations A and B (V_{AB}) can be determined as follows:

$$V_{AB} = d_{AB}/(t_2 - t_1)$$

where d_{AB} is the distance between points A and B (ft).

Visually, the average speed of the vehicle is determined by the slope of the line connecting the two points as shown in Fig. 1.1. The instantaneous speed of the vehicle, $v(t)$, i.e., the speed for a very small portion of the trajectory, is the respective slope of the line connecting the two points. The steeper the slope of the line, the higher the speed of the vehicle is, as it travels a longer distance in a shorter amount of time. As shown in Fig. 1.1, the speed of the vehicle near location A is lower than the average speed between points A and B, as the slope of that line (labeled instantaneous speed) is smaller.

Mathematically, the instantaneous speed of the vehicle at time t, $v(t)$, can be provided by:

$$v(t) = \mathrm{d}x/\mathrm{d}t$$

where dx is the change in distance and dt the change in time.

The distance covered by the vehicle in time dt can be determined as:

$$v(t)\mathrm{d}t = \mathrm{d}x$$

By integrating both parts of this equation, we have:

$$\int_{t_0}^{t} \mathrm{d}x = \int_{t_0}^{t} v(t)\mathrm{d}t \Rightarrow$$
$$x(t) - x(t_0) = \int_{t_0}^{t} v(t)\mathrm{d}t \Rightarrow$$
$$\boxed{x(t) = x(t_0) + \int_{t_0}^{t} v(t)\mathrm{d}t} \tag{1.1}$$

This is the general equation for determining the distance traveled as a function of speed. Given any speed function, this equation can provide the distance traveled in time t, $x(t)$.

The change in speed per unit of time is called acceleration (a). Mathematically, it is described as a function of speed as follows:

$$a(t) = \mathrm{d}v/\mathrm{d}t$$

When the final speed of the vehicle is lower than its initial speed, the sign of $\mathrm{d}v$ is negative and the vehicle is decelerating. Throughout this book, we will be referring to both types of movements as acceleration, and we will be using acceleration functions.

Given the relationship between speed and distance provided above, acceleration can also be described as:

$$a(t) = d^2x/\mathrm{d}t^2$$

The speed function can be determined as follows:

$$a(t)\mathrm{d}t = \mathrm{d}v \Rightarrow$$
$$\int_{t0}^{t} \mathrm{d}v = \int_{t0}^{t} a(t)\mathrm{d}t \Rightarrow$$
$$v(t) - v(t_0) = \int_{t0}^{t} a(t)\mathrm{d}t \Rightarrow$$
$$\boxed{v(t) = v(t_0) + \int_{t0}^{t} a(t)\mathrm{d}t} \tag{1.2}$$

This is the general equation for determining the final speed of the vehicle, or the speed at the end of a time interval as a function of its initial speed, and the acceleration or deceleration function.

The change in acceleration per unit of time is called jerk (j). Jerk is also referred to as jolt. Mathematically, it is described as follows:

$$j(t) = \mathrm{d}a/\mathrm{d}t = d^2v/\mathrm{d}t^2 = d^3x/\mathrm{d}t^3$$

In traffic movement, jerk is important to consider, as high values are uncomfortable to passengers. Thus, particular values of jerk may be used in traffic analysis as an upper limit when modeling vehicle movement.

The Case of Constant Speed

If the speed $v(t)$ is constant, the acceleration is zero:

$$v(t) = \text{constant} \Rightarrow a(t) = 0$$

Using Eq. (1.1), we can obtain the distance function for this case as follows:

$$x(t) = x(t_0) + \int_{t_0}^{t} v(t)\mathrm{d}t \Rightarrow$$
$$x(t) = x(t_0) + \int_{t_0}^{t} vt \Rightarrow$$
$$\boxed{x(t) = x_0 + v(t - t_0)} \tag{1.3}$$

Assuming that the initial distance $x(t_0)$ and the initial time t_0 are zero, the equation becomes:

$$\boxed{x(t) = vt} \tag{1.4}$$

Figure 1.2 illustrates graphically the case of constant speed. Figure 1.2a shows the speed of the vehicle over time, while Fig. 1.2b shows the trajectory of the vehicle in a time–space diagram.

Since integration of a function provides the area under the curve, the distance can also be estimated as the area under the speed curve.

The Case of Constant Acceleration

As indicated earlier, acceleration is provided as a function of speed as follows:

$$a(t) = \mathrm{d}v/\mathrm{d}t \Rightarrow$$
$$\mathrm{d}v = a(t)\mathrm{d}t$$

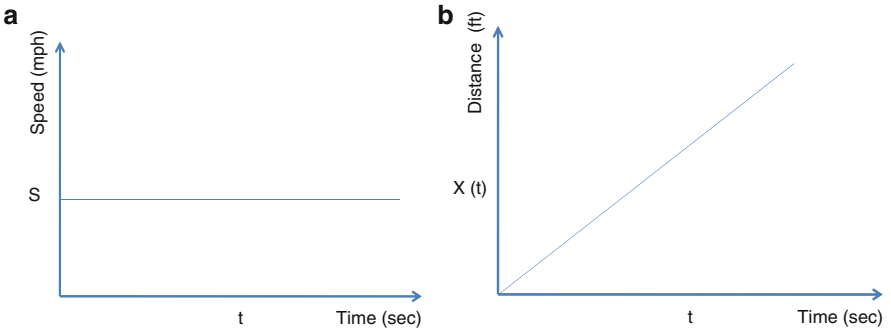


Fig. 1.2 The case of constant speed

By integrating both parts of the equation:

$$\int_{t0}^{t} \mathrm{d}v = \int_{t0}^{t} a(t)\mathrm{d}t \Rightarrow$$

$$v(t) - v(t_0) = \int_{t0}^{t} a(t)\mathrm{d}t \Rightarrow$$

$$v(t) = v(t_0) + \int_{t0}^{t} at \Rightarrow$$

$$\boxed{v(t) = v_0 + a(t - t_0)} \tag{1.5}$$

Assuming that the initial speed $v(t_0)$ and the initial time t_0 are zero, the equation becomes:

$$\boxed{v(t) = at} \tag{1.6}$$

The distance can be calculated using Eq. (1.3) by inserting the speed function of Eq. (1.5):

$$x(t) = x_0 + \int_{t0}^{t} [v(t_0) + a(t - t_0)]\mathrm{d}t \Rightarrow$$

$$x(t) = x_0 + v_0(t - t_0) + \int_{t0}^{t} a(t - t_0)\mathrm{d}t \Rightarrow$$

$$x(t) = x_0 + v_0(t - t_0) + \int_{t0}^{t} a(t)\mathrm{d}t - \int_{t0}^{t} a(t_0)\mathrm{d}t \Rightarrow$$

$$x(t) = x_0 + v_0(t - t_0) + \int_{t0}^{t} (at)\mathrm{d}t - \int_{t0}^{t} (at_0)\mathrm{d}t \Rightarrow$$

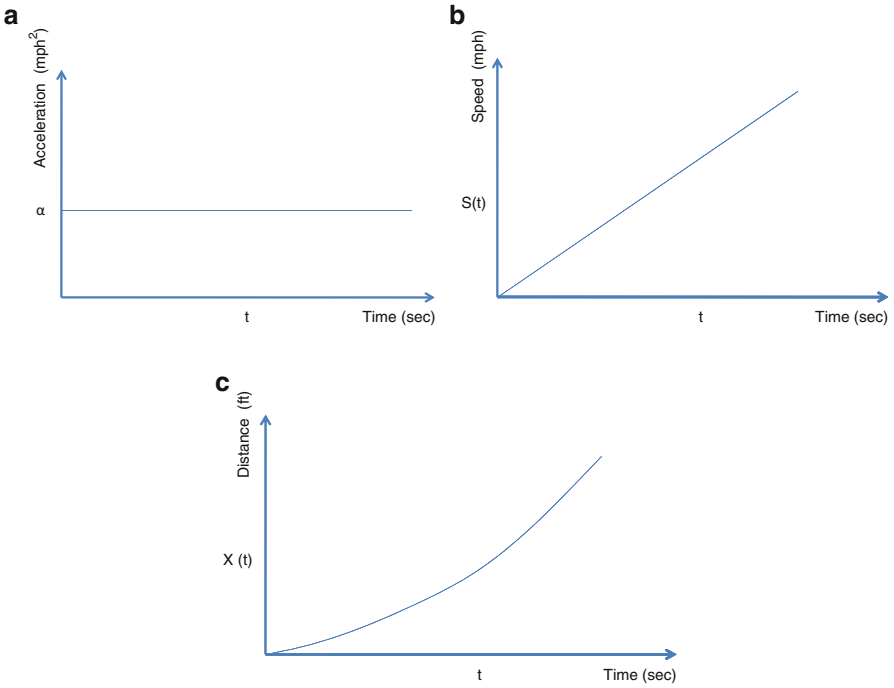


Fig. 1.3 The case of constant acceleration

$$x(t) = x_0 + v_0(t - t_0) + \int_{t_0}^{t} \frac{1}{2} a\, t^2 - \int_{t_0}^{t} a t_0 t$$

$$\boxed{x(t) = x_0 + v_0(t - t_0) + \frac{1}{2}\alpha\left(t^2 - t_0^2\right) - \alpha t_0 (t - t_0)} \tag{1.7}$$

Figure 1.3 illustrates graphically the case of constant acceleration. Figure 1.3a shows the acceleration of the vehicle over time. Figure 1.3b shows the speed of the vehicle, while Fig. 1.3c shows the trajectory of the vehicle in a time–space diagram.

The Case of Varying Acceleration

Realistically, acceleration (and its negative, deceleration) is not constant in time. Acceleration capabilities decrease as the speed of the vehicle increases [2]. As shown in Eq. (1.2), the speed of a vehicle can be obtained based on the acceleration function as follows:

$$v(t) = v(t_0) + \int_{t0}^{t} a(t)\mathrm{d}t$$

Let us assume that the acceleration function is linear and is given by the following equation:

$$\alpha(t) = y - zt$$

Then, the speed at time t, $v(t)$, can be estimated by substituting this function into Eq. (1.2) and completing the integration:

$$\begin{aligned}
v(t) &= v(t_0) + \int_{t0}^{t} (y - zt)\mathrm{d}t \Rightarrow \\
v(t) &= v(t_0) + \int_{t0}^{t} (y)\mathrm{d}t - \int_{t_0}^{t} (zt)\mathrm{d}t \Rightarrow \\
v(t) &= v(t_0) + \int_{t_0}^{t} yt - \int_{t_0}^{t} \left(zt^2\right)/2 \Rightarrow \\
v(t) &= v(t_0) + y(t - t_0) - \frac{z(t - t_0)^2}{2}
\end{aligned}$$

As in the case of constant acceleration, the distance can be calculated using Eq. (1.3):

$$\begin{aligned}
x(t) &= x(t_0) + \int_{t0}^{t} v(t)\mathrm{d}t \Rightarrow \\
x(t) &= x(t_0) + \int_{t_0}^{t} \left[v(t_0) + y(t - t_0) - \frac{z\,(t - t_0)^2}{2} \right] \mathrm{d}t \Rightarrow \\
x(t) &= x(t_0) + v(t_0)t + \frac{y(t - t_0)^2}{2} - \frac{z(t - t_0)^3}{6}
\end{aligned}$$

Assuming the initial time and distance are zero, the distance can be estimated as:

$$x(t) = v(t_0)t + \frac{y(t)^2}{2} - \frac{z(t)^3}{6}$$

Figure 1.4 illustrates graphically the case of linearly decreasing acceleration. Figure 1.4a shows the acceleration of the vehicle over time. Figure 1.4b shows the speed of the vehicle, while Fig. 1.4c shows the trajectory of the vehicle in a time–space diagram. For the case of varying acceleration, the shapes of these graphs would change based on the acceleration function.

Equations that estimate stopping sight distance, braking distance, and other such design elements are based on a constant acceleration. For example, the AASHTO

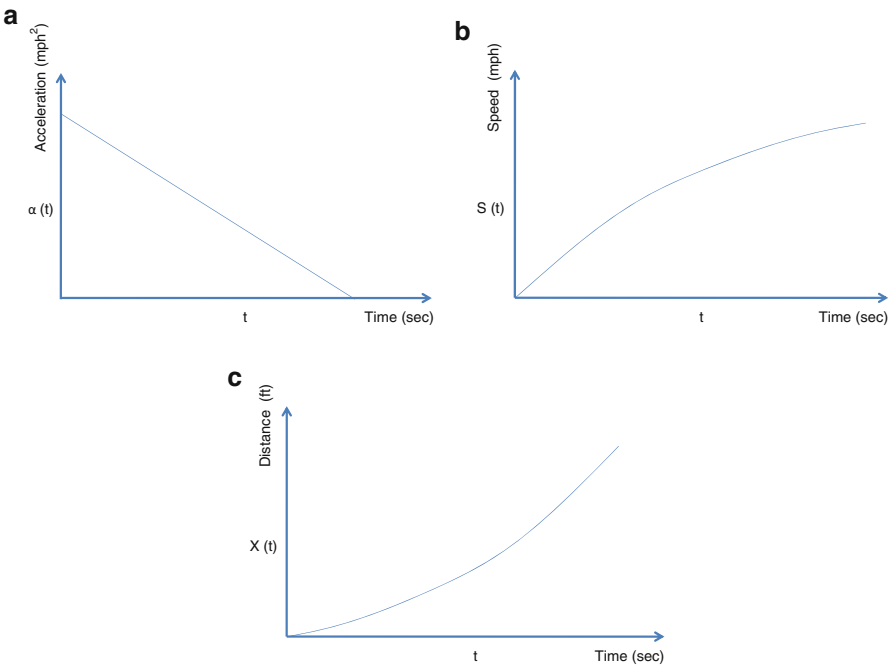


Fig. 1.4 The case of linearly decreasing acceleration

Green Book [3] provides the following equation for obtaining the stopping sight distance:

$$\boxed{\text{SSD} = 1.47vt + 1.075\,\frac{v^2}{a}} \tag{1.8}$$

where

SSD is stopping sight distance, in ft
v is design speed, in mph
t is brake reaction time, in s (recommended value is 2.5 s)
α is acceleration rate, in ft/s^2

The first component in Eq. (1.8) represents the distance traveled during reaction time, while the second component represents the braking distance. Coefficients 1.47 and 1.075 are used for unit conversions.

A discussion of realistic acceleration functions for various types of vehicles is provided later in the chapter.

Example 1.1 A vehicle approaches a traffic signal at a speed of 45 mph. When the vehicle is 200 ft away from the approach stop bar, the signal indication turns yellow

for 3 s and then red. Assuming that the driver's reaction time is 1 s, and the vehicle's deceleration function is $\alpha(t) = -12 - 1.5$ t, provide the following:

(a) If the driver decides to apply the brakes, will the vehicle be able to stop safely at the signal?
(b) If the driver decides to proceed through the intersection at the vehicle's initial speed of 45 mph, how long will it take the vehicle to clear the signal? Will it be able to cross without running the red light?
(c) Calculate part (a) assuming the vehicle's deceleration is constant and equal to -17 ft/s^2. What are the effects of this assumption on the estimated stopping distance?

Solution to Example 1.1
Part (a)
Using the equations of motion presented earlier, we first obtain the speed function that corresponds to the acceleration function provided. The speed function is estimated using Eq. (1.2) as follows:

$$v(t) = v(t_0) + \int_{t0}^{t} a(t)\mathrm{d}t$$
$$v(t) = v(t_0) + \int_{t0}^{t} (-12 - 1.5\,t)\mathrm{d}t \Rightarrow$$
$$v(t) = v(t_0) + \int_{t0}^{t} -12\ \mathrm{d}t - \int_{t_0}^{t} 1.5t\ \mathrm{d}t \Rightarrow$$
$$v(t) = v(t_0) + \int_{t_0}^{t} -12\mathrm{d}t - \int_{t_0}^{t} 1.5t^2/2 \Rightarrow$$
$$v(t) = v(t_0) - 12t - \left(1.5\,t^2\right)/2$$

We can solve this equation for $v(t) = 0$ mph to determine how much time it would take for the vehicle to stop with the given acceleration function. The resulting equation is of quadratic form:

$$(1.5/2)t^2 + 12\,t - 45(5,280/3,600) = 0$$

This equation has two solutions ($t_1 = -20.3$ s and $t_2 = 4.3$ s) with only the second one being viable. Thus, with the given acceleration function, the vehicle will stop in 4.3 s.

Next, we can obtain the distance function so that we can determine what distance the vehicle would travel during the time of 4.3 s. The distance function is estimated using Eq. (1.1) as follows:

$$x(t) = x(t_0) + \int_{t_0}^{t} v(t)\mathrm{d}t$$

$$x(t) = x(t_0) + \int_{t_0}^{t} \left[v(t_0) - 12t - \left(1.5t^2\right)/2\right] \mathrm{d}t$$

$$x(t) = x(t_0) + v(t_0)t - \frac{12(t)^2}{2} - \frac{1.5(t)^3}{6}$$

Using this distance function, we can estimate the distance traveled during the 4.3 s:

$$x(t) = 45(5,280/3,600) \times 4.3 - \frac{12(4.3)^2}{2} - \frac{1.5\ (4.3)^3}{6} = 153.0\,\text{ft}$$

This is the distance traveled after the vehicle starts decelerating. We also need to account for the distance the vehicle will travel during the driver's reaction time and before applying the brakes. Based on Eq. (1.4), the vehicle will travel at its initial speed of 45 mph for a distance of

$$x(t) = vt = 45(5,280/3,600) \times 1 = 66\,\text{ft}$$

Thus, the vehicle will travel for a total of 66 + 153 = 219 ft before stopping, which is longer than the initial distance of 200 ft. Figure 1.5 plots the vehicle's deceleration, speed, and distance as a function of time for the subject vehicle. In the graphs, $t = 0$ s is the time deceleration starts, while $t = -1$ is the time the signal turns yellow. In the speed graph, initial speed is constant and equal to 66 ft/s (=45 mph) during the driver's reaction time; it starts to drop after $t = 0$. In the time–space diagram, $x = 0$ is the location where the vehicle begins decelerating, and $x = -66$ ft is the initial location of the vehicle. As estimated above and shown in Fig. 1.5, the vehicle will not be able to stop safely with the given acceleration function.

Part (b)
In this case, the vehicle travels through the intersection without decelerating. The distance covered during the yellow interval is estimated using Eq. (1.4):

$$x(t) = vt = 45 \times (5,280/3,600) \times 3 = 198\,\text{ft}$$

The vehicle will narrowly miss the yellow interval and will cross the intersection during the red. We can also estimate the time the vehicle will take to reach the stop bar as follows:

$$x(t) = vt \Rightarrow$$
$$t = x(t)/v \Rightarrow$$
$$t = 200/[45 \times (5,280/3,600)] = 3.03\ \ s$$

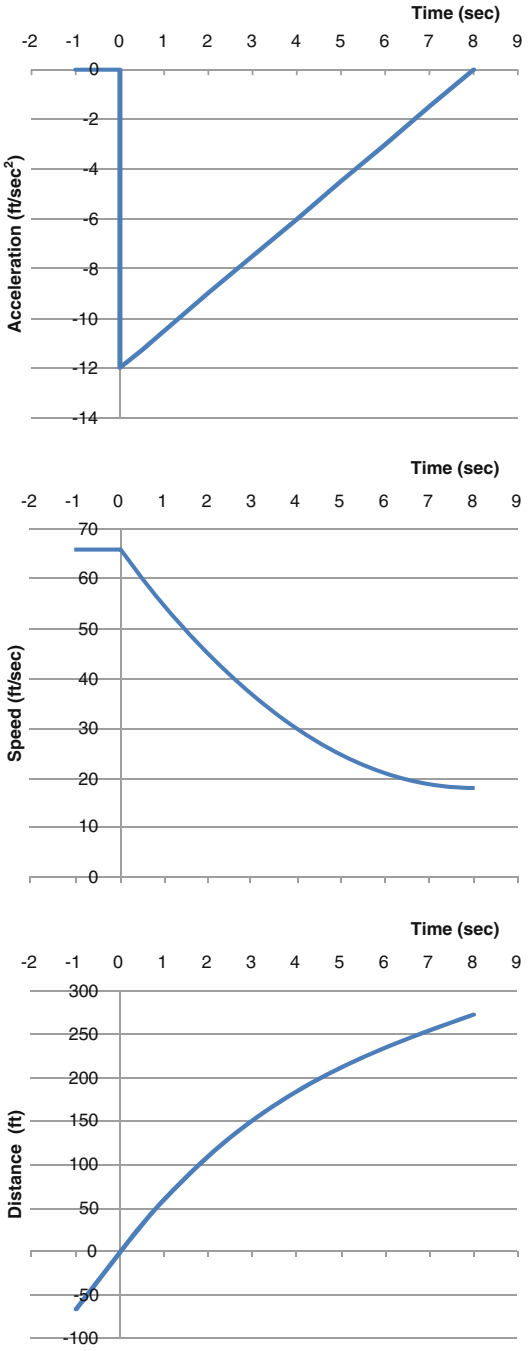


Fig. 1.5 Acceleration function, speed function, and vehicle trajectory for Example 1.1 part (a)

Part (c)
In this case, the vehicle will travel at its initial speed for 1 s, and then, it will decelerate at a constant deceleration rate. To answer this question, we need to estimate the distance the vehicle will travel during the driver's reaction time (1 s) and the time and distance it will travel before stopping.

The distance the vehicle will travel at its initial speed is obtained using Eq. (1.4):

$$x(t) = vt = 45 \times (5,280/3,600) \times 1 = 66\,\text{ft}$$

The speed function for the case of constant acceleration is obtained from Eq. (1.5). We can solve this equation for $v(t) = 0$ mph to determine how much time it would take for the vehicle to stop with the given acceleration rate:

$$\begin{gathered} v(t) = v_0 + a(t - t_0) \Rightarrow \\ 0 = 45 \times (5,280/3,600) - 17t \Rightarrow \\ t = 3.9s \end{gathered}$$

The distance the vehicle will travel while decelerating at a constant rate is obtained using Eq. (1.7) for $t = 3.9$ s and $t_0 = 0$:

$$x(t) = x_0 + v_0(t - t_0) + \frac{1}{2}\alpha\left(t^2 - t_0^2\right)$$

$$x(t) = 45(5,280/3,600) \times 3.9 - \frac{17(3.9)^2}{2} = 128.1\,\text{ft}$$

This is the distance traveled after the vehicle starts decelerating. We also need to account for the distance the vehicle will travel during the driver's reaction time and before applying the brakes. That distance was estimated in part (a) to be 66 ft. Thus, the vehicle will travel for a total of 66 + 128.1 = 194.1 ft before stopping, which is shorter than the initial distance of 200 ft. Thus, the vehicle will stop safely in this case. However, the deceleration rate indicated (-17 ft/s^2) is very high for most vehicle types.

Equations of Motion as a Function of Distance and Speed

Equations of motion can also be derived as a function of distance or speed. For example, we can obtain time as a function of distance and speed as follows:

$$t(x) = t(x_0) + \int_{x0}^{x} \frac{\mathrm{d}x}{v(x)}$$

We can also obtain distance as a function of speed and acceleration as follows:

$$x(v) = x(v_0) + \int_{v0}^{v} \frac{v}{a(v)} \mathrm{d}v$$

These functions are the inverse of the ones presented earlier and result in the same trajectories. However, occasionally it is essential or convenient to use an independent variable other than time. For example, in automobiles acceleration is a function of the vehicle's speed: it is higher for lower gears and lower for higher gears. In this case, acceleration cannot be provided as a function of time; it can only be provided as a function of speed, and thus, it is essential to use the speed-based equations. For additional information on the development of such functions, consult [1] which provides a very thorough discussion.

Vehicle Trajectories and Traffic Performance

When the trajectory of a vehicle is known, we can obtain from it a complete set of performance measures and related statistics. For example, we can obtain the average speed of the vehicle over a given distance; we can obtain its average acceleration; and we can obtain the delay a vehicle encounters by comparing an ideal travel time to the actual travel time of the vehicle (additional discussion of specific performance measures and their definitions is provided in Chap. 5). Other useful measures that can be obtained are the standard deviation of speed, the standard deviation of the acceleration, the speed distribution, etc.

As indicated earlier, the vehicle trajectory is the basic building block in understanding traffic operations. When the vehicle trajectories of a traffic stream are known, we can estimate any of the parameters listed above for the entire traffic stream. The development of trajectories for interacting vehicles is discussed in Chap. 2. The remainder of the chapter discusses vehicle, driver, and environment characteristics as these affect vehicle trajectories and traffic operations.

Effects of Vehicle Characteristics on the Motion of a Single Vehicle

Vehicle characteristics are a key component in traffic flow theory and traffic operations. Most importantly, it is the size and weight of the vehicle as well as its engine characteristics that affect the vehicle acceleration function which in turn affects vehicle speed. Another set of important characteristics for a vehicle are its length and trailer coupling which affect the amount of space it occupies, as well as its turning capability.

There are several different types of vehicles. The Federal Highway Administration (FHWA) distinguishes 13 different vehicle classes [4, 5]. These are described below and are also illustrated in Fig. 1.6:

Motorcycles: These are two- or three-wheeled motorized vehicles, typically steered by handlebars rather than steering wheels. This class includes motorcycles, motor scooters, mopeds, motor-powered bicycles, and three-wheeled motorcycles.

Passenger cars: This class includes all sedans, coupes, and station wagons used for carrying passengers; they include those passenger cars pulling recreational or other light trailers.

Other two-axle, four-tire single-unit vehicles (*SUVs*): This class includes all other passenger cars with two axles and four tires. It includes pickups, vans, campers, motor homes, ambulances, and minibuses.

Buses: These are vehicles used for carrying passengers, and they typically have two axles and six tires or three or more axles. This category includes only traditional buses (including school buses) functioning as passenger-carrying vehicles. Modified buses should be considered to be a truck and should be appropriately classified.

Two-axle, six-tire, single-unit trucks: This class includes all vehicles on a single frame including trucks, camping and recreational vehicles, motor homes, etc. with two axles and dual rear wheels.

Three-axle single-unit trucks: This class includes all vehicles on a single frame including trucks, camping and recreational vehicles, motor homes, etc. with three axles.

Four or more axle single-unit trucks: This class includes all trucks on a single frame with four or more axles.

Four or fewer axle single-trailer trucks: This class includes all vehicles with four or fewer axles consisting of two units, one of which is a tractor or straight truck power unit.

Five-axle single-trailer trucks: This class includes all five-axle vehicles consisting of two units, one of which is a tractor or straight truck power unit.

Six or more axle single-trailer trucks: This class includes all vehicles with six or more axles consisting of two units, one of which is a tractor or straight truck power unit.

Five or fewer axle multi-trailer trucks: This class includes all vehicles with five or fewer axles consisting of three or more units, one of which is a tractor or straight truck power unit.

Six-axle multi-trailer trucks: This class includes all six-axle vehicles consisting of three or more units, one of which is a tractor or straight truck power unit.

Seven or more axle multi-trailer trucks: This class includes all vehicles with seven or more axles consisting of three or more units, one of which is a tractor or straight truck power unit.

The movement of a vehicle is dictated by two opposing forces: the tractive effort and the resistance. Tractive effort is the force generated by the vehicle engine to

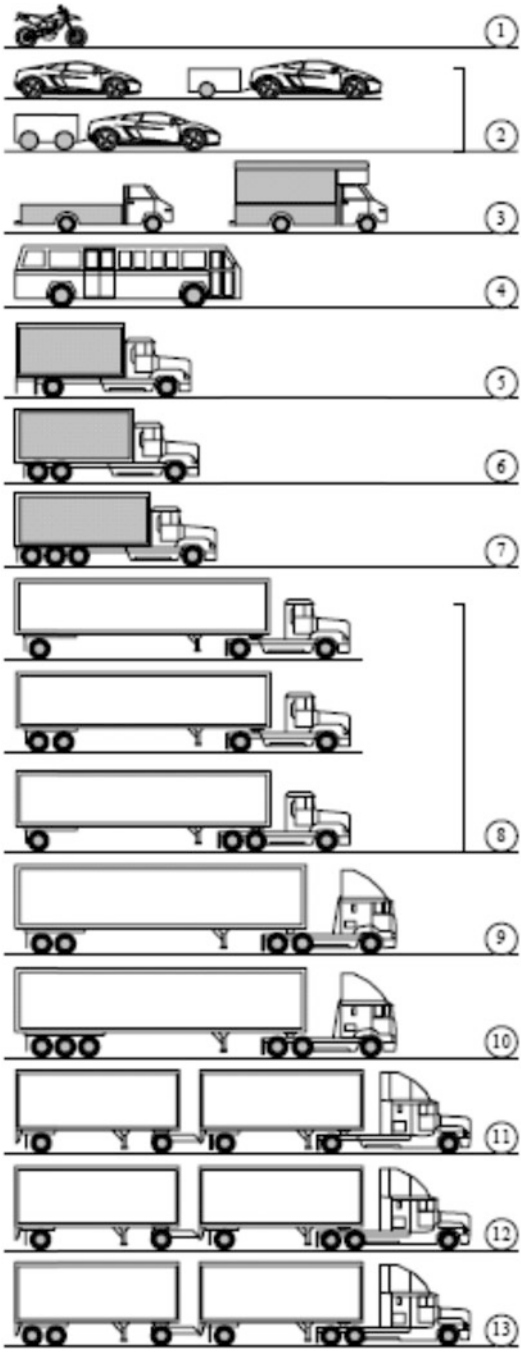


Fig. 1.6 FHWA vehicle types (From Jazar, R.N., *Vehicle Dynamics Theory and Applications*, Figure 1.21, page 27; Reproduced with permission of Springer-Verlag GmbH)

accelerate the vehicle, and it depends on the engine's horsepower and the vehicle's weight. Resistance is the force that hinders vehicle motion, and its sources can be aerodynamic, rolling (roadway–tire interaction), or gravitational. We will not discuss these in detail here, but we will briefly discuss specific vehicle characteristics that affect traffic operations. The interested reader can consult [6] for a more extended discussion of these forces as they apply to vehicle movement.

The vehicle characteristics that mostly affect traffic operations are:

Wt/Hp (weight-to-horsepower ratio): The Wt/Hp provides a measure of the vehicle load to the engine power of the vehicle. It affects the maximum speed a vehicle can attain on steep upgrades (crawl speed), as well as its acceleration capabilities, both of which have a significant impact on the vehicle's trajectory and the traffic stream's operations. Trucks with heavier loads and less engine power generally operate at lower acceleration rates, particularly on steep upgrades. Slower-moving vehicles are particularly detrimental to traffic operational quality when there are minimal passing opportunities for other vehicles in the traffic stream. For passenger cars this is not a very important characteristic because there is not high variability between different types of passenger cars. However, there is significant variability in truck performance.

Width, length, and trailer coupling: The width of a vehicle would affect its speed, particularly when traveling in narrower lanes. Furthermore, it may affect traffic operations in adjacent lanes by forcing other vehicles to slow down when passing. The width, length, and trailer coupling affect the off-tracking characteristics of a vehicle (i.e., the degree to which the path of its rear wheels does not follow the path of the front wheels). They also affect lane design and required lane widths, particularly along horizontal curves. The encroachment of heavy vehicles in adjacent lanes affects their usability by other vehicles and thus has an impact on operations.

Braking and deceleration capabilities: Braking is a complex function which is dictated by the vehicle's braking system (torque of the brakes, presence of antilock system, etc.) as well as resistance forces [2]. The deceleration capability of a vehicle decreases with increasing size and weight.

Frontal area cross section: Aerodynamic drag affects the acceleration of a vehicle.

Vehicle height: The vehicle height, even though not typically included in traffic analysis procedures, may affect the sight distance for following vehicles and thus may affect the resulting spacing and time headways and, ultimately, the capacity of a highway facility.

There is wide variability in the performance of different vehicles even for those within the same class. This makes the estimation and prediction of trajectories in a traffic stream a challenge. Consider that the most important performance characteristic (Wt/Hp) varies even for the same truck when that is loaded versus when it is empty!

Also, acceleration is higher at lower speeds, and it decreases gradually as the vehicle reaches higher speeds. The acceleration fluctuates significantly, and thus modeling it presents a challenge.

Table 1.1 Vehicle characteristics' default values (*Source*: [9])

Vehicle type	Length (ft)	Maximum speed on level grade (mi/h)	Maximum acceleration starting from zero speed (ft/s^2)	Maximum comfortable deceleration (ft/s^2)	Jerk (ft/s^3)
Passenger car	14	75	10	15	7
Single-unit truck	35	75	5	15	7
Semi-trailer truck	53	67	3	15	7
Double-bottom trailer truck	64	61	2	15	7
Bus	40	65	5	15	7

Average acceleration rates for passenger vehicles are typically in the range between 3 and 5 ft/s^2, while those of trucks are lower, and depending on the type of truck can be as low as 0.15 ft/s^2 [7].

The 13 classes described above are often combined based on the purposes of the classification. For example, the second and third classes of vehicles are often combined into one category named "passenger cars" because automated vehicle classifiers have difficulty distinguishing between them. The Highway Capacity Manual (HCM 2010 [8]), which is the preeminent reference for traffic operational analysis in the USA, distinguishes between only three categories: passenger cars, trucks, and buses/RVs. In this case, the analysis only considers the performance of a "typical" truck (200 Ht/Hp); vehicle length or other characteristics are not considered explicitly. For design purposes, AASHTO groups vehicles into four categories: passenger cars, buses, trucks, and recreational vehicles [3]. Reference [9] groups vehicles into five categories and provides default values for their characteristics that are significant in traffic analysis. Table 1.1 provides these five categories along with the respective characteristics (additional information on the use of these characteristics is provided in Chap. 7—Simulation). Advances in vehicle design and technology are likely to affect these characteristics; thus, it is advisable to reevaluate such default values based on vehicle changes and traffic stream composition. Advances related to vehicle interaction and vehicle-to-infrastructure interaction are also very important, and those are discussed in Chap. 2.

The vehicle characteristics that are important in traffic operations are not necessarily the same ones that are important for highway design purposes. In highway design, the analyst is primarily interested in identifying the design elements that would be adequate or functional for a specific facility based on the types of vehicles using it. Thus, the objective is to identify the "design vehicles," i.e., the vehicle or group of vehicles that should be accommodated within the facility. The designer is interested in the braking and deceleration of the vehicle to estimate components such as stopping sight distances, as well as in the turning characteristics of the vehicle to estimate the width of the turning path (for a detailed discussion of the truck characteristics affecting highway design, consult [10]). In this case, the percent of each vehicle type in the traffic stream is not as important;

what is important is to correctly identify the design vehicles for the subject facility. However, for evaluating traffic operations, we are interested not only in the vehicle types and their characteristics but also in the percent of various vehicles in the traffic stream and in the variability of their characteristics. The significance of this variability will become more apparent in later chapters of this book.

Effects of Driver Characteristics and Behavior on the Motion of a Single Vehicle

The vehicle characteristics discussed above provide a range of possible movements for a given vehicle; however, individual driver characteristics, actions, and choices are important aspects in the formation of a vehicle trajectory and ultimately in the traffic operational quality in a traffic stream. The specific movement of a vehicle is to a large degree affected by driver ability and choices, within the respective vehicle capabilities and for the overall roadway environment. Personal abilities (such as reaction time and vision capability) and preferences (such as speed selection, acceleration, and deceleration) significantly affect the trajectory as well as the overall travel times of the subject vehicle and the following vehicles. Therefore, in traffic operational analysis, it is important to understand and be able to model the range of drivers' capabilities, attitudes, and behaviors as these relate to vehicle motion.

There are several documents that have been developed to provide insight into driver characteristics and behavior and their impact on safety, highway design, and traffic operations. The Highway Safety Manual (HSM [11]) provides insights on the interactions of drivers and roadways from the perspective of roadway safety. The main driver characteristics discussed include driver attention and information processing ability, vision capability, perception–reaction time, and speed choice:

Attention and information processing: Even though humans can rapidly process an enormous amount of information, there is a limited amount they can process each second (according to [11], on average, humans can consciously recognize 16 U of information in 1 s). Therefore, highway designers should consider the workload, i.e., the amount of information presented to drivers. Designing the roadway environment in accordance to driver expectations facilitates information processing; therefore, providing consistency between successive design elements is very important. Presenting information sequentially, rather than all at once, also helps distribute the workload and facilitates information processing.

Vision: According to [11], 90 % of the information drivers use is visual. The following aspects of vision are important: visual acuity, the ability to see well at a distance; contrast sensitivity, the ability to see differences in luminance between an object and its background; peripheral vision, the ability to detect objects outside the area of most accurate vision; movement in depth, the ability

to estimate the speed of another vehicle; and visual search, the ability to search the rapidly changing environment from within a moving vehicle and collect pertinent information.

Perception–reaction time: Perception–reaction time includes the time to detect an object, process the information, decide whether and how to respond, and initiate action. Perception–reaction time depends on a variety of factors, including the driving environment, the object detected, and the driver characteristics. According to [11], a recent study concluded that a perception–reaction time of 2.0 s seems to be the upper limit for all subjects and conditions tested. However, for conditions of low light and low contrast, perception–reaction time may be much longer. (The website http://www.mrmont.com/games/brakingdistance.html provides an interactive way to test your reaction time using a vehicle braking simulator.)

Speed choice: The speed which drivers select when uninhibited by the presence of other vehicles is a function of both the roadway design and the environment.

Another document [12] provides human factors principles and guidance for considering specific design elements such as sight distance, horizontal alignment, and vertical alignment. For example, in discussing sight distance, it lists the human factors that affect perception–reaction time: vision capabilities, age, driver expectations, driver experience, object size and height, lighting conditions, and information overload. The document also provides guidelines for the type and placement of signs and markings. For example, it discusses countermeasures for improving pedestrian conspicuity at crosswalks, including the application of flashing lights and beacons to alert drivers to the presence of pedestrians.

Reference [13] discusses human factors from a traffic flow theory perspective. The elements discussed include perception–reaction time and control movement time (i.e., the amount of time it takes to make the desired move, such as applying the brakes or adjusting the steering wheel). The report discusses the reaction time and distance to signs and signals, obstacle and hazard detection and identification, and responses to other vehicles (vehicle ahead and vehicle alongside). In contrast to [11, 12], it also discusses the variability of driver behavior, which is an important element in traffic operations. This variability stems from differences in factors such as gender, age, vision capabilities, and driver impairment. These affect acceleration, deceleration, speed choice, overtaking and passing, and gap acceptance.

The degree of aggressiveness is another significant factor that has been incorporated in traffic operational analysis. For example, [14] conducted three focus groups to investigate drivers' intended actions along a freeway-ramp-merging segment. The study used "selfishness" as a criterion to develop three behavioral categories: aggressive, average, and conservative. It also found that the degree of aggressiveness of each driver varies as a function of their task and the prevailing traffic conditions. Reference [15] collected driver behavior-related data to develop a new lane-changing modeling framework for urban arterials. The study used focus groups and in-vehicle data collection and classified drivers into four groups based on their risk-taking behavior and their desire to gain speed advantage. The degree of

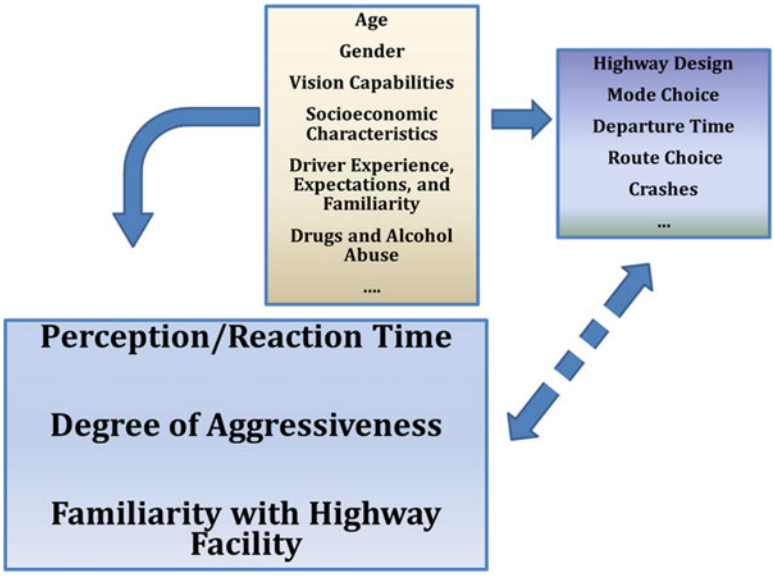


Fig. 1.7 Driver behavior and driver characteristics

aggressiveness defines a variety of traffic characteristics such as desired speed, acceleration rate, and passing decisions, which then affect traffic operational quality.

In considering human factors, there is an important distinction between highway design and traffic flow considerations. In highway design, we are primarily interested in identifying the design elements that would be adequate or functional for the vast majority of drivers, and thus, we are interested in worst-case scenarios. However, in studying traffic flow, we are also very interested in the variability of driver behavior.

In summary, the most important elements of driver behavior that affect traffic operations and have been incorporated in their analysis relate to the perception and reaction process, the level of aggressiveness, and familiarity with the highway facility (or travel time information availability regarding the facility). These behaviors can be correlated to a variety of personal and socioeconomic characteristics, as shown in the orange box of Fig. 1.7. As shown, these same driver characteristics are used to model a variety of other transportation-related behaviors related to design, route and mode choice, and safety. The dashed double arrow signifies that these decisions may affect driver behavior (e.g., the highway design affects driver familiarity), and driver behavior may also affect the system safety and design (e.g., the degree of aggressiveness affects crashes).

The perception and reaction time is important in establishing the drivers' reaction to external stimuli (recall that in Example 1.1, the perception and reaction time was used to determine the start of the deceleration, which ultimately affects the

stopping distance). This parameter is important in many aspects of transportation, including the determination of safe stopping distance, the duration of the yellow interval, and the design of horizontal and vertical curves along highways.

The degree of aggressiveness is an important element in traffic simulation modeling (discussed in more detail in Chap. 7—Simulation). Commercially available simulators adjust various types of parameters considering this characteristic: speed choice, lane-change decisions, passing decisions, etc.

The third component relates to the familiarity of the driver with the facility, but also the availability of travel time and other information. The familiarity of the driver with the facility can affect the drivers' choice of speed, choice of lane (e.g., when the driver has information regarding an incident and its exact location), etc. Drivers on a recreational trip or visitors to an area would be likely to be searching for their target destination or next turn and thus are likely to be traveling at a lower speed. On the other hand, commuters are more likely to be familiar with the facility, anticipate congestion from on-ramps and off-ramps, and be familiar with sources of information in the area. The HCM 2010 [8] provides adjustment factors to account for driver familiarity; however, their values have not been thoroughly validated using field data.

The importance of driver behavior on traffic operations has only recently been given much attention, and thus, research on this topic in the engineering literature has been limited. There is however a significant volume of research from the psychology and sociology perspectives that relates driver characteristics to driver behavior elements, and there are likely many opportunities for using and expanding such research to improve traffic flow models.

Effects of the Driving Environment on the Motion of a Single Vehicle

The driving environment is the third important component that affects vehicle motion. The driving environment includes the location and surroundings (e.g., urban, suburban, rural, etc.), the type of facility (e.g., freeway, urban arterial, rural highway, etc.), the highway design, the presence and type of control and its characteristics, as well as other elements such as rain or snow, incidents, and work zones. Figure 1.8 graphically illustrates the impact of these five categories of environmental influences on vehicle trajectories.

The driving environment can be perceived differently by different types of drivers and may result in different actions by different driver types. Low visibility may cause one driver to reduce their speed by 10 mph, while it may cause another driver to reduce their speed by 30 mph. The driving environment may also have a different impact for vehicles with different characteristics. A steep upgrade may significantly reduce the speed of a heavy vehicle, while having no impact on a passenger vehicle. Conversely, the design characteristics of a facility are set as a

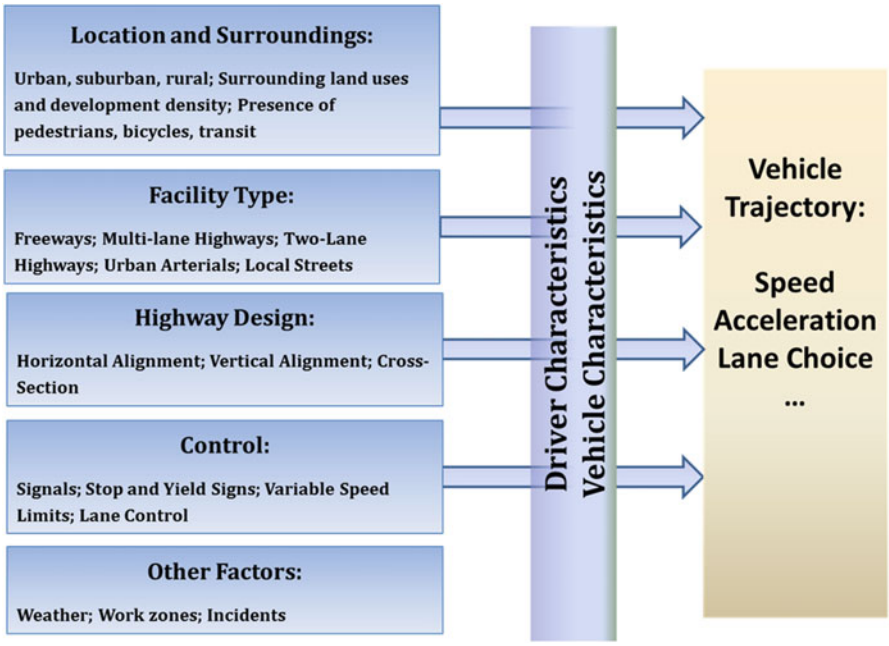


Fig. 1.8 The impact of the driving environment on vehicle trajectories

function of vehicle size and performance and driver abilities. Thus, the impact of the driving environment interacts with the driver and vehicle characteristics to produce the resulting vehicle trajectories. The remainder of this section briefly discusses the five categories of environmental characteristics that affect vehicle motion.

Location and Surroundings

The overall surroundings of the highway facility include the adjacent land uses and density as well as the presence of other transportation modes (e.g., bicycles and pedestrians). Generally, more dense urban environments result in lower speeds, while less dense rural environments increase speeds. The presence of pedestrians and bicycles generally result in lower speeds and more conservative driver actions. The location and surroundings may also be correlated with the purpose of the trip, which also affects vehicle trajectories. Facilities located in tourist areas would attract a higher percent of recreational trips, which are generally characterized by lower speeds and less aggressive driving. Commuter routes on the other hand would mostly cater to drivers familiar with the facility who are more likely to travel at faster speeds and drive more aggressively.

Facility Type

Facilities can be categorized as freeways, multilane highways, two-lane highways, urban streets, and local streets [8]. Freeways and multilane highways are characterized by fewer access points and wider cross section, which generally result in overall higher speeds. Urban and local streets are distinguished by signal and sign control, as well as presence of buses, bicycles, pedestrians, and parking, all of which generally result in lower speeds. They also have driveways and cross streets and thus a higher percent of turning movements which increases the amount of lane changing along the facility.

Highway Design

Highway design takes into consideration driver and vehicle capabilities, and it is primarily based on the type of highway facility [3]. Highways can be classified based on their primary function which is related to the degree to which they provide access and mobility. The four main highway types are:

Limited access facilities
Arterials
Collectors
Local streets

The design characteristics of a facility are based on a selected design speed. The design speed affects the desired speed of each driver and thus the free-flow speed of a facility (i.e., the average speed of the traffic stream under low-demand conditions). There are three main categories of design characteristics that define the geometry of a highway section based on the selected design speed [3]:

Horizontal alignment: This element refers to the horizontal curvature of the highway facility. The higher the selected design speed, the larger the curve radius needs to be so that high speeds can be accommodated. An important measure considered in the design of horizontal curves is the minimum stopping sight distance (estimated using equations of motion, assuming constant acceleration). The selected horizontal alignment and its characteristics affect the free-flow speeds along the facility. Generally, speeds along horizontal curves tend to be lower than those along tangent sections: the lower the radius of the curve, the lower the speeds would be. Vehicles typically decelerate when negotiating sharp horizontal curves and then accelerate as they depart the curve and approach a tangent section.

Vertical alignment: This element refers to the vertical curvature (sag or crest) as well as the grade and length of grade along the highway facility. As in horizontal alignment, the stopping sight distance is a key consideration in the design of crest vertical curves. In terms of traffic operational quality, the characteristic that impacts it the most is grade. Steep grades can result in lower speeds, particularly

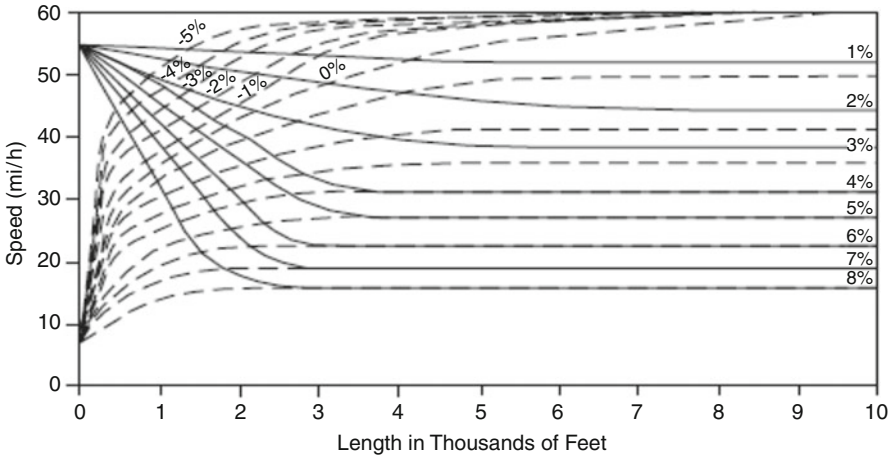


Fig. 1.9 Acceleration and crawl speed for trucks with 200 Ht/Hp (From *Highway Capacity Manual 2010*. Copyright, National Academy of Sciences, Washington, D.C., Exhibit 11-A1, p. 11–45; Reproduced with permission of the Transportation Research Board)

for heavy trucks with low-performance characteristics. The longer the upgrade or downgrade section, the more pronounced the impact on traffic would be. Crawl speeds (maximum speeds that can be achieved for a given terrain and a given vehicle type) can be determined as a function of grade. Figure 1.9 illustrates the effect of grade on speed. The solid lines in the figure indicate the speed of the truck as a function of the degree of the upgrade and length of grade. The dashed lines provide the same information for downgrade sections.

Cross section: This element refers to the number of lanes, lateral clearance, and superelevation of the facility. The two elements of cross section that have been shown to affect traffic operations are the lane width and the lateral clearance. The HCM 2010 indicates that the free-flow speed of a freeway would be reduced for lane widths less than 12 ft by as much as 6.6 mph and for lateral clearances less than 6 ft by as much as 3.6 mph. It also indicates a reduction in free-flow speed to account for the number of lanes: free-flow speed for two-lane freeways is estimated to be 4.5 mph lower than that for five-lane freeways. However, the research on the topic is sparse as it has proven very difficult to collect suitable data and isolate the effect of each specific element of the cross section and their interactions on the free-flow speed.

Generally, highway design affects mostly the speed profile along the facility, as it is directly related to the free-flow speed of the traffic stream. As indicated earlier, the design characteristics often interact with the driver and vehicle characteristics, and thus, their impact is not uniform to all drivers and vehicles. Also, each design element affects each facility type in a different way. For example, the cross section of a facility is more likely to affect speed at freeways than at urban streets. The speed along urban streets is more likely to be affected by signalization, the number of access points, and interactions with other modes.

Control

Traffic control consists primarily of traffic signals and traffic signs (YIELD and STOP signs), which interrupt the flow of traffic. Traffic signals can be placed either at intersections or at on-ramps. Intersection signals allocate the right-of-way sequentially to conflicting groups of movements or approaches (additional information is provided in Chap. 9), while ramp signals restrict the flow of traffic as it enters a freeway (additional information is provided in Chap. 8).

YIELD signs are most commonly placed either at roundabouts (see Chap. 10) or on-ramps (see Chap. 8). STOP signs are placed at unsignalized intersection approaches, and they can function either as a T-intersection, a two-way stop intersection, or as an all-way stop intersection (these are discussed in more detail in Chap. 10).

Another form of traffic control that has recently been introduced is the variable speed limit sign, which aims to controls traffic by reducing the speed limit (and thus the free-flow speed) of the facility. Such signs are typically installed upstream of freeway bottlenecks and work zones to harmonize and reduce speeds in anticipation of downstream congestion (additional information is provided in Chap. 8).

Lastly, lane control is another type of control used in traffic management. Lane control in a freeway environment may consist of lane use policies for trucks, high-occupancy vehicle (HOV) lanes, high-occupancy toll (HOT) lanes, reversible lanes, or the temporary use of hard shoulder (additional information is provided in Chap. 8). Along urban streets, lane control (or channelization) designates certain lanes for specific movements.

Traffic control affects the individual vehicle trajectories as it requires decelerating and accelerating at specific locations. For some types of traffic control (such as demand-actuated signals or YIELD signs at roundabouts), the trajectory of a vehicle and the operations of the respective approach are a function of the demand of the other approach(es), resulting in significant interaction between the conflicting traffic movements. Finally, lane control and channelization affect traffic operations as it adjusts the per lane demand and subsequently the traffic stream operations.

Other Factors

Other factors that affect vehicle trajectories and ultimately traffic operations include weather conditions, quality of the pavement, incidents, work zones, advanced traveler information, and vehicle-to-infrastructure technologies.

With respect to weather, the literature provides a significant body of research which provides estimates of adjusted free-flow speed as a function of rainfall, snow, and reduced visibility conditions. The HCM 2000 indicates that free-flow speed may be reduced by as much as 10 mph for heavy rain and as much as 26 mph for snow [16]. The HCM 2010 indicates potential reductions in capacity more than

17 % for heavy rain and more than 27 % for heavy snow conditions [8]. The impact of pavement quality is not as clearly documented, even though there is general recognition that poor pavement quality would reduce free-flow speeds.

The likelihood and presence of incidents and work zones and their impact on traffic flow are recently receiving increased attention, as they can adversely affect traffic operational quality. These are more extensively discussed in Chap. 5, in connection with travel time reliability analysis.

Advanced technologies include information systems which provide the driver with information regarding optimal route choice and congestion locations. They also include technologies currently under development which provide communication capabilities between vehicles as well as between vehicles and the infrastructure. Even though such technologies are not yet widespread, they are likely to affect our way of driving in the near future as more of these are being developed and implemented in the field.

References

1. Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
2. Gillespie TD (1992) Fundamentals of vehicle dynamics. Society of Automotive Engineers, Inc., Warrendale, PA
3. American Association of State Highway and Transportation Officials (2011) A policy on geometric design of highways and streets, 6th edn. American Association of State Highway and Transportation Officials, Washington, DC
4. Jazar RN (2008) Vehicle dynamics: theory and application. Springer, New York, NY
5. Federal Highway Administration. http://www.fhwa.dot.gov/policy/ohpi/vehclass.htm. Accessed 6 May 2011
6. Mannering F, Washburn S, Kilareski W (2009) Principles of highway engineering and traffic analysis, 4th edn. Wiley, New York, NY
7. May AD (1990) Traffic flow fundamentals. Prentice Hall, Englewood Cliffs, NJ
8. Transportation Research Board, National Academies of Science (2010) Highway capacity manual 2010. Transportation Research Board, National Academies of Science, Washington, DC
9. Dowling R, Skabardonis A, Alexiadis V (2004) Traffic analysis toolbox, volume III: Guidelines for applying traffic microsimulation software, Federal Highway Administration, Publication no. FHWA-HRT-04-040, Washington, DC, July 2004
10. Harwood DW, Torbic DJ, Richard KR, Glauz WD, Elefteriadou L (2003) Review of truck characteristics as factors in geometric design, NCHRP report 505. Transportation Research Board
11. Highway Safety Manual, 1st edn, vol 1, American Association of State Highway and Transportation Officials (AASHTO) (2010)
12. Transportation Research Board (2012) Human Factors Guidelines for Road Systems, 2nd edn, NCHRP report 600. Transportation Research Board, Washington, DC
13. TRB special report 165; revised monograph on traffic flow theory, Web document http://www.fhwa.dot.gov/publications/research/operations/tft/
14. Kondyli A, Elefteriadou L (2009) Driver behavior at freeway-ramp merging areas: focus group findings, Transportation research record 2124. National Academy Press, pp 157–166

15. Sun D, Elefteriadou L (2010) Research and implementation of lane changing model based on driver behavior, Transportation research record: Journal of the Transportation Research Board, no. 2161. Transportation Research Board of the National Academies, Washington, DC, pp 1–10
16. Transportation Research Board, National Academies (2000) Highway capacity manual 2000. Transportation Research Board, National Academies, Washington, DC

Problems

1. Solve Example 1.1 assuming that the vehicle's reaction time is 1.5 s and its initial speed is 50 mph. How do the results compare to those of Example 1.1?
2. A driver is approaching a traffic signal at a speed of 40 mph. The driver's perception and reaction time is 1.5 s. When the vehicle is at a distance of 300 ft prior to the stop bar, the signal turns to a yellow indication which has duration of 3 s.

   (a) If the vehicle's deceleration function is $-(9.5\ t + 16)$, will the vehicle be able to stop safely at the signal before it turns red?
   (b) Plot the driver's trajectory and discuss the adequacy of the yellow interval duration.
   (c) Recalculate part (a) with a perception and reaction time of 5 s. How does the driver's perception and reaction time affect your response to questions (a) and (b)?

3. Identify in transportation-related manuals and applications (Green Book, HCM, etc.), three applications of the equations of motion. Provide the derivation of the final equation used, and discuss any assumptions employed.
4. Two vehicles approaching an uncontrolled intersection perpendicularly to each other start braking simultaneously to avoid a collision. The driver of vehicle A has a reaction time of 2 s, while the driver of vehicle B has a reaction time of 1 s. The distance of vehicle A from the potential collision point is 350 ft, while that of vehicle B is 200 ft. The initial speed of vehicle A is 50 mph, and its maximum deceleration is -10 ft/s^2. The initial speed of vehicle B is 35 mph, and its maximum deceleration is -8 ft/s^2. Will the two vehicles collide? (similar problem provided in [1])
5. A crash occurred between a bus and a passenger vehicle at a stop-controlled T-intersection. It was determined that the passenger vehicle was traveling along the major roadway when the bus entered the traffic stream from the side street and the passenger vehicle rear-ended it. At the time of the impact, the speed of the passenger vehicle was 20 mph. Assuming that the maximum deceleration of the passenger vehicle is -10 ft/s^2, what was the initial speed of the vehicle when it started decelerating? Pavement marks indicate that the braking distance was 40 ft. Given that the speed limit along the major roadway is 40 mph, was the passenger vehicle speeding?

6. Conduct a literature review on traffic psychology and summarize your findings regarding the relationship between individual driver characteristics and driving performance.
7. Obtain the most recent Green Book equation which estimates stopping sight distance for design purposes and discuss it in the context of the equations of motion presented above. Should this equation consider varying acceleration rather than constant acceleration?
8. Find in the literature one acceleration function for a passenger car and one for a heavy vehicle (truck or bus). Discuss their relative performance characteristics and the shape of their respective acceleration functions.

Chapter 2
Modeling Vehicle Interactions and the Movement of Groups of Vehicles

The previous chapter focused on the movement of an individual vehicle and provided equations of motion assuming no interaction with adjacent vehicles. This chapter examines these interactions between vehicles, which is at the heart of traffic flow theory, as it is these which produce the observed traffic operational conditions. Important traffic operational characteristics we are interested in include the capacity of a facility (i.e., the maximum amount of traffic that can pass through a point or section, in vehicles or other units of traffic per unit of time) and its operating speed (i.e., the speed at which the facility operates under a given set of prevailing conditions, including the demand, the highway design, etc.). Those concepts are discussed in more detail in Part II.

Vehicle interactions can be defined in terms of three basic relationships: car-following, lane changing, and gap acceptance. Car-following is the process by which a vehicle follows another vehicle in close proximity. Generally, car-following occurs when the speed of the lead vehicle affects the speed of the following vehicle. Car-following algorithms provide the trajectory of the following vehicle as a function of the lead vehicle. Car-following affects both the capacity of a facility and its speed. The closer the vehicles follow each other, the higher the capacity of a facility (more on that in Part II). Also, the behavior of the vehicles following another vehicle affects the speed of the facility: more aggressive car-following generally leads to higher (although not necessarily safer) overall speeds.

Lane changing is the process by which a vehicle decides to change lanes, and it generally involves the requirement or decision to change lanes, the selection of a target lane (when it is relevant), and the selection of a suitable gap. Lane changing is thus also related to the third process, gap acceptance. Gap acceptance involves the selection of a suitable gap (usually defined as the time headway between the rear end of the lead vehicle and the front end of the following vehicle) to change lanes, or to cross a conflicting traffic stream, as in the case of a STOP-controlled approach. Similarly to car-following, lane changing and gap acceptance affect both the capacity and the operating speed of a facility. For example, more frequent lane changing to gain speed advantage typically leads to higher operating speeds.

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_2,

When drivers generally accept smaller gaps they are able to traverse the stop bar faster, leading to higher capacities.

These three processes and the respective algorithms provide the basic set of vehicle interactions and they are the key components of traffic microsimulation models, which have become increasingly popular as computer power has increased. These models replicate on a computer the movement of individual vehicles in highway networks in order to conduct experiments and evaluate various alternative improvements (see Chap. 7 for additional information on traffic simulation). This chapter first discusses car-following, presents a historical overview of car-following models, and summarizes the most important algorithms currently used. The second section summarizes lane-changing models, while the third one presents gap acceptance principles.

Car-Following

Let us consider two vehicles, one traveling behind the other in a single highway lane. The movement of the first vehicle is primarily defined by the principles discussed in Chap. 1. The movement and trajectory of the second vehicle, however, are based on additional factors. The following vehicle must take care not to collide with the vehicle in front. Thus it needs to maintain a suitable spacing and speed. Let us assume that the lead vehicle travels uninhibited from traffic ahead and thus its speed corresponds to the driver's desired speed considering the prevailing driving environment. If the following vehicle's desired speed is lower than that of the lead vehicle then their spacing will keep increasing, and the following vehicle will not be constrained by the position and speed of the lead vehicle. If however the following vehicle's desired speed is higher, then sooner or later it will enter a "car-following" state with the lead vehicle, where its spacing, speed, and acceleration will be dictated by those of the lead vehicle. The closer the following vehicle is to the lead vehicle, the more sensitive the reactions of the following vehicle would be to the actions of the lead vehicle. This sensitivity also increases with speeds.

Similarly to the discussion regarding the movement of a single vehicle (which would be the uninhibited lead vehicle in this case), the trajectory of the following vehicle depends to a significant degree on its driver and vehicle. More aggressive drivers tend to have higher desired speeds and to follow at shorter distances. With respect to vehicle characteristics, heavier trucks, because of their braking requirements, keep longer distances to the vehicles in front of them.

Figure 2.1 presents the time–space diagram and trajectories of vehicles A and B. As shown, the second vehicle (B) is initially traveling at a higher speed than the first vehicle (A). At time t_1, after it approached and decelerated, it continued to travel as dictated by the speed of vehicle A. In that figure, the horizontal distance between the two vehicles represents their time headway (h), while the vertical distance represents their space headway (or spacing, s). The time headway is defined as the time difference of successive vehicle crossings taken at a given location. The space

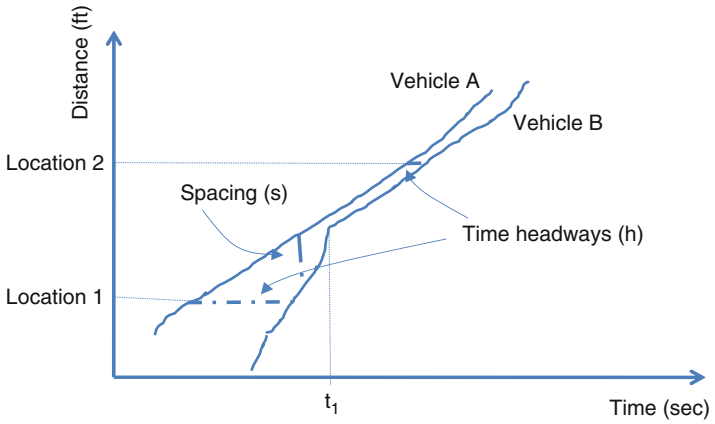


Fig. 2.1 Time–space diagram with two vehicle trajectories

headway is defined as the distance between two vehicles, measured usually from the front of the lead vehicle to the front of the following vehicle, at a given time. These are called microscopic traffic characteristics, because they consider the movement of individual vehicles and their relative time and space headways. Together with speed they are the three fundamental microscopic characteristics of traffic.

The time headway varies in space; thus if an observer is standing at Location 1, the time headway measured would be very different than that measured downstream at Location 2: the time headway at Location 2 is significantly smaller. Similarly, the spacing varies in time between these two vehicles.

Figure 2.2 presents a time–space diagram with several vehicles traveling along a single-lane roadway segment. Group X is separated by Group Y by a significant time and space gap. This occurs because the leading vehicle of Group Y desires to travel at a lower speed, and thus vehicles behind it must lower their speeds accordingly. As discussed earlier, the time headway between vehicles is represented by the horizontal distance between vehicles (h), while the space headway, or spacing, is represented by the vertical distance (s). Mathematically, the flow (F) at point P is

$$\boxed{F(\text{veh/h}) = 3,600/h_{\text{avg}}} \tag{2.1}$$

where h_{avg} is the average time headway in s.

In other words, Eq. (2.1) indicates that within an hour (or 3,600 s) the highway can process F vehicles with an average time headway h_{avg}. Similarly, within the analysis interval T of Fig. 2.2, the highway lane can process:

$$\text{Flow(vehicles)} = T\,\text{s}/h_{\text{avg}}$$

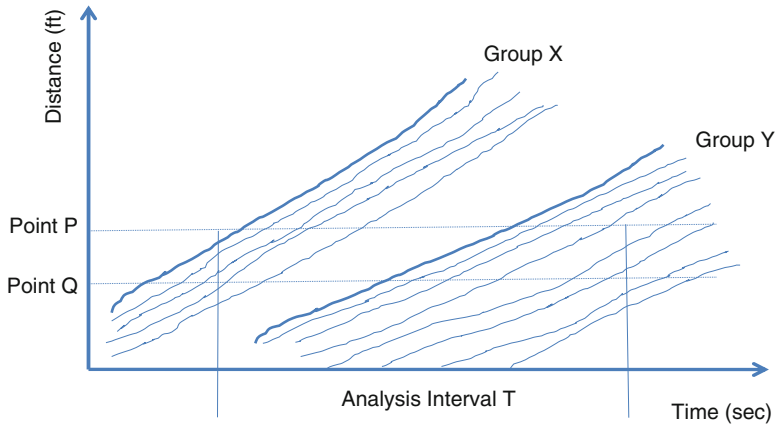


Fig. 2.2 Time–space diagram with groups of vehicles

This volume can be converted into flow by extrapolating the interval T into a full hour:

$$F(\text{veh/h}) = \text{Volume}\,(3,600/T) = \left(T\,\text{s}/h_{\text{avg}}\right)(3,600/T) = 3,600/h_{\text{avg}}$$

As h_{avg} is reduced, the total volume (or flow) increases. Conversely, as it increases, the total volume that can be processed decreases. Thus, the manner in which one vehicle follows another and their respective time headways are crucial in terms of the number of vehicles that a facility carries.

In Figure 2.2, the total number of vehicles an observer at Point P would count in time T is 9. During the same interval T, an observer at Point Q would count eight vehicles. Because of this variability in the manner in which vehicles follow one another (both between vehicles and for a specific vehicle in space) there is variability in our measurements of volume and flow. Considering the equations developed thus far, we can now examine the maximum amount of traffic a facility can carry, i.e., capacity. In theory the capacity of a particular highway can be obtained as follows:

$$\text{Capacity} = 3,600/(h_{\text{min}})$$

In practice however this minimum varies widely in time and space, as shown in Fig. 2.2. The minimum time headway at point P is not the same as that at point Q. Furthermore, this minimum is not representative of the time headways present during a particular time interval, and thus the actual maximum flow would be lower than that estimated using the above equation. Lastly, the minimum headways vary on a daily basis as a function of various factors including driver characteristics and behavior as well as vehicle capabilities for the traffic stream. Thus, a better

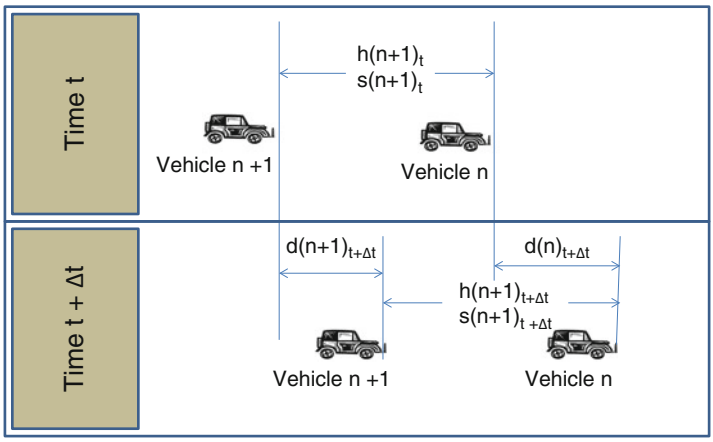


Fig. 2.3 Conceptual diagram and notation for car-following

understanding of the car-following process can help us appreciate the extent and impact of variability on traffic operations and the manner in which car-following affects our observations at the macroscopic level.

Let us now turn our attention to modeling the car-following process. Car-following algorithms determine the movement of the following vehicle at time $t + \Delta t$, as a function of its relationship to the lead vehicle at time t. Conceptually, the movement of the following vehicle (Vehicle $n + 1$) relative to the lead vehicle (Vehicle n) can be described as indicated in Fig. 2.3. At time t, vehicle $n + 1$ follows vehicle n at a time headway $h(n + 1)_t$ and a spacing $s(n + 1)_t$. Within time Δt, vehicle n has traveled a distance $d(n)_{t+\Delta t}$, while vehicle $n + 1$ has traveled a distance $d(n + 1)_{t+\Delta t}$. If the lead vehicle is unconstrained by traffic or other conditions ahead, its speed will be dictated by the geometry of the highway as well as its respective driver and vehicle characteristics, and its trajectory will follow the rules discussed in Chap. 1. The movement of the following vehicle however will largely depend on the trajectory of the lead vehicle, in addition to the characteristics of the follower driver and vehicle. At time $t + \Delta t$ its time headway from the lead vehicle will be $h(n + 1)_{t+\Delta t}$, and its spacing will be $s(n + 1)_{t+\Delta t}$.

Various car-following algorithms have been developed and are reported in the literature, and a variety of performance measures have been used to predict the trajectory of the following vehicle. A car-following algorithm might be based on predicting the acceleration of the following vehicle, another might be based on its speed, and another might be based on its space headway from the lead vehicle. The next subsection provides an historical overview of car-following algorithm development and implementation. Next, there is a discussion on some of the most widely cited algorithms, followed by an overview of efforts to compare car-following models to field data. The last section provides come concluding comments on car-following algorithms.

A Historical Overview of Car-Following Algorithms

Car-following models were initially developed in the 1950s [1], as traffic was increasing and researchers sought to understand what affects highway capacity. Since then, several models have been developed to replicate the time and space headways in car-following mode and generally to replicate car-following behavior. These models generally describe the trajectory of the following vehicle as a function of the trajectory of the first vehicle and the vehicles' distance/time headway. This section describes a few representative car-following models in chronological order starting from the earliest ones and concluding with some of the currently used models. The discussion in this section is not exhaustive in any way, and it only aims to familiarize the reader with some of the basic relationships developed to describe car-following.

Early Models

In one of the first papers to address car-following, Pipes [2] examined mathematically the dynamics of a line of traffic composed of n vehicles. His model assumed that the distance of the following vehicle to the lead vehicle is equal to one vehicle length for every 10 miles per hour of speed, plus a minimum distance between vehicles at standstill. This assumption was based on the California Vehicle Code which stated that "A good rule for following another vehicle at a safe distance is to allow yourself the length of a car (about 15 ft) for every 10 miles per hour you are travelling." For this rule to be satisfied, the coordinates $x(n)$ and $x(n + 1)$ of two successive vehicles, n and $n + 1$, must satisfy the following equation:

$$\boxed{x(n) = x(n+1) + b + L(n) + s_{\text{MIN}}(n+1)} \tag{2.2}$$

where

$x(n)$, $x(n + 1)$ are the coordinates of vehicles n and $n + 1$, measured from the front of each vehicle
b is the distance between vehicles when they are at a standstill, in ft
$L(n)$ is the length of vehicle n, in ft
$s_{\text{MIN}}(n + 1)$ is the minimum distance between vehicles based on the California Vehicle Code, in ft

The quantity $s_{\text{MIN}}(n + 1)$ can also be written as follows:

$$\boxed{s_{\text{MIN}}(n+1) = Tv(n+1)} \tag{2.3}$$

where

$v(n + 1)$ is the speed of vehicle $n + 1$, in ft/s
T is the time it takes to travel this minimum distance, in s

For example, for a vehicle traveling at 10 mph (≈14.67 ft/s), and assuming a vehicle length of 15 ft, T is

$$T = s_{\text{MIN}}(n+1)/v(n+1) = (15\,\text{ft})/(14.67\,\text{ft/s}) = 1.023s$$

If a vehicle travels at a speed of 50 mph, its length is 15 ft, and the distance between vehicles at standstill is 6 ft, then it is expected that the vehicle will keep a distance of 5 × 15 + 6 = 81 ft from the lead vehicle.

Combining Eqs. (2.2) and (2.3) results in

$$x(n) = x(n+1) + b + L(n) + Tv(n+1)$$

Differentiating with respect to time:

$$\dot{x}(n) = \dot{x}(n+1) + T\dot{v}(n+1) \Rightarrow$$

$$\boxed{v(n) = v(n+1) + Ta(n+1), \quad n = 1, 2, 3 \ldots k \text{ vehicles in the line of traffic}} \tag{2.4}$$

Equation (2.4) are the dynamical equations that describe the movement of k vehicles. Using these, Pipes developed differential equations for providing the velocity and acceleration of a line of vehicles, as a function of the movement of the lead vehicle. He solved these equations for specific initial conditions of the lead vehicle and recognized that the complexity of vehicle movement in terms of velocity and acceleration functions of the lead vehicle would require the use of a computer in order to solve them.

Aside from the difficulty in solving those equations analytically, the Pipes model assumes that the target spacing is achieved instantaneously and does not consider the reaction time of the following vehicle [3]. Also, the model requires the use of a theoretical spacing which cannot be violated, as it does not allow for deviations from the "following distance" law assumed. In reality, drivers have a time lag in reacting to changes in the lead vehicle's trajectory and position themselves with considerable deviation from this law, with significant variability in their spacing, both between vehicles, and longitudinally for a particular driver.

Regardless of these limitations, the Pipes model opened the door to significant development in the area of car-following. Starting from the system of Eq. (2.4), one can obtain:

$$v(n) = v(n+1) + Ta(n+1) \Rightarrow$$

$$\boxed{a(n+1) = \frac{v(n) - v(n+1)}{T}} \tag{2.5}$$

In this form, the Pipes model is similar to the stimulus-response type models developed by Gazis, Herman, Rothery, and others (also known as GHR models or GM models, because they were developed and tested at the General Motors

laboratory). These models [1, 3–6] considered the sensitivity of the reaction to the lead vehicle, and they were of the form:

$$a(n+1)_{t+\Delta t} = \lambda_{m,l}\left[v(n+1)_t\right]^m \frac{v(n)_t - v(n+1)_t}{\left[x(n)_t - x(n+1)_t\right]^1} \tag{2.6}$$

where Δt is the next time interval (it can also be interpreted as the reaction time), $\lambda_{m,l}$ represented a sensitivity parameter, and m and l are calibration exponents. As shown in Eq. (2.6) a major difference between the GHR model and the Pipes model is that the GHR model incorporates time into the equation and provides the acceleration of the following vehicle at time $t + \Delta t$ as a function of the vehicles' status at time t. The general form of Eq. (2.6) is

$$\text{Response}\,(t+\Delta t) = \text{Function}\{\text{Sensitivity},\ \ \text{Stimulus}\,(t)\}$$

where response is the acceleration of the following vehicle at time $t + \Delta t$, stimulus is the speed difference between the two vehicles at time t, and the remaining factors represent the *s*ensitivity of the model.

For $m = 0$ and $l = 0$, the model in Eq. (2.6) reduces to

$$a(n+1)_{t+\Delta t} = \lambda_{0,0}\left[v(n)_t - v(n+1)_t\right]$$

This was the first form of the GHR model which was tested and calibrated by conducting car-following experiments. According to [1] in those initial tests the parameter value for λ ranged between 0.17 and 0.74.

Subsequent work produced several different calibrations of those factors. References [1, 6] provide a very thorough discussion of the development and calibration of the GHR models along with optimal values for the m and l parameters. An important finding of those investigations was that different conditions resulted in different combinations of m and l. Thus, for example, it was suggested that non-congested and congested conditions should use different calibration parameters [6, 7].

From Eq. (2.6), one of the shortcomings of the GHR models is that when the vehicles travel at the same speed (speed difference equals zero and thus the stimulus becomes zero), the model cannot estimate the reaction of the following vehicle. Also, the speed difference determines whether the following vehicle will accelerate or decelerate, regardless of the vehicles' spacing: even when the vehicles are far apart, the following vehicle will always decelerate as a response of the lead vehicle's deceleration. Finally, when the following vehicle's speed increases, the sensitivity of the following vehicle also increases: thus, at high speeds, the model becomes too sensitive to the stimuli from the lead vehicle. In practical applications, minor changes in speed differences result in large acceleration values. These issues, along with the difficulty in reaching consensus as to the appropriate set of l and m factors that should be used under different conditions, have inhibited the use of those models. For a detailed discussion on the GM models along with applications, consult [1, 2].

Early Microsimulator Implementation: The PITT Model

One of the first car-following models to be developed and implemented in a microsimulator environment was the PITT model (developed at the University of Pittsburgh). The model was initially implemented into INTRAS, a microsimulator developed in the late 1970s by the Federal Highway Administration (FHWA) to evaluate freeway control and management strategies [8]. Subsequently INTRAS was replaced by FRESIM and then CORSIM both of which have been widely used and evaluated (additional information on CORSIM is provided in Chap. 7).

Similarly to the Pipes model, the PITT model is also based on the assumption that the follower vehicle attempts to maintain a fixed time headway between its front bumper and the rear bumper of the lead vehicle [8]. It is assumed that this distance is proportional to the speed of the follower and the speed difference between the leader and the follower. Mathematically:

$$\boxed{s(n+1)_t = L(n) + B(n) + kv(n+1)_t + bk\left[v(n)_t - v(n+1)_t\right]^2} \tag{2.7}$$

where

$L(n)$ is the length of the lead vehicle (n)
$B(n)$ is the buffer between vehicles at a standstill
k and b are car-following sensitivity parameters
$v(n)_t$ and $v(n+1)_t$ are the speeds of the lead and following vehicles, respectively, at time t

Similarly, at time $t + \Delta t$, the equation becomes

$$\boxed{s(n+1)_{t+\Delta t} = L(n) + B(n) + kv(n+1)_{t+\Delta t} + bk\left[v(n)_{t+\Delta t} - v(n+1)_{t+\Delta t}\right]^2} \tag{2.8}$$

Equations (2.7) and (2.8) provide the required equilibrium conditions for two vehicles in car-following. However, in a car-following model we need to estimate the acceleration of the following vehicle at time $t + \Delta t$, as a function of the positions and speeds of the two vehicles at time t. To introduce the following vehicle's acceleration at time $t + \Delta t$ into the equations, we express the spacing as a function of the respective coordinates as follows:

$$s(n+1)_{t+\Delta t} = x(n)_{t+\Delta t} - x(n+1)_{t+\Delta t}$$

The last term of the equation, which provides the distance the following vehicle travels during Δt, can be estimated assuming that the vehicle travels with a constant acceleration, and the equation becomes

$$\boxed{s(n+1)_{t+\Delta t} = x(n)_{t+\Delta t} - \left[x(n+1)_t + v(n+1)_t \Delta t + a(n+1)_{t+\Delta t}\frac{\Delta t^2}{2}\right]} \quad (2.9)$$

We also express $v(n+1)_{t+\Delta t}$ as a function of acceleration as follows:

$$\boxed{v(n+1)_{t+\Delta t} = v(n+1)_t + a(n+1)_{t+\Delta t}\Delta t} \quad (2.10)$$

Next, we equate the right parts of Eqs. (2.8) and (2.9), replacing $v(n+1)_{t+\Delta t}$ from Eq. (2.10), and we solve for $a(n+1)_{t+\Delta t}$. Also, because the last part of Eq. (2.8) is very small, we replace $v(n+1)_{t+\Delta t}$ with $v(n+1)_t$. The final equation is

$$\boxed{a(n+1)_{t+\Delta t} = 2\frac{\left[x(n)_{t+\Delta t} - x(n+1)_t - L(n) + B(n) - v(n+1)_t(k+\Delta t) - bk\left(v(n)_{t+\Delta t} - v(n+1)_t\right)^2\right]}{\Delta t^2 + 2k\Delta t}} \quad (2.11)$$

Equation (2.11) represents the basic car-following equation for the PITT model. According to [8], the parameter b is a constant which takes the value 0.1 when $v_n < v_{n+1}$, or 0 otherwise; in essence, when the lead vehicle is traveling faster than the follower their speed difference does not affect the acceleration of the following vehicle.

Contrary to the GM models, the PITT car-following model considers the final speed of the lead vehicle, i.e., the speed at the end of Δt, rather than the respective conditions at time t. In other words, the acceleration during Δt is estimated as a function of where the lead vehicle will be at the *end* of Δt. In implementing this model the perception and reaction time of the driver need to be considered separately and after the acceleration is calculated [8, 9]. According to [9], the PITT car-following model does not replicate traffic oscillations very well, and any traffic disturbances diminish quickly. This is important in that breakdown does not occur randomly, as is the case in the field, but only after demand exceeds capacity (see Chap. 4 for a discussion of capacity and breakdown). Finally, to implement the model it is necessary to introduce several constraints to ensure that the trajectory of the following vehicle is reasonable; application of the model has shown that those constraints determine the acceleration most of the time [9].

Currently Used Models: The Gipps Model

The Gipps model [10] was developed in Australia and is one of the most widely used and cited car-following models. It is currently used in the AIMSUN microsimulator [11]. This model can be categorized as a multi-regime model

because it considers the desired speed of the following vehicle, as well as whether the following vehicle is in breaking mode or car-following mode. The model calculates two speeds: the speed of the following vehicle under non-constrained conditions and the speed that would result if the following vehicle is constrained by the lead vehicle. The minimum of these two speeds is selected as the follower vehicle speed. The model is the following:

$$v(n+1)_{t+\Delta t} = \min \left\{ \begin{array}{l} v(n+1)_t + 2.5a(n+1)_{\mathrm{MAX}}\Delta t\left(1 - \frac{v(n+1)_t}{v(n+1)_{\mathrm{DES}}}\right)\left(0.025 + \frac{v(n+1)_t}{v(n+1)_{\mathrm{DES}}}\right)^{1/2} \\ \hline b(n+1)\Delta t + \sqrt{\left(b(n+1)^2\Delta t^2 - b(n+1)\left[2\left[x(n)_t - L(n) - x(n+1)_t\right] - v(n+1)_t\Delta t - \frac{v(n)_t^2}{\widehat{b}_n}\right]\right)} \end{array} \right. \tag{2.12}$$

where

$v(n+1)_{t+\Delta t}$ is the speed of vehicle $n+1$ at time $t+\Delta t$
Δt is the apparent reaction time, a constant for all vehicles
$v(n+1)_t$ is the speed of vehicle $n+1$ at time t
$a(n+1)_{\mathrm{MAX}}$ is the maximum acceleration which the driver of vehicle $n+1$ wishes to undertake
$v(n+1)_{\mathrm{DES}}$ is the speed at which the driver of vehicle $n+1$ wishes to travel
$b(n+1)$ is the actual most severe deceleration rate that the driver of vehicle $n+1$ wishes to undertake ($b(n+1) < 0$)
$x(n)_t$ is the location of the front of vehicle n at time t
$x(n+1)_t$ is the location of the front of vehicle $n+1$ at time t
$L(n)$ is the effective size of vehicle n; that is the physical length plus a margin into which the following vehicle is not willing to intrude even when at rest
$v(n)_t$ is the speed of vehicle n at time t
$\widehat{b}_n$ is the most severe deceleration rate that vehicle $n+1$ estimates for vehicle n

The first term of Eq. (2.12) estimates the following vehicle's speed under non-constrained conditions and was developed based on field data. The equation ensures that the speed of vehicle $n+1$ will not exceed the driver's desired speed. It also ensures that the acceleration of vehicle $n+1$ will reach zero when the desired speed is reached. Finally, it modifies the acceleration of vehicle $n+1$ when it is not constrained by vehicle n, so that it increases with decreasing speed and vice versa.

The second term in Eq. (2.12) was developed to allow the following vehicle to be able to stop safely even for the most severe braking of the lead vehicle. The equation is based on the most severe braking of the lead vehicle and estimates the stopping distance of the following vehicle considering the driver's reaction time plus a safety margin.

One of Gipps' goals in developing this model was to be able to relate each of its parameters to vehicle or driver characteristics that the driver of the following vehicle could estimate or perceive. Thus, instead of using the most severe deceleration of vehicle n, he used the respective estimate of the following driver for that quantity $\left(\widehat{b}_n\right)$. Calibrating this model is relatively easy, as one can measure or estimate its parameters based on vehicle and driver characteristics.

The results this model produces have generally been found reasonable [6, 12]. One of the criticisms of this model is that it is based on a "safe" headway, which may not necessarily be valid in real traffic [6]. Drivers may in practice be willing to accept shorter headways. Also, researchers have suggested that considering the traffic conditions downstream may result in a more realistic model.

Example 2.1 A vehicle is in car-following mode, following a lead vehicle with the trajectory shown in Table 2.1. Assuming that the desired maximum speed of the following vehicle, $v(n+1)_{\mathrm{DES}}$, is 75 mph, the maximum acceleration which the following vehicle wishes to undertake, $a(n+1)_{\mathrm{MAX}}$, is 6.5 ft/s^2, the actual most severe deceleration that the follower wishes to undertake, $b(n+1)$, is -9.5 ft/s^2, the most severe deceleration rate that vehicle $n+1$ estimates for vehicle n, $\widehat{b}_n$, is -11.5 ft/s^2, and the effective vehicle length, $L(n)$, is 25 ft, calculate the trajectory of the following vehicle using the Gipps car-following model.

Solution to Example 2.1

Table 2.2 calculates the detailed trajectories of both vehicles. The left side of the table provides the data related to the movement of the lead vehicle. The acceleration of the lead vehicle for each time interval is calculated using the speeds at time t and $t+1$. For example, the average acceleration during the interval 2–3 s is (70.40–74.73)/1 s $= -4.33$ ft/s^2. The location of the lead vehicle is estimated based on the equations of motion assuming constant acceleration during each 1-s interval.

The speed of the following vehicle is estimated using Eq. (2.12). The column labeled "Speed 1" estimates the first part of the Gipps car-following equation, while the one labeled "Speed 2" estimates the second part. The minimum of these two speeds is the estimated speed of the following vehicle. The acceleration, location, and spacing of the following vehicle are estimated using the equations of motion similarly to those for the lead vehicle. Figure 2.4 provides the speed vs. time plot for both the lead and following vehicles, while Fig. 2.5 provides the trajectories of the two vehicles.

Table 2.1 Lead vehicle trajectory and follower starting position

Lead vehicle		Following vehicle	
Time (s)	Speed (mph)	Speed (mph)	Spacing (ft)
1.00	52.37	54.3	120
2.00	50.95		
3.00	48.00		
4.00	46.00		
5.00	44.00		
6.00	42.00		
7.00	41.00		
8.00	39.00		
9.00	37.00		
10.00	35.00		
11.00	33.00		
12.00	31.00		
13.00	29.00		
14.00	25.00		
15.00	22.00		
16.00	19.00		
17.00	21.00		
18.00	24.00		
19.00	26.00		
20.00	29.00		
21.00	31.00		
22.00	33.00		
23.00	36.00		
24.00	39.00		
25.00	42.00		
26.00	45.00		
27.00	49.00		
28.00	50.91		
29.00	51.50		
30.00	50.67		

Other Currently Used Models

The MITSIM Model

The car-following model used by the MIT Simulator (MITSIM [13]) is a multi-regime version of the GHR model. Multi-regime models are those which consider various conditions (or regimes) for the car-following vehicle and specify different models for each condition.

The MITSIM model estimates the acceleration of the following vehicle at the end of $t + \Delta t$ as a function of the headway and the relative speed between the lead and the following vehicle. Depending on the size of the headway between the lead and the

Table 2.2 Lead and following vehicle trajectory data for Example 2.1

Lead vehicle					Following vehicle						
Time (s)	Speed (mph)	Speed (ft/s)	Acceleration (ft/s^2)	Location (ft)	Speed 1 (mph)	Speed 2 (mph)	Min of speeds 1 and 2	Speed (ft/s)	Location (ft)	Spacing (ft)	Acceleration (ft/s^2)
1.00	52.37	76.81	Not known	0.00	54.30			79.64	−120.00	120.00	Not known
2.00	50.95	74.73	−2.08	75.77	56.95	46.39	46.39	68.04	−46.16	121.93	−11.60
3.00	48.00	70.40	−4.33	148.33	49.78	45.89	45.89	67.30	21.51	126.82	−0.74
4.00	46.00	67.47	−2.93	217.26	49.32	43.99	43.99	64.53	87.42	129.84	−2.78
5.00	44.00	64.53	−2.93	283.26	47.58	42.83	42.83	62.81	151.09	132.17	−1.71
6.00	42.00	61.60	−2.93	346.33	46.50	41.59	41.59	61.00	213.00	133.33	−1.82
7.00	41.00	60.13	−1.47	407.20	45.35	40.28	40.28	59.08	273.04	134.16	−1.92
8.00	39.00	57.20	−2.93	465.86	44.13	39.72	39.72	58.25	331.71	134.16	−0.83
9.00	37.00	54.27	−2.93	521.60	43.60	38.30	38.30	56.18	388.92	132.68	−2.07
10.00	35.00	51.33	−2.93	574.40	42.27	36.83	36.83	54.02	444.02	130.38	−2.16
11.00	33.00	48.40	−2.93	624.26	40.88	35.31	35.31	51.79	496.93	127.34	−2.24
12.00	31.00	45.47	−2.93	671.20	39.44	33.73	33.73	49.47	547.56	123.64	−2.31
13.00	29.00	42.53	−2.93	715.20	37.93	32.11	32.11	47.09	595.84	119.36	−2.39
14.00	25.00	36.67	−5.87	754.80	36.37	30.43	30.43	44.63	641.70	113.10	−2.46
15.00	22.00	32.27	−4.40	789.26	34.75	27.28	27.28	40.01	684.02	105.25	−4.63
16.00	19.00	27.87	−4.40	819.33	31.67	24.73	24.73	36.27	722.15	97.18	−3.73
17.00	21.00	30.80	2.93	848.66	29.15	22.11	22.11	32.44	756.51	92.16	−3.84
18.00	24.00	35.20	4.40	881.66	26.53	22.78	22.78	33.42	789.43	92.23	0.98
19.00	26.00	38.13	2.93	918.33	27.21	24.57	24.57	36.04	824.16	94.17	2.62
20.00	29.00	42.53	4.40	958.66	29.00	25.96	25.96	38.08	861.22	97.45	2.04
21.00	31.00	45.47	2.93	1,002.66	30.37	28.29	28.29	41.49	901.00	101.67	3.41
22.00	33.00	48.40	2.93	1,049.60	32.66	29.99	29.99	43.98	943.73	105.87	2.50
23.00	36.00	52.80	4.40	1,100.20	34.32	31.75	31.75	46.57	989.01	111.19	2.59
24.00	39.00	57.20	4.40	1,155.20	36.03	34.37	34.37	50.40	1,037.50	117.70	3.83
25.00	42.00	61.60	4.40	1,214.60	38.54	37.05	37.05	54.34	1,089.87	124.73	3.94
26.00	45.00	66.00	4.40	1,278.40	41.09	39.78	39.78	58.35	1,146.21	132.18	4.01
27.00	49.00	71.87	5.87	1,347.33	43.66	42.55	42.55	62.41	1,206.59	140.74	4.06
28.00	50.91	74.67	2.80	1,420.60	46.24	46.17	46.17	67.72	1,271.66	148.94	5.31

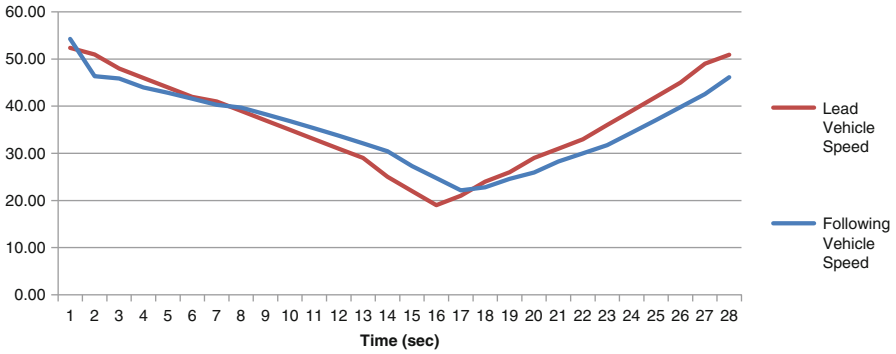


Fig. 2.4 Lead and following vehicle speeds for Example 2.1

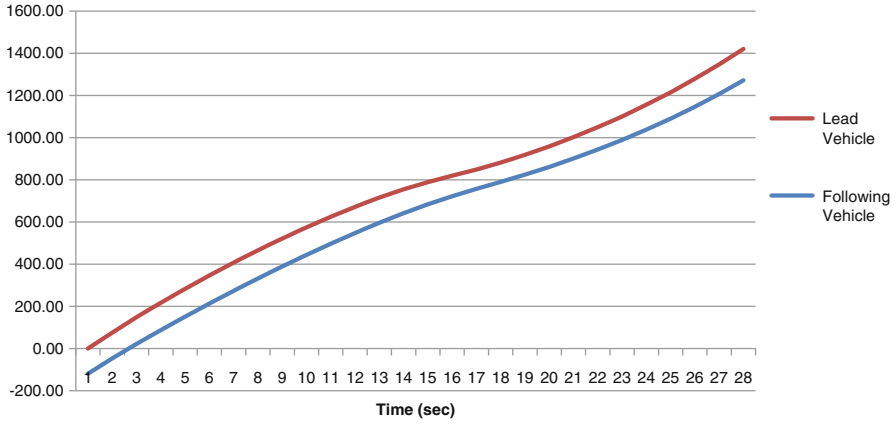


Fig. 2.5 Lead and following vehicle trajectories

following vehicles, the following vehicle is classified to be into one of three regimes: free following, emergency decelerating, and car-following.

Free flowing regime: If the time headway is larger than a predetermined threshold h^{upper}, the vehicle acceleration is not estimated as a function of the lead vehicle trajectory. If the vehicle's current speed is lower than its maximum speed, it accelerates at the maximum acceleration rate to achieve its desired speed as quickly as possible. If its current speed is higher than the maximum speed, the vehicle decelerates with the normal deceleration rate.

Emergency regime: If a vehicle has headway smaller than a predetermined threshold h^{lower}, it is in emergency regime. In this case, the vehicle uses an appropriate deceleration rate in order to avoid colliding with the lead vehicle.

Car-following regime: If a vehicle has headway between h^{lower} and h^{upper}, it is in car-following. In this case the acceleration rate is calculated using the GHR model [13, 14]:

$$\alpha(n+1)_{t+\Delta t} = \alpha^{\pm} \frac{v(n+1)_t^{\beta\pm}}{s(n+1)_t^{\gamma\pm}} \left[v(n)_t - v(n+1)_t\right] \tag{2.13}$$

where

$a(n+1)_{t+\Delta t}$ is the acceleration of vehicle $n+1$ at time $t+\Delta t$
$v(n)_t$ is the speed of vehicle n at time t
$v(n+1)_t$ is the speed of vehicle $n+1$ at time t
$s(n+1)_t$ is the spacing between the follower and the lead vehicle
α, β and γ are model parameters that can be calibrated as a function of driver behavior. Positive values of the parameters correspond to acceleration, while negative values to deceleration.

Equation (2.13) is identical to Eq. (2.6), only with slight changes in notation (α, β and γ instead of $\lambda_{m,l}$, m, l). The major difference between the MITSIM and the GHR models is that MITSIM explicitly considers the different regimes the following vehicle may belong in and sets appropriate limits in the application of the car-following equation.

The Wiedemann Model

The Wiedemann model is also a multi-regime model, which uses specified thresholds for anticipated changes in driver behavior. The model is also defined as a psychophysical model, because the thresholds used to define different regimes are based on driver perceptions and actions. For example, instead of using deterministic thresholds for spacing in order to apply a particular model, the Wiedemann approach considers the distance at which drivers are able to perceive relative velocity between their vehicle and the lead vehicle; when they cannot perceive it, the following vehicle is not in car-following any more.

Figure 2.6 displays the thresholds and regimes of Wiedemann model. Four driving modes (or regimes) can be distinguished [14]:

1. Free driving: no influence of leading vehicles. In this mode the follower vehicle seeks to reach and maintain her/his individually desired speed.
2. Approaching: when passing the approaching point (SDV) threshold. This regime consists of the process of adapting the driver's own speed to the lower speed of the lead vehicle. While approaching, a driver applies a deceleration so that the speed difference of the two vehicles is zero in the moment she/he reaches her/his desired safety distance.
3. Following: the thresholds SDV, ABX (desired minimum following distance at low speed differences), SDX (the maximum following distance), and OPDV (the

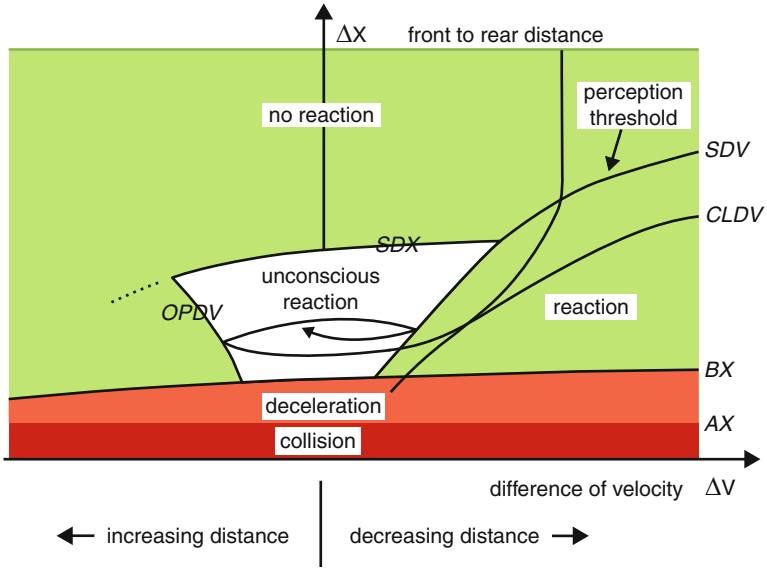


Fig. 2.6 Thresholds and regimes in the Wiedemann car-following model (From PTV VISSIM Users Manual; Reproduced with permission of PTV AG)

increasing speed difference) constitute this regime. The driver follows the preceding car without any conscious acceleration or deceleration. She/he keeps the safety distance more or less constant.

4. Braking or emergency regime: when the front to rear distance is smaller than ABX the follower adopts the emergency regime. The driver applies medium to high deceleration rates if the distance falls below the desired safety distance. This can occur if the preceding car changes speed abruptly or if a third car changes lanes in front of the lead vehicle.

For each regime, the acceleration is described as a result of speed, speed difference, distance, and the individual characteristics of driver and vehicle. The driver switches from one mode to another as soon as she/he reaches a certain threshold that can be expressed as a combination of speed difference and distance.

Evaluations of Car-Following Algorithms Using Field Data

Various techniques have been used to compare car-following models to field data. Some of the earlier attempts [15, 16] compared the GM model results to data collected by wire-linked vehicles. More recently [17], researchers used an aerial data collection technique to collect car-following data. Recent advances in GPS and sensor technology have led to an increasing number of studies that compare car-following models to field data. A few studies have used GPS [18], while others

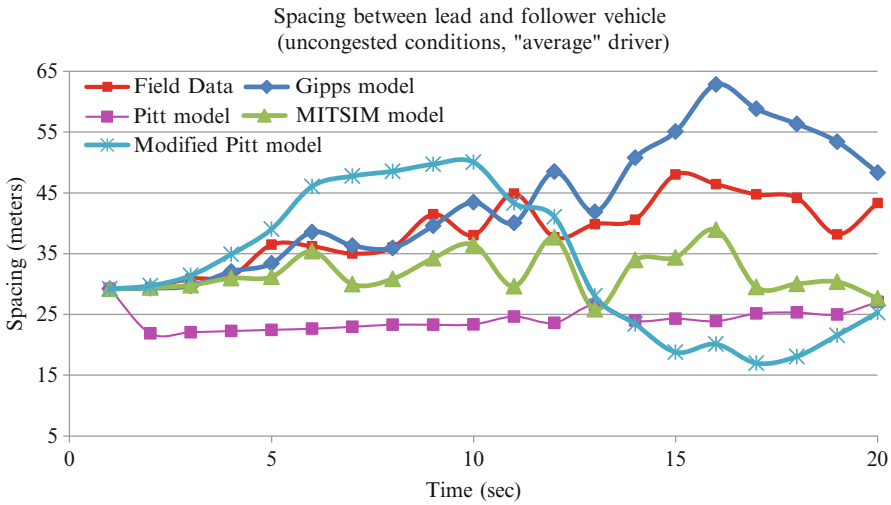


Fig. 2.7 Comparison of selected car-following models to field data [12]

[12, 19] have used an instrumented vehicle to record the lead and following vehicle trajectories as well as driver information. One of the often cited conclusions is the significance of accurate measurements in obtaining speeds and distances of the subject vehicles.

Figure 2.7 provides an illustration of a car-following trajectory compared to estimated trajectories obtained using an instrumented vehicle [12] The graph compares car-following models to a trajectory by an "average" driver (i.e., neither too aggressive nor too conservative) and for non-congested conditions. The red line indicates the field-measured spacing. As shown, each model predicts a slightly different spacing throughout the car-following event. The authors of the research reported that the models generally predict speed more accurately than spacing. However, the best variable to be used in calibration is spacing: calibrating by spacing minimizes the errors that can be accumulated and can distort the final trajectory. Generally, car-following for congested scenarios has been found to be most accurately predicted because there is much less freedom and variability in driver behavior during congested conditions [12]. For non-congested conditions the driver type variability is much more significant, and calibration needs to take into consideration driver aggressiveness.

Concluding Remarks on Car-Following Models

Car-following is a complex process, which depends significantly on driver decisions and actions. There are several recent advances both to develop more

accurate models for replicating car-following and to control car-following behavior through the development of advanced vehicle technologies.

Various car-following models have been developed and continue to be developed in light of tremendous advances in traffic simulation. Multi-regime models appear to provide the most logical approach to car-following, particularly those that consider driver behavior and its variability. With respect to future developments on the topic, one of the most recent research efforts which appears promising used fuzzy logic to describe driver behavior in car-following [20]. Such research is based on the recognition that the reactions of the following vehicle to the lead vehicle are not based on a deterministic relationship, but rather on a set of random or approximate driving behaviors. Such models are at their infancy and have not been extensively calibrated or tested.

More generally, the following factors may affect car-following behavior, but have not yet been thoroughly evaluated to quantify their impact: highway design elements such as grade or lane width; vehicle characteristics such as braking ability; driver characteristics such as reaction time; and adverse weather and other environmental conditions. Thus, there is still quite a bit of research that needs to be completed to be able to replicate car-following behavior reasonably well for a variety of conditions and cases and considering variability in driver behavior.

An essential part of car-following model development is the ability to obtain good quality field data and evaluate or calibrate following vehicle trajectories as estimated by various types of models. Existing data collection capabilities (video data collection, sensors, etc.) afford much greater opportunities than previously available to accomplish this. It is important to be able to accurately observe the trajectories of both the lead and the following vehicle and also to measure and evaluate driver behavior and characteristics, so that car-following behavior can be modeled.

An important concept that has been discussed since the infancy of car-following models is the stability of the traffic stream, i.e., its ability to absorb a perturbation [3, 21]. A traffic system is referred to as stable when the fluctuation in acceleration of the lead vehicle does not cause an increasingly fluctuating acceleration and spacing in the following vehicle or vehicles. Local stability refers to the acceleration and spacing of only one car-following vehicle; asymptotic stability refers to the acceleration and spacing of a series of vehicles traveling one behind another. Although a series of vehicles may be stable locally, the system may not be stable when each vehicle amplifies the perturbation which travels upstream with increasing magnitude [3]. It has been shown [16] that when a car-following vehicle considers information related to the car downstream from the lead vehicle the traffic stream stability increases. It is speculated [3] that when such information is lacking in the real world, as for example when there is reduced visibility, the traffic stream becomes highly unstable and thus there is a high probability of rear-end collisions. With respect to car-following models, it is thus important that factors related to two or more vehicles downstream of the car-following vehicle be incorporated, so that the resulting traffic stream is as stable as a similar one in the field.

As researchers understood the importance of car-following, research has also sought to develop vehicle-based technologies that can control it in order to improve safety and increase capacity. The concept of automated highways where vehicles follow one another under specified rules has been around for quite some time. More recent efforts by vehicle manufacturers to develop adaptive cruise control (ACC) and cooperative adaptive cruise control (CACC) mechanisms for use in passenger vehicles have the potential not only to significantly increase safety, but also to result in significant changes in traffic flow and car-following. Studies have shown [22] that ACC may reduce congestion even at a market penetration of 20 % of vehicles in the traffic stream. Thus it is possible that future work in car-following focuses on the development of appropriate algorithms that can be used within vehicles such that they will be able to maintain stable headways, thus increasing safety and reducing congestion.

Lane Changing

Compared to car-following models, in which the behavior of the lead vehicle is relatively unaffected by the lag one, the lane-changing process depends on many parameters, and hence it is more complex. Drivers have many different reasons for changing lanes, and their lane change maneuver is likely to be affected by its urgency (e.g., it may be imperative to change lanes in order to get positioned for an upcoming turn, while it may not be as pressing to pass a slightly slower vehicle ahead). Also, drivers may change lanes with or without cooperation from the vehicles in the target lane. Therefore the characteristics of lane changing are several and diverse. However, it has not been studied as extensively as car-following. Its examination has intensified relatively recently and particularly with the increased use of microsimulation.

There are no analytical relationships that encompass the entire lane-changing process. Instead, it is typically modeled as a sequence of several decision-making steps, such as those shown in Fig. 2.8 [23]. In Step 1 the driver considers whether a lane change is necessary or desirable. When a lane change is necessary (e.g., because of an upcoming turn) the lane change is called mandatory; otherwise it is called discretionary. Discretionary lane changes are performed to gain some advantage, such as to increase the vehicle's speed, or improve its position in the queue. Different drivers have different criteria on whether to attempt a discretionary lane change, and thus driver characteristics and behavior become very important in the lane-changing process. In modeling this step, one may develop criteria or thresholds that, once met, would trigger an attempt for a lane change. These criteria, or thresholds, may differ for different types of drivers and under different conditions, and they could take into consideration various factors related to the traffic conditions and the driving environment.

Next, the driver determines the target lane for the lane change (Step 2). This decision is more complex for discretionary lane changes, where the driver needs to select a lane based on a set of criteria (e.g., queue length or operating speed in the

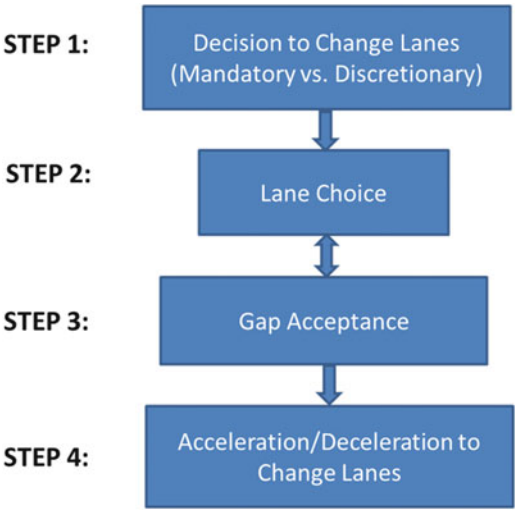


Fig. 2.8 The four steps typically involved in lane changing [23]

target lane). Again, those criteria differ widely for different types of drivers. An approach that has been used in target lane determination is discrete choice (or discrete outcome) models [24, 25]. Such models are based on the concept of utility maximization, i.e., they assume that the driver will select the lane that will provide the most utility, and they estimate the utility provided by each alternative. The probability that driver x will select lane i is

$$\boxed{P_{x,i} = \text{prob}[V_{x,i} > V_{x,n}] \quad \text{for all} \quad n \neq i} \tag{2.14}$$

where

$V_{x,i}$ is the total utility of lane i to traveler x
$V_{x,n}$ is the total utility of all other lanes to traveler x

The total utility can be written as follows:

$$V_{x,i} = U_{x,i} + \varepsilon_{x,i} = \sum_j \alpha_{i,j} y_{x,i,j} + \varepsilon_{x,i}$$

where

$U_{x,i}$ is the specifiable nonrandom component of utility for lane i and traveler x
$\varepsilon_{x,i}$ is the unspecifiable component of utility for lane i and traveler x, assumed to be random
$\alpha_{i,j}$ is the coefficient estimated using field data for lane i and roadway/environment characteristic j
$y_{x,i,j}$ is the roadway/environment characteristic for lane i and traveler x

Assuming that the random unspecified component of utility $\varepsilon_{x,i}$, generalized extreme value distributed, we obtain the logit model formulation [24, 25]:

$$P_{x,i} = \frac{e^{U_{x,i}}}{\sum_n e^{U_{x,n}}} \tag{2.15}$$

where e is the base of the natural logarithm ($e = 2.718$).

In Step 3 the driver evaluates the size of possible gaps in the target lane(s). The driver may accept a gap, reject a given gap, and attempt to find an acceptable one or reevaluate whether to perform a lane change. This step involves the gap acceptance process, which is discussed in more detail in the next section. Steps 2 and 3 may occur simultaneously, as for example when a driver is attempting to find a suitable gap in either of two adjacent lanes during a discretionary lane change. In Step 4, the vehicle moves into the target lane by adjusting its speed as necessary (accelerating or decelerating).

Many previous studies have focused on lane-changing behavior along freeways [26, 27], while fewer have studied lane changes in urban arterials, where the possible lane-changing reasons and occasions are much more numerous [23]. Gipps [28] developed one of the first lane-changing models for microsimulation tools, focusing on urban streets. In his model Gipps considered (a) whether it is physically possible and safe to change lanes; (b) the location of permanent obstructions; (c) the presence of transit lanes; (d) the driver's intended turning movement, (e) the presence of heavy vehicles; and (f) speed. The model consists of a flowchart to replicate decision making related to changing lanes and uses mathematical expressions to answer specific questions in the decision process. It was developed to satisfy the conflicting objectives of being positioned in the correct lane for an upcoming turn (long-term objective) while taking advantage of opportunities to gain speed (short-term objective). It assumes that the closer the vehicle is to the upcoming turn, the lower the probability of making lane changes for speed advantage.

In subsequent work, Hidas [29] developed a lane-changing algorithm which incorporated forced and cooperative lane-changing maneuvers. A forced lane change is defined as one where the lag vehicle in the target lane is forced to decelerate in order to create a suitable gap for the lane-changing vehicle. A cooperative lane-changing maneuver is one where the lag vehicle willingly decelerates to accommodate the lane-changing vehicle. These maneuvers are more likely to occur during congested conditions.

Example 2.2 The following utility functions have been developed using data from three-lane freeways to indicate driver preference for aggressive drivers for each of the three lanes:

$$U_{x,\mathrm{L}} = 23.5 - 1.65\,F_{\mathrm{TL}} + 0.18 v_{\mathrm{L}}$$
$$U_{x,\mathrm{M}} = 71.87 - 1.85 F_{\mathrm{TM}} + 0.21 v_{\mathrm{M}}$$
$$U_{x,\mathrm{R}} = 259.14 - 2.27 F_{\mathrm{TR}} + 0.25 v_{\mathrm{R}}$$

where

$U_{x,\mathrm{L}}$, $U_{x,\mathrm{M}}$, $U_{x,\mathrm{R}}$ are the utilities of the left, middle, and right lanes, respectively
F_{TL}, F_{TM}, and F_{TR} are the truck flows of the left, middle, and right lanes, respectively
v_{L}, v_{M}, and v_{R} are the operating speeds of the left, middle, and right lanes, respectively

If during the peak hour the truck flows are $F_{\mathrm{TL}} = 20$ vph, $F_{\mathrm{TM}} = 45$ vph, and $F_{\mathrm{TR}} = 120$ vph and the speeds are $v_{\mathrm{L}} = 60$ mph, $v_{\mathrm{M}} = 58$ mph, and $v_{\mathrm{R}} = 53$ mph, estimate the probabilities that aggressive drivers will select each of the three lanes.

Solution to Example 2.2
The nonrandom component of utility for each of the three lanes is as follows:

$$U_{x,\mathrm{L}} = 23.5 - 1.65F_{\mathrm{TL}} + 0.18v_{\mathrm{L}} = 23.5 - 1.65 \times 20 + 0.18 \times 60 = 1.3$$
$$U_{x,\mathrm{M}} = 71.87 - 1.85F_{\mathrm{TM}} + 0.21v_{\mathrm{M}} = 0.8$$
$$U_{x,\mathrm{R}} = 259.14 - 2.27F_{\mathrm{TR}} + 0.25v_{\mathrm{R}} = -0.01$$

From Eq. (2.15), the probabilities that drivers will select each of the three lanes are as follows:

$$P_{x,\mathrm{L}} = \frac{e^{U_{x,\mathrm{L}}}}{\sum_n e^{U_{x,n}}} = \frac{e^{1.5}}{e^{1.5} + e^{0.8} + e^{-0.01}} = \frac{3.699}{3.699 + 2.23 + 0.09} = \frac{3.699}{6.919} = 0.535$$
$$P_{x,\mathrm{M}} = \frac{e^{U_{x,\mathrm{M}}}}{\sum_n e^{U_{x,n}}} = \frac{e^{0.8}}{e^{1.5} + e^{0.8} + e^{-0.01}} = \frac{2.23}{3.699 + 2.23 + 0.99} = \frac{2.23}{6.919} = 0.322$$
$$P_{x,\mathrm{R}} = \frac{e^{U_{x,\mathrm{R}}}}{\sum_n e^{U_{x,n}}} = \frac{e^{-0.01}}{e^{1.5} + e^{0.8} + e^{-0.01}} = \frac{0.99}{3.699 + 2.23 + 0.99} = \frac{0.99}{6.919} = 0.143$$

Gap Acceptance

Gap acceptance models are used to determine the number of vehicles or units of traffic that can pass through a conflicting traffic stream, when the conflicting traffic has to evaluate the size of the gap and make decisions regarding its acceptance. A gap is usually defined as the time headway between the rear end of the lead vehicle and the front end of the following vehicle; however, some references also define it as the time headway between successive vehicle arrivals. We will use here the former definition, and we will refer to the second one simply as the time headway. We also refer to the lead gap and the lag gap, as illustrated in Fig. 2.9. The lead gap is the gap to the lead vehicle in the target lane, while the lag gap is the gap to the lag vehicle in the target lane.

Gap acceptance modeling seeks to predict whether a particular size gap will be accepted or rejected by a driver or group of drivers under a given set of prevailing conditions. The driver arriving from a conflicting traffic stream, or changing lanes

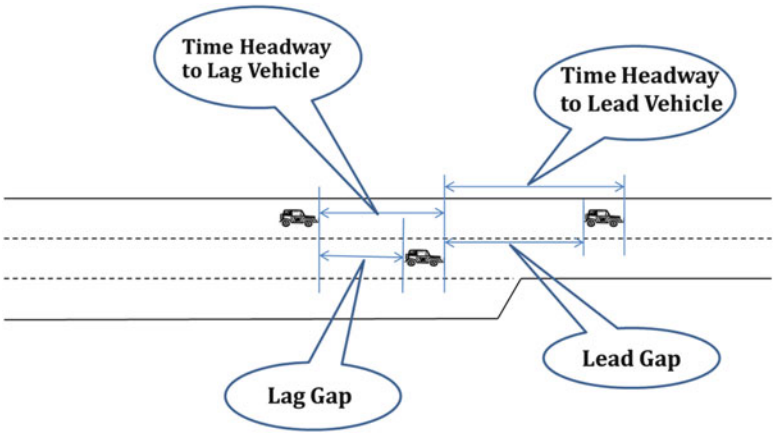


Fig. 2.9 Lane changing with consideration of lead and lag gaps

into the target lane, must evaluate the size of the gap and decide whether to accept it or not. Critical gap is defined as the minimum gap that a driver is willing to accept. The size of the critical gap varies significantly depending on the type of maneuver (lane changing along a freeway segment vs. making a left turn out of a stop-controlled approach) and the driver characteristics. Also, the size of the gap a driver would accept varies based on the amount of time a driver has already been waiting; the longer this interval is, the more likely that the driver would accept a shorter gap.

A major difficulty in gap acceptance modeling is that this critical gap cannot be directly measured. It must be inferred based on the gaps drivers accept and those they reject. Critical gap estimation can be accomplished using a variety of methods. Reference [30] provides an excellent overview of critical gap estimation methods for unsignalized intersections. Even though critical gap may be estimated for a variety of maneuvers, the first gap acceptance models focused on unsignalized intersection operations and thus the majority of the literature to date discusses gap acceptance for that movement.

One of the most popular methods for critical gap estimation is the maximum likelihood estimation. This method uses the maximum rejected gap and the accepted gap for each driver (there is only one accepted gap we can observe) to estimate the probability that the critical gap (t_c) is between those two values. If the cumulative distribution of accepted gaps a_i for a particular driver i is $F(a_i)$, and the cumulative distribution of the maximum rejected gaps r_i is $F(r_i)$, then the probability that this driver's critical gap t_c is between r_i and a_i is $F(a_i) - F(r_i)$. Figure 2.10a provides an example of the probability density function (pdf) of the accepted and maximum rejected gaps for a particular driver. As expected, as the gap size increases, the probability that the driver will accept the gap increases, while the probability that it will reject it decreases. Figure 2.10b shows the cumulative density function (cdf) for the two distributions along with the

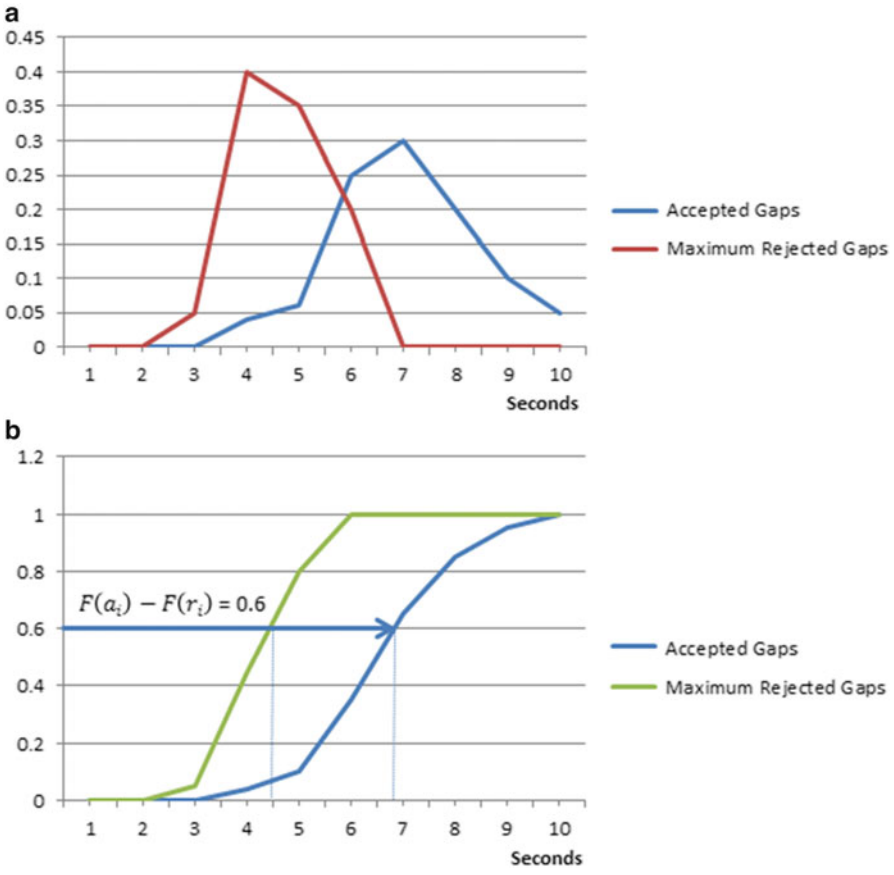


Fig. 2.10 Pdf and Cdf of the accepted and maximum rejected gaps for driver *i*

probability that the driver's critical gap t_c is between r_i and a_i. For the example of Fig. 2.10 the probability that the critical gap is between 4.5 and 6.8 s is 0.6.

The likelihood that a sample of n drivers having an accepted gap and a largest rejected gap of (a_i, r_i) is [30]:

$$L^* = \prod_{i=1}^{n} [F(a_i) - F(r_i)] \tag{2.16}$$

The logarithm, L, of the likelihood is

$$L = \sum_{i=1}^{n} \text{In}[F(a_i) - F(r_i)] \tag{2.17}$$

The likelihood L^* is maximized when its logarithm L is maximized. Next, the user must assume a given type of distribution of critical gaps $F(t_c)$ for the prevailing driver population; it is also assumed that all drivers are consistent (they will always accept a gap of size x and reject a gap of size y). The parameters of the critical gap distribution (mean and variance) are estimated by setting the partial derivatives of L with respect to these parameters to zero:

$$\boxed{\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} \frac{\frac{\partial F(a_i)}{\partial \mu} - \frac{\partial F(r_i)}{\partial \mu}}{F(a_i) - F(r_i)} = 0} \tag{2.18}$$

$$\boxed{\frac{\partial L}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{\frac{\partial F(a_i)}{\partial \sigma^2} - \frac{\partial F(r_i)}{\partial \sigma^2}}{F(a_i) - F(r_i)} = 0} \tag{2.19}$$

These two equations have to be solved using iterative procedures.

Most gap acceptance models have assumed that drivers are both consistent and homogeneous (all drivers behave identically and will accept the same size gaps). These assumptions are not realistic, but they have proven to be acceptable from a model accuracy perspective [31].

References

1. May AD (1990) Traffic flow fundamentals. Prentice Hall, Englewood Cliffs, NJ
2. Pipes LA (1953) An operational analysis of traffic dynamics. J Appl Phys 24(3):274–281
3. Gazis DC (2002) Traffic Theory. Kluwer's international series. Kluwer academic publishers, ISBN: 0-306-48217-7, p259
4. Gazis D, Herman R, Rothery RW (1961) Non-linear follow-the Leader models of traffic flow. Oper Res 9:545–567
5. Edie LC (1960) Car following and steady state theory for non-congested traffic. Oper Res 9:66–76
6. Brackstone MA, McDonald M (1999) Car-following: a historical review. Transp Res Part F 2:181–196
7. Ceder A, May AD (1976) Further evaluation of single and two regime traffic flow models. Transp Res Rec 567:1–30
8. Roess RP, Ulerio JM (1997) NCHRP Report 385, Comparison of the 1994 highway capacity manual's Ramp analysis procedures and the FRESIM model, Transportation Research Board, Washington, DC
9. Cohen SL (2002) Application of car-following systems in microscopic time-scan simulation models, Transportation Research Record 1802, Transportation Research Board, Washington, DC, pp 239–247
10. Gipps PG (1981) A behavioural car-following model for computer simulation. Transp Res Part B 15B:105–111
11. AIMSUN 5.0 User's Guide (2010) Transport simulation systems - TSS, Barcelona, Spain
12. Soria I (2010) Assessment of car-following models using field data. MS thesis, University of Florida, May 2010

13. Yang Q, Koutsopoulos HN (1996) A microscopic traffic simulator for evaluation of dynamic traffic management systems. Transp Res Part C 4(3):113–129
14. VISSIM 5.10 User's Guide (2010) PTV Planung Transport Verkehr AG, Germany
15. Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car-following. Oper Res 6(2):165–184
16. Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car-following. Oper Res 7(1):86–106
17. Ozaki H (1993) Reaction and anticipation in the car-following behavior. In: Proceedings of the 13th international symposium on traffic and transportation theory, p 366
18. Chundury S, Wolshon B (2000) Evaluation of CORSIM car-following model by using Global Positioning System field data. Transp Res Rec 1710(-1):114–121
19. Panwai S, Dia H (2005) Comparative evaluation of microscopic car-following behavior. IEEE Trans Intell Transp Syst 6(3):314–325
20. Chakroborty P, Kikuchi S (1999) Evaluation of the general motors based car-following models and a proposed fuzzy inference model. Transp Res Part C Emerg Technol 7(4):209–235
21. Traffic flow theory, a monograph (Web Document: http://www.tfhrc.gov/its/tft/tft.htm), Chapter 4: Car-following
22. Martin B (2010) Evaluating the impacts of advanced driver assistance systems using a driving simulator: an exploratory analysis. MS Thesis, University of Florida, Dec 2010
23. Sun D, Elefteriadou L (2010) Research and implementation of lane changing model based on driver behavior, Transportation Research Record: Journal of the Transportation Research Board, No 2161, Transportation Research Board of the National Academies, Washington, DC, pp 1–10
24. Mannering FL, Washburn SS, Kilareski WP (2009) Principles of highway engineering and traffic analysis, 4th edn. Wiley, Hoboken, NJ
25. Washington SP, Karlaftis MG, Mannering FL (2003) Statistical and econometric methods for transportation data analysis, 2nd edn. CRC, Boca Raton, FL
26. Ahmed KI (1999) Modeling divers' acceleration and lane changing behavior. Sc.D. Dissertation, Massachusetts Institute of Technology, Department of Civil and Environmental Engineering, Cambridge, MA
27. Laval JA, Daganzo CF (2006) Lane-changing in traffic streams. Transp Res Part B 40 (3):251–264
28. Gipps PG (1986) A model for the structure of lane changing decisions. Transp Res 20B:403–414
29. Hidas P (2002) Modeling lane changing and merging in microscopic traffic simulation. Transp Res Part C 10:351–371
30. Brilon W, Koenig R, Troutbeck RJ (1999) Useful estimation procedures for critical gaps. Transp Res Part A 33:161–186, Pergamon
31. (1992) Revised monograph on traffic flow theory, Federal Highway Administration. http://www.fhwa.dot.gov/publications/research/operations/tft/

Problems

1. Solve Example 2.1 assuming the following: desired maximum speed of the following vehicle, $v(n+1)_{\mathrm{DES}}$, is 80 mph, the maximum acceleration which the following vehicle wishes to undertake, $a(n+1)_{\mathrm{MAX}}$, is 8 ft/s^2, the actual most severe deceleration that the follower wishes to undertake, $b(n+1)$, is -10.5 ft/s^2, the most severe deceleration rate that vehicle $n+1$ estimates for vehicle n, $\widehat{b}_n$, is -12.5 ft/s^2, and the effective vehicle length, $L(n)$, is 25 ft. How do the results differ from those of Example 2.1. Explain any differences observed.

2. Solve Example 2.1 assuming that the car-following model is GHR with parameters $m = 1$, $l = 1$, and $\lambda_{m,l} = 2.6$. How do the results compare to those of Example 2.1?
3. Conduct a literature review of car-following models. What parameters are used in addition to the ones described in this chapter?
4. Conduct a literature review of lane-changing models for freeways and arterials.
5. The following utility functions have been developed using data from four-lane freeways to indicate driver preference for truck drivers for each of the four lanes:

$$U_{x,L1} = 181.25 - 0.18F_{L1} + 1.27v_{L1}$$

$$U_{x,L2} = 341.3 - 0.27F_{L2} + 1.35v_{L2}$$

$$U_{x,L3} = 40.53 - 0.05F_{L3} + 0.35v_{L3}$$

$$U_{x,L4} = 18.26 - 0.02\,\mathrm{F_{L4}} + 0.07\,\mathrm{v_{L4}}$$

where $U_{x,L1}$, $U_{x,L2}$, $U_{x,L3}$, $U_{x,\mathrm{L4}}$ are the utilities of each of the four lanes.

F_{L1}, $F_{L2,}$ F_{L3}, and F_{L4} are the passenger vehicle flows of each of the four lanes.

v_{L1}, v_{L2}, v_{L3}, and v_{L4} are the operating speeds of each of the four lanes.

If during the peak hour the passenger vehicle flows are $F_{L1} = 1{,}380$ vph, $F_{L2} = 1{,}560$ vph, $F_{L3} = 1{,}250$ vph, and $F_{L4} = 1{,}320$ vph and the speeds are $v_{L1} = 55$ mph, $v_{L2} = 60$ mph, $v_{L3} = 62$ mph, and $v_{L4} = 62$ mph, estimate the probabilities that truck drivers will select each of the four lanes.

6. Conduct a literature review of gap acceptance models for unsignalized intersections.
7. Conduct a literature review of gap acceptance models for merging at freeway junctions.

Part II
The Traffic Stream: Traffic Flow Performance Characteristics

The second part of this book discusses the movement of groups of vehicles and their overall performance when viewed as a traffic stream. The traffic stream performance was one of the first areas of research in traffic flow theory, and thus, there is a wealth of research studies on the development of traffic stream models. This part of this book focuses on describing the traffic stream as a whole; it discusses the variables that are typically used to assess the traffic operational performance of a traffic stream along a highway facility and the relationships between some of these key performance measures.

Chapter 3 discusses the three fundamental variables of interest in assessing traffic stream performance: flow, speed, and density. It also presents and discusses the fundamental diagram which provides the theoretical relationship between these three variables, as well as traffic stream models that have been developed to fit field data. Such models have been developed and are being used to evaluate the overall performance of a highway facility.

Chapter 4 focuses on capacity definition, measurement, and estimation. Capacity is one of the key performance characteristics, and its importance lies in that when it is exceeded, conditions become congested. The traffic stream models as well as many of the predictive models have very different shapes and trends when conditions are non-congested versus when they become congested.

Chapter 5 provides an overview of performance measures used to assess traffic operational performance. The chapter provides definitions as well as methods of measurement and estimation for some of the key performance measures used in traffic analysis. Measures discussed include travel time, travel time reliability, delay, and queue length.

Chapter 3
Flow, Speed, Density, and Their Relationships

Flow, speed, and density are the three primary characteristics of traffic and are used to describe various aspects of operations of a highway facility. When describing and assessing traffic operations, we are often concerned with the movement of a group of vehicles, or the traffic stream as a whole, rather than the movement of each vehicle. In those cases, it is more convenient to describe traffic operations in terms of macroscopic measures of traffic.

This chapter first defines each of these measures along with their microscopic expressions and then illustrates their definition in the context of a time–space diagram. The last section presents the fundamentals of the relationship between them, which constitutes what is referred to as the fundamental diagram.

Flow and Time Headway

Flow, Capacity, and Demand

Flow is defined as the rate at which vehicles travel through a particular point or highway segment; it is the inverse of the time headway ($=3{,}600/h_{\text{avg}}$). Flow is expressed in units of traffic per unit of time, typically vehicles per hour (vph). The time headway is often referred to as the microscopic expression of flow, and it is given in units of time over unit of traffic, typically seconds per vehicle (s/veh). Volume is typically expressed in units of traffic, and it represents the count of the units of traffic within a specified time interval. For example, if the volume measured at a particular location is 1,400 vehicles within 15 min, then the flow is 5,600 vph ($=1{,}400 \times 4$ 15-min intervals). Both of these measures may be expressed on a per lane basis, i.e., in units of vehicles per hour per lane (vphpl).

In practice, we often measure volume in 15-min intervals. Then, within the hour of analysis, we identify the 15-min interval with the highest volume, we extrapolate to flow (e.g., we multiply the 15-min peak volume by 4 to obtain the hourly rate),

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_3,

and we use this flow to conduct our analysis. Alternatively, if only hourly volumes are available, we divide them by the peak hour factor (PHF) so that we can obtain and analyze our highway network based on the flow equivalent to the highest volume 15-min interval. The PHF is defined as follows:

$$\boxed{\text{PHF} = \frac{V_{\text{H}}}{4 \times V_{15}}} \tag{3.1}$$

where

V_{H} is the volume during the full hour
V_{15} is the volume during the peak 15-min period in the hour

The PHF expresses the variability of volumes between 15-min periods (or smaller time periods) within the hour. The value of the PHF ranges between 0.25 and 1. When all the volume is concentrated in one time interval and there is no traffic in the other three intervals, the PHF is 0.25. When the volume is uniformly distributed throughout the hour, the PHF is 1. In reality, the PHF does not reach these extreme values, and it typically ranges between 0.70 and 0.90. Values of the PHF are often assumed based on the type of location studied or even based on local agency estimates. Of course, when 15-min volumes are available, the peak hourly flow can be calculated directly, and there is no need to select or assume a PHF.

Generally, smaller intervals of data collection show higher fluctuations and variability, and they can result in extremely high or extremely low flow rates. For example, 1-min data collection intervals will have a much larger range of flows than 15-min intervals. The analysis interval selected during data collection is a function of the purpose of the study. Typical traffic operational analyses use 15-min intervals. However, some traffic management techniques and algorithms (such as ramp metering) may be based on 20-s or 1-min intervals.

Two key traffic measures expressed in units of vph are capacity and demand. Even though these are expressed in the same units as flow or volume, the terms should not be used interchangeably. Capacity provides a measure of maximum throughput, and it represents the maximum amount of traffic a facility can handle under a given set of conditions (Chap. 4 provides a more detailed discussion on capacity). Demand represents the amount of traffic that desires to use a particular highway facility. Two types of demand can be distinguished: the demand that arrives to use the facility and the demand that desires to use the facility but does not because travelers anticipate congestion at that location and/or during that particular time interval. This latter type of demand is termed "latent demand," and it is very difficult to measure or estimate. In this book, when we refer to demand, we do not consider latent demand; we focus only on the demand that arrives to use the facility.

During non-congested conditions, when there are no constraints in the highway facility, the demand arriving to use the facility is equal to the volume (or flow) at

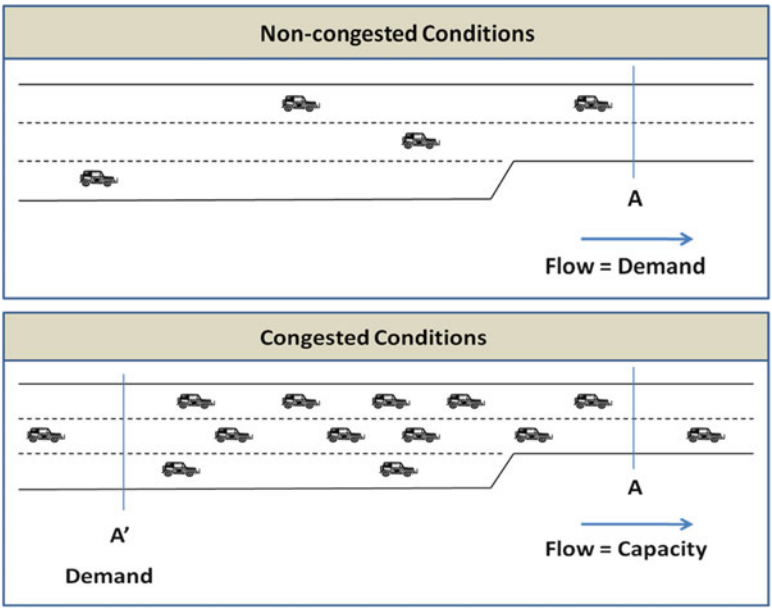


Fig. 3.1 Demand and capacity for non-congested and congested conditions

that location. The top part of Fig. 3.1 shows traffic flowing through a bottleneck while conditions are uncongested. Flow can be measured anywhere throughout this segment, and it will be found to be approximately the same. However, during congested conditions, shown in the bottom part of Fig. 3.1, the demand arriving to use the facility cannot be processed and queues form upstream of the bottleneck location. In that case, the demand is equal to the amount of traffic joining the back of the queue per unit of time (point A′), while the throughput is equal to the capacity of the bottleneck (point A).

Example 3.1 For the facility shown in Fig. 3.2 and for the volumes shown in Table 3.1, identify the peak 15-min period at A, estimate the peak flow at A, and estimate the PHF for the time period 17:00–18:00. What is the capacity of the facility, and what is the maximum queue that would be observed during the entire analysis period?

Solution to Example 3.1

The volumes at points A and A′ are similar for the first two 15-min intervals. However, starting with the third interval, the two values deviate significantly. This occurs because conditions become congested, and the demand upstream of the bottleneck exceeds its capacity.

The peak 15-min period at A is 17:45–18:00, with a volume of 538 vehicles.

The peak flow at A is $538 \times 4 = 2{,}152$ vph.

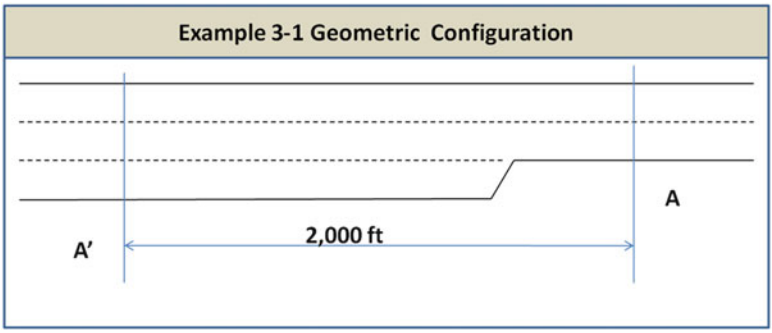


Fig. 3.2 Geometric configuration for Example 3.1

Table 3.1 Traffic volume data for Example 3.1

Time	Point A′ (vehicles)	Point A (vehicles)
17:00–17:15	170	152
17:15–17:30	421	447
17:30–17:45	723	520
17:45–18:00	684	538
18:00–18:15	585	496
18:15–18:30	406	503
18:30–18:45	250	492
18:45–19:00	210	251
19:00–19:15	204	221

The PHF for the time period 17:00–18:00 is estimated using Eq. (3.1):

$$\text{PHF} = \frac{V_{\text{H}}}{4 \times V_{15}} = \frac{152 + 447 + 520 + 538}{4 \times 538} = 0.77$$

Note that the PHF during the period 17:30–18:30 when traffic is much more uniform is 0.96.

The capacity of the facility can be estimated using the discharge volumes at point A. The discharge volumes are those occurring when the demands at point A′ are significantly higher, indicating vehicle accumulation between those two points. The capacity can be estimated using the volumes between 17:30 and 18:15. During this time, the average 15-min volume is (520 + 538 + 496)/3 = 518 volume. This corresponds to a capacity flow of 518 × 4 = 2,072 vph.

Alternatively, capacity can be assumed to correspond only to the highest discharge throughput, i.e., 538 vehicles. In this case, the capacity is estimated to be 538 × 4 = 2,152 vph. Detailed discussion on capacity and its estimation are provided in Chap. 4. Currently, there is no guidance in the HCM 2010 [1] or

Table 3.2 Queue estimates for Example 3.1

Time	Point A′ (vehicles)	Point A (vehicles)	Queue
17:00–17:15	170	152	
17:15–17:30	421	447	
17:30–17:45	723	520	203
17:45–18:00	684	538	349
18:00–18:15	585	496	438
18:15–18:30	406	503	341
18:30–18:45	201	492	50
18:45–19:00	240	310	−20
19:00–19:15	232	221	

elsewhere regarding the best approach for obtaining capacity in this case; thus, both answers can be considered correct, depending on the overall study objectives.

The maximum queue can be estimated based on the vehicle accumulation between points A and A′. Table 3.2 duplicates the volumes provided, along with an estimate of the queue for each time period. The queue starts accumulating during the time period 17:30–17:45, and it is estimated as the difference between the incoming volume at point A′ and the outgoing volume at point A ($723 - 520 = 203$ vehicles). For the following intervals, the queue is estimated in a similar manner, but it includes the vehicles accumulated during the previous intervals. For example, for the time period 17:45–18:00, the queue is $684 - 538 + 203 = 349$. The queue has completely dissipated during the interval 18:45–19:00, as the estimate of the queue is negative! The maximum queue observed is during the interval 18:00–18:15, and its value is 438 vehicles.

Time Headway

The inverse of flow, time headways, is of interest not just in modeling the movement of individual vehicles but also as expressions of the operational quality of the traffic stream. Time headway distributions are used in various applications (gap acceptance, percent time spent following), as they allow us to consider microscopic traffic characteristics for the entire traffic stream. In gap acceptance, we are interested in the number and frequency of usable gaps which we can easily obtain if we know the distribution of time headways. Another application relates to percent time spent following, defined as the percent of time vehicles spend following another vehicle (used in evaluating the performance of two-lane highways). The frequency of short headways is a good approximation of this performance measure, and it can be obtained through the time headway distribution.

The time headways in the field have been shown to follow these distributions: For low flow, the exponential distribution has been shown to represent time headways. Note that the exponential interarrival times correspond to Poisson-

Table 3.3 Field data for Example 3.2

Time headway (s)	Frequency (before)	Frequency (after)
(0–2)	198	158
(2–4)	156	146
(4–6)	148	125
(6–8)	110	98
(8–10)	127	102
(10–12)	85	117
(12–14)	46	93
(14–16)	48	80
(16–18)	62	72
(18–20)	48	75
(20–22)	71	66
(22–24)	53	58
(24–26)	77	49
(26–28)	64	58
(28–∞)	102	113

distributed arrivals. For high flow, when vehicles tend to travel closer together, the normal distribution has often been used. Other distributions commonly used in traffic operations include Cowan's M3 and the shifted negative exponential distribution.

Example 3.2 Time headway data have been collected along the northbound direction of a two-lane highway before and after the installation of a passing lane upstream of the site. Table 3.3 provides a summary of the time headways and their frequencies for the before and after conditions. Has the percent time following changed after the installation of the passing lane? Has the average time headway been increased? Assume that a vehicle is in following mode when its time headway is less than 4 s.

Solution to Example 3.2
Figure 3.3 plots the data provided in Table 3.3. The data show that the frequency of smaller headways (less than 10 s) is reduced, while the frequency of the midsize headways (10–20 s) has increased after the installation of the passing lane. The total number of headways less than 4 s was 198 + 156 = 354 in the before condition and 158 + 146 = 304 in the after condition. The total number of headways in the before sample was 1,395, and in the after sample 1,410. Therefore, the percent time spent following in the before condition was 354/1,395 = 0.25, or 25 %, while in the after condition, it is 304/1,410 = 0.22 or 22 %. Therefore, the percent time spent following has been reduced.

Table 3.4 provides the calculations for estimating the average time headway before and after the installation of the passing lane. The leftmost column provides the midpoint of each time headway interval, which is used to represent the interval in the calculations. The two rightmost columns include the detailed calculations.

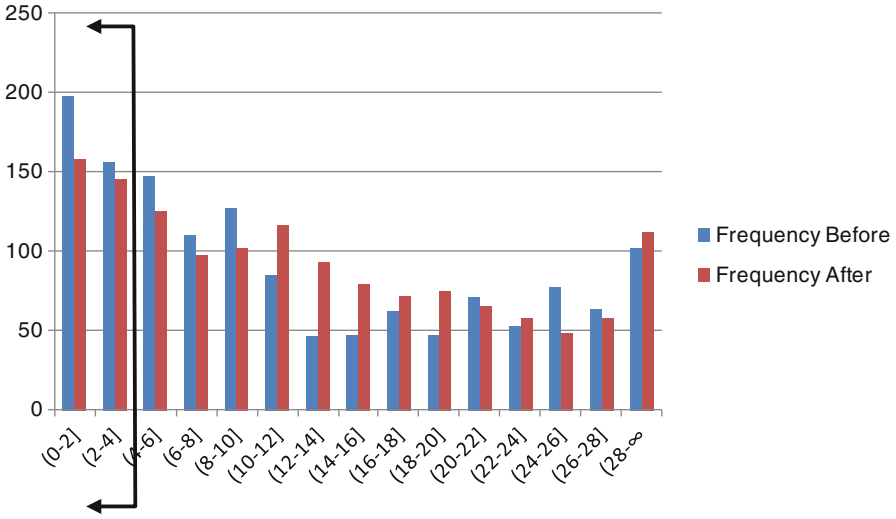


Fig. 3.3 Histogram of data from Example 3.2

Table 3.4 Calculations and solution to Example 3.2

Midpoint of the time headway interval	Time headway (s)	Frequency (before)	Frequency (after)		Calculations (before)	Calculations (after)
1	(0–2)	198	158		198	158
3	(2–4)	156	146		468	438
5	(4–6)	148	125		740	625
7	(6–8)	110	98		770	686
9	(8–10)	127	102		1,143	918
11	(10–12)	85	117		935	1,287
13	(12–14)	46	93		598	1,209
15	(14–16)	48	80		720	1,200
17	(16–18)	62	72		1,054	1,224
19	(18–20)	48	75		912	1,425
21	(20–22)	71	66		1,491	1,386
23	(22–24)	53	58		1,219	1,334
25	(24–26)	77	49		1,925	1,225
27	(26–28)	64	58		1,728	1,566
29	(28–∞)	102	113		2,958	3,277
	Total (veh)	1,395	1,410	Mean (s)	12.09	12.74

The number in each cell is the multiple of the midpoint by the respective frequency. The average time headway for each condition is shown at the bottom of the table.

A statistical test (z-test) is conducted next to determine whether the change in the percent time spent following is statistically significant. The null hypothesis, H_0, is that the average time headway has not changed and that the mean of the first sample is equal to the mean of the second sample ($\mu_1 - \mu_2 = 0$). The alternate hypothesis, H_α, is that the average time headway has changed ($\mu_1 - \mu_2 \neq 0$). The test statistic is [2]:

$$z = \frac{(m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.2}$$

where m_1 and m_2 are the estimates of the first and second sample means, respectively, σ_1 and σ_2 are the two standard deviations, and n_1 and n_2 are the sample sizes. The value of the test statistic is -1.89. Based on the Standard Normal Table (Appendix A) and for a 95 % confidence interval, the rejection region is for $z < -1.96$. Since $-1.96 < -1.89$, the null hypothesis cannot be rejected, and it can be concluded that statistically the average time headway has not changed since the installation of the passing lane.

Measurement Techniques for Flow and Time Headways

Volume can be measured using several different types of equipment. The most frequently used devices are in-pavement detectors (a.k.a. loop detectors), microwave-based sensors, and video-based sensors. Each type of technology has its own strengths and weaknesses with respect to environmental and traffic conditions, as well as installation and maintenance [3]. In-pavement loop detectors are often used to record volumes and also to serve as vehicle detection devices in actuated signal control systems. Their main advantages are that they are very accurate compared to alternative methods, and they are not sensitive to adverse weather conditions. However, they can be costly to install and maintain as they require lane closures and pavement cutting.

Another type of technology used extensively in freeway management systems is microwave radar. Remote Traffic Microwave Sensors (RTMS) are widely used to collect volume (as well as speed and occupancy) data. These types of devices are becoming very popular as they are relatively easy to install and maintain. They are also not affected by adverse weather conditions. However, they may not be accurate unless they are very well calibrated. Continuous wave (CW) Doppler sensors in particular cannot detect stopped vehicles, and thus, data may not be accurate for congested conditions.

Video detection is another popular data collection method, which provides increased flexibility and ease of maintenance. Such systems provide increased

flexibility, as one can easily modify the placement of virtual detectors on an image of the intersection approach and adjust their location and type. On the other hand, video data collection can be costly, and it typically requires specialized training to maintain. Such systems can be susceptible to adverse weather conditions, particularly rain, fog, and snow, and high winds. They need to be installed 30–50 ft above ground to minimize occlusion.

Other technologies include laser radar, ultrasonic, acoustic, and others. Reference [3] provides a thorough assessment and discussion of various traffic detector technologies.

Speed

Speed is measured in units of distance per unit of time, typically miles per hour (mph). The same measure is used for describing the movement of an individual vehicle (microscopic analysis) as well as the movement of the traffic stream (macroscopic analysis).

There are two ways in which average speeds can be obtained. The observer could measure instantaneous speeds at a particular location and obtain the average of those instantaneous speeds at the particular location. In this case, the average speed is estimated as:

$$v_{\text{avg-time}} = \frac{\sum_1^n v_i}{n} \tag{3.3}$$

where

v_i are the instantaneous speeds
n are the total number of instantaneous speeds in the sample

This method, which is based on instantaneous speeds, yields the time-mean speed at a particular location. Alternatively, the observer could measure the travel time of each vehicle between two particular locations and obtain each vehicle's speed as the inverse of their travel time. In this case, the average speed is estimated as:

$$v_{\text{avg-space}} = \frac{d}{\sum_1^n \frac{t_i}{n}} \tag{3.4}$$

where

d is the distance over which travel times were measured
t_i are the travel times observed
n are the total number of travel times measured

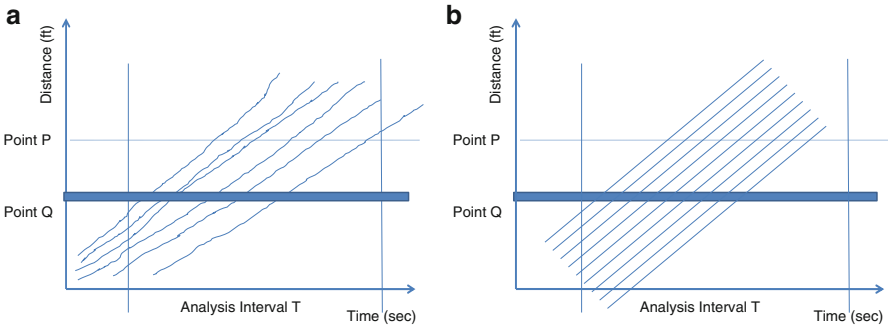


Fig. 3.4 Time-mean speed and space-mean speed with different vehicle trajectories

This second equation estimates the space-mean speed, or the harmonic mean of the speeds. The time-mean speed estimates the average speed for vehicles that travel an equal amount of time, while the space-mean speed estimates the average speed for vehicles that travel an equal amount of distance. Figure 3.4a presents a time–space diagram with vehicle trajectories, illustrating the concepts of time-mean speed versus space-mean speed. The time-mean-speed is estimated using the measurements dx/dt at location Q, while the space-mean speed is estimated using the travel times between locations Q and P. The harmonic mean is always lower than the arithmetic mean, as it weighs more heavily the lower elements. Thus, the space-mean speed is less than the time-mean speed, because slower vehicles tend to stay on the subject segment longer and thus are weighted more heavily. In Fig. 3.4b, all vehicles travel at the same speed. In this case, the time-mean speed and the space-mean speed are equal.

The approximate relationship between time-mean speed and space-mean speed is [4]:

$$v_{\text{avg-time}} = v_{\text{avg-space}} + \frac{\sigma^2_{\text{space}}}{v_{\text{avg-space}}} \tag{3.5}$$

Where σ^2_{space} is the variance of the space-mean speed.

Similarly, the approximate relationship between space-mean speed and time-mean speed is [4]:

$$v_{\text{avg-space}} = v_{\text{avg-time}} - \frac{\sigma^2_{\text{time}}}{v_{\text{avg-time}}} \tag{3.6}$$

where σ^2_{time} is the variance of the time-mean speed.

In addition to the average speed, the analyst might be interested in the entire speed distribution or in a specific measure of the distribution. For example, the standard deviation of speed is of interest because it has been linked to the crash

Table 3.5 Calculations and solution to Example 3.3

Vehicle	Speed (mph)	Travel time (h)	Travel time (s)
1	60.00	0.0167	60.00
2	50.00	0.0200	72.00
3	40.00	0.0250	90.00
Average	50.00	0.0206	74.00

potential and incident levels. Also, the speed limit is theoretically set based on the 85th percentile speed. Speed is an extremely useful performance measure. It can be directly used to indicate whether the facility is congested or non-congested. In the HCM 2010 [1], it is used to define the quality of service for two lane highways and urban streets.

There are several different speed-related terms that are used in traffic operational analysis. Some of the most frequently used are:

Free-flow speed (*FFS*): This is the speed of a facility at low flows (for freeways, it can be measured for flows less than 1,300 vphpl). It reflects the design features of the facility and the driver behavior and preferences when unconstrained by traffic conditions.

Operating speed: This is defined as the speed of the facility for the prevailing conditions. At low flows, this is equal to the FFS.

Design speed: This is the speed for which the facility was designed. It reflects design elements such as superelevation, stopping sight distance, and passing sight distance, etc.

Example 3.3 An observer records the instantaneous speed of three vehicles at the beginning of a 1-mile segment to be 60, 50, and 40 mph. If it is assumed that the vehicles travel at constant speeds throughout the 1-mile segment, calculate the time-mean speed and the space-mean speed for this sample of three vehicles.

Solution to Example 3.3

Table 3.5 shows the speeds of the three vehicles along with their respective estimated travel times in hours and in seconds. According to Eq. (3.3), the time-mean speed is

$$v_{\text{avg}} = \frac{\sum_1^n v_i}{n} = \frac{60 + 50 + 40}{3} = 50\,\text{mph}$$

From Eq. (3.4), the space-mean speed is:

$$v_{\text{avg}} = \frac{d}{\sum_1^n \frac{t_i}{n}} = \frac{1}{\frac{0.0167 + 0.0200 + 0.0250}{3}} = \frac{1}{0.0206} = 48.65\,\text{mph}$$

As indicated earlier, the space-mean speed is lower than the time-mean speed, unless all speeds are equal. The difference in speeds between the first two vehicles

Table 3.6 Field data for Example 3.4

Speed (mph)	Frequency (before)	Frequency (after)
(30–32)	2	4
(32–34)	17	21
(34–36)	28	21
(36–38)	49	53
(38–40)	95	199
(40–42)	173	230
(42–44)	308	436
(44–46)	351	283
(46–48)	267	194
(48–50)	108	51
(50–52)	54	30
(52–54)	35	27
(54–56)	24	10
(56–58)	5	5
(58–60)	8	1

is equal to the difference between the last two (10 mph); however, the differences in travel times are 12 s and 18 s, respectively. The slower speeds weigh more heavily when estimating the space-mean speed shown in Table 3.6.

Example 3.4 A sample of speeds was collected before and after the installation of a new sign which lowered the speed limit from 50 mph to 45 mph. Did the new speed limit sign reduce the average speed at this location? Were the speeds before the installation of the sign normally distributed? Are they normally distributed after the installation of the sign?

Solution to Example 3.4
Table 3.7 provides the calculations for estimating the average speed before and after the installation of the sign. To calculate the average value, each frequency observation is multiplied by the midpoint of the respective speed interval. The results of this calculation are shown in the last two columns of Table 3.7. The sum of the entire column is then divided by the total number of the respective observations to obtain the average speed for each condition.

A z-test of means is conducted first to evaluate whether the speed limit reduced average speeds. The null hypothesis, H_0, is that the average speed has not changed and that the mean of the first sample is equal to the mean of the second sample ($\mu_1 - \mu_2 = 0$). The alternate hypothesis, H_α, is that the speed has changed ($\mu_1 - \mu_2 \neq 0$). The test statistic is estimated according to Eq. (3.2), and it is found to be equal to 8.78. Based on the Standard Normal Table (Appendix A) and for a 95 % confidence interval, the rejection region is for $z > 1.96$. Since $1.96 < 8.78$, the null hypothesis is rejected, and it can be concluded that statistically the average speed has changed.

Figure 3.5 presents the speed distributions before and after the installation of the speed limit sign. As shown, in several of the lower-speed bins, the frequencies

Table 3.7 Average speed calculations for Example 3.4

Midpoint of the speed interval	Speed (mph)	Frequency (before)	Frequency (after)		Calculations (before)	Calculations (after)
31	(30–32)	2	4		62	124
33	(32–34)	17	21		561	693
35	(34–36)	28	21		980	735
37	(36–38)	49	53		1,813	1,961
39	(38–40)	95	199		3,705	7,761
41	(40–42)	173	230		7,093	9,430
43	(42–44)	308	436		13,244	18,748
45	(44–46)	351	283		15,795	12,735
47	(46–48)	267	194		12,549	9,118
49	(48–50)	108	51		5,292	2,499
51	(50–52)	54	30		2,754	1,530
53	(52–54)	35	27		1,855	1,431
55	(54–56)	24	10		1,320	550
57	(56–58)	5	5		285	285
59	(58–60)	8	1		472	59
	Total (veh)	1,524	1,565	Mean (mph)	44.48	43.23

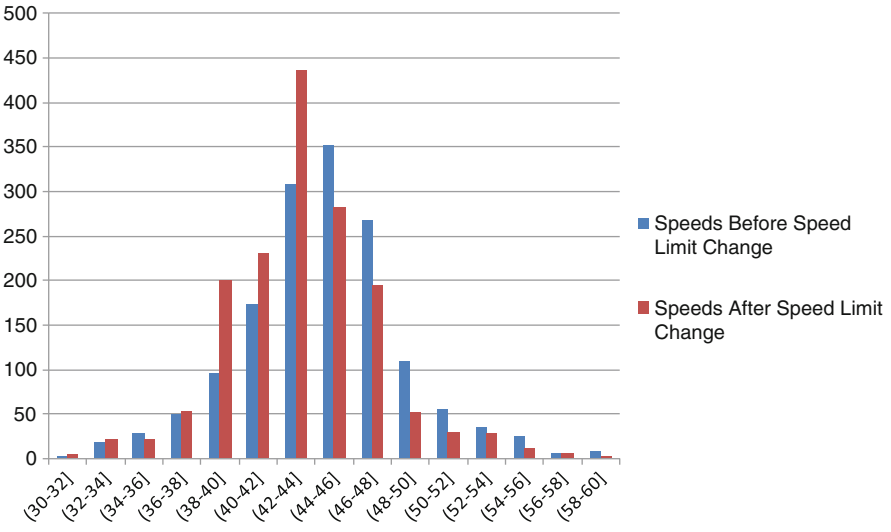


Fig. 3.5 Speed distributions before and after the installation of the speed limit sign

increased, while in several of the higher-speed bins, frequencies decreased. The mode of the distribution (the most frequent value) was reduced from (44–46) to (42–44). The standard deviation of the before distribution is calculated to be 4.12 mph, while the standard deviation after is 3.73 mph.

A statistical test is conducted to determine whether the two samples are from normal distributions. The χ^2 (Chi-Square) test for distributions is used for this evaluation. In the χ^2 test, we first obtain the frequencies of the observed distribution and the theoretical frequencies of the test distribution (in this case the normal distribution). Next, we determine their differences and calculate the following quantity:

$$\chi^2 \text{ statistic} = \sum_{i=1}^{n} \frac{(f_0 - f_t)^2}{f_t}$$

This quantity is compared to the respective table value (Appendix B). For $n = (15 - 1) - 2 = 12$ degrees of freedom, the rejection region for 95 % confidence level (5 % significance level) is $\chi^2 > 21.0$. The calculations to compare the before data to a normal distribution using the χ^2 test are shown in Table 3.8. The theoretical frequency, f_t, for each cell is estimated as follows. First, we calculate the z value for the upper limit of each speed interval ($z = (x - \mu)/\sigma$). Next, using the Standard Normal Table (Appendix A), we obtain the respective probability of occurrence for values below each calculated z value. Then, the probability of occurrence for each interval is calculated by subtracting the probabilities of the respective upper limits of successive intervals. We obtain the theoretical frequency for each interval by multiplying the respective probability by the total number of data points in the entire sample. The last column of the table provides the calculations for the χ^2 test statistic. As shown, for the before data, that value is 164.61, which is significantly larger than the table value. Thus, the hypothesis that the sample is from a normal distribution is rejected. Similarly, we obtain the χ^2 test statistic for the after data. That value is 122.55, and the null hypothesis is rejected in this case as well.

Measurement Techniques for Speed

Speed can be measured using various methods. The most common method for measuring time-mean speed is using dual-loop detectors, which use the same technology as the single-loop detectors used to measure volumes. Dual-loop detectors estimate the time it takes for a vehicle to traverse the distance that separates the two single-loop detectors. RTMS and video detection are also used to measure time-mean speeds. When sampling a relatively small amount of vehicles, radar guns can also be used. These devices can also trace a given vehicle and provide a time series of its speed.

Measurement of space-mean speed, which is based on travel times, is more challenging, as it requires the monitoring of a vehicle over a longer section of the highway facility. Travel time measurement methods are discussed in Chap. 5.

Table 3.8 The χ^2 test for the before data of Example 3.4

Speed (mph)	Observed frequency f_0 (before)	z-Statistic	Probability of speed lower than upper value of interval	Probability of speed within interval	Theoretical frequency for normal distr. f_t (before)	$f_0 - f_t$	$\frac{(f_0 - f_t)^2}{f_t}$
(30–32)	2	−3.02	0.0013	0.0013	2.0	0	0.00
(32–34)	17	−2.4	0.0055	0.0042	6.4	11	17.55
(34–36)	28	−2.06	0.0197	0.0142	21.6	6	1.87
(36–38)	49	−1.57	0.0582	0.0385	58.7	−10	1.60
(38–40)	95	−1.09	0.1379	0.0797	121.5	−26	5.77
(40–42)	173	−0.60	0.2743	0.1364	207.9	−35	5.85
(42–44)	308	−0.12	0.4522	0.1779	271.1	37	5.02
(44–46)	351	0.37	0.6443	0.1921	292.8	58	11.59
(46–48)	267	0.85	0.8023	0.1580	240.8	26	2.85
(48–50)	108	1.34	0.9099	0.1076	164.0	−56	19.11
(50–52)	54	1.82	0.9656	0.0557	84.9	−31	11.24
(52–54)	35	2.31	0.9896	0.0240	36.6	−2	0.07
(54–56)	24	2.79	0.9974	0.0078	11.9	12	12.34
(56–58)	5	3.28	0.9995	0.0021	3.2	2	1.01
(58–60)	8	3.76	1.0000	0.0005	0.8	7	68.75
N =	1,524				1,524	χ^2 Statistic =	164.61

Density and Space Headway

Density is expressed in units of traffic per unit of distance, typically vehicles per mile (vpm), or in vehicles per mile per lane (vpmpl). Density is the inverse of space headway ($=5{,}280/s_{avg}$). Density is very difficult to measure directly with the technology available today; however, it is a very useful measure of performance. An important application of density is its use in the HCM 2010 [1] to assess the quality of service of basic freeway sections, ramp junctions, and weaving segments (additional discussion regarding density and its use in assessing freeway performance is provided in Chap. 8).

The maximum density (D_{max}) that can be achieved on a highway facility can be estimated considering the minimum spacing, s_{min}. The minimum spacing s_{min} consists of the average vehicle length plus the minimum gap between vehicles, i.e., the gap when vehicles are stopped. Assuming an average vehicle length of 20 ft and an average gap between vehicles of 5 ft, the maximum density is:

$$D_{max} = \frac{5,280}{s_{min}} = \frac{5,280}{20+5} \approx 211 \text{ vpmpl}$$

A surrogate often used for density is occupancy. Occupancy is defined as the time a roadway sensor is occupied, i.e., a vehicle occupies its zone of influence, and it is usually expressed as a percent.

Traffic Stream Characteristics in Time and Space

Figure 3.6 illustrates a time–space diagram with a series of trajectories. Based on the definition of density provided earlier, the density at time t_1 is the number of vehicles present between points A and B. This definition is adequate for practical purposes, but mathematically point measurements cannot be defined. Thus, density is estimated over a very small time, dt, as follows:

$$\boxed{D_{AB} = \frac{n\mathrm{d}t}{d_{AB}\mathrm{d}t}} \tag{3.7}$$

where

D_{AB} is the density between points A and B
n is the number of vehicles located between points A and B
d_{AB} is the distance between points A and B

The denominator of Eq. (3.7) is the area of the light-shaded vertical rectangle in Fig. 3.6. If dt is adequately small, we can assume that no vehicles arrive or depart

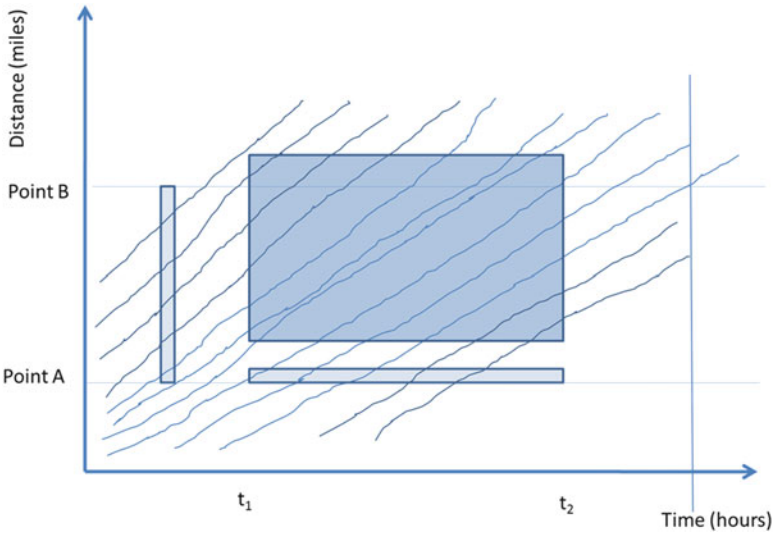


Fig. 3.6 Traffic stream characteristics and the time–space diagram

during that time. In this case, the nominator of Eq. (3.7) represents the total travel time of all vehicles traversing the region. Then, Eq. (3.7) can be rewritten as [5]:

$$\boxed{D_{\mathrm{AB}} = \frac{\text{Total travel time (veh-hours)}}{\text{Total time} - \text{space region (mile-hours)}}} \tag{3.8}$$

This expression of density can be generalized to any region, such as the large dark-shaded rectangle of Fig. 3.6. This approach, originally proposed by Edie [6], allows us to obtain density using very different measurement techniques and also allows us to compare different traffic streams observed for different times and across different distances [5].

Similarly, we can define flow, as illustrated in the light-shaded horizontal rectangle of Fig. 3.6:

$$\boxed{F_{t_{1-2}} = \frac{m\,\mathrm{dd}}{t_{1-2}\mathrm{dd}} = \frac{\text{Total travel distance (veh-miles)}}{\text{Total time} - \text{space region (mile-hours)}}} \tag{3.9}$$

where

$F_{t_{1-2}}$ is the flow between times t_1 and t_2
m is the total number of vehicles passing between times t_1 and t_2
dd is a very small distance
t_{1-2} is the time between t_1 and t_2

The ratio of Eqs. (3.9) to (3.8) provides us with speed, defined as the ratio of the total distance traveled divided by the total time:

$$\boxed{v\ (\text{miles per hour}) = \frac{\text{Total travel distance (veh-miles)}}{\text{Total travel time (veh-hours)}}} \tag{3.10}$$

Additional details on these definitions are provided in [5, 6].

Traffic Stream Models

Mathematically, the three fundamental traffic flow characteristics (flow, speed, density) are related as follows:

$$\text{Flow} = \text{Speed} \times \text{Density}$$

or

$$\boxed{F = v \times D} \tag{3.11}$$

In terms of units,

$$\text{vph} = \text{mph} \times \text{vpm}$$

Thus, when two of these parameters are known the third can be estimated.

Since the beginnings of traffic flow theory, researchers have been attempting to model the relationship between these three key traffic flow measures. Such relationships provide important information regarding the performance of a facility at different demand levels. For example, we can estimate how the facility would operate if 500 vehicles per hour are added. Such models, also referred to as traffic stream models, have been used in the HCM to define levels of service for freeway and multilane highway segments. A multitude of models describing the relationships between each pair of these parameters have been developed over the years.

The remainder of this subsection presents and discusses some of the most prevalent traffic stream models reported and used in the literature, discusses the HCM models and the importance of location in developing or calibrating traffic stream models, and relates mathematically these macroscopic traffic stream models to the microscopic car-following models.

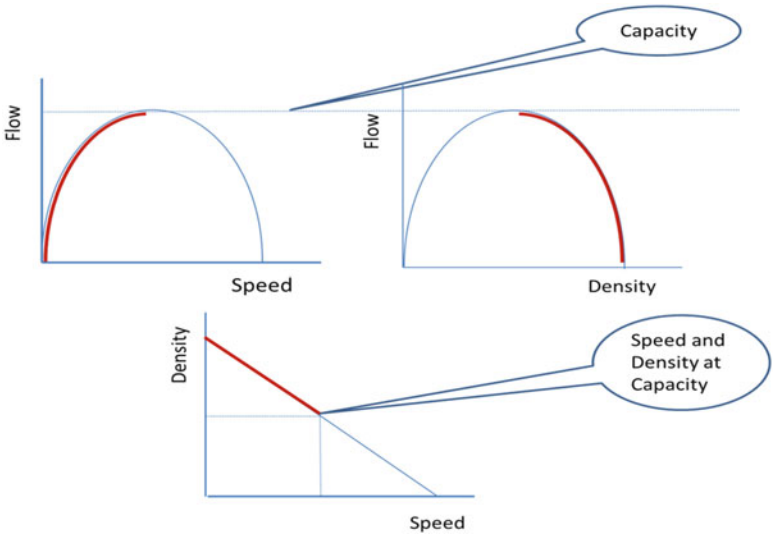


Fig. 3.7 Flow, speed, density relationships: the Greenshields model

The Greenshields Model

The first traffic stream model was developed by Greenshields [7] who developed a linear speed–density relationship based on field data. Figure 3.7 provides a sketch of the Greenshields model, which consists of the flow–speed, flow–density, and speed–density relationships. As indicated earlier, because of the mathematical relationship between the three measures, when one of the relationships is defined, the other two are defined as well. Thus, the linear speed–density relationship results in parabolic flow–speed and flow–density curves.

In Fig. 3.7, the portions of the curves shown in blue correspond to non-congested conditions, while the portions shown in red correspond to congested conditions. For the flow–speed diagram and starting from the rightmost part of the curve, when flow is very low, speed is at its highest level (FFS) as drivers are able to travel at their desired speed. As flows increase, speed gradually decreases. The highest flow, or capacity, is shown to occur at the transition between non-congested to congested conditions. The speed at capacity is often referred to as optimum speed. On the left side of the curve with increasing congestion flow decreases, and so does speed. For very severe congestion, both flow and speed approach zero.

Similarly, in the flow–density curve, maximum flow corresponds to optimum density, while severely congested conditions with flow approaching zero correspond to jam density (D_j), i.e., density when vehicles are at a standstill. Both the flow–speed and flow–density curves are based on the assumption that the speed–density relationship is linear. The original research conducted by Greenshields calibrated these relationships to field data where the maximum

speed (FFS, or v_f) was approximately 43.8 mph. Even though this speed is quite low for today's highways, the overall model and relationships between the three measures are still relevant.

Greenshields derived the parabolic speed–flow relationship based on the assumption of a linear speed–density relationship and using Eq. (3.11). If we assume that the speed–density relationship is linear, then:

$$\boxed{v = v_f - \frac{v_f}{D_j} D} \tag{3.12}$$

where

v_f is free-flow speed
D_j is jam density

Since flow $F = v \times D$, we can multiply both sides of Eq. (3.12) by density, D, to estimate flow:

$$\boxed{F = v \times D = v_f D - \frac{v_f}{D_j} D^2} \tag{3.13}$$

Since $\frac{dF}{dD} = 0$ when D approaches D_{opt}, if we differentiate with respect to D for $D \to D_{opt}$:

$$\frac{dF}{dD} = 0 = v_f - 2\frac{v_f}{D_j} D_{opt} \Rightarrow$$

$$\boxed{D_{opt} = \frac{D_j}{2}} \tag{3.14}$$

Of course, the relationship of Eq. (3.14) is true only if we assume a linear speed–density relationship. Note that for a vehicle length of 20 ft, and a distance between vehicles at standstill of 5 ft, jam density, D_j, is 5,280/25 = 211 vehicles per mile per lane. According to the HCM 2010, the density at capacity is 45 vehicles per mile per lane; therefore, based on recent field measurements, the Greenshields relationship does not replicate this boundary condition very well.

The speed–flow relationship can be derived by substituting $F/v = D$ in Eq. (3.12) and then solving for F:

$$v = v_f - \frac{v_f}{D_j} \frac{F}{v} \Rightarrow$$

$$\boxed{F = \frac{D_j}{v_f}\left(v_f\, v - v^2\right)} \tag{3.15}$$

For $F = F_{max}$ (i.e., at capacity), Eq. (3.15) becomes:

$$F = v_{opt} D_{opt} = \frac{D_j}{v_f}\left(v_f\, v_{opt} - v_{opt}^2\right) \Rightarrow$$

$$\boxed{v_{opt} = \frac{v_f}{2}} \tag{3.16}$$

Substituting Eqs. (3.14) and (3.16), into Eq. (3.11) for F_{max}:

$$\boxed{F_{max} = \frac{D_j\, v_f}{4}} \tag{3.17}$$

The Greenshields model has been used extensively in transportation analysis because it is easy to use in analytical models; however, field observations do not support its shape.

Overview of Other Traffic Stream Models

After the Greenshields model, several other models have been proposed attempting to fit the field data to an analytical model (reference [7] provides a more comprehensive overview). For example, Greenberg [8] proposed a logarithmic relationship between speed and density:

$$\boxed{v = v_{opt} \mathrm{In}\left(\frac{D_{max}}{D}\right)} \tag{3.18}$$

where

v is the space-mean speed (mph)
v_{opt} is the optimum speed (mph)
D_{max} is the jam density (vplpm)

A main drawback of this model is that when density approaches zero, speed approaches infinity. Thus, this model is not appropriate for predicting speeds at lower densities.

Underwood [9] developed the following model:

$$\boxed{v = v_{FF} \cdot \mathrm{e}^{-D/D_{opt}}} \tag{3.19}$$

where

v_{FF} is the free-flow speed
D_{opt} is the optimum density, i.e., the density corresponding to capacity

The drawback of this model is that speed becomes zero only when density reaches infinity. Thus, this model is not appropriate for predicting speeds at high densities.

Pipes [10] proposed a model shown by the following equation:

$$\boxed{v = v_f \cdot \left[1 - \left(\frac{D}{D_{max}}\right)^n\right]} \tag{3.20}$$

where n is a parameter which allows for a more generalized modeling approach. When n is set to one, Pipes' model is identical to the Greenshields model.

Van Aerde [11] proposed the following four parameter model, which can be simplified to revert to the Greenshields model:

$$\boxed{D = \frac{1}{c_1 + c_2/(v_f - v) + c_3 v}} \tag{3.21}$$

where

c_1 is the fixed distance headway constant, $c_1 = mc_2$

c_2 is the first variable distance headway constant, $c_2 = \frac{1}{D_{max}(m+1/v_f)}$

c_3 is the second variable distance headway constant, $c_3 = \frac{-c_1 + v_{opt}/q_{max} - c_2/(v_f - v_{opt})}{v_{opt}}$

m is a constant used to solve for the three headway constants $= \frac{2v_{opt} - v_f}{(v_f - v_{opt})^2}$

The Van Aerde model reduces to the Greenshields model when $c_1 = c_3 = 0$ and it allows additional flexibility to match field data.

The above models are called single-regime models, as they are based on the assumption that the same speed–density relation is valid for the entire range of densities seen in traffic streams. However, field observations show that different relations may occur at different range of densities. Several articles in the literature discuss the discontinuity between the non-congested and congested conditions, and the two-capacity phenomenon (one value corresponding to non-congested conditions and another to congested conditions). Reference [12] discusses some of the multi-regime models that have been proposed in the literature and provides a thorough review of traffic stream models developed before 1990. Reference [13] provides a detailed discussion of traffic stream models and issues related to their calibration.

Most articles focus on traffic stream models for uninterrupted flow facilities (such as freeways and multilane highways), while a few others discuss interrupted flow facilities (i.e., those where traffic is interrupted by traffic signals or signs).

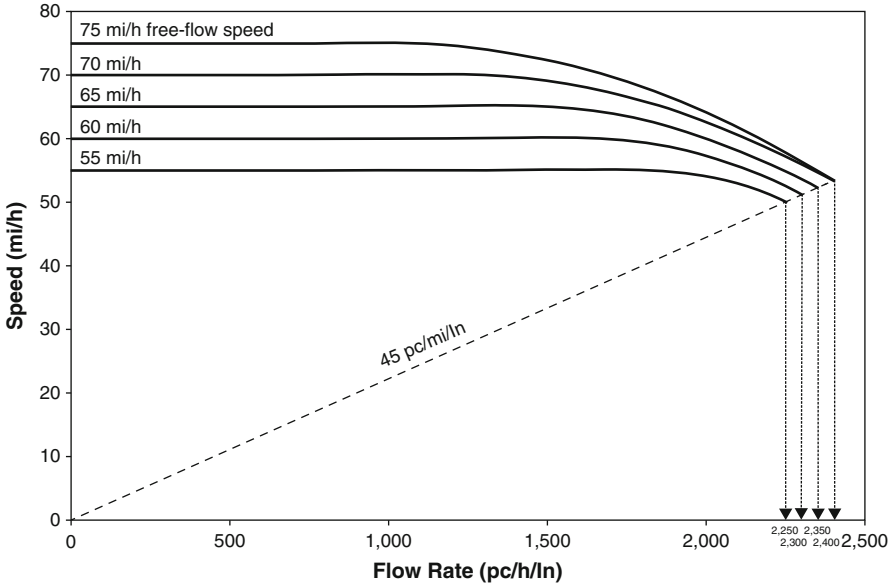


Fig. 3.8 Speed flow curves for freeways (From *Highway Capacity Manual 2010*. Copyright, National Academy of Sciences, Washington, D.C., Exhibit 11-2, p. 11–3; Reproduced with permission of the Transportation Research Board)

The HCM Models

The HCM has historically (beginning with HCM 1965) provided relationships between the primary traffic characteristics (speed, flow, and density) that have been the basis of highway capacity analysis procedures, particularly for uninterrupted flow facilities. Figure 3.8 presents a series of speed–flow curves that are provided in the HCM 2010 [1] and illustrates the relationship between speed and flow for basic freeway segments and for various free-flow speeds (FFS) ranging from 55 to 75 mph. As shown in the figure, speed remains constant for low flows and begins to decrease as flow reaches 1,300–1,750 pc/h/ln. The figure indicates that capacity values vary by free-flow speed. It also clearly illustrates the assumption used in the development of these curves that capacity is reached when density is 45 passenger cars per mile per lane (pc/mi/ln). These curves were developed using data throughout the US, based on the following findings and assumptions:

1. At low flows speed remains constant, and this portion of the curve is not calibrated, but it remains equal to the free-flow speed; this finding was supported by the field data.
2. The capacity points were held constant and equal to the values provided in the HCM 2000; given the difficulty in arriving at a single capacity number and the observed variability around capacity, no change was warranted from this new analysis (more discussion on the issue of capacity is provided in Chap. 4).

3. The form of the equation used to calibrate the curvilinear portion of the curves is as follows:

$$v = \frac{\alpha}{1 + \left[\frac{\alpha}{d(c - F + \mathrm{BP})}\right]} \tag{3.22}$$

where

v is the operating speed
α is a speed parameter
d is the length of the segment
c is the capacity of the segment
F is the flow of the segment
BP is the flow breakpoint where the constant portion of the speed ends.

Additional information on the development of the curves is provided in the HCM 2010 ([1], vol. 4), and in various documents maintained by the Highway Capacity and Quality of Service Committee of TRB (http://www.ahb40.org).

The HCM 2010 provides speed–flow relationships for undersaturated (i.e., non-congested) flow only. When demand exceeds the capacity of the facility, the facility will become oversaturated, with queues forming upstream of the bottleneck location. There are currently no relationships developed for oversaturated conditions at freeways because research has not been conclusive on this topic. One of the complicating factors is that for congested conditions operations depend on the duration of congestion, not just the demand level. Thus, when conditions are congested and assuming the same demand, the average speed would decrease as the duration of that level of demand increases.

Figure 3.9 shows the speed–flow relationship for two-lane highways [1]. In this case, the free-flow speed ranges from 45 to 65 mph, and the relationship between speed and flow is linear. Note that the figure provides the relationship for directional flow, not for the two-lane two-way flow.

Example 3.5 Using Fig. 3.8, determine the operating speed of a facility with a demand of 1,950 pc/h/ln, if the facility free-flow speed is 70 mph. What is the density of the facility at that demand?

Solution to Example 3.5
Figure 3.10 provides a graphical solution to the problem. Starting from the horizontal axis and for a demand of 1,950 pc/h/ln, we draw a vertical line to intersect the curve for freeway segments with 70 mph free-flow speed. Then from that point, we draw a horizontal line to the speed axis, and we determine that the operating speed under those conditions is expected to be 64 mph. The density can be calculated using Eq. (3.4):

$$F = S \times D \Rightarrow D = F/S \Rightarrow D = 1{,}950/64 = 30.5 \ \text{pc/mi/ln}$$

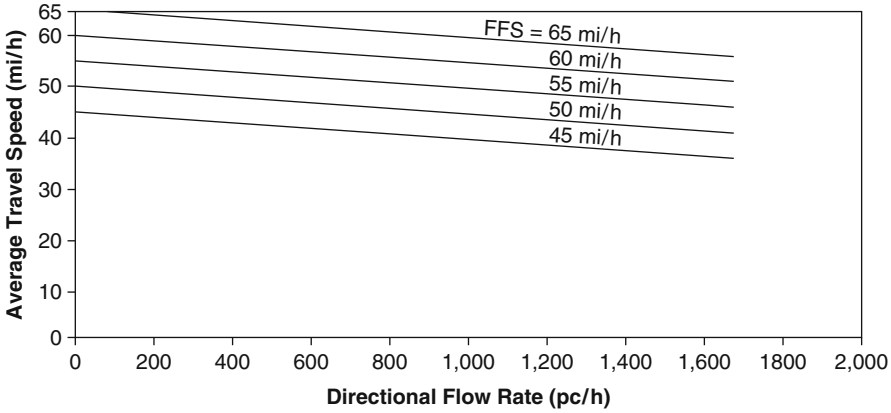


Fig. 3.9 Speed flow curves for two-lane highways (From *Highway Capacity Manual 2010*. Copyright, National Academy of Sciences, Washington, D.C., Exhibit 15-2, p. 15–6; Reproduced with permission of the Transportation Research Board)

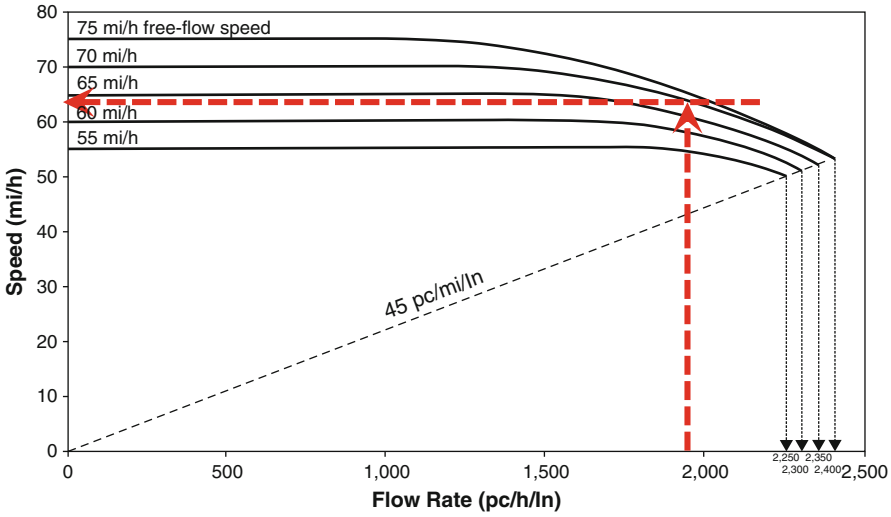


Fig. 3.10 Solution to Example 3.5 (From *Highway Capacity Manual 2010*. Copyright, National Academy of Sciences, Washington, D.C., Exhibit 11-2, p. 11–3; Reproduced with permission of the Transportation Research Board)

Data Collection Location and the Speed–Flow Density Relationships

When collecting field data to obtain speed–flow density relationships, one cannot gather data for the entire speed–flow density curve at a single location. The exact location where data are collected, and its relationship to the nearest bottleneck,

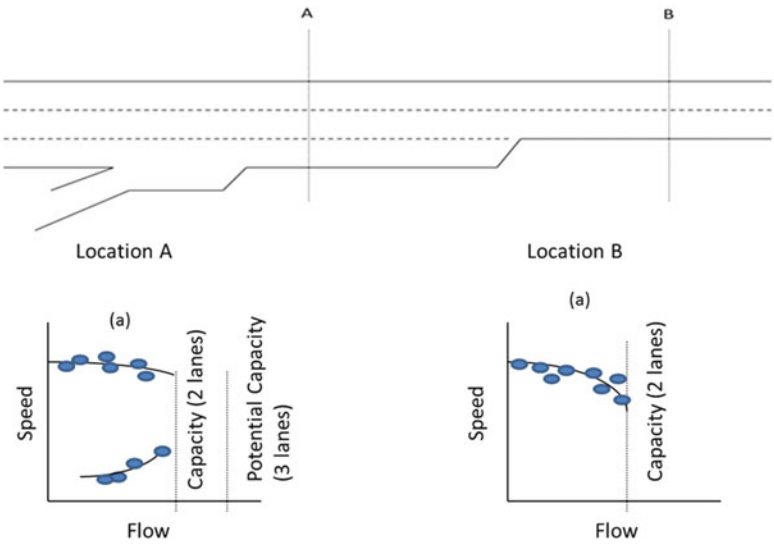


Fig. 3.11 Measurement locations and the corresponding speed–flow relationships

dictates which portion of the curves will be represented. Figure 3.11 provides a sketch of a freeway facility with two consecutive bottlenecks (one merge and one lane drop) along with the speed–flow relationships expected to be obtained at locations A and B. If we are collecting data at both locations A and B, at low flows, speeds will be high, and the data will likely be similar in magnitude. As flows increase steadily, they will reach the capacity of point B, and the bottleneck at location B will result in queue backup into location A. The downstream bottleneck location (i.e., location B) represents the location where capacity flows can actually be measured. At that location, the maximum throughput would be equivalent to a two-lane segment capacity, and it can be easily measured, assuming there is no downstream bottleneck or spillback into this section. However, at location A, the "potential" capacity is that of a three-lane segment, which, however, cannot be attained due to the downstream bottleneck. As soon as location B reaches capacity and breaks down, the queue created spills back into location A, which also becomes congested (for additional discussion, see [12]). Therefore, at location A, one will be able to obtain portions of the non-congested side of the curve as well as portions of the congested side of the curve, but not data close to capacity. At location B, one will be able to obtain the entire portion of the non-congested side of the curve including capacity, but will not be able to obtain any portion of the congested side.

Relationship to Car-Following Models

In the late 1950s, researchers [14] determined that there was a one-to-one relationship between the GHR car-following models (discussed in Chap. 2) and the

macroscopic models developed around the same time. For example, the GHR model for $m = 0$ and $l = 1$ corresponds to the Greenberg macroscopic model [12, 14]. Thus, when we model the microscopic aspects of traffic in essence, we also define the macroscopic relationship of speed–flow and density. The detailed derivation of the relationship is illustrated below.

According to Eq. (2.6), the third GHR microscopic car-following model ($m = 0$, $l = 1$) is given by the following equation:

$$\boxed{a(n+1)_{t+\Delta t} = \lambda_{0,1} \frac{v(n)_t - v(n+1)_t}{x(n)_t - x(n+1)_t}} \tag{3.23}$$

Integrating with respect to t yields:

$$v(n+1) = \lambda_{0,1} \mathrm{In}[x(n) - x(n+1)] + C_1$$

If v is substituted for $v(n + 1)$ and $1/D$ is substituted for $x(n) - x(n + 1)$, then:

$$v = \lambda_{0,1} \mathrm{In}\left(\frac{1}{D}\right) + C_1$$

Let $\lambda_{0,1}$ In C_2 be substituted for C_1, then:

$$v = \lambda_{0,1} \mathrm{In}\left(\frac{1}{D}\right) + \lambda_{0,1} \mathrm{In} C_2$$

$$\boxed{v = \lambda_{0,1} \mathrm{In}\left(\frac{C_2}{D}\right)} \tag{3.24}$$

When $D = D_{\max}$, $v = 0$ (vehicles are at a standstill), therefore the equation becomes:

$$0 = \lambda_{0,1} \mathrm{In}\left(\frac{C_2}{D_{\max}}\right)$$

Solving for C_2, we can get:

$$\mathrm{In}\left(\frac{C_2}{D_{\max}}\right) = 0 \Rightarrow \left(\frac{C_2}{D_{\max}}\right) = 1 \Rightarrow C_2 = D_{\max}$$

Substituting $D_{\max}$ for C_2 in Eq. (3.19) gives:

$$\boxed{v = \lambda_{0,1} \mathrm{In}\left(\frac{D_{\max}}{D}\right)} \tag{3.25}$$

Since $F = v \times D$, multiplying each side of Eq. (3.20) by D gives:

$$F = D\lambda_{0,1}\text{In}\left(\frac{D_{\max}}{D}\right)$$

Differentiating $\frac{\mathrm{d}F}{\mathrm{d}D}$ as $D \rightarrow D_{\text{opt}}$ (optimum density), we can get

$$\frac{\mathrm{d}F}{\mathrm{d}D} = 0 = \lambda_{0,1}\text{In}\left(\frac{D_{\max}}{D_{\text{opt}}}\right) + \lambda_{0,1}D_{\text{opt}}\left(\frac{D_{\text{opt}}}{D_{\max}} \cdot \left(-\frac{D_{\max}}{D_{\text{opt}}^2}\right)\right) \Rightarrow$$

$$0 = \lambda_{0,1}\text{In}\left(\frac{D_{\max}}{D_{\text{opt}}}\right) - \lambda_{0,1}$$

Solving for D_{opt}, we get:

$$\text{In}\left(\frac{D_{\max}}{D_{\text{opt}}}\right) = 1 \Rightarrow \left(\frac{D_{\max}}{D_{\text{opt}}}\right) = e \Rightarrow D_{\text{opt}} = \frac{D_{\max}}{e}$$

Substituting v_{opt} for v and $\frac{D_{\max}}{e}$ for D in Eq. (3.20), we get:

$$v_{\text{opt}} = \lambda_{0,1}\text{In}\left(\frac{D_{\max}}{D_{\max}/e}\right) = \lambda_{0,1}\text{In}e = \lambda_{0,1}$$

Then substituting v for $\lambda_{0,1}$ in Eq. (3.20), we can get

$$v = v_{\text{opt}}\text{In}\left(\frac{D_{\max}}{D}\right)$$

This is the macroscopic Greenberg traffic stream model given earlier in Eq. (3.14).

Pedestrian Traffic Stream Models

Traffic models have also been developed for pedestrian traffic streams. In traffic analysis, pedestrian movement is considered either for pedestrian-only facilities or in combination with other traffic (vehicles, bicycles). For the analysis of pedestrian-only facilities, there are three broad categories for replicating pedestrian flow: macroscopic models, cellular automata, and social forces models. Macroscopic pedestrian stream models are similar to vehicular traffic stream models, but they use parameters such as pedestrian space, speed, and flow per width. In the HCM [1], the LOS for pedestrian facilities is obtained as a function of pedestrian space

(in square feet per pedestrian), which is the inverse of density. Pedestrian space is estimated as:

$$A_P = \frac{v_p}{F_p}$$

where

A_p is the pedestrian space, in ft^2/p
v_p is the pedestrian speed, in ft/min
F_p is the pedestrian flow per unit width, in p/ft/min

The discrete cellular automata approach segments the walkway into individual cells and assigns pedestrians to specific cells. Pedestrians move forward/backwards, left/right, or diagonally within the walkway in order to reach their destination. Conditional bounds are set so that pedestrians do not share the same cell [15].

In the third category of pedestrian traffic stream models, pedestrians are motivated by social forces, including location, velocity, mass, direction, and repulsive forces of other pedestrians. These models are much more flexible in the types of conditions that can be modeled. This flexibility allows them to replicate complex scenarios, such as the evacuation of a building during an emergency. Force-based models tend to require large amounts of processing power when modeling larger areas and many pedestrians [16, 17].

References

1. Transportation Research Board, National Academies (2010) Highway Capacity Manual, Transportation Research Board, National Research Council, Washington, DC
2. Mendenhall W, Wackerly DD, Scheaffer RL (1990) Mathematical statistics with applications, 4th edn. PWS-KENT, Boston, MA
3. FHWA traffic detector handbook, 3rd edn, Publication no. FHWA-HRT-06-108, Oct 2006.
4. Gerlough DL, Huber MJ (1975) Traffic flow theory: a monograph, TRB special report 165. National Research Council, Washington, DC
5. Daganzo CF (1997) Fundamentals of transportation and traffic operations. Pergamon, Oxford
6. Edie LC (1963) Discussion of traffic stream measurements and definitions. In: Almond J (ed) Proceedings of the second international symposium on the theory of traffic flow, OECD, Paris, pp 139–154
7. Greenshields B (1935) A study of Traffic Capacity, Highway Research Board. In: Proceedings of the annual meeting of the Highway Research Board, vol 14, pp 448–477
8. Greenberg H (1959) An analysis of traffic flow. Oper Res 7:78–85
9. Underwood RT (1961) Speed, volume, and density relationships: quality and theory of traffic flow. Yale Bureau of Highway Traffic, New Haven, Connecticut, pp 141–188
10. Pipes LA (1967) Car following models and the fundamental diagram of road traffic. Transp Res 1:21–29
11. Van Aerde M (1995) Single regime speed-flow-density relationship for congested and uncongested highways. Presented at the 74th TRB annual conference, Paper no. 950802, Washington, DC
12. May AD (1990) Traffic flow fundamentals. Prentice Hall, Englewood Cliffs, NJ

13. Roess RP, Prassas ES, McShane WR (2011) Traffic engineering, 4th edn. Pearson Education, Inc., Upper Saddle River, NJ
14. Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady state flow. Oper Res 7(4):499–505
15. Blue VJ, Adler JL (2001) Cellular automata microsimulation for modeling bi-directional pedestrian walkways. Transp Res Part B Methodol 35(3):293–312
16. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. Phys Rev E 51:4282–4286
17. Fellendorf M, Schönauer R, Huang W (2012) Social force based vehicle model for two-dimensional spaces. Presented at the transportation research boards 91st annual meeting, Washington, DC

Problems

1. For the headway data shown below, conduct the following analyses:
   (a) Plot the probability density function and the cumulative density function.
   (b) Determine the sample mean, mode, variance, and 85th percentile.
   (c) Estimate the hourly flow during the data collection.
   (d) Test the hypothesis that these data are from a negative exponential distribution.

Headway data (in s)					
4	5	11	3	14	22
1.5	14	182	20	13	8
8	22	16	10	10	23
16	7	4	12	5	2.5
4.5	12	2	5	5	2
2	8	8	1.5	3	24
3	22	25	4	2	25
2	3.5	14	3	3	8
19	24	5	12	21	7
10	8	40	38	18	21
17	13	19	28	19	12
23	6	15	39	5	4
31	2	2	17	5	3
9	3	6.5	3	2	15
5.5	4	28	12	2	23
17	5	12	26	6	19
42	1.5	8	39	29	3
29	12	7.5	18	17	8
19	9	13	31	7	27
20	4	34	15	3	20
8	2.5	4	4	31	15
2	12	15	4	8	17

2. Conduct a literature review on statistical tests for distributions and summarize your findings. Describe which test is most appropriate for specific conditions.
3. A study of freeway flow at a particular site has resulted in a calibrated speed–density relationship as follows:

$$U = 65.4\ (1 - 0.0075\ \mathrm{K})$$

From this relationship

(a) Find the free-flow speed and jam density.
(b) Derive equations describing flow versus speed and flow versus density.
(c) Determine the capacity of the site mathematically.
(d) Sketch the speed–density, flow–speed, and flow–density curves.
(e) How does this relationship handle boundary conditions (minimal flow, capacity, and jam density conditions)?

4. Conduct a literature review on traffic stream models and for each one identified state their advantages and disadvantages.
5. Collect or obtain speed–flow data from a freeway location and plot them. Discuss the location of the data collection site versus the plot obtained. Is your site located upstream or downstream of a bottleneck?

Chapter 4
Capacity

How much traffic can a facility carry? This is one of the fundamental questions designers and traffic engineers have been asking since highways have been constructed. The term "capacity" has been used to quantify the traffic-carrying ability of transportation facilities. The value of capacity is used when designing or rehabilitating highway facilities to determine their geometric design characteristics such as the desirable number of lanes, it is used to design the traffic signalization schemes of intersections and arterial streets, it is used in evaluating whether an existing facility can handle the traffic demand expected in the future, and it is also used in the operations and management of traffic control systems (ramp metering algorithms, congestion pricing algorithms, signal control optimization, incident management, etc.).

Traditionally, the capacity of a particular facility has been treated as a single number; however, its definition and numerical value have evolved over time. The Highway Capacity Manual [1] is the publication used most often to estimate capacity. The current published version of the HCM 2010 defines the capacity of a facility as "...the maximum sustainable hourly flow rate at which persons or vehicles reasonably can be expected to traverse a point or a uniform section of a lane or roadway during a given time period, under prevailing roadway, environmental, traffic, and control conditions."

With respect to uninterrupted flow facilities, traffic engineers have long recognized the inadequacy and impracticality of this capacity definition. First, the expression "maximum...that can reasonably be expected..." is not specific enough for obtaining an estimate of capacity from field data. Second, field data-collection efforts have shown that the maximum flow at a given facility varies from day to day; therefore, a single value of capacity does not reflect real-world observations. Third, field data have shown that the maximum throughput may be different during non-congested and congested conditions. Fourth, this definition implicitly assumes that capacity occurs immediately prior to breakdown, i.e., the transition to congested conditions; however, field data have shown that the maximum throughput usually occurs much earlier than breakdown. Given the importance of capacity in the planning, design, and operations of highway facilities, significant

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_4,

resources have been devoted to developing an adequate definition of capacity and understanding the mechanism breakdown.

With respect to interrupted flow facilities (i.e., signalized and unsignalized intersections), even though capacity is identically defined, its measurement and estimation is very different. The capacity of a signalized intersection approach is a function of the signalization scheme and most importantly the percent of time that this particular approach is given the green. The capacity of a stop-controlled approach on the other hand is a function of the availability of gaps on the main traffic stream. Because of the significant differences in their operation, the estimation and measurement of capacity for these facilities is discussed in the respective chapters (Chaps. 9 and 10).

Capacity is also important in other types of transportation facilities. For example, when analyzing off-street pedestrian facilities, the HCM 2010 uses the following capacity values to evaluate them [1]: walkways with random flow, 23 p/min/ft; walkways with platoon flow (average over 5 min), 18 p/min/ft; cross-flow areas (sum of both flows), 17 p/min/ft; and stairways (up direction), 15 p/min/ft. In highway transportation, we usually use units of vehicles per unit of time, while specific applications (e.g., comparisons in the capacities of a freeway lane and an exclusive transit lane) may also use units of persons per unit of time. Of significant interest is also the capacity of airports, railroads, and ports. Each of those modes has its own specific factors that affect capacity; however, the principles of estimating capacity are the same regardless of the mode. For example, the literature provides methods for estimating airport runway capacity, which is defined as the maximum number of operations (arrivals, departures, and touch and go) that can take place in an hour [2]. Similarly to highway capacity, the capacity of a runway is the inverse of the minimum interarrival times of aircraft. However, those minimum interarrival times are a function of runway use configuration, type and weight of aircraft, etc. [2].

This chapter provides an overview of the state of the art in the measurement and estimation of highway capacity focusing on uninterrupted flow facilities. This chapter first provides a historical perspective on the definition of capacity in the Highway Capacity Manual (HCM), followed by an overview of recent research findings and recommendations related to defining and measuring capacity. The third section discusses the measurement of capacity in the context of uninterrupted flow facilities.

Capacity in the HCM: A Historical Perspective

The HCM is the publication used most often to estimate capacity as a function of the prevailing characteristics of a highway facility. The first edition of the HCM (1950) defined three levels of roadway capacity: basic capacity, possible capacity, and practical capacity. Basic capacity was defined as “the maximum number of passenger cars that can pass a point on a lane or roadway during 1 h under the most nearly ideal roadway and traffic conditions which can possibly be attained.” Possible capacity was “the maximum number of vehicles that can pass a given

point on a lane or roadway during 1 h under prevailing roadway and traffic conditions." Practical capacity was a lower volume chosen "without the traffic density being so great as to cause unreasonable delay, hazard, or restriction to the drivers' freedom to maneuver under prevailing roadway and traffic conditions."

The second edition of the HCM (1965) states the following: "The confusion that has existed regarding the meaning and shades of meaning of many terms ... has contributed ... to the wide differences of opinion regarding the capacity of various highway facilities. . .. In fact, the term which is perhaps the most widely misunderstood and improperly used ... is the word 'capacity' itself." In other words, the three terms related to capacity allowed for various interpretations by different traffic analysts. To rectify this problem, the HCM 1965 defined a single capacity, similarly to the possible capacity of the HCM 1950. The term basic capacity was replaced by the term capacity under ideal conditions, whereas the term practical capacity was replaced by a series of service volumes to represent traffic operations at various levels of service (LOS). In the HCM 1965, the definition of capacity was revised to read as follows: "Capacity is the maximum number of vehicles which has a reasonable expectation of passing over a given section of a lane or a roadway in one direction (or in both directions for a two-lane or three-lane highway) during a given time period under prevailing roadway and traffic conditions." This definition includes the term "reasonable expectation," which acknowledges that there is variability in the numerical value of the maximum number of vehicles. Subsequent editions and updates of the HCM (1985, 1994, 1997, and 2000) define capacity in a similar manner, with the most recent definition (HCM 2010) as indicated in the introduction to this chapter. This most recent definition indicates that there is an expected variability in the maximum volumes, but it does not specify when, where, and how capacity should be measured, nor does it discuss the expected distribution, mean, and variance of capacity.

Capacity values provided in the HCM have increased over time. For example, the HCM 1950 indicated that the capacity of a basic freeway segment lane is 2,000 pc/h/ln, whereas HCM 2010 indicates that capacity may reach 2,400 pc/h/ln for certain freeway facilities. As was discussed in the previous chapter, the capacity for facilities with FFS 75 mph is 2,400 pc/h/ln and decreases with decreasing FFS. For example, the capacity of a basic freeway segment with FFS of 55 mph is expected to be 2,250 pc/h/ln.

In summary, the definition of capacity within the HCM has evolved over time. There has been an implicit effort to consider the expected variability of maximum volumes in the capacity definition, but there is still no specific guidance in that document on where, when, and how capacity should be measured at a highway facility.

The State of the Art in Defining and Measuring Capacity

When attempting to measure the capacity of specific facilities, practitioners and researchers have frequently encountered various difficulties. Let us review some field data so that we can better understand the issues encountered.

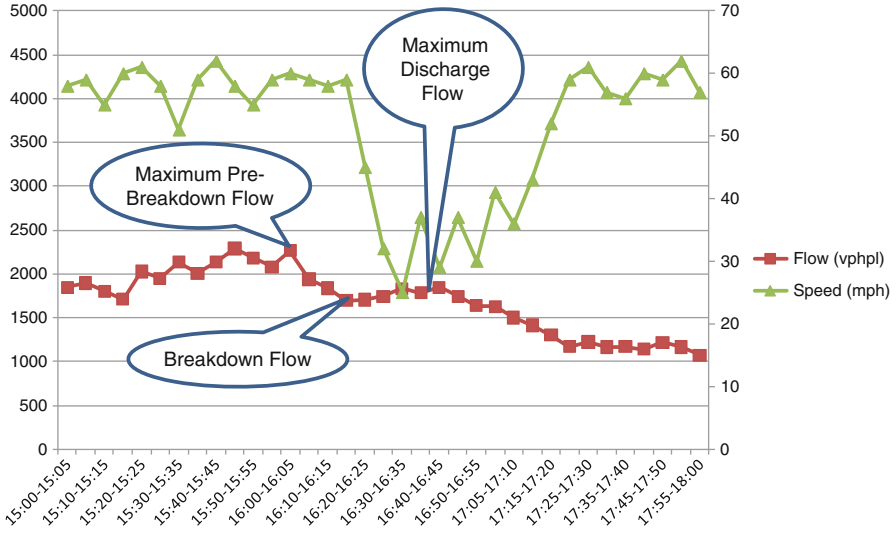


Fig. 4.1 Example of a time series of flow-speed data (adapted from [3])

Figure 4.1 illustrates a time series of flow-speed data at a given freeway site. The horizontal axis indicates the time, the green line indicates the average speed in mph (right-side vertical axis), and the red line indicates the corresponding per lane flows for this three-lane facility (left-side vertical axis). Each data point represents a 5-min data collection period. As shown, the speed is relatively stable (with the exception of one data point at 15:30) and remains around 60 mph until about 16:15 h. The sudden drop in speed that is shown at that time period is typical of a breakdown, i.e., the transition of operations from non-congested to congested conditions [3]. The lower speeds last until about 17:15, and after that speed slowly recovers to the range of 60 mph.

During non-congested conditions, the flows fluctuate between approximately 1,600 and 2,300 vphpl. The highest value observed during non-congested conditions is 2,291 vphpl, and it occurs at 15:45, much earlier than the time of breakdown. In other words, this maximum number does not necessarily coincide with the transition from non-congested to congested conditions. The question researchers and practitioners are attempting to answer is: should capacity be defined as the transition point or as the maximum flow prior to this transition?

Maximum Throughput Values

As Fig. 4.1 shows, there are at least three different time periods of interest with regard to defining capacity on a freeway: prior to the breakdown of flow (drop in speeds), the interval immediately preceding breakdown, and the extended interval

during congested conditions. The flow during the time period immediately prior to breakdown is 5,088 vph, while the maximum flow observed during congested conditions is 5,540 vph. Each of these values provides the maximum throughput under a specific set of conditions, and thus, any of them can be characterized as "capacity." Additional measures could also be developed: for example, one might obtain the average discharge flow, i.e., the average discharge during congested conditions. Which one should we use? There is still no definitive answer to this question, though the literature leans toward using different measures for different project objectives (i.e., one value for design and another for traffic management). Reference [4] provides a more detailed discussion on this topic. The remainder of the chapter refers to these three values as a group as *capacity*, with references to a specific one as appropriate.

Maximum Throughput Distributions

Typically, each day of data collection produces a graph similar, but not identical, to that of Fig. 4.1. The numerical value of each of these three parameters varies, and their range is relatively large, on the order of several hundred veh/h/ln. Their relative value also varies significantly on a daily basis. Table 4.1 lists ranges of values for the parameters identified in Fig. 4.1 as obtained from several freeway facilities around the USA [5]. As shown, each parameter varies significantly, in the order of several hundred vphpl. Should an average of those numbers be used, or should the 85th percentile of the distribution be obtained? Furthermore, what sample size should be used in the determination of capacity? Are the data points provided in Table 4.1 an adequate sample?

Definitions of Breakdown

An issue of particular importance in the definition of capacity is that it has to be somehow tied to breakdown (i.e., the onset of congestion). If we don't have breakdown on the facility, how do we know that capacity has been reached? Thus, to define capacity, breakdown has to be uniquely identified and defined as well.

To identify breakdown, first we need to identify the bottleneck location, i.e., the location where congestion starts. Congestion due to a downstream bottleneck is not considered to be a breakdown event, as congestion in this case is caused by queue spillback, rather than excess demand at the subject location. In this case the throughput is dictated by the carrying ability of the downstream section, rather than that of the subject section.

Once the bottleneck location of interest has been identified, one needs to identify the time of breakdown quantitatively. Usually, this is accomplished by quantifying the required amount of speed drop as well as the minimum duration of that speed drop that would signify the transition to congested conditions. Research studies have defined breakdown in different ways. One study [6] identified and defined

Table 4.1 Capacity measures [5]

			Range of values			
Location	Lanes	Data points	Breakdown flow (vphpl)	Maximum pre-breakdown flow (vphpl)	Ave. flow for 10 min before breakdown (vphpl)	Average discharge flow (vphpl)
Minneapolis, MN	2	35	1,350–2,370	1,920–2,610	1,614–2,271	1,435–1,896
Portland, OR	2	32	1,530–2,565	1,935–2,565	1,296–2,118	1,414–2,025
Toronto, Canada	3	56	1,420–2,520	1,840–2,560	1,512–2,264	1,580–2,046
Sacramento, CA	3	35	1,440–2,280	1,860–2,460	1,597–2,154	1,183–1,843
Sacramento, CA	4	40	630–2,100	1,680–2,265	948–1,962	1,100–1,756
San Diego, CA	4	39	1,305–2,175	1,860–2,310	1,611–1,979	1,422–1,890
San Diego, CA		34	1,392–2,076	1,812–2,076	1,660–1,866	1,473–1,741

breakdown at freeway merge areas as a speed drop below 56 mph (90 km/h) with duration of at least 15 min. Another study on weaving areas [7] used a value of 50 mph (80 km/h). A third study [8] defined breakdown as a reduction of average speed within one 5-min interval to below a threshold of 43.5 mph (70 km/h). A most recent study [9] provided algorithms for obtaining the breakdown based on changes in speed or changes in occupancy or changes in the flow–occupancy correlation. The authors of that study concluded that speed was the measure that can identify breakdown faster than the other two.

In summary, the location where breakdown can be observed can be identified relatively easily. However, the quantitative identification of the exact time of breakdown is not as straightforward. There have been various thresholds and performance measures that have been used in the past. It is very likely that different facilities have different thresholds at which speed is likely to drop when congestion starts. Additional research will need to be conducted to identify these.

Breakdown Probability Models

Breakdown probability models have been developed more recently to take advantage of our new understanding of randomness in capacity and to use this knowledge to improve traffic management. These models provide the probability of breakdown as a function of incoming flows or demands. Figure 4.2 provides an example of such a set of models developed for a ramp-merge bottleneck. The freeway has two lanes per direction and the ramp is metered. The horizontal axis provides the freeway demand, the vertical axis provides the probability of breakdown, and each curve represents the probability of breakdown under the specified ramp metering (or ramp volume) level. For example, for a freeway flow of 5,000 vph and a metering rate of 23 veh/min (or 1,380 vph), the probability of breakdown over a 1-min interval is approximately 29 %. Conversely, if an agency wants to limit the probability of breakdown, they could restrict the ramp flow using ramp metering, such that the probability of breakdown stays below a given threshold (20 % in Fig. 4.2). Furthermore, and in light of the difficulty in selecting an appropriate value for capacity, one could select such values based on a "tolerable probability of breakdown." Using Fig. 4.2 as an example, and assuming a tolerable probability of breakdown equal to 20 %, capacity would be the sum of the ramp and freeway flows at each of the points intersecting the red line. For example, for a ramp metering rate of 25 veh/min (ramp flow $= 25 \times 60 = 1{,}500$) and a mainline flow of 4,200 vehicles, the capacity would be 1,500 + 4,200 = 5,700 vph.

Breakdown prediction models such as the ones shown in Fig. 4.2 are developed for the "critical" bottlenecks, i.e., those where congestion starts recurrently due to merging operations, rather than as a result of a downstream bottleneck. Critical bottlenecks are generally identified by congestion (e.g., low speeds) propagating upstream, whereas the downstream area is free-flowing (or near free-flowing).

The most recent method used to develop the breakdown probability models is the product limit method (PLM). The PLM is based on the work by Kaplan and Meier [10],

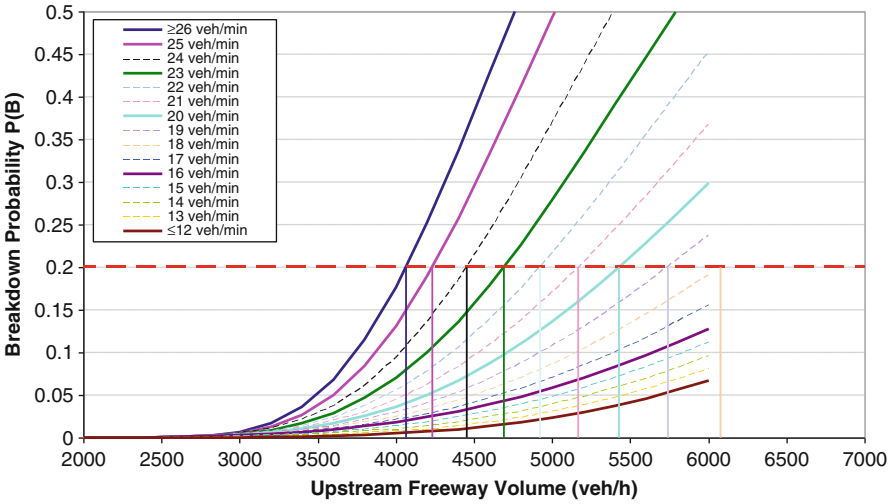


Fig. 4.2 Probability of breakdown models for a ramp-merge junction [9]

which uses lifetime data analysis techniques for estimating the time until failure of mechanical parts or the duration of human life. In lifetime data analysis, the samples obtained are incomplete, in the sense that one doesn't have the values of the random variable itself, but the observed lifetimes of the variable. In other words, observations are cut short by an unrelated event (in the case of medical data, death; in the case of freeway flows, breakdown). The lifetime distribution function is given by [9]:

$$F(t) = 1 - S(t) \tag{4.1}$$

where

$F(t) = p(T \leq t)$ is the distribution function of a lifetime, T is lifetime duration
$S(t) = p(T > t)$ is the survival function

The product-limit estimator of the survival function is given by:

$$\hat{S}(t) = \prod_{t_j < t} \frac{n_j - \delta_j}{n_j} \tag{4.2}$$

where

n_j is the number of items with lifetime $T \geq t_j$
δ_i is the number of failures or deaths at time t_j

In the context of freeway breakdown, the event of a failure at time t is the event of a breakdown at volume q, and the lifetime is the breakdown volume. Thus,

Eq. (4.1) can be modified to calculate the distribution of breakdown probability, as shown in the following relationship:

$$F(q) = p(q_i \leq q) = 1 - p(q_i > q) \tag{4.3}$$

where

$F(q)$ is the breakdown probability distribution
q is the observed traffic volume (veh/h/ln)
q_i is the traffic volume in interval i, which is the one prior to the drop in speeds, defined as the breakdown flow (veh/h/ln)
$p(q_i > q)$ is the probability that the breakdown volume is greater than the observed volume (i.e., the probability that no breakdown will occur up to that volume)

Equation (4.2) can then be written as follows:

$$\hat{p}\,(q_i > q) = \hat{S}\,(q) = \prod_{i:q_i \leq q} \frac{k_i - d_i}{k_i}, \quad i \in B \tag{4.4}$$

where

q is the observed traffic volume (veh/h/ln)
q_i is the traffic volume in interval i, which is the one prior to the drop in speeds, defined as the breakdown flow (veh/h/ln)
k_i is the number of intervals with a traffic volume of $q \geq q_i$
d_i is the number of breakdowns at a volume of q_i
B is the set of breakdown intervals $\{B_1, B_2, B_3, \ldots\}$

The corresponding breakdown probability function is

$$F(q) = 1 - \prod_{i:q_i \leq q} \frac{k_i - d_i}{k_i}, \quad i \in B \tag{4.5}$$

If each observed volume that causes breakdown is considered separately (i.e., only one observation of breakdown for every maximum pre-breakdown volume q_i, $d_i = 1$), then (4.5) becomes:

$$F(q) = 1 - \prod_{i:q_i \leq q} \frac{k_i - 1}{k_i}, \quad i \in B \tag{4.6}$$

Additional information in applying this technique is provided in [9].

Example 4.1 Data were collected immediately downstream of a bottleneck that regularly experiences breakdown. Table 4.2 provides the sorted list of 1-min volumes for the 10 min prior to breakdown for all breakdowns observed. Develop the probability of breakdown model for this site using the PLM method.

Table 4.2 Field data for Example 4.2

Sorted upstream 1-min volumes	Breakdown occurrence δ_i	Sorted upstream 1-min volumes	Breakdown occurrence δ_i	Sorted upstream 1-min volumes	Breakdown occurrence δ_i	Sorted upstream 1-min volumes	Breakdown occurrence δ_i	Sorted upstream 1-min volumes	Breakdown occurrence δ_i	Sorted upstream 1-min volumes	Breakdown occurrence δ_i
920	0	1,600	0	1,760	0	1,860	0	1,980	0	2,060	0
1,080	0	1,600	0	1,760	0	1,860	0	1,980	0	2,080	1
1,160	0	1,600	0	1,760	0	1,860	0	1,980	0	2,080	0
1,180	0	1,620	0	1,760	0	1,860	0	1,980	0	2,080	0
1,280	0	1,620	0	1,780	0	1,860	0	1,980	0	2,080	0
1,280	0	1,620	0	1,780	0	1,880	0	1,980	0	2,080	1
1,300	0	1,620	0	1,780	0	1,880	0	2,000	0	2,080	1
1,320	0	1,620	0	1,780	0	1,880	0	2,000	0	2,100	0
1,360	0	1,620	0	1,780	0	1,880	0	2,000	0	2,100	1
1,380	1	1,640	0	1,780	0	1,880	0	2,000	0	2,100	0
1,380	0	1,640	0	1,780	0	1,880	0	2,000	0	2,100	1
1,400	0	1,640	0	1,800	0	1,880	0	2,000	0	2,120	0
1,400	0	1,640	0	1,800	0	1,880	0	2,000	0	2,120	0
1,400	0	1,640	0	1,800	0	1,880	0	2,000	0	2,120	0
1,400	0	1,640	0	1,800	0	1,900	0	2,000	1	2,120	0
1,440	0	1,640	0	1,800	0	1,900	1	2,000	1	2,120	1
1,440	0	1,660	0	1,800	0	1,900	0	2,000	0	2,120	0
1,440	0	1,660	0	1,800	0	1,920	0	2,000	1	2,120	0
1,460	0	1,660	0	1,800	0	1,920	0	2,020	0	2,120	1
1,480	0	1,660	0	1,800	0	1,920	0	2,020	0	2,140	0

1,480	0	1,680	0	1,820	0	1,920	0	2,020	0	2,140	1
1,480	0	1,680	0	1,820	0	1,920	1	2,020	1	2,140	1
1,520	0	1,680	0	1,820	0	1,920	1	2,020	1	2,160	1
1,540	0	1,680	0	1,820	0	1,920	0	2,020	1	2,160	1
1,540	0	1,700	0	1,820	0	1,940	0	2,020	0	2,160	1
1,540	0	1,700	0	1,820	0	1,940	0	2,020	0	2,160	1
1,540	0	1,700	0	1,820	0	1,940	0	2,020	0	2,180	0
1,540	0	1,700	0	1,820	0	1,940	0	2,020	0	2,180	1
1,540	0	1,720	0	1,820	0	1,940	0	2,020	0	2,200	1
1,540	0	1,720	0	1,820	0	1,940	0	2,020	0	2,200	1
1,560	0	1,720	0	1,820	0	1,940	0	2,020	0	2,220	0
1,560	0	1,720	0	1,840	0	1,940	0	2,040	0	2,220	0
1,560	0	1,720	0	1,840	0	1,960	0	2,040	1	2,220	1
1,580	0	1,720	0	1,840	0	1,960	1	2,060	0	2,260	0
1,580	0	1,740	0	1,840	0	1,960	0	2,060	0	2,260	1
1,580	0	1,740	0	1,840	0	1,960	0	2,060	0	2,280	0
1,580	0	1,740	0	1,840	0	1,960	0	2,060	0	2,280	1
1,580	0	1,740	0	1,860	0	1,960	0	2,060	1	2,300	1
1,590	0	1,760	0	1,860	0	1,980	0	2,060	1	2,320	1
1,600	0	1,760	0	1,860	0	1,980	0	2,060	1	2,320	1

Table 4.3 Sample calculations for Example 4.1

Sorted upstream 1-min volumes	δ_i	k_i	$(k_i - d_i)/k_i$	$\Pi(k_i - d_i)/k_i$	$F(q)$
920	0	240	1.0000	1.0000	0.000
1,080	0	239	1.0000	1.0000	0.000
1,160	0	238	1.0000	1.0000	0.000
1,180	0	237	1.0000	1.0000	0.000
1,280	0	236	1.0000	1.0000	0.000
1,280	0	235	1.0000	1.0000	0.000
1,300	0	234	1.0000	1.0000	0.000
1,320	0	233	1.0000	1.0000	0.000
1,360	0	232	1.0000	1.0000	0.000
1,380	1	231	0.9957	0.9957	0.004
1,380	0	230	1.0000	0.9957	0.004
1,400	0	229	1.0000	0.9957	0.004
1,400	0	228	1.0000	0.9957	0.004
1,400	0	227	1.0000	0.9957	0.004
1,400	0	226	1.0000	0.9957	0.004
1,440	0	225	1.0000	0.9957	0.004
1,440	0	224	1.0000	0.9957	0.004
1,440	0	223	1.0000	0.9957	0.004
1,460	0	222	1.0000	0.9957	0.004
1,480	0	221	1.0000	0.9957	0.004

Solution to Example 4.1
Equation (4.5) is used to calculate the probability of breakdown. Table 4.3 provides the calculations for the first few volumes. The last column in that table provides the probability of breakdown at each volume level.

Table 4.4 provides the complete list of estimated breakdown probabilities, while Fig. 4.3 provides the plot of these values as a function of 1-min volumes.

Summary of the State of the Art in Defining and Measuring Capacity

The answers to the questions above have not yet been answered fully. The literature to date has concluded the following [4]:

- Capacity is a random variable: several studies have shown that there is variability in the maximum sustained flows observed in the range of several hundred vehicles per hour per lane.
- The transition from non-congested to congested flow is probabilistic and may occur at various flow levels.

Table 4.4 Breakdown volumes and probabilities

Breakdown volumes	$F(q)$
1,380	0.004
1,900	0.024
1,920	0.024
1,920	0.034
1,960	0.042
2,000	0.059
2,000	0.074
2,000	0.088
2,020	0.104
2,020	0.119
2,020	0.135
2,040	0.153
2,060	0.173
2,060	0.192
2,060	0.212
2,080	0.232
2,080	0.254
2,080	0.276
2,100	0.299
2,100	0.322
2,120	0.349
2,120	0.379
2,140	0.410
2,140	0.441
2,160	0.472
2,160	0.503
2,160	0.534
2,160	0.565
2,180	0.599
2,200	0.632
2,200	0.665
2,220	0.707
2,260	0.756
2,280	0.817
2,300	0.878
2,320	0.939
2,320	1.000

- There are (at least) three time periods of interest with regard to capacity on a freeway: prior to the breakdown of flow (drop in speeds), the interval immediately preceding breakdown, and the extended interval during congested conditions.
- The maximum values for each of these are random variables, possibly normally distributed.
- Regardless of which one of the three periods of interest are used to define and determine the capacity of a facility, the entire distribution should be obtained or a probability distribution function estimated, over a large number of days.

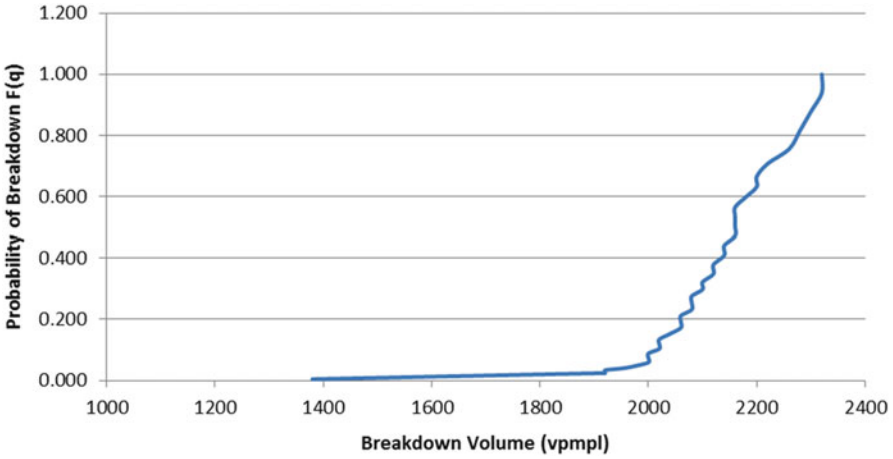


Fig. 4.3 Probability of breakdown plot for Example 4.1

- If both the breakdown flow and the queue discharge flow are random variables, then not only is it possible for the respective value to exceed breakdown flow on a particular day, but one could even estimate that probability if both distributions were known. Hence, treatment of either of these possible measures of capacity as a random variable removes what has been a point of debate in the literature: Is there a capacity drop after breakdown? The answer is "sometimes" (see [11] for a discussion of the "two-capacity" phenomenon).

Capacity of Uninterrupted Flow Facilities

This section provides procedures for obtaining maximum throughput (i.e., capacity) estimates along uninterrupted flow facilities. It provides guidance on observing and measuring (1) maximum pre-breakdown throughput, (2) breakdown flow, and (3) maximum discharge flow. The last section discusses briefly the HCM 2010 capacity estimates for freeway facilities.

Field Data Collection

The four important elements that should be considered when observing breakdown and maximum throughput are site selection and measurement location and definition of breakdown, time interval, and sample size.

Regarding site selection, the site should be regularly experiencing congestion and breakdown as a result of high demands and not as a result of a downstream bottleneck. For example, the site illustrated in Fig. 4.4 (previously discussed in Chap. 3) has two consecutive bottlenecks (one merge and one lane drop). In this case, the downstream bottleneck location (i.e., location B) should be the data-collection point.

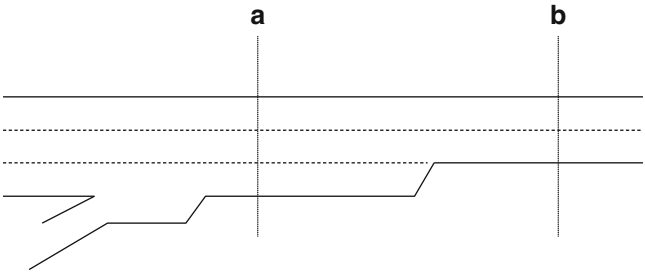


Fig. 4.4 Freeway facility with two consecutive bottlenecks

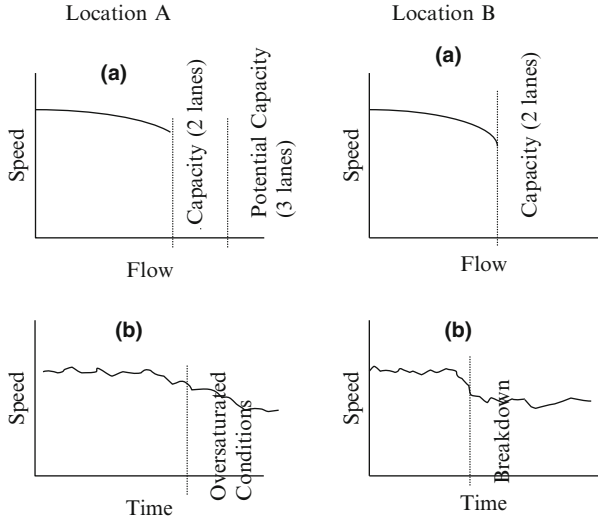


Fig. 4.5 Speed-flow curves and time series for freeway locations A and B

To graphically illustrate the importance of site and location selection, Fig. 4.5 provides the speed-flow relationships and the time series of speed and flow for locations A and B across the freeway facility shown in Fig. 4.4. Figure 4.5 illustrates that the maximum throughput at location B is equivalent to its potential capacity, that is, two-lane segment capacity. At location A, the potential capacity is that of a three-lane segment, which, however, cannot be attained because of the presence of the downstream bottleneck. As soon as location B reaches capacity and breaks down, the queue created spills back into location A, which also becomes oversaturated. The speed time series at location B shows the breakdown occurrence, which typically occurs with a relatively steep speed drop. At location A, speed drops gradually as a result of the downstream breakdown.

The second important element when measuring maximum throughput is the definition and identification of breakdown. As discussed in the previous section, it is important to use a consistent definition of breakdown, determined based on the site characteristics. Given that speed drop and its respective duration can uniquely identify the presence of breakdown, it is recommended that breakdown be defined quantitatively using these two parameters.

The third element that is important in clearly defining maximum throughput is the selection of an appropriate time interval. Time intervals that are typical in traffic operational analysis range between 5 and 15 min. Previous research has demonstrated that maximum throughput increases for smaller intervals owing to general flow variability. Longer intervals result in averaging the flow over a longer time period, and thus, the respective maximum throughput is lower.

The last element to be considered is the required sample size. Given the inherent variability of maximum throughput, it is important to observe an adequate number of breakdown events and the respective maximum pre-breakdown and maximum discharge flows. Sample size determination equations should be used to establish the required number of observations for the desired precision in the maximum throughput estimate.

Once these four elements are established, flow and speed data can be collected at the selected site(s), and time series plots can be prepared (such as the one depicted in Fig. 4.1) for each breakdown event. The breakdown can be identified based on the definition selected, and the respective breakdown flow can be obtained from the time series. Next, the maximum pre-breakdown flows and the maximum discharge flows can be obtained for each breakdown event.

Note that the process for obtaining capacity estimates in a simulator is similar to that of obtaining it in the field. Additional information is provided in Chap. 7 Simulation.

Capacity Estimates in the HCM 2010

The HCM 2010 provides capacity estimates for (1) basic freeway segments, (2) ramp-merge segments, (3) weaving segments, (4) multilane highways, and (5) two-lane highways. It provides, for each segment type and set of geometric conditions, a single value of capacity. For example, for freeway facilities, capacity values are given as 2,250 passenger cars per hour per lane (pc/h/ln) for freeways with free-flow speeds of 55 mph and up to 2,400 pc/h/ln when the free-flow speed is 75 mph (ideal geometric and traffic conditions). These values represent average conditions at similar sites around the USA, obtained based on general trends of maximum flows observed at various freeway locations. The HCM 2010 capacity definition is more closely aligned with the definition of maximum pre-breakdown flow. The HCM 2010 does not define breakdown flows and maximum discharge flows, nor does it provide estimates for these at various facility types.

References

1. Transportation Research Board, National Academies (2010) Highway capacity manual 2010. Transportation Research Board, National Academies, Washington, DC
2. Airport capacity and delay, US DOT, FAA, Advisory Circular no. 150.5060-5, 1983
3. Elefteriadou L, Lertworawanich P (2003) Defining, measuring and estimating freeway capacity. Transportation Research Board Meeting, Washington, DC
4. Elefteriadou L, Hall F, Brilon W, Roess R, Romana M (2006) Revisiting the definition and measurement of capacity. In: 5th International symposium on highway capacity and quality of service, Yokohama, 25–29 July 2006
5. Lu C, Elefteriadou L (2011) An investigation of freeway capacity before and during incidents. Presented at the transportation research board annual meeting, Washington, DC, January 2011
6. Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining capacity and breakdown, Transportation Research Circular E-C018. In: Proceedings of the 4th international symposium on highway capacity, 27 June–1 July 2000, pp 84–95
7. Lertworawanich P, Elefteriadou L (2003) A methodology for estimating capacity at ramp weaves based on gap acceptance and linear optimization. Transp Res Part B Methodol 37B(5): 459–483
8. Brilon W (2005) Randomness and reliability in freeway traffic flow. TRAIL Research School, Delft
9. Elefteriadou L, Kondyli A, Brilon W, Jacobson L, Hall F, Persaud B (2009) Proactive ramp management under the threat of freeway-flow breakdown, NCHRP 3-87. Transportation Research Board, Washington, DC
10. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
11. Banks JH (1991) The two-capacity phenomenon: some theoretical issues, Transportation Research Record 1320. Transportation Research Board, Washington, DC, pp 234–241

Problems

1. Conduct a literature review to identify previous studies that quantify capacity-related measures. Provide numerical values of capacity estimates for various locations around the USA and abroad.
2. Using Fig. 4.2, and assuming that capacity is defined to occur when the probability of breakdown is 0.25, determine the following: (a) What is the capacity of the ramp if the demand from the mainline is 4,500 vph? (b) What is the capacity of the mainline freeway if the metered ramp supplies 1,380 vph? Can you relate the numbers in the graph to the capacities provided in the HCM 2010 method for freeways?
3. Two "critical" freeway-ramp junctions (1 and 2) experience breakdown events independently. Assume that the ramps are metered, but there are no data available regarding ramp metering rates. Assume also that based on freeway data collected at the two segments, the breakdown volume is Weibull distributed with parameters:

   (a) $A1 = 18.36 \quad \beta 1 = 2,269$
   (b) $A2 = 17.85 \quad \beta 2 = 2,105$

Find the expected value $E(q)$ and the standard deviation (σ) of the estimated Weibull breakdown volume. For any time interval i, what is the probability of breakdown at each ramp junction if the mainline volume exceeds 2,000 veh/h/ln?

4. Conduct a literature review on the two-capacity phenomenon. Should our analysis tools provide two different capacity values, one before breakdown and one after?
5. Your State Department of Transportation is interested in estimating what the implications of such a finding are. Assuming that the per lane capacity is 2,300 pcph before the breakdown, and the maximum throughput drop is 10 %, calculate the following: (a) What is the difference in throughput under the two scenarios over an hour for a three-lane facility? (b) If the expected demand of the facility is 2,500 pcph over the peak hour, what is the expected difference in maximum queue length, total delay, and average delay per vehicle at the end of the hour? (c) What is the percent difference in the average delay per vehicle between the two scenarios?

Chapter 5
Traffic Operational Performance Measures

The previous two chapters discussed the fundamental traffic stream parameters. This chapter discusses additional performance measures often used in traffic operational assessment, such as travel time measures, delay measures, and queue measures. Multimodal performance measures as well as measures related to growth management, the environment, and sustainability are discussed briefly. The last section identifies the performance measures used to define the level of service, or LOS, in the HCM. These performance measures are called measures of effectiveness, or MOEs.

Travel Time

Travel time is defined as the time it takes to travel from a given origin to a given destination, and it is the reciprocal of speed. Mathematically, the travel time (TT) for a given distance d is:

$$\mathrm{TT(h)} = \frac{d}{v_{\mathrm{avg}}}$$

where

d is the distance from a given origin to a given destination (miles)
v_{avg} is the average speed throughout the trip (mph)

Figure 5.1 presents a time space diagram with a vehicle trajectory, indicating its travel time from point A to point B. The horizontal distance along any trajectory indicates its travel time.

Travel time is a particularly useful performance measure as travelers can readily relate to it and typically plan their travel around it. Even though travel time can be easily tracked by the traveler, the transportation analyst cannot measure it as easily for large groups of vehicles, because it requires the identification and

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_5,

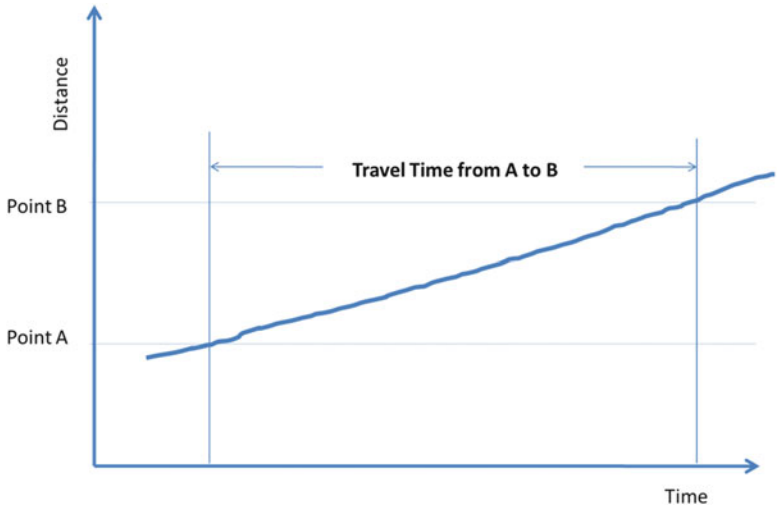


Fig. 5.1 Travel time in a time–space diagram

matching of each specific vehicle at two distinct points in the study network. The problem of measuring travel times is compounded when there are multiple origins and destinations, and the analyst attempts to match multiple vehicles at origins to multiple destinations.

Another difficulty with measuring travel times is when the entire trajectories of the vehicles are not known, and there is a possibility vehicles stop briefly at intermediate destinations. Consider, for example, the graph in Fig. 5.2, which shows the trajectory of a single vehicle. The vehicle briefly stops twice as it travels from point A to point B. Its total travel time includes the time spent stopped. This vehicle trajectory includes three different running times, i.e., times when the vehicle was actually traveling. There are instances when the analyst is interested in both the running time and the stopped time, as when analyzing the performance of a signalized arterial. In that case, the number of stops and the time being stopped are directly related to the performance of the system as well as to emissions estimates. There are other instances when the analyst is only interested in the running times. For example, an analyst may be interested in the performance of a freeway facility, but does not want to include in that travel time any stops at rest areas, as these are not related to the freeway performance.

Travel time is much easier to define and measure when using a simulator or when observing travel times (as well as stopped times and running times) from within a vehicle. Travel time and delay studies along arterial streets often use a vehicle probe to identify and measure the number and location of stops, as well as the travel time along a corridor. In such studies an observer within a vehicle records the details of the trip, including the presence of queues, the stops at traffic signals, and the location and time of midblock delays.

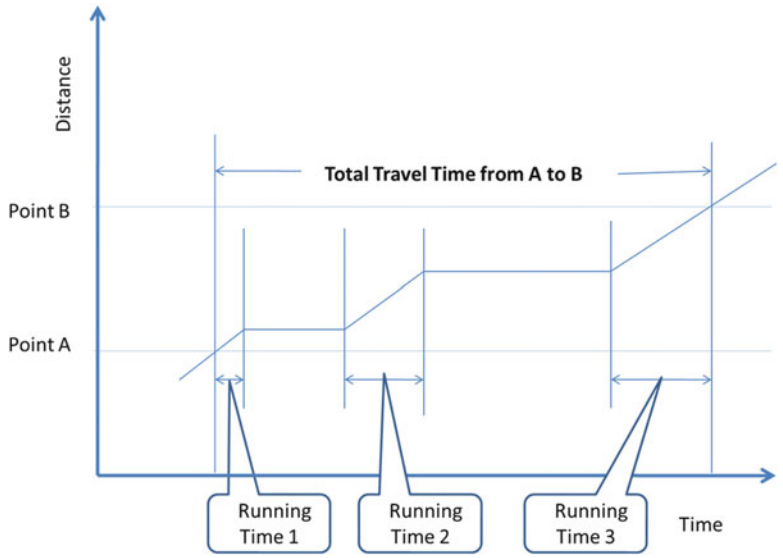


Fig. 5.2 Running time and travel time

Travel Time During Non-congested and Congested Conditions

Figure 5.3 illustrates conceptually the relationship between demand and travel time along a highway link upstream of a bottleneck. The first part of the curve indicates that for low demands, when conditions are non-congested, travel time stays fairly constant. This part of the curve corresponds to the stable portion of the speed–flow curve (presented in Figs. 3.6 and 3.7). At low demands, speed (and its reciprocal travel time) remains fairly flat, with a slight reduction (or travel time increase) as flows approach capacity. The second part of the curve represents congested conditions, during which travel time increases with increasing demand. This part of the curve is the most complex to understand and model, because its shape and value depend on the absolute and relative values of the demand and the capacity, as well as the duration of the analysis period. Furthermore, the demand fluctuates (the capacity may also fluctuate due to special events and work zones, as well as due to the randomness discussed in Chap. 4), and thus, the compounded effect of this fluctuating demand is difficult to estimate. Another related factor in estimating the travel time along the link is the proportion of the link that is congested as a function of time. The longer the demand exceeds capacity, the longer the queue and the higher the travel time. Previous research has suggested estimates of travel time during congested conditions as a function of the relationship between the capacity and the demand, the duration of congested conditions, and the link length [1, 2]. The sketch of Fig. 5.3 only provides one potential shape of the relationship, assuming steadily increasing demand and constant capacity along a highway link.

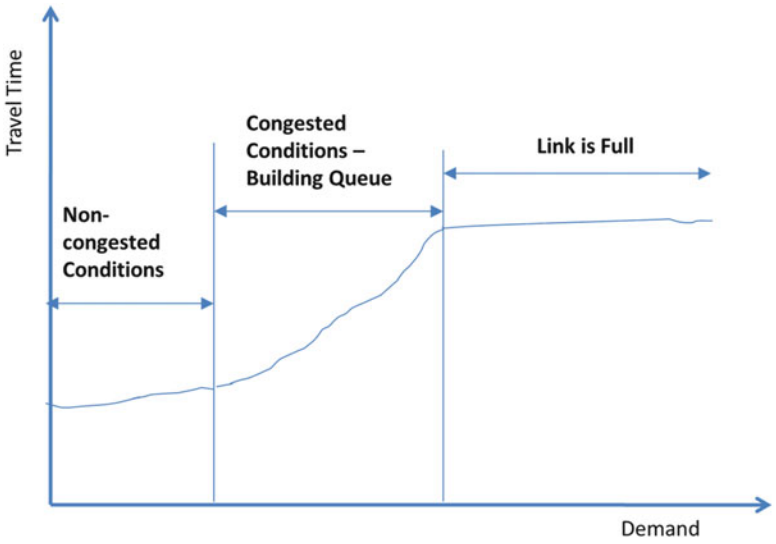


Fig. 5.3 Travel time as a function of demand

The third part of the curve represents conditions when the link is full, and thus, the travel time along the link is fairly constant. Under those conditions, the queue may be increasing in the upstream link, and the overall travel time may be increasing, but the travel time in our study link will be constant as long as the downstream capacity remains unchanged.

The Distribution of Travel Time and Travel Time Reliability

The distribution of travel times over a long period of time (say a year) is relevant when the analyst is interested in assessing the performance of a facility and how various policies and strategies may affect it. The travel time distribution allows us to examine operations beyond those of the typical peak hour and consider the entire distribution of demands throughout the year. This approach takes into account the utilization of the facility over a longer period of time rather than its performance during a single design hour. It also allows agencies to evaluate the robustness of existing or planned traffic management policies (such as those related to incident removal, work zones, and special events).

Figure 5.4 shows a sketch of a typical travel time distribution for a freeway segment. Such a distribution provides the number of intervals that have exhibited a particular travel time over a long period of time, often 1 year. The left side of the distribution corresponds to lower travel times (i.e., higher speeds) and non-congested conditions, while the right side corresponds to intervals with longer travel times (and lower speeds). The exact shape of the distribution varies

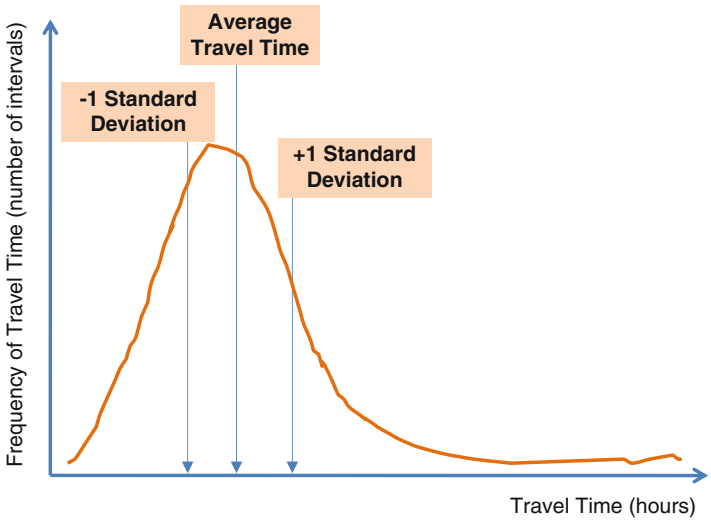


Fig. 5.4 A sample travel time distribution

depending on the time interval(s) evaluated. For example, if the travel time distribution is created for the evening peak hours throughout a year, the frequency of the congested intervals will be relatively high, and the distribution will be flatter; if the travel time distribution is created for all hours of the day throughout the year, the distribution will have a sharper peak, as the variability in travel times will be much greater. When such a distribution is obtained, one can also obtain the values that correspond to, say, one standard deviation on either side of the average travel time. It is often these values (or some such set of parameters) that are reported as the range of expected travel time.

The concept of reliability is relatively new in traffic engineering, but it is becoming increasingly important as part of the engineering design process. There is one definition of reliability generally accepted in various areas of engineering and manufacturing: reliability is "the probability that a component or system will perform a required function for a given period of time when used under stated operating conditions. It is the probability of non-failure over time" [3]. In the context of travel time, reliability can be defined as the probability that travel time does not exceed a previously defined threshold. Figure 5.5 illustrates a distribution of travel time and its reliability. In this example, travel time reliability is 87 %; in other words, 87 % of the observations had travel time less than the predefined threshold.

Travel time reliability has been used extensively to evaluate other transportation modes. For example, US DOT evaluates the reliability of travel for airlines using the percentage of on-time performance. A flight is considered to be "on time" if it is not delayed more than 15 min beyond the scheduled arrival (http://www.transtats.bts.gov/HomeDrillChart.asp). In the Transit Capacity and Quality of

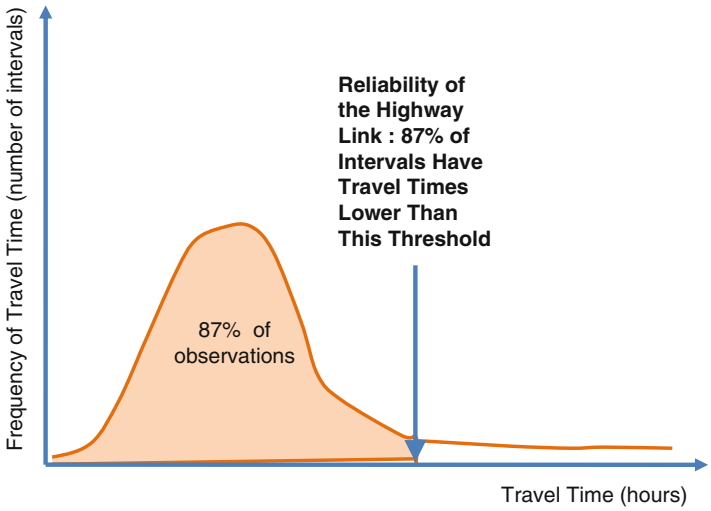


Fig. 5.5 Example of a travel time reliability estimate

Service Manual (TCQSM), "on-time" performance is defined to occur when a trip is less than 5 min late relative to the scheduled arrival time. The same concept has become more significant in transportation operations, primarily through the Strategic Highway Research Program 2 (SHRP2), which identified travel time reliability as one of the key performance measures highway agencies should monitor. An example where reliability is used to assess freeway performance is the I-95 Express in Miami, FL, which provides the reliability of its high-occupancy toll (HOT) lanes as the percent of time they operate above 45 mph [4].

Despite the fact that the definition of reliability is widely accepted in engineering, in the area of highway transportation, this is not the case. There are several performance measures that have been developed which in essence measure the variability of travel times. These are also referred to as "travel time reliability," and in those cases, the terms variability and reliability are used interchangeably.

The definitions that relate to travel time variability focus on the traveler-perceived unpredictability of travel times. Such measures include the Travel Time Index and the Buffer Time Index. The Travel Time Index is defined as follows:

$$\text{Travel Time Index} = \frac{\text{Mean travel time during peak hours}}{\text{Mean travel time during free-flow conditions}}$$

The Buffer Time Index is defined as follows:

$$\text{Buffer Time Index} = \frac{95\,\%\ \text{travel rate-average travel rate}}{\text{Average travel rate}} \times 100\,\%$$

In using these performance measures, the analyst should have a clear understanding of their advantages and limitations. For example, the traditional definition

of reliability (as shown in Fig. 5.5) is very useful for transportation agencies that desire to monitor their transportation facilities. Freeway management agencies may use it to determine the percent of time a particular facility operates below a particular operating speed. On the other hand, such a measure may not be as useful to highway travelers, as they most likely would desire a quantitative estimate of travel time, and most probably a range of travel times so that they can plan their upcoming trip (as shown in Fig. 5.4).

The Travel Time Index (TTI) provides a good measure for evaluating the deviation of a particular analysis interval from "ideal conditions." The Buffer Time Index (BTI) focuses purely on the travel time variability, and it can be a useful measure for travelers. However, it should not be used to track the overall operational performance of a facility, as lower variability would result both when the facility is severely congested and when it is completely non-congested.

Travel Time for Traveler Information Purposes

Since travel time is a performance measure that users of the transportation system can easily understand, providing estimated travel times can be very useful for travelers planning their route and selecting a preferred mode of transportation. The question is, how should that travel time be estimated, and how should it be reported to travelers? One approach is to measure the travel times of vehicles traversing a particular section and report the average of that travel time. The issue with this approach is that the time reported is for the previous time interval $(n - 1)$, and it applies for the vehicles that have already completed the route; the travel time during the next time interval (n) will not necessarily be the same or even similar to that of the $n - 1$ interval. The magnitude of the difference between the two travel times depends on the length of the study section, the duration of the time interval studied, and the probability of incident occurrence and congestion along the route. A second approach is to develop travel time estimates by time of day for specific links and report travel time based on historical data. Such an approach could use a travel time distribution such as the one shown in Fig. 5.4 and report the range of travel times within one standard deviation.

Several methods have been reported in the literature on predicting and reporting travel time for traveler information systems (e.g., see [5]). The most successful of these systems are based on a combination of the two approaches (real-time information from the previous interval plus historical trends). Such methods rely on the distribution of travel time during a particular interval and report the expected travel time considering a given margin of error. For example, travel time may be reported to be between 10 and 13 min for a particular origin–destination.

Recently, private entities have developed tools for obtaining and reporting travel time along specific segments. Companies such as TomTom (http://www.tomtom.com), Inrix (http://www.inrix.com), and Navteq (http://www.navteq.com) provide traffic data as well as navigation considering real-time traffic. Such tools can provide

traveler information for specific routes; however, the algorithms behind those engines are not publicly available, and it is not immediately clear what assumptions are made in providing these travel time estimates .

Delay

Delay is defined as the excess travel time it takes to traverse a particular point or segment. Figure 5.6 graphically illustrates the concept of delay. As shown, delay is defined relative to a known, ideal travel time. Delay is not easy to measure at the traffic stream level, as it requires information related to the trajectory of each individual vehicle.

There are several different types of delay:

Stopped delay is the delay that a vehicle experiences while completely stopped or when its speed is below a pre-specified threshold.

Travel time delay is the difference between the actual travel time and an "ideal" travel time, typically related to the free-flow speed or the speed limit of the facility.

Control delay is the delay experienced due to the presence of traffic control devices such as signals and STOP signs. In essence it represents the time spent in queue plus the delay due to acceleration and deceleration.

Control delay is the performance measure used to evaluate intersections (signalized, unsignalized, and roundabouts) in the Highway Capacity Manual [6].

These definitions of delay are depicted in Fig. 5.7.

Delay is typically measured or estimated in units of seconds/vehicle. Aggregate delay may also be obtained in units of vehicle seconds, vehicle minutes, or vehicle hours. Practically, delay is difficult to measure in the field. Existing methods to approximate it are based on queue measurement throughout the cycle [7]. Chapter 9 of this book provides a derivation of Webster's delay equation, which is related to the uniform delay portion of control delay in the HCM .

Queue Length

Queue length is typically defined as the number of vehicles waiting to be served. In traffic operations, queue length is an important performance measure, as it can significantly affect the quality of traffic flow. For example, queue spillback from a left turn lane blocks through traffic. Similarly, queue spillback at an interchange ramp terminal affects traffic operations along the mainline freeway.

Queuing theory studies various types of queuing systems (including transportation but also manufacturing, industrial, banking, and other systems), and attempts to provide answers to questions such as how many vehicles are expected to be in queue and how long will a traveler wait for service. Queuing analysis is discussed in more detail in Chap. 6 of this book.

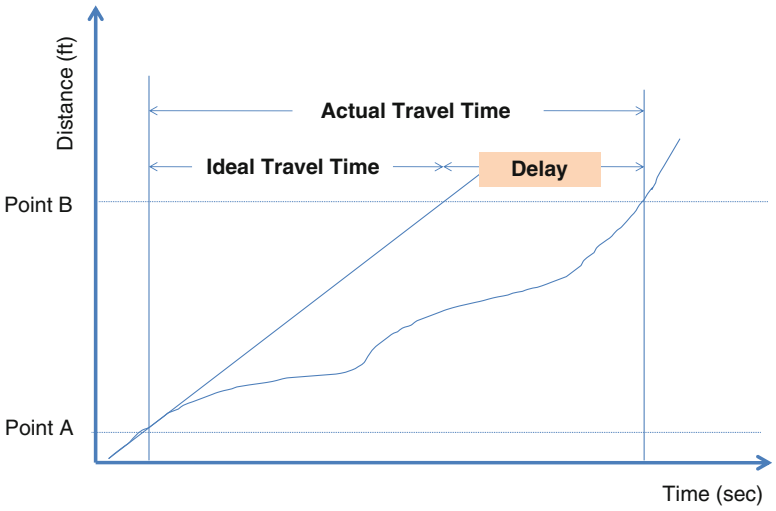


Fig. 5.6 Graphical illustration of delay

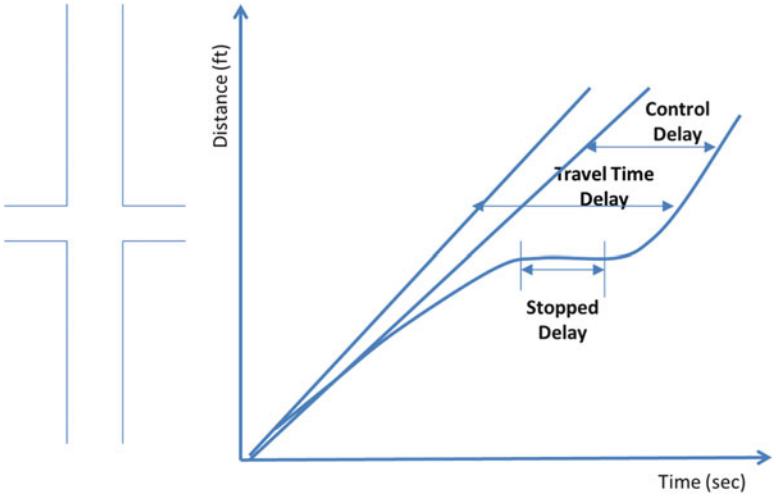


Fig. 5.7 Definitions of delay

Queue length is used as a performance measure when spillback is important, such as in the case of left- and right-turning pockets or at metered ramps. Queue length is more difficult to define and measure for uninterrupted flow facilities (such as freeways and two-lane highways), and we mostly use it when analyzing interrupted flow facilities.

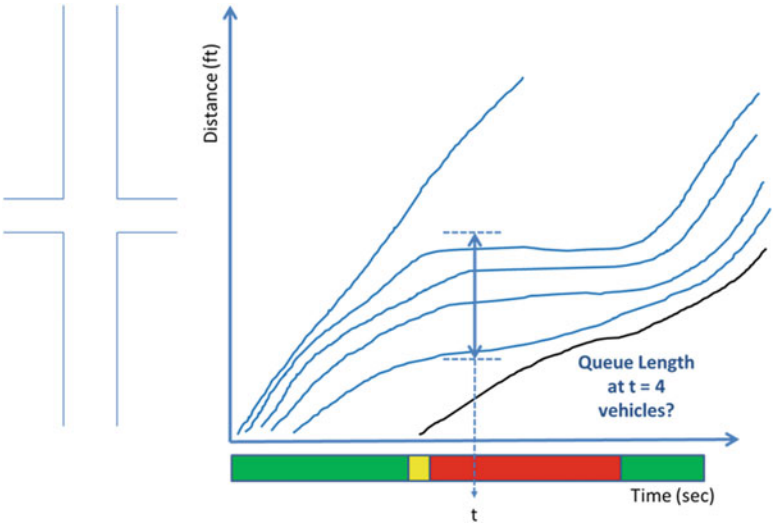


Fig. 5.8 Queuing at a signalized intersection approach

There are several measures that can be used to indicate the extent of queuing:

- Percentile queue: the queue length expected to be present for a selected percentile of time (e.g., 85 % queue, 95 % queue, etc.)
- Average queue: the expected value of the queue length
- Queue storage ratio: the expected value or a selected percentile of the queue divided by the available storage

Queue length measurement in the field can be tricky. Figure 5.8 illustrates the trajectories of a group of vehicles, which decelerate while approaching a red signal and then accelerate when the signal turns green. One can simply say that the number of vehicles queued at time t is determined by counting the number of vehicles stopped at the intersection at a particular time, t. However, an important complication is the definition of a "queued" vehicle. When should we consider that the vehicle is part of a queue? Should it be when its speed is 0 mph, or should it be when its speed is lower than a given threshold? If it is the latter, which threshold should we use? 5 mph? 10 mph? In the example of Fig. 5.8, the first three vehicles are either stopped or have very low speeds at time t. The fourth vehicle appears to be traveling at a higher speed. Thus, depending on our speed threshold, this fourth vehicle may or may not be considered as part of the queue. Similarly, the fifth vehicle appears to be affected by the presence of the queued vehicles, but depending on our definition of a queued vehicle, it may or may not be part of the maximum queue at the intersection approach. The same difficulty in the measurement of a queue can be encountered in the context of traffic simulation models (additional discussion on this issue is provided in Chap. 7).

The issue becomes even more complex when considering the slowly moving queues that typically develop along freeways and highways with access control. In those cases, queues are even more difficult to define and measure, as traffic may operate in "stop-and-go" mode over significant distances.

Another issue with respect to queue length estimation is the consideration of vehicle length. Queuing analysis models typically consider each unit of traffic as a point. However, vehicular queues include the length of the vehicle as well as a safety distance between successive vehicles. Thus, the overall queue length in distance units should consider the length of the vehicles typically present as well as the expected distance between vehicles.

Other Mobility-Related Performance Measures

In addition to the performance measures discussed so far, there are several other measures that are related to mobility. Some of the most commonly used ones are:

v/*c*: It is the ratio of the demand-to-capacity, and it is a measure of capacity utilization. It is used to evaluate signalized intersection operations. It can also be used to distinguish between non-congested and congested conditions. The HCM [6] uses it to define capacity and the upper boundary of LOS E.

Number of stops: This measure is typically used to evaluate the quality of progression along an arterial. It is often expressed as a stop rate, i.e., the number of stops per number of vehicles served.

VMT: It is the total vehicle miles traveled, and it is used as a measure of facility utilization. It can also be used in growth management to estimate the additional VMT to be generated by a new development.

VHT: It is the total vehicle hours traveled, and it incorporates delay considerations.

Person-capacity: It is a measure used in the evaluation of transit facilities; it is a function of the transit vehicle's capacity, the agency policy regarding loading standards, and the frequency of transit vehicles over an hour.

Other performance measures related to mobility include measures related to infrastructure (e.g., lane-mile- and land-use-related measures), accessibility (e.g., availability of facilities and ease of access), and sustainability (e.g., emissions). For a detailed discussion of such performance measures, consult [8, 9]. The selection and use of performance measures is very important in the evaluation of transportation systems, as different measures present different perspectives of the bigger picture. The practitioner should select an appropriate set of performance measures as the basis of analysis as a function of the specific study objectives.

Measures of Effectiveness and Level of Service

The HCM uses specific performance measures to define the LOS of each type of facility. According to the HCM [6], the LOS is ". . .a quantitative stratification of a performance measure or measures that represent quality of service. It facilitates the

Table 5.1 MOEs in the HCM 2010: uninterrupted flow facilities

Type of facility		MOEs	Comments
Basic freeway segments		Density (pc/mi/ln)	LOS F occurs if density >45; assumes that this density corresponds to capacity
Freeway weaving segments		Density (pc/mi/ln)	LOS F occurs if density >35; assumes that this density corresponds to capacity
Freeway merge and diverge segments		Density (pc/mi/ln)	LOS F occurs if density >35; assumes that this density corresponds to capacity
Two-lane highways	Class I highways	Percent time spent following (PTSF) (%) and average travel speed (mi/h)	LOS boundaries vary by highway class; it is assumed that the upper boundary of LOS E represents capacity
	Class II highways	Percent time spent following (PTSF) (%)	
	Class III highways	Percent of free-flow speed (PFFS) (%)	
Multilane highways		Density (pc/mi/ln)	LOS boundaries vary with free-flow speed (FFS); assumes density at LOS F corresponds to capacity; density at capacity varies from 40 to 45 as a function of FFS
Bicycle mode for two-lane highways and multilane highways		Bicycle LOS score	This score is a traveler perception index based on five variables including vehicular demands and pavement condition

presentation of results through the use of an A (best) to F (worst) scale. It is defined by one or more service measures that both reflect the traveler perspective and are useful to operating agencies." The LOS concept was developed in order to communicate to non-transportation professionals the quality of traffic operations.

The performance measures which are used in the LOS designation are called service measures or measures of effectiveness (MOEs). Table 5.1 lists the MOEs used in the HCM 2010 for uninterrupted flow facilities, while Table 5.2 lists the MOEs used for interrupted flow facilities. As shown in Table 5.1, density is the primary MOE for freeways and highways. Even though speed and travel time are of primary concern to drivers, they do not vary much along the range of possible volumes for non-congested conditions. Therefore, they cannot be used to define different LOS, and density is used instead. For two-lane highways, there are different classes (based on whether their primary function is access or mobility), and each class uses different MOEs and LOS criteria. For uninterrupted flow facilities, it is assumed that the upper boundary of LOS E coincides with capacity, which is not necessarily the case, as discussed in Chap. 4.

Table 5.2 MOEs in the HCM 2010: interrupted flow facilities

Types of facilities		MOEs	Comments
Urban street facilities	Auto mode	Travel speed as percentage of base free-flow speed (%) and *v/c*	*v/c* considered is the largest of those for through movements at each intersection along the facility
	Pedestrians	Pedestrian LOS score and avg. pedestrian space (ft^2/p)	LOS score is length-weighted average of LOS scores of each segment along the facility
	Bicycles and transit	LOS score based on traveler perception assessment	
Urban street segments	Auto mode	Travel speed for through vehicles as percentage of base free-flow speed (%) and *v/c*	*v/c* is for through movement at downstream boundary intersection
	Pedestrian mode	Pedestrian LOS score and avg. pedestrian space (ft^2/p)	LOS score based on LOS scores for the link and the intersection
	Bicycles and transit	LOS score based on traveler perception assessment	
Signalized intersections	Auto mode	Control delay (s/veh) and *v/c*	LOS can be computed for each lane group, each intersection approach, and for the entire intersection; for approaches and intersection-wide assessment, LOS is defined solely by control delay
	Pedestrian and bicycle modes	LOS score based on traveler perception assessment	LOS score for pedestrians estimated for each crosswalk and intersection corner; LOS score for bicycles is estimated for each approach
Interchange ramp terminals		Control delay (s/veh), *v/c* ratio, queue-to-storage ratio, R_Q	LOS is estimated for each origin–destination (O–D); LOS F occurs when either the *v/c* or queue-to-storage ratio for any of the lane groups that contain this O–D exceeds 1.0

(continued)

Table 5.2 (continued)

Types of facilities		MOEs	Comments
Two-way stop-controlled intersections (TWSC)	Auto mode	Control delay (s/veh) and *v/c*	LOS is determined for each minor-street movement, not for the major-street approaches or the overall intersection; LOS F occurs if *v/c* exceeds 1.0, regardless of the control delay
	Pedestrian mode	Control delay (s/pedestrian)	LOS is defined for pedestrians crossing a traffic stream not controlled by a STOP sign and for midblock pedestrian crossings
All-way stop-controlled intersections (AWSC)		Control delay (s/veh) and *v/c*	LOS F occurs if *v/c* exceeds 1.0, regardless of the control delay; for approaches and intersection-wide assessment, LOS is defined solely by control delay
Roundabouts		Control delay (s/veh)	LOS F occurs if *v/c* exceeds 1.0, regardless of the control delay; for approaches and intersection-wide assessment, LOS is defined solely by control delay
Off-street pedestrian and bicycle facilities	Exclusive walkways/ stairways	Space(ft^2/p)	Space is the primary MOE; related measures include flow rate (p/min/ft), average speed (ft/s), *v/c*
	Pedestrians on shared-use paths	Weighted event rate/h	An event is a bicycle meeting or passing a pedestrian; related measure is bicycle service volume per direction (bicycles/h)
	Exclusive and shared bicycle facilities	Bicycle LOS score	The score incorporates meetings per minute, active passing per minute, presence of a center line, path width, and delayed passing

Regarding interrupted flow facilities, the LOS for urban facilities and segments is based on travel speed as a percent of base free-flow speed, while the LOS for intersections (signalized, unsignalized, roundabouts) is primarily based on control delay. For interrupted flow facilities, there is the recognition that capacity does not coincide with a particular value of speed or control delay, and thus, the v/c is used to define LOS F.

Generally, the concept of LOS in the HCM has been intended to describe how well a transportation facility operates from a traveler's perspective. However, not much research was undertaken to relate traveler perception to LOS boundaries when these were originally developed. More recently, such research has shown that travelers consider a much broader set of criteria, including facility esthetics, safety, comfort, convenience, etc. [10]. Some of these measures considered by travelers are not under the control of traffic engineers, and they are difficult to quantify, thus are not included as part of the HCM analyses. However, the most recent HCM [6] considers these traveler perceptions to establish LOS scores for bicycles, pedestrians, and transit. The interrupted flow facilities methods consider bicycles, pedestrians, and transit more extensively, and their LOS is considered separately using various LOS scores, which are determined using a variety of inputs based on traveler perception.

Even though the LOS structure has been useful in communicating quality of service aspects to the public and decision-makers, it has also created some challenges. First, quality is subjective, and thus, it is not possible to provide broad recommendations on a "desirable" LOS. For example, a delay of 60 s/veh in an urban area would seem trivial, while in a rural area, it would be considered significant. Therefore, transportation agencies need to develop their own rules and guidelines on what is the target LOS for the design of various facilities considering the costs and benefits of alternative policies. Second, the LOS structure creates discontinuities around the boundaries between two levels of service. For example, a very small increase in delay may result in a worse LOS, when that value is around an LOS boundary; however, a larger change in delay may not result in an LOS change. Lastly, the LOS framework was developed before congestion became a frequent occurrence, and thus, its stratification focuses on non-congested conditions. When demand exceeds capacity, the LOS is F, and there is no distinction between different levels of congestion. However, in today's environment, it is important to be able to assess how severe congestion is. To accomplish that, one needs to obtain performance measures such as delay, travel time, and queue length, rather than the prevailing LOS.

References

1. Akçelik R (1991) Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function. Aust Road Res 21(3):49–59
2. Chrysikopoulos G (2010) Freeway travel time estimation as a function of demand. Master's thesis, University of Florida, Gainesville, FL

3. Ebeling CE (1997) Introduction to reliability and maintainability engineering. McGraw-Hill Companies Inc., New York, NY
4. http://www.sunguide.org/index.php/tmc_reports/. Accessed 30 Apr 2012
5. Kothuri SM, Tufte KA, Ahn S, Bertini RL (2007) Using archived ITS data to generate improved travel time estimates. In: Proceedings of the 86th annual meeting of the Transportation Research Board, Washington, DC, 21–25 January 2007
6. Transportation Research Board, National Academies (2010) Highway Capacity Manual, National Research Council, Washington, DC
7. Transportation Research Board, National Academies (2000) Highway Capacity Manual (Chapter 16, Appendix A). Washington, DC
8. Cambridge Systematic, Inc., Dowling Associates, Inc., System Metrics Group, Inc., and Texas Transportation Institute (2008) Cost-effective performance measures for travel time delay, variation, and reliability. NCHRP Report 618. Transportation Research Board, Washington, DC
9. Zietsman J, Ramani T (2011) Sustainability performance measures for state DOTs and other transportation agencies, NCHRP Project no. 08-74
10. Dowling RG, Reinke DB, Flannery A, Ryus P, Vandehey M, Petrisch TA, Landis BW, Rouphail NM, Bonneson JA (2008) NCHRP Report 616: multimodal level of service analysis for urban streets. Transportation Research Board of the National Academies, Washington, DC

Problems

1. Use the HCM 2010 speed–flow–density curves to estimate travel time for non-congested conditions. Produce a graph showing travel time as a function of demand for a freeway segment.
2. Use Akcelik's formula [1] to predict travel time and plot travel time versus demand.
3. Conduct a literature review to identify previous studies that estimate travel time for freeways and arterials.
4. A set of travel time data have been collected for a year during the evening peak (4–6 pm) along a freeway route and are provided below. The route length is 3 miles, and the speed limit along this facility is 65 mph. What is the reliability of the facility if on-time arrivals occur for speeds of at least 50 mph? What is the Travel Time Index, and what is the Buffer Time Index? What can you conclude about the operations of this facility during the evening peak hour?

Travel time (s)	Frequency
$\leq$150	2
(150–180)	139
(180–210)	226
(210–240)	167
(240–270)	93
(270–300)	65
(300–330)	25
(330–360)	8
>360	5

Part III
Traffic Operational Analysis Techniques

This part presents and discusses various analysis techniques available for assessing the traffic operational quality.

Chapter 6 discusses mathematical models (shockwave analysis and queuing analysis) as well as empirical models (HCM-type techniques).

Chapter 7 presents traffic simulation modeling principles, discusses simulation programming and model building, and presents some of the commercially available simulation models.

Chapter 8 provides an overview of other techniques used in traffic operational analysis, including time series analysis, fuzzy logic, and genetic algorithms.

Chapter 6
Mathematical and Empirical Models

Traffic models are very useful for various purposes. First, they can help in the design and operations of traffic systems since they can predict traffic operational conditions at some time in the future under various sets of design, traffic, and control characteristics. Traffic engineers and designers can make decisions regarding facility modifications or traffic management improvements based on the expected impact of those improvements in the transportation system. Second, they can help in the evaluation of existing systems and in the development of priorities for improvement. Mathematical models are those that describe a physical system mathematically. Such models describe specific relationships. For example, Flow = Speed × Density is a mathematical model. Empirical models are those based on field observations (empirical observations) rather than on relationships that can be mathematically described. Empirical models predict how a system behaves rather than explaining how its components interact. Empirical models can be very useful when the mathematical relationship is unknown or very difficult to express. Examples of empirical models are the traffic stream relationships discussed in Chap. 3.

This chapter provides an overview of mathematical and empirical models currently used in traffic analysis. The first part of this chapter discusses shockwave analysis, while the second part of the chapter presents the fundamentals of deterministic queuing analysis.

Shockwave Analysis

In previous chapters, we studied trajectory plots, and we analyzed the movement of individual vehicles and small groups of vehicles. Let us now consider plots of trajectories for a much longer time period (say, the entire peak hour) to observe and analyze traffic movement in a broader context. Figure 6.1 illustrates a plot of trajectories through a signalized approach as the queue builds up and then dissipates.

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_6,

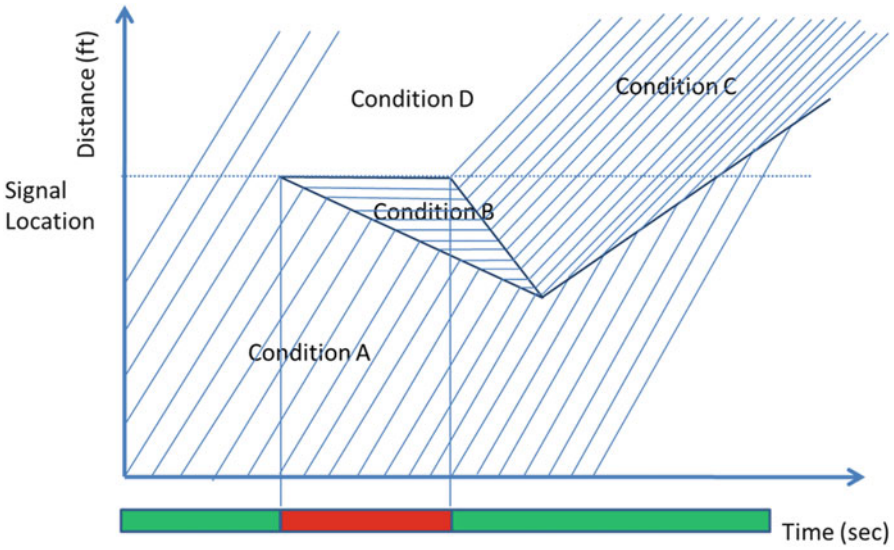


Fig. 6.1 Vehicle trajectories through a signalized approach

As shown, there are several changes in conditions (i.e., discontinuities) observed as vehicles approach the signal, stop at the red, and then accelerate to depart. There is a discontinuity in the set of trajectories as vehicles approach and join the end of the queue (from condition A to condition B); there is another discontinuity as vehicles depart the queue and travel through the bottleneck (from condition B to condition C). These discontinuities are changes in density and flow which travel through traffic at a certain speed as vehicles travel through the intersection. These discontinuities are called shockwaves, and they are defined as a propagation of a change in flow and density [1]. Their observation, measurement, and estimation can be very useful in traffic analysis as they can describe traffic operations in space and time. By observing the shockwaves created and their respective speeds, we can determine what the impacts of various disturbances are both in time and in space.

We can develop equations for estimating the speeds of the shockwaves using equations of motion. The total number of vehicles that cross the boundary between two adjacent conditions (say A and B in Fig. 6.1) in time t is:

$$N = (v_A - v_w)D_A t = (v_B - v_w)D_B t \tag{6.1}$$

where

N is the total number of vehicles
v_A is the speed of vehicles in region A
v_w is the speed of the shockwave
D_A is the density in region A
v_B is the speed of vehicles in region B
D_B is the density in region B

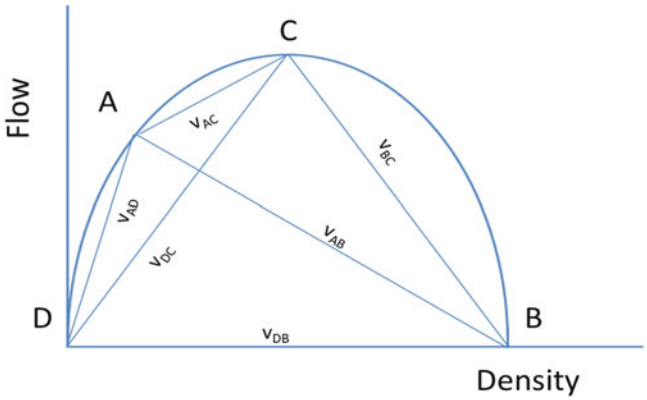


Fig. 6.2 Shockwaves through a signalized approach

Since the flow at each of the two regions is $F_A = D_A\ v_A$ and $F_B = D_B\ v_B$, Eq. (6.1) can be written in the following form:

$$\begin{aligned}(v_A - v_w)D_A &= (v_B - v_w)D_B \Rightarrow \\ v_A D_A - v_w D_A &= v_B D_B - v_w D_B \Rightarrow \\ F_A - v_w D_A &= F_B - v_w D_B \Rightarrow \\ v_w &= (F_A - F_B)/(D_A - D_B)\end{aligned} \tag{6.2}$$

Equation (6.2) provides the speed of the shockwave, i.e., the speed with which disturbances are propagated in the traffic stream. Shockwave speed is a function of the flow and density differences in regions A and B.

Since the discontinuity and its speed are expressed in terms of flow and density, the shockwaves can also be plotted in a flow–density diagram. For the example of Fig. 6.1, there are a total of six shockwaves. Figure 6.2 shows these shockwaves in a flow–density diagram. Each of the conditions of Fig. 6.1 is shown along the flow–density diagram. The slope of the line connecting two conditions is the ratio of their differences in the flow over the respective differences in density, and thus, it represents the shockwave speed, as indicated in Eq. (6.2).

There are several types of shockwaves that are distinguished based on their direction as well as based on whether they result in queue formation or dissipation. For example, the shockwave between conditions A and B is a backward forming wave, while the shockwave between conditions B and C is a backward recovery wave [2].

Example 6.1 Estimate the speed of all shockwaves in Fig. 6.1; assuming that vehicles approach the signal to join the queue at a speed of 45 mph and at a flow of 540 veh/h (Condition A), the jam density is 130 veh/mi/ln (Condition B), the free-flow speed is 70 mph, and the speed–density relationship is linear.

Solution to Example 6.1
The speeds of the following shockwaves will be estimated: v_{AD}, v_{AB}, v_{BD}, v_{DC}, v_{BC}, v_{AC}.

The speed of shockwave v_{AD} is:

$$V_{AD} = (F_A - F_D)/(D_A - D_D)$$

Since the flow and density at condition D are both zero, the equation becomes:

$$V_{AD} = F_A/D_A = V_A = 45\,\text{mph}$$

The speed of shockwave v_{AB} is:

$$V_{AB} = (F_A - F_B)/(D_A - D_B)$$

Since flow at condition B is zero, the equation becomes:

$$V_{AB} = F_A/(D_A - D_B)$$

Since $D_A = F_A/v_A = 540/45 = 12$ veh/mi/ln:

$$V_{AB} = F_A/(D_A - D_B) = 540/(12 - 130) = -4.6\,\text{mph}$$

The negative sign indicates that the shockwave is moving backward relative to the flow of traffic (backward moving shockwave).

The speed of shockwave v_{BD} is:

$$V_{BD} = (F_B - F_D)/(D_B - D_D)$$

Since flows for both conditions B and D are zero, the speed of the shockwave is zero, i.e., the shockwave is stationary:

$$V_{BD} = (F_B - F_D)/(D_B - D_D) = 0$$

The speed of shockwave v_{DC} is:

$$v_{DC} = (F_D - F_C)/(D_D - D_C) = F_C/D_C = v_C$$

Since condition C represents capacity (i.e., optimal speed and density), and since we have assumed a linear relationship between speed and density, the speed at condition C is half the free-flow speed (v_{FF}). Therefore,

$$v_{DC} = v_C = v_{FF}/2 = 70/2 = 35\,\text{mph}$$

The speed of shockwave v_{BC} is:

$$V_{BC} = (F_B - F_C)/(D_B - D_C) = -F_C/(D_B - D_C)$$

As indicated earlier, the negative sign signifies that the shockwave is backward moving (backward recovery, in this case). The flow at condition C is:

$$F_C = v_C \times D_C = v_C \times (D_J/2) = 35 \times 130/2 = 2,275\,\text{veh/h/ln}$$

Therefore,

$$V_{BC} = -F_C/(D_B - D_C) = -2,275/(130 - 65) = -35\,\text{mph}$$

The speed of shockwave v_{AC} is:

$$V_{AC} = (F_A - F_C)/(D_A - D_C) = (540 - 2,275)/(12 - 65) = 32.7\,\text{mph}$$

Cumulative Curves and Queuing Analysis

Graphically one can consider the cumulative (or total) number of vehicles that have arrived and the total number of vehicles that have departed, and the difference between the two represents the queue accumulation at each time. An example of such a graph is provided in Fig. 6.3. The horizontal axis of the graph provides the analysis time, while the vertical axis provides the cumulative number of vehicles that have traversed a particular point up to the respective time. The arrivals curve shown indicates the cumulative number of arrivals, while the departures curve indicates the cumulative number of departures. At the beginning of the analysis period, the same number of vehicles arrives and departs. When congestion starts, the number of vehicles departing is lower and the two curves deviate. The horizontal distance between the two curves indicates the delay that vehicle i experiences between its arrival and its departure. The vertical distance indicates the number of vehicles queued at a particular time t. Toward the end of the analysis period, the two curves meet again, indicating the queue has dissipated. Geometrically, the area between the two curves is the total delay experienced by the vehicles in queue.

Figure 6.4 provides the cumulative arrival and departure curves at a signalized approach. When the signal is red, vehicles arrive (blue line) at a rate λ, but are stopped, and there are no departures. The vertical difference between arrivals and departures represents the queue length at time t. When the signal turns green, vehicles discharge at their maximum rate (saturation flow rate). The queue clears some time during the green interval and then the departures are equal to the arrivals.

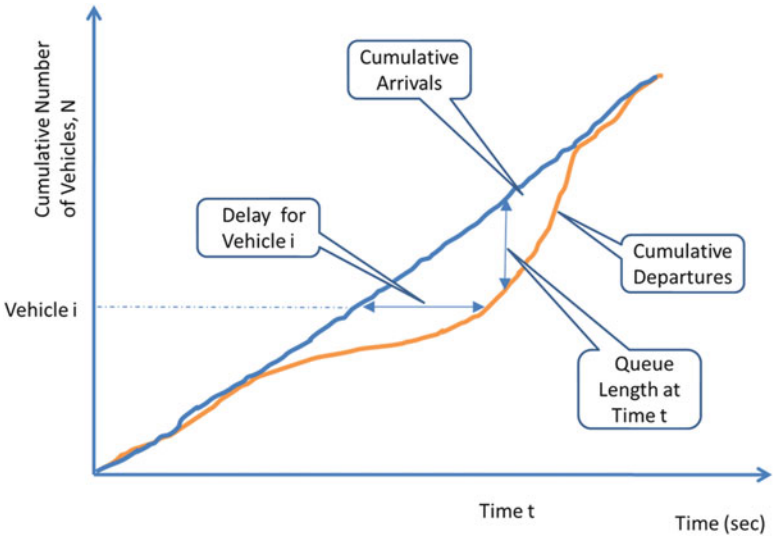


Fig. 6.3 Cumulative curves, delay, and queue

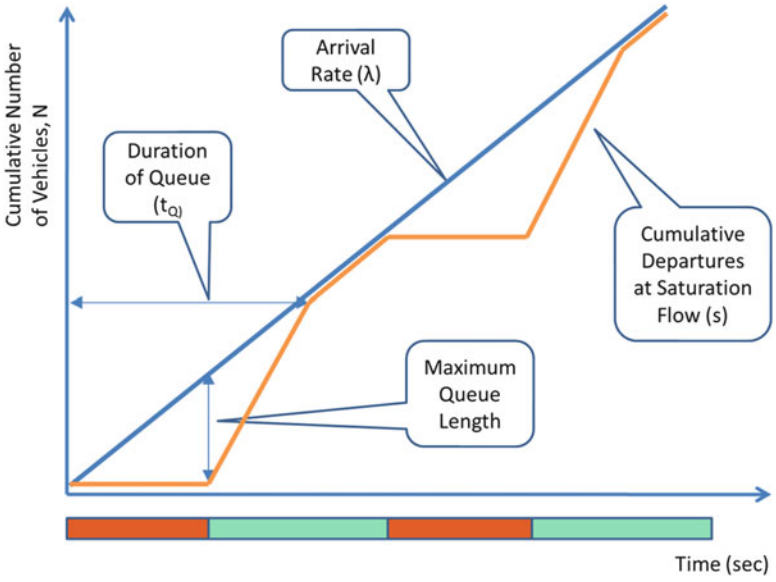


Fig. 6.4 Cumulative curves, delay, and queue at a signalized approach

According to the graph, the maximum queue occurs at the end of the red interval (with duration r), and it can be estimated as [2]:

$$Q_{\max} = \lambda r$$

The queue is present for time t_Q, which can be estimated as follows:

$$\begin{aligned} \lambda t_Q &= s(t_Q - r) \Rightarrow \\ \lambda t_Q &= s t_Q - sr \Rightarrow \\ t_Q(s - \lambda) &= sr \Rightarrow \\ t_Q &= sr/(s - \lambda) \end{aligned}$$

The total number of vehicles queued, N_Q are estimated as:

$$N_Q = \lambda t_Q$$

The average queue while the queue is present is:

$$Q_{PA} = Q_{\max}/2$$

The average queue throughout the cycle is:

$$Q_{PA} = Q_{PA}(t_Q/C) = Q_{\max} t_Q/2C$$

where C is the cycle length.

References

1. Gerlough DL, Huber MJ (1975) Traffic flow theory: a monograph. In: Transportation Research Board Special Report 165, National Academies, Washington, DC
2. May AD (1990) Traffic flow fundamentals. Prentice Hall, Englewood Cliffs, NJ

Problem

1. Conduct a literature review and discuss comparisons between shockwave analysis and queuing analysis. Do the two analysis methods provide consistent results? What are differences between the two methods?

Chapter 7
Simulation Modeling

Simulation is generally defined as an imitation of a system or process, while computer simulation is the replication of a system or a process on a computer. Simulation has been used in many fields in order to understand interactions between system components or evaluate alternative designs. It is routinely used in various and very diverse environments, including the training of pilots using flight simulators, in weather prediction, in the design of communications networks, as well as in entertainment (e.g., video games).

In transportation, simulation is used to study various aspects of the system, including port, airport, and rail operations, demand modeling, interactions between land use and transportation, and traffic operations. The use of computer simulation models has become particularly prevalent among transportation practitioners and researchers. Such models typically replicate the movement of units of traffic (automobiles, buses, pedestrians, etc.) along a simulated network, considering the interactions between the environment, the vehicle, and the driver. Simulation can be very helpful in evaluating alternative solutions for transportation systems where analytical techniques cannot be applied or are not available, and it can consider the effects of microscopic characteristics such as individual driver behavior and vehicle characteristics.

There are many different types of simulators, depending on their scope, scale, and approach. As indicated in the Introduction to Part III, traffic simulators can be categorized as being micro-, meso-, or macroscopic, depending on the level at which traffic flow is being represented. Microscopic models imitate the movement of every vehicle by taking into account attributes such as its speed and acceleration as a function of surrounding vehicles and highway environment [1]. Microscopic models are developed based on car-following, lane-changing, and gap acceptance theories [2]. Typically, vehicles enter a transportation network assuming an arrival distribution, and their path is tracked through the network every time step (e.g., every second). Based on these, the simulator calculates aggregate performance measures, such as travel time and delay, for all vehicles (often reported by highway segment and time period).

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_7,

Macroscopic models replicate the movement of groups of vehicles (e.g., platoons) and do not analyze individual vehicle movement. Macroscopic simulation models are based on the deterministic relationships of the flow, speed, and density of the traffic stream. Mesoscopic models are a hybrid of microscopic and macroscopic models as they typically model the movement of clusters or platoons of vehicles and use equations that indicate how these clusters of vehicles interact.

Generally, simulators can be defined as time based or event based. In time-based simulation, the simulator keeps track of vehicles every time step and collects statistics on that basis. With event-based simulation, the status and location of each unit of traffic are updated only when an event occurs (e.g., when a vehicle arrives or when a traffic signal changes from red to green). Most commercially available traffic simulators are time based, because algorithms such as those for car-following require updating of vehicle positions at small time steps.

Another criterion for categorization of traffic simulation models is the use of stochastic elements. A model is termed deterministic if no element of the model is subject to randomness, i.e., every run will produce the same result. When the model includes one or more stochastic elements, it is called stochastic. In a stochastic model, random variables are used to determine specific components (such as interarrival times) or the actions that should be taken (such as the probability of a lane change).

Simulators can also be categorized as normative or descriptive. A descriptive model seeks to describe how traffic will behave in a given situation, for example, how a facility will operate with a lane closure. A normative model on the other hand seeks to optimize a given objective, for example, to minimize travel time through an arterial [2].

Lastly, traffic models can be categorized as static or dynamic. Static are those where the inputs of the model (e.g., the demands on each link) are not affected by the passing of time or by the conditions of the network, while dynamic are those models that evolve in time as a function of various elements. For additional discussion on simulation tools for traffic operational analysis, consult [2, 3] (particularly Chap. 6: HCM and Alternative Tools).

This chapter focuses on stochastic microsimulation, because it is the most often used type of simulation in traffic operational analysis. The chapter first discusses some of the basic principles of stochastic microsimulation as it pertains to traffic operations, while the second section provides an overview of the three key algorithms within microsimulators (car-following, lane changing, and gap acceptance). The third section outlines the principles involved in using a commercially available microsimulation package, and the fourth section discusses issues related to building a microsimulator using a simulation language. The fifth section provides information and examples regarding various commercially available simulators, while the sixth section summarizes the pros and cons of simulation as a traffic analysis tool and discusses cases when simulation would be the preferred approach.

Principles of Stochastic Microsimulation: An Example

Let us consider a very simple stochastic microsimulator, which replicates the movement of vehicles through a parking garage booth (Fig. 7.1) and collects appropriate statistics. Vehicles arriving to the booth provide their parking stub and pay the parking assistant the appropriate fee. We are interested in obtaining the queue length distribution, so that we can design the storage space between the booth and the parking spaces inside the garage to minimize potential blockage.

A microsimulator for this system would replicate the vehicle arrivals, the vehicle movement through the system (e.g., through the use of car-following models), and the interactions between the vehicle and the environment (in this case the parking booth). In essence the microsimulator would move each vehicle sequentially based on its prespecified rules of motion; it would monitor the movement of each vehicle and generate vehicle trajectories. Based on these vehicle trajectories, the simulator then calculates various performance measures, including queue length, travel speed, delay, etc.

To replicate our example system in a microsimulator, there are several components that need to be carefully considered and modeled accordingly:

- The arrival pattern, or interarrival distribution. The interarrival distribution may be a mathematical expression, or it may be a distribution defined by a series of x and y pairs, where x is the interarrival time interval and y is the corresponding frequency.
- The follow-up time and the respective distribution. The follow-up time is the time it takes for a vehicle to move from its position at the front of the queue to its final position for payment.

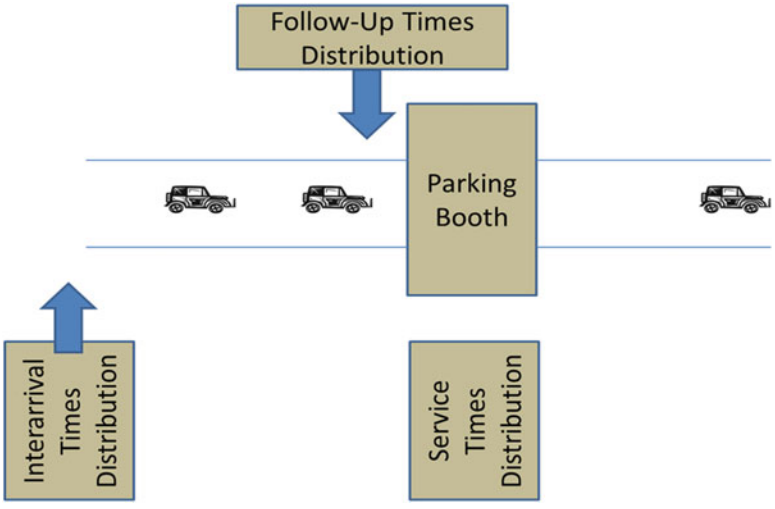


Fig. 7.1 Simulation of a parking garage booth

- The rules for acceleration, deceleration, and car-following of each type of vehicle. If trucks or other heavier vehicles are to be present in the traffic stream, we need to consider differences in the performance of various vehicle populations.
- The service pattern, or the service distribution. The payment type (cash, card, automated payment) would dictate the type of service distribution to be used. Similarly to the interarrival distribution, the service distribution may be a mathematical expression, or it may be a distribution defined by a series of x and y pairs.
- The collecting of statistics to obtain queue length measures, as well as other pertinent information such as travel time and delay information.

In this example, if the arrival and service distributions can be described mathematically and if we are willing to make some simplifying assumptions regarding the follow-up time (e.g., by incorporating it into the service time), the system can be analyzed using queuing techniques (as discussed in Chap. 6), and there is no need for developing a microsimulation algorithm. If however the arrival and service distributions cannot be mathematically described, then an analytical method would not be feasible; if we need to consider follow-up times for different types of vehicles, or for different types of drivers, or if we need to consider varying demands within our analysis interval, we can arrive at a solution much more easily using microsimulation techniques. Lastly, if our study system is expanded to consider the adjacent surface street network, there may be delays to vehicles attempting to exit the garage. These may cause spillback into the parking booth and may further increase queue lengths inside the garage. Such a condition cannot be easily considered, if at all, using analytical techniques.

Let us assume that for our example simulation, the interarrival distribution probability density function (pdf) is determined to be the one shown in Fig. 7.2a. The respective cumulative distribution is shown in Fig. 7.2b. In the cumulative distribution, the vertical axis shows the probability that a particular value will be below the respective distribution value shown along the horizontal axis. The cumulative probability values range between 0 and 1 for all distributions. If we select a random number between 0 and 1, we can enter the cumulative distribution graph and identify a specific interarrival headway. We can use this technique to obtain the interarrival times of the vehicles that would arrive during our analysis period (as long as the interarrival times follow this distribution within 1 h). We can accomplish this by randomly obtaining numbers between 0 and 1 (uniformly distributed), and using the cdf, to assign an interarrival time to each vehicle. For example, if our first random number is 0.3, then our first interarrival time is approximately 8 s, if our second random number is 0.7, then our second interarrival time is 27 s, and so on. In this way, we can randomly generate a list of interarrival times for our experiment, so that they follow the cdf of Fig. 7.2b.

The process of selecting random numbers can be accomplished on the computer through the use of random number generators (RNGs). These provide a sequence of "pseudo" random numbers based on a seed number given initially. They are called "pseudo" random number sequences because given the same seed, they always produce the same sequence of random numbers. For example, the Excel spreadsheet

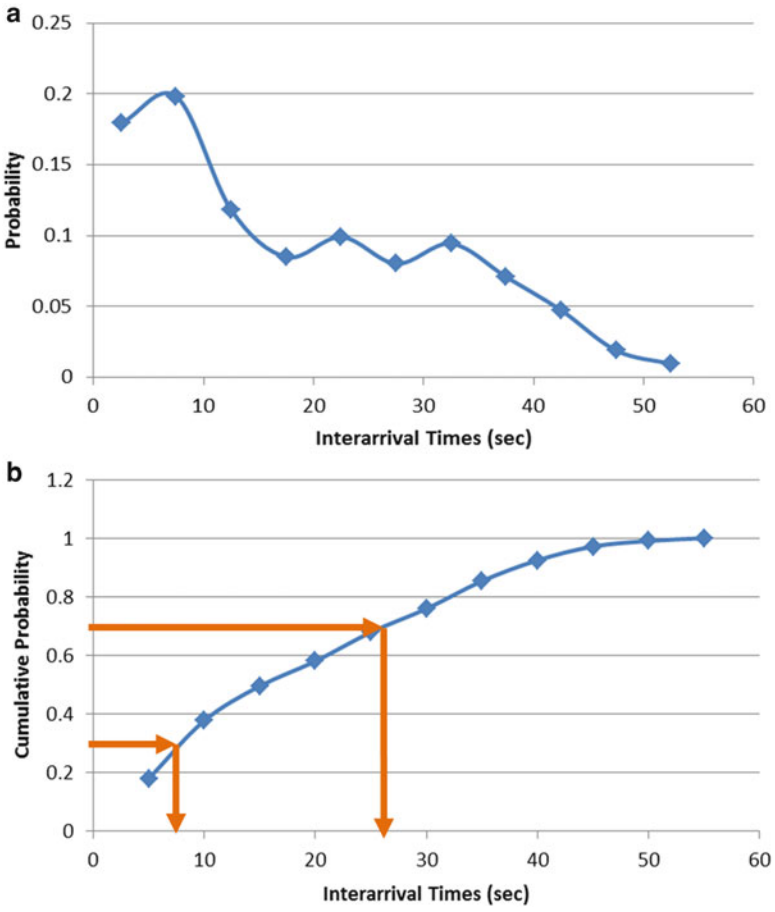


Fig. 7.2 Interarrival distribution for parking booth example. (**a**) Interarrival times probability density function (pdf). (**b**) Interarrival times cumulative probability density function (cdf)

has a function that can be used to produce random numbers from a uniform distribution: RAND(). There is significant research that has been conducted on the topic of random number generators, as well as tests that can be conducted regarding the validity of random number generators (for additional discussion, see [4, 5]).

Service times can be similarly obtained for each vehicle using the respective cdf. Simulation allows us to consider various types of payment types and the resulting distribution. For example, we can experiment with proportions of vehicles paying cash, via credit card and via electronic payment, and we can use three different types of service distributions for each payment type. If we assume that cash payments represent 40 % of vehicles, credit card payments 25 %, and electronic tag payments 35 %, we can flag each arriving vehicle with the respective type of payment. Again, we can use RNDs to accomplish this so that we can randomly tag

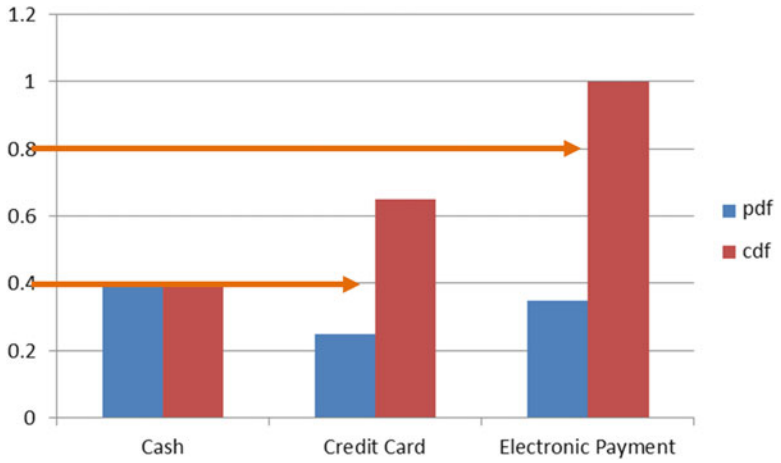


Fig. 7.3 Payment types for parking booth example

vehicles as they arrive. Figure 7.3 provides the pdf and cdf for the payment type distribution. If our first random number is 0.5, then the vehicle will be flagged to pay by credit card; if our second random number is 0.9, the vehicle will be flagged to pay electronically; and so on.

This example can be modeled in a spreadsheet application so that we can collect queue lengths and travel times. However, this can be accomplished much more easily using a simulation language such as GPSS or ARENA (see Example 7.1).

Key Components of Traffic Microsimulators

Commercially available traffic microsimulators function in a manner similar to that described in the example above. However, they are much more complex as they consider a variety of vehicle and driver characteristics, as well as infrastructure elements. They also keep track of many different operational parameters and report a variety of performance measures. This section discusses the key algorithms and elements of traffic microsimulators.

Algorithms Used for Vehicle Traffic Movement

There are three basic algorithms that are the key components in a traffic microsimulator (these were also discussed from an analytical perspective in Chap. 2):

Car-Following These algorithms estimate the trajectory of the following vehicle as a function of the lead vehicle. Chapter 2 discussed the history of car-following

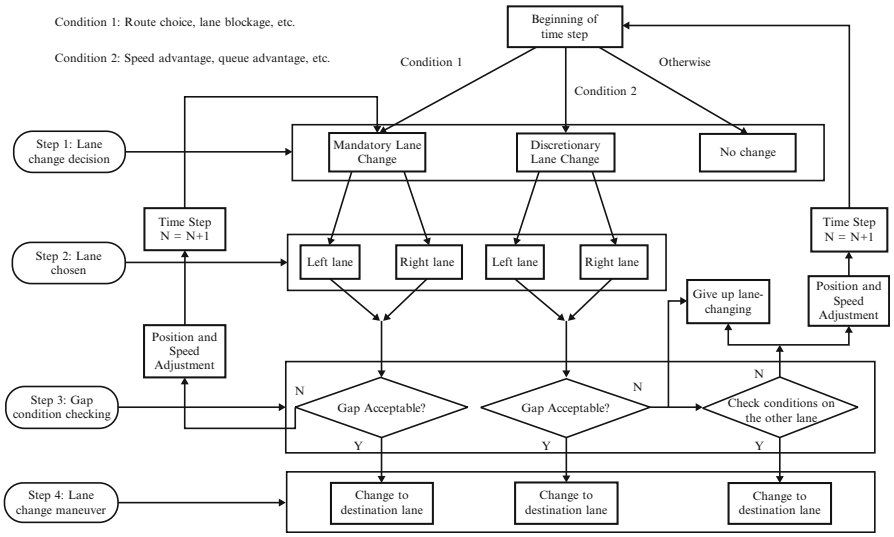


Fig. 7.4 A general lane-changing model for arterial streets [6]

models and described some of the most prevalent ones in currently available microsimulators. Car-following algorithms generally determine the following vehicle's acceleration or speed at time $t + \Delta t$ as a function of its relative position, speed, and acceleration at time t. In addition to the equations shown in Chap. 2, simulators generally employ additional rules that prescribe the conditions under which a vehicle is in car-following mode versus free mode (i.e., uninhibited by traffic ahead). Car-following rules or equations may differ for different types of vehicles and different types of drivers.

Lane Changing Lane-changing algorithms replicate the process of lane changing, from the decision to make a lane change until the time the lane change is completed. The process typically involves a series of rules and assumptions rather than a series of equations. Figure 7.4 provides a flowchart illustrating the process of lane changing for urban arterial streets [6]. Lane changing can be categorized as mandatory or discretionary. The first one refers to lane changing occurring in such cases as when there is a lane closure or for lane positioning in anticipation of a turn downstream; the second one refers to lane changes for speed or queue advantage. Lane changing is also related to gap acceptance, described below.

Gap Acceptance Gap acceptance refers to the process of selecting and using a gap such as when crossing a major street, when merging into the freeway, or when changing lanes. As shown in Fig. 7.4, gap acceptance is integral to the process of lane changing. Figure 7.5 presents a conceptual description of the merging process

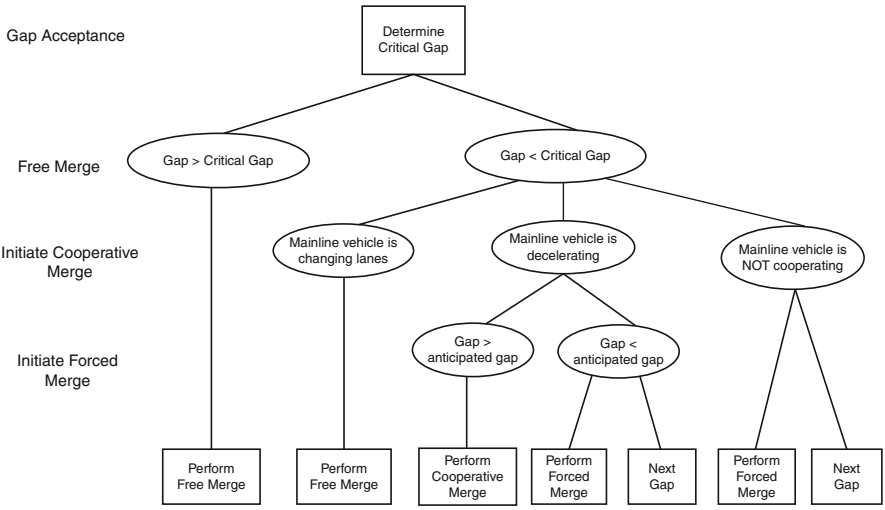


Fig. 7.5 Conceptual description of the merging process [7]

which considers the lane-changing and gap acceptance characteristics at freeway merges [7].

Network Representation

In order to accurately model the interaction and crossing of traffic streams, micro-simulators first establish rules for representing the transportation network. One of the first and simpler ways to replicate a network is the node/link configuration. In this configuration, currently used by CORSIM [8], each crossing of traffic streams is represented by a node. Nodes are connected via links. Various properties can be associated with each node and each link. For example, nodes may be signalized intersections, unsignalized intersections, or freeway merges. Link properties may include free-flow speed, number of lanes, lane width, etc.

In AIMSUN (http://www.aimsun.com), a traffic network is composed of a set of sections and nodes. The sections are connected via nodes. A section is a group of contiguous lanes with the same traffic direction, and it is characterized by attributes such as maximum speed, theoretical capacity, visibility distance, length, and slope. A node has one or more origin sections and one or more destination sections.

In VISSIM (http://www.vissim.com), the traffic network is represented by a link–connector structure. The connectors are used to tie two links to represent merging, crossing, and splitting.

Infrastructure Elements

There are several infrastructure-related elements that have an effect on vehicle motion (see Chap. 1 for a discussion of the driving environment) and can be considered in a traffic microsimulator.

- Highway design (lane width, presence of channelized right turns, etc.)
- Control (actuated signals, ramp metering)
- Other factors (presence of work zones, incidents, etc.)

Several of these elements are considered in most packages (e.g., actuated signals, incidents, etc.). Others can be considered indirectly. For example, the effect of lane and shoulder width on a freeway may be replicated by adjusting the free-flow speed, or desired speed. To account for such impacts, we need data and models indicating the correlation between lane width and free-flow speed, taking into account other aspects of the environment and potential correlations between these.

There are some other elements which can be replicated, but there are no available models to describe the relationship between that element and the impact on vehicle movement. An example of such a case is the impact of a specific incident type on traffic operations: we know that an incident affects various aspects of driver behavior and vehicle movement, but we do not yet have good models to replicate the impact for different types of incidents.

Drivers, Travelers, and Vehicles

One of the strengths of stochastic microsimulators is that they can consider a variety of driver and vehicle types and account for the variability that is actually present in the traffic stream. Most traffic microsimulators consider a range of driver types. For example, CORSIM considers 10 driver types [8], while VISSIM can consider a user-defined number of vehicle–driver unit types [9], and PARAMICS has 64 different driver profiles [10]. At this time, little has been done to study the impact of specific driver characteristics and to accurately model driver behavior. References [6, 7, 11–13] provide some insight into categorization of different types of drivers and their impact on traffic operations. These references suggest that three or four driver types may be a reasonable number of groups for traffic operational purposes. Additional such studies are needed to evaluate these recommendations.

One key aspect that is related to driver types is the desired speed, which is the speed at which each driver chooses to travel when uninhibited. This should ideally be related to the speed limit or the design speed of the facility, but vary by driver type. Desired speed is used to determine whether a driver would attempt a discretionary lane change. It is also typically used to determine delay, estimated as the difference between the actual travel time and the travel time corresponding to this desired speed. The analyst should be aware that an increase in this desired speed

may result in an increase in delay: if the desired speed of the driver is higher, then delay increases even if there are no changes in the actual operating conditions!

The behavior of pedestrians and bicyclists and their interaction with the traffic stream is another area of interest in simulation, especially as the transportation community is increasingly interested in implementing and studying high-density urban networks with large numbers of travelers using alternative transportation. Existing commercial simulators consider pedestrians and bicyclists to varying degrees. However, for the packages that do consider them, it is not clear whether their models are based on field studies and to what extent these have been validated using field data.

The vehicle types considered within a simulator and their corresponding performance are a very important element in traffic simulation. Different simulators provide varying degrees of flexibility in replicating the performance of various types of vehicles. The analyst should be aware that simulators have a set of built-in types of vehicles which are used as defaults and should be familiar with these in order to assess their suitability for various types of analysis.

When generating vehicles and drivers, simulators use or assume an arrival distribution. They assign an initial vehicle speed to all vehicles entering the network, as well as a driver type, a vehicle type, and the entering lane if relevant. The simulator typically assumes a minimum interarrival headway. This headway then corresponds to the maximum flow or volume that can enter the simulation. It is important to be aware of the value of this parameter, as it affects the amount of traffic that can enter the network at various entering points. If the demand exceeds the capacity of these points, there may be vehicles "queued" outside the simulated network. No statistics are collected on these vehicles, and they are not considered part of any queue. Thus, it may appear that the network is undersaturated and speeds are relatively high, while vehicles are queued outside the network.

Performance Measures

A key aspect of traffic simulation models is the estimation of specific performance measures, including travel time, delay, average speed, maximum queue, etc. Traffic simulators collect these data in a specific, predefined way. For example, to obtain queue length, one must define what consists of a queued vehicle. Should we only consider vehicles with zero speed? Should we consider vehicles with speed less than 3 mph or 5 mph? The answer to these questions affects the resulting queue estimates. Similar questions can be posed for travel time and delay measurements. For example, should we estimate the travel time of a link based on all vehicles on a link or only for those that enter and exit the link within the analysis period? Should delay at a signalized intersection approach include only the delay encountered within the approach link, or should we also consider the delay past the signal (which is more consistent with the HCM 2010 control delay definition)?

Each simulator defines these performance measures differently. It is important that the analyst be aware of those definitions, to be able to understand the results provided by the package and also to compare these to field measurements for purposes of calibration (see next section: Using Microsimulation).

Other Elements

Animation and the ability to visualize traffic conditions is one of the great strengths of microsimulation. Agency representatives as well as the general public can quickly grasp the impacts of various transportation alternatives, as vehicles, buses, and pedestrians are animated on a computer screen. Most of the microsimulators also provide 3D animation, and they are able to superimpose the animation on a realistic rendering of the actual study site. The danger in relying entirely on animation is that the underlying assumptions of the simulation package are not visible, and thus, it is easy to forget about them or misjudge their impact on the results. Seemingly minor assumptions in programming can make a huge difference in the results. For example, the selection of the free-flow speed or desired speed affects the computation of delay; simply reducing the free-flow speed will result in lower delays, with no other change in the network.

Two important and related components of microsimulation are the driver reaction time and the simulation time step. The reaction time is the time it takes a driver to react to traffic conditions and initiate some action when necessary. Ideally, this reaction time should be equal to the simulation time step so that if there is a need for action at step n, it can be initiated by the driver at step $n + 1$. In practice, reaction time and simulation step are set to be equal; however, not all drivers have the same reaction time. This variability in reaction times is very difficult, if not impossible to replicate in a time-based microsimulator. Some microsimulators allow users to set the driver reaction time for the network. Lower reaction times result in more swift driver actions and thus generally lower delays and travel times. Conversely, longer reaction times result in more sluggish operations and increased delays.

Using Microsimulation

Simulation models require many assumptions in order to replicate the transportation network examined, as well as its environment, traffic control, driver characteristics, and vehicle characteristics. The simulated network generated should be such as to ensure that it adequately represents the real world for the purposes of the particular study. It is very important that the model is designed such that it can capture and evaluate alternative designs of interest and can produce the performance measures needed. This section describes a five-step procedure ([1] provides an eight-step

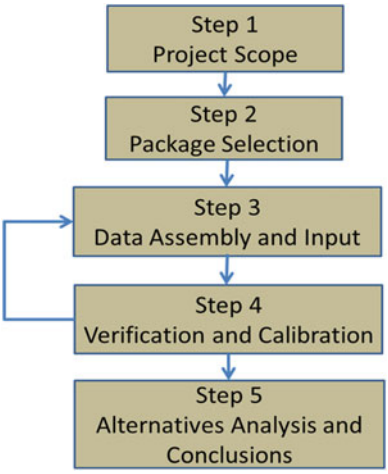


Fig. 7.6 Process for using microsimulation

procedure, while [14] provides a seven-step procedure; however, the sequence of tasks is essentially the same as the one provided here) for selecting and applying appropriate simulation model(s) for operational analyses. These steps are graphically presented in Fig. 7.6 and discussed below.

Step 1: Project Scope

The first step is to identify the problem and to define the purpose of the study, including the alternatives to be evaluated. Important elements to consider during this step are:

1. *The size and extent of the network to be evaluated and the types of facilities it involves*. Does it involve freeways, or urban streets? Are pedestrians important? How many intersections and highway links are currently affected or might be affected by one of the alternative solutions?
2. *The network geometry and traffic control*. Is the replication of horizontal and vertical curves important in the project? Should the package used be able to replicate ramp metering? Should it be able to replicate actuated control?
3. *Other network characteristics of importance*. Should buses and bus routes be considered in the analysis? Are work zones to be evaluated? Is the simulation of incidents an important consideration? Should the package be able to consider ITS components?
4. *The vehicle and driver characteristics*. Are there any particular types of vehicles that need to be modeled? Are there any particular requirements with respect to driver types and their characteristics?
5. *Analysis period duration and characteristics*. Is multi-period analysis required? What characteristics and conditions (geometric, traffic, signal control) need to vary by time period? Should the analysis results be reported in 15-min periods or in 1-h periods?

6. *Types of alternatives to be considered*. What elements do the alternatives to be considered involve? Do they involve changes to pedestrian paths? Do they involve the addition of ITS components? Are conditions in the study network likely to become oversaturated for part of the analysis period?
7. *Outputs and performance measures*. What performance measures are of interest in the analysis, and which simulation packages can provide them? What is the required level of detail? Should flows and speeds be provided by lane? Should performance measures be provided for specific turning movements? Should queue length be provided?

The output of this step should be a clear description of the project needs with respect to elements that need to be considered in the simulation, as well as a description of the required performance measures and outputs.

Step 2: Package Selection

This step involves the selection of an appropriate package to address the project needs identified in Step 1, as well as other considerations. The following criteria need to be considered in the package selection:

1. *Package capabilities relative to the specific project scope*. Each simulation package has specific capabilities and strengths with respect to the types of facilities and characteristics it can consider, as well as in terms of its output. Specific project requirements identified in Step 1 need to be very carefully considered, so that the selected package can replicate all elements required by the project scope. For example, various packages have limitations in the maximum number of lanes considered within a link; it is important for the analyst to consider the project scope requirements (including the alternatives to be evaluated) relative to the selected package capabilities.
2. *Algorithms and models employed within the package*. As indicated earlier in this chapter, each package uses a different set of algorithms for various traffic processes. Even though this aspect of simulation rarely is of interest in selecting a package, it does become of interest for specific applications. For example, in replicating operations in roundabouts, gap acceptance modeling and the manner in which a specific package handles it are very important.
3. *Ease of use and familiarity*. Often the project scope can be addressed by any of a number of commercially available microsimulation packages. In those cases, ease of use and familiarity also becomes a factor in selecting a simulator, as these packages tend to be expensive and can be cumbersome to use and understand for first-time users.

The output of this step is the selection of the package to be used in the project.

Step 3: Data Assembly and Input

This step consists of the data needs identification, data assembly, and input into the selected package. Typically, the analyst first identifies sources of existing data and then develops a plan for collecting any missing information and data, such as volumes and queue lengths. Sources of existing data would typically consist of roadway design plans as well as signal control and other traffic management information. Data collection includes input data for entering into the microsimulator, as well as calibration data for comparing the performance of the simulator to the field conditions (see Step 4). Input data typically include demands and/or turning volumes for various types of vehicles, presence and frequency of bicycles and pedestrians, and free-flow speeds. Calibration data typically include queue lengths for selected approaches, travel times for specific origin–destinations, and prevailing speeds. To collect this information, a data collection plan should be developed, to consider data collection locations, data collection and reduction methods, equipment and materials needs, time(s) and duration of the data collection, personnel assignments, coordination with local agencies, etc. In some cases, a pilot data collection may be advisable, to ensure that the data collection plan is workable before conducting a full-scale data collection. In general, microscopic packages require more intensive and detailed data than mesoscopic and macroscopic ones.

When all required data have been assembled or collected, the input files are created according to the input format of the selected package.

Step 4: Verification and Calibration

Once all data are input, the package needs to be run and debugged to ensure it is operating as expected. This process is slightly different for each commercially available package; some provide real-time feedback when an error is detected, while some provide error information and the respective code(s) after the run has been initiated.

The verification process evaluates whether the inputs have been entered correctly and the package replicates the network as intended. It involves the checking of inputs and outputs in a qualitative manner, to ensure that all functions have been properly incorporated in the package. During this process, animation is a very powerful tool that can be used very effectively in identifying discrepancies. For example, the analyst can visually check the formation of queues at specific locations or the function of actuated traffic signals.

After verification, the analyst begins the calibration. Calibration refers to the process by which the results of the package are quantitatively compared to the field data, and if necessary, assumptions in the model are adjusted to ensure that it replicates reality adequately. It is important to emphasize that the process of calibration should in no way alter any of the field-collected data which the analyst

has confirmed to be accurate. The calibration process seeks to modify only specific assumptions within the simulator so that with a given input data set (such as geometry and volumes), one obtains the same result as in the field data. Examples of such assumptions may be related to the lane-changing algorithms, car-following, driver population impacts, etc. Another set of assumptions may relate to missing or erroneous data. The assumptions reevaluated in the calibration process should consider the actual field operations and should be modified such that they match field conditions and observations as much as possible.

To perform the calibration, one or more performance measure(s) should first be selected to compare the field results to the simulated results. For example, travel time along a corridor might be selected as the primary performance measure, with queue lengths at selected links or movements as secondary measures. As indicated in Step 3, input data and calibration data should be collected simultaneously so that the observed performance corresponds to the respective set of demands and field conditions. Exact matches of multiple performance measures are nearly impossible, because simulators are not exact representations of reality. Thus, matching a performance measure to the field data does not guarantee that all performance measures provided will match exactly. Quite the contrary, the analyst needs to be aware as much as possible of assumptions built into the particular simulator, so that he/she can select an appropriate calibration measure for the purposes of the particular project, and also to assess the reason and significance of any discrepancies observed between the simulator and the field data.

During calibration, it is very important for the analyst to understand the principles of traffic flow as well as the algorithms and assumptions of the package, so that the adjustments made are reasonable. Sometimes, modifying a link or movement parameter might have unexpected consequences because of flaws in the package. The analyst should know how to interpret the simulation results and draw inferences from them. It is often the case that the field data and observations need to be reexamined to confirm the simulation results. On occasion it is also necessary to collect some additional information on specific geometric characteristics, or environmental conditions, which were not initially considered (see arrow from Step 4 to Step 3 in Fig. 7.6).

Stochastic simulators require multiple runs, as each run represents a single sample. During the calibration process, the required sample size needs to be determined so that the appropriate number of runs can be performed. The required number of runs can be estimated based on the central limit theorem (CLT) assuming a desired level of confidence in the estimate of the mean, as well as an acceptable error. The equation used to determine the required sample size is:

$$N \geq \frac{z^2}{e^2} s^2$$

where

N is the sample size
z is the standard normal value for the desired level of confidence (two-tailed value)
e is the acceptable error
s is the standard deviation of the performance measure's distribution

Since the standard deviation of the simulated performance measure is typically not available in advance, one can obtain an estimate of it from a preliminary set of runs (say ten runs) and then estimate whether this preliminary number of runs provides an adequate sample for the purposes of the study. Multiple runs should be conducted in this manner for purposes of calibration, as well as for alternative scenario analysis.

At the beginning of the simulation, the network is empty, and it takes some time to fill it up with vehicles. The time required for the network to fill up is the initialization or warm-up period. During this time, no statistics should be collected, as conditions are not representative of the analysis period. The required initialization period varies depending on the size of the network, and it should be at least as long as it takes for vehicles to travel from one end of the network to the other.

Microsimulators typically allow for consideration of multiple time periods with varying conditions (e.g., varying demands). The number and duration of time periods vary based on network conditions and the needs of the study.

Typically, statistical analysis is required to evaluate whether the simulated network replicates the field conditions. The analyst may perform tests of means (such as the *z*-test or the *t*-test), or a test of distributions (such as the chi-square test or the *K*-*S* test), to compare one or more simulated performance measures to field-measured ones.

Additional information on the process of calibration is provided in [14].

Step 5: Alternatives Analysis and Conclusions

After the process of calibration, the package is ready to be used in the evaluation of alternatives. As indicated earlier, those alternatives might consist of a combination of geometric, traffic, and control modifications, and these should be determined from the very early stages of the project as they might affect the project scope and package selection.

As during calibration, statistical analysis would likely be required to evaluate whether one or more alternative solutions result in significant differences in one or more performance measures. The analyst may perform tests of means or tests of distributions. One particularly useful technique in comparing different alternatives in the simulator is the use of paired samples. To minimize variability due to the use of random numbers, the analyst may use the same set of random number seeds for the alternatives to be tested and then use the paired *t*-test to compare the results.

Developing a Microsimulator

Occasionally, in traffic analysis, we may be interested in simulating conditions for which there is no commercially available simulator. In this case, one can use one of the several available simulation languages, which have been developed explicitly

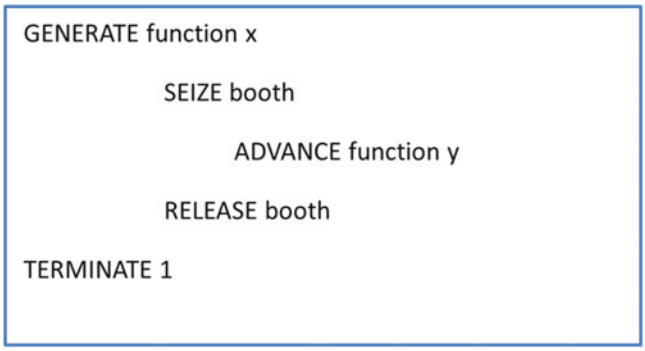


Fig. 7.7 GPSS Example 1

for this purpose. An example of such a language is GPSS (General Purpose Simulation System), which was originally developed by IBM in the 1960s [15]. There is a free student version currently available at http://www.minutemansoftware.com/downloads.asp. Other languages include SIMSCRIPT (http://www.simscript.com) and ARENA (http://www.arenasimulation.com). Such tools are applied in various aspects of manufacturing, communication networks, financial analysis, and inventory control. This section will briefly describe and use the GPSS language to demonstrate microsimulation model development specifically for traffic applications.

GPSS Concepts

GPSS can be used to replicate any type of queuing system where units of traffic are generated; they join a queue, obtain service, and then depart the system. In GPSS, transactions, or units of traffic, are generated through a simple command and move through a series of blocks, which represent specific actions or delays encountered. The program keeps track of various statistics such as the location of each transaction, its travel time through the system, the contents of each queue in the system, etc. The analyst has significant flexibility in the assumptions employed in the development of the simulation system as well as in the types of statistics collected.

Figure 7.7 provides a very simple GPSS example program. The program replicates the operations of a parking booth such as the one shown in Fig. 7.1. The first block is "GENERATE," which creates transactions based on the rules provided following the command (function x). This function represents the interarrival distribution, and it could be a constant, a mathematically expressed function, or a user-defined function. The block "SEIZE" replicates the action of seizing the facility named "booth." Each transaction generated in the previous command proceeds to seize the booth. Each transaction that is generated seizes the booth and receives service

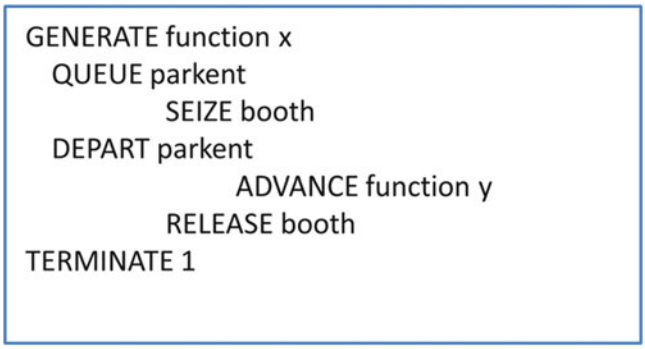


Fig. 7.8 GPSS Example 1 with queue statistics

which has duration that follows the "function *y*" distribution. The "ADVANCE" block replicates the service time for the booth using this function by delaying the progress of the transaction through the program. Once service is completed, the transaction moves to the "RELEASE" block and releases the booth. The block "TERMINATE" removes the transaction from the simulation; the number 1 is used to decrement a counter every time a transaction exits the system. Once the booth is available, the next transaction moves to that block to receive service.

The blocks SEIZE and RELEASE are used together to define the seizing and releasing of a facility. Those block names are used for a single server system, i.e., when there is only one booth. The operants "function *x*" and "function *y*" may define an existing mathematical function (e.g., the normal distribution) or a user-defined function or a constant. The interarrival times with which transactions are generated as well as the service times are randomly obtained, using in-built random number generators. The interested user may consult the GPSS user's guide for additional details on the correct syntax to replicate specific functions (http://www.minutemansoftware.com), as well as the use of random number generators in GPSS.

Figure 7.8 provides a slightly enhanced version of the program, to allow for collection of queue statistics. As shown, the only difference is the addition of the blocks "QUEUE" and "DEPART." The queue entity "parkent" defines the queue which forms to seize the booth. The two blocks are used to provide a time stamp on when a transaction joined the queue and when it began to receive service. Each transaction joins the queue as soon as it is generated. If the booth is available, the transaction immediately moves to seize the booth; otherwise, it waits in the queue block. As soon as the booth is available, the transaction moves to that block, departs the queue, and then receives service at the "ADVANCE" block.

In contrast to commercially available traffic microsimulators, GPSS is an event-based simulation system. The simulated clock advances to the next event (e.g., to the time when the next transaction arrives or to the end of a service time). Also, in GPSS the movement of transactions from one block to the next is instantaneous. Each transaction moves down the list of commands as far as it can,

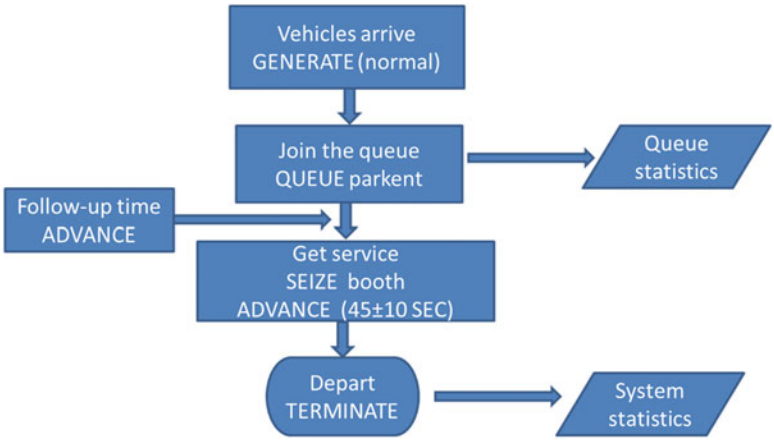


Fig. 7.9 Flowchart for Example 7.1

until it encounters a delay (e.g., when a facility is occupied). To simulate vehicle movement in GPSS, one needs to account for the time it takes a driver to react, as well as the acceleration, positioning, and deceleration components of the vehicle movement.

Example 7.1 Replicate the system shown in Fig. 7.1 using GPSS assuming that the interarrival times of vehicle arrivals are normally distributed with a mean of 50 s and a standard deviation of 8 s and that service times are 45 ± 10 s, uniformly distributed. Provide the following:

(a) The parking booth utilization in 15 min of simulation
(b) The average and maximum queue length
(c) The average time in the queue for each vehicle

Assume that the follow-up time for vehicles in the lead position of the queue to position themselves at the parking booth for service is 5 ± 2 s, uniformly distributed.

Solution to Example 7.1
Figure 7.9 provides a flowchart for replicating the system shown in Fig. 7.1, while Fig. 7.10 provides the GPSS program that describes it. The flowchart provides an overview of the path followed by each unit of traffic.

The code provided in Fig. 7.10 was developed using GPSS version 5.2.2. The GENERATE indicates use of the normal function; the first number of the operand indicates the random number generator number to be used for this function, while the second and third numbers provide the mean and standard deviation of the distribution, respectively. Two sequential ADVANCE blocks are used; the first one corresponds to the follow-up time, while the second one corresponds to

```
*********************************************
*     Example 7-1 - Entrance to parking
*
*********************************************

          GENERATE   (Normal(1,50,8))
          QUEUE      parkent
          SEIZE      booth
          DEPART     parkent
          ADVANCE    5,2              ; follow-up time
          ADVANCE    45,10            ; receive service
          RELEASE    booth
          TERMINATE


**********Timer
          GENERATE   900
          TERMINATE  1
```

Fig. 7.10 GPSS program for Example 7.1

the service time. The last few lines of the code simulate the function of the timer. A single transaction is generated 900 s into the simulation, and it immediately exits the program. The operand of the TERMINATE statement is 1, while that of the other "TERMINATE" statements is zero. Thus, the program will run for 900 s irrespective of the actual units of traffic that go through the system.

Figure 7.11 provides the output of a single run. The first part of the output provides the total simulation time, the total number of blocks in the program, and the number of facilities. Then to the right of each block, the program provides the number of transactions that entered, as well as the current (at time 900 s) content of the queue.

The bottom part of the report provides statistics related to this run:

(a) The utilization of the booth was 0.903.
(b) The maximum queue was 1, while the average queue was 0.206.
(c) The average time in the queue was 10.9 s, while the average time in queue for vehicles that had to wait was 15.4 s.

Next, we need to consider how many runs should be completed, obtain results for each run, and obtain the average values for each of the parameters requested. Since we do not have the standard deviation of any of the parameters requested, we will run the simulation program for ten times and obtain an estimate of the standard deviation for each parameter. Table 7.1 provides the results from ten runs, along with the respective means and standard deviations. If the analyst is interested in a specific level of precision, additional runs could be performed to achieve a desirable level. For example, if our desirable tolerance for maximum queue estimation is ± 0.25, then we would need to complete 15 runs.

```
                GPSS World Simulation Report - Example 7-1.17.1


                    Wednesday, April 25, 2012 16:16:34

              START TIME            END TIME  BLOCKS  FACILITIES  STORAGES
                   0.000             900.000    10        1          0


                  NAME                          VALUE
              BOOTH                           10001.000
              PARKENT                         10000.000


 LABEL              LOC  BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                    1    GENERATE            17              0       0
                    2    QUEUE               17              0       0
                    3    SEIZE               17              0       0
                    4    DEPART              17              0       0
                    5    ADVANCE             17              0       0
                    6    ADVANCE             17              1       0
                    7    RELEASE             16              0       0
                    8    TERMINATE           16              0       0
                    9    GENERATE             1              0       0
                   10    TERMINATE            1              0       0

FACILITY          ENTRIES  UTIL.   AVE. TIME AVAIL. OWNER PEND INTER RETRY DELAY
 BOOTH              17     0.903      47.830  1       18     0    0     0       0

QUEUE              MAX CONT. ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 PARKENT             1     0     17       5     0.206     10.894     15.433   0

FEC XN    PRI         BDT       ASSEM  CURRENT  NEXT  PARAMETER    VALUE
    19     0        915.864      19       0       1
    18     0        926.498      18       6       7
    20     0       1800.000      20       0       9
```

Fig. 7.11 GPSS program output for Example 7.1

Table 7.1 GPSS program results for ten runs for Example 7.1

Measures	Results of ten runs	Mean	Standard deviation
Utilization of the booth	0.903, 0.891, 0.913, 0.947, 0.95, 0.934, 0.912, 0.891, 0.949, 0.901	0.919	0.024
Maximum queue	1, 1, 1, 2, 2, 1, 1, 1, 2, 1	1.3	0.483
Average queue	0.206, 0.124, 0.085, 1.148, 0.763, 0.289, 0.118, 0.056, 0.737, 0.123	0.365	0.379
Average time in queue	10.9, 6.6, 4.5, 57.4, 36.2, 14.5, 6.2, 3.0, 36.8, 6.2	18.221	18.612
Average time in queue for vehicles that had to wait	15.4, 11.2, 8.5, 60.8, 38.2, 17.4, 9.6, 5.6, 39.0, 8.5	21.421	18.314

The interested user may use GPSS to plot a variety of measures such as queue content in real time, as the simulation is running. There is also the ability to stop and restart the simulation in order to observe each step and monitor the progress of each transaction. The GPSS website (http://www.minutemansoftware.com/) provides a tutorial and many examples for a variety of applications.

Commercially Available Simulators

A literature review easily shows there has been a wealth of microsimulators that have been used for transportation-related applications. One of the first such simulators was INTRAS [16] which was developed to replicate freeway operations and eventually evolved into the FRESIM package. NETSIM, which simulated arterial networks, and FRESIM were developed by the US Federal Highway Administration. They were eventually combined into the CORSIM package [17]. CORSIM is currently maintained and distributed by the University of Florida McTrans Center. TEXAS [18] and EVIPAS [19] were developed to replicate operations at isolated intersections.

As of this writing, AIMSUN, CORSIM, PARAMICS, and VISSIM are some of the most popular commercially available microsimulators. They can all replicate freeway and arterial networks. They use different underlying assumptions, algorithms, and performance measures, and thus, their results cannot be easily compared. Table 7.2 provides a brief overview of the characteristics and scope of selected microsimulators.

Is Simulation the Right Tool?

Simulation is a very powerful technique that has been used extensively in traffic analysis. However, its use requires significant time and resources, and thus, it is not the best approach for all types of studies. Generally, simulation is most valuable when evaluating traffic systems that have congested conditions, extended queues and spillback, or varying conditions during the analysis period. Simulation can consider a variety of different drivers and vehicles and their impact on operations. It is most useful in analyzing conditions for which analytical techniques are not available, or are not adequate. Simulation can be used to assist us in understanding and assessing new or unusual designs, novel traffic control algorithms, and the interactions of specific elements in the traffic system, prior to implementation in the field.

On the other hand, if there are analytical tools available (e.g., HCM methods) to solve a particular problem, then the use of simulation is not necessarily warranted, as it is likely to be more time-consuming and expensive. The process of calibration, to ensure that the model fits the field conditions, can be particularly time-consuming. For each of the facilities it considers, the HCM 2010 [3] provides guidance on the types of conditions for which simulation or other alternative tools are warranted.

Table 7.2 Characteristics of selected microsimulators

Simulator	Development	Characteristics	Scope
AIMSUN	TSS, Barcelona, Spain (http://www.aimsun.com)	Using the Gipps car-following model	Freeways and arterial streets; can be customized through Application Programming Interface (API)
CORSIM	Originally FHWA (consists of NETSIM and FRESIM components) Currently McTrans at the University of Florida	Uses the PITT car-following model	Freeways, arterial streets, two-lane highways; can interface with external applications through run-time extension (RTE)
EVIPAS	Bullen et al.	It evaluates phasing and provides optimum signal timing settings	Isolated intersections
MITSIM	MIT, 1996 (http://its.mit.edu/software/mitsimlab)	Object-oriented, uses Herman's car-following model [20]	Developed to evaluate Advanced Traffic Management Systems (ATMS) and Advanced Traveler Information Systems (ATIS) at the operational level; focuses on the automobile mode
PARAMICS	Quadstone, UK (http://www.paramics-online.com)	Based on the psychophysical car-following model by Fritzsche [21]	Freeways and arterial streets; can be customized through Application Programming Interface (API)
SIDRA	Akcelik and Associates, Pty, Ltd. Melbourne, Australia, (http://www.sidrasolutions.com)	Methodology is based on critical gaps	Intersection and roundabout analysis tool
TEXAS	University of Texas	Considers actuated traffic signal control	Intersections and interchanges
TRARR	Australian Road Research Board	Simulates uninterrupted traffic	Two-lane rural roads with occasional passing lanes
TWOPAS	Midwest Research Institute	Simulates uninterrupted traffic	Two-lane rural roads
VISSIM	PTV Transworld AG (http://www.vissim.com)	Uses the Wiedemann car-following model	Very flexible; can replicate many different transportation modes including transit and pedestrians

References

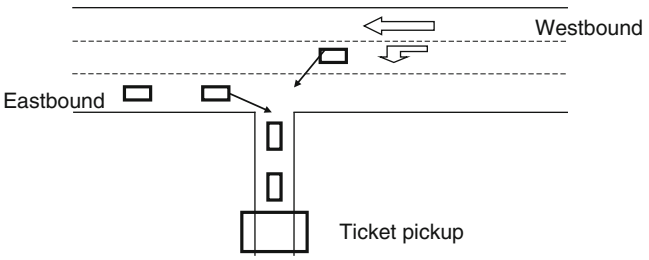
1. Elefteriadou L, List G, Leonard J, Lieu H, Thomas M, Giguere R, Brewish R, Johnson G (1999) Beyond the highway capacity manual: a framework for selecting simulation models in traffic operational analyses. Transportation Research Record 1678. National Academy Press, pp 96–106
2. Alexiadis V, Jeannotte K, Chandra A (2004) Traffic analysis toolbox: volume I: traffic analysis tools primer. Report FHWA-HRT-04-038, Federal Highway Administration, Washington, DC, June 2004
3. Transportation Research Board, National Academies of Science (2010) Highway Capacity Manual, Transportation Research Board, National Research Council, Washington, DC
4. Knuth D (1997) Chapter 3: random numbers. In: The art of computer programming, volume 2. Semi numerical algorithms, 3rd edn. ISBN: 0321751043
5. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Chapter 7: random numbers. In: Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, New York, NY, ISBN 978-0-521-88068-8
6. Sun DJ (2009) A lane-changing model for urban arterial streets. Ph.D. Dissertation, University of Florida
7. Kondyli A, Elefteriadou L (2011) Modeling driver behavior at freeway-ramp merges. Transportation Research Record: Journal of the Transportation Research Board of the National Academies, No 2249, Washington, DC, Jan 2011, pp 29–37
8. CORSIM User's Guide (2006) FHWA US Department of Transportation. Office of Operations Research, Development and Technology, McLean, VA
9. VISSIM Website. http://www.ptvag.com/software/transportation-planning-traffic-engineering/services/vissim-faq/
10. PARAMICS Website. http://www.paramics-online.com/paramics-features.php
11. Sun D, Elefteriadou L (2011) Information categorization based on driver behavior for urban lane changing maneuvers. Transportation Research Record: Journal of the Transportation Research Board of the National Academies, No 2249, Washington, DC, Jan 2011, pp 86–94
12. Sun D, Elefteriadou L (2010) Research and implementation of lane changing model based on driver behavior. Transportation Research Record: Journal of the Transportation Research Board, No 2161, Transportation Research Board of the National Academies, Washington, DC, pp 1–10
13. Kondyli A, Elefteriadou L (2009) Driver behavior at freeway-ramp merging areas: focus group findings. Transportation Research Record 2124. National Academy Press, pp 157–166
14. Dowling R, Skabardonis A, Alexiadis V (2004) Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation software. Report FHWA-HRT-04-040, Federal Highway Administration, Washington, DC, June 2004
15. Schriber TJ (1991) An introduction to simulation using GPSS/H. Wiley, New York, NY
16. Wicks DA, Lieberman ED (1976) Development and testing of INTRAS a microscopic freeway simulation model, volume I. Program design and parameter calibration, KLD Associates, FHWA-RD-76- 75, Final Report, pp 146
17. Halati A, Lieu H, Walker S (1997) CORSIM—Corridor traffic simulation model. In: Traffic congestion and traffic safety in the 21st century: challenges, innovations, and opportunities. American Society of Civil Engineers, pp 570–576
18. Rioux T (1977) TEXAS—a microscopic traffic simulation package for isolated intersections. University of Texas, Center for Highway Research, Austin, TX
19. Bullen AGR, Hummon N, Bryer T, Nekmat R (1987) EVIPAS: a computer model for the optimal design of a vehicle actuated traffic signal. Transp Res Rec 1114:103–110
20. Yang Q, Koutsopoulos HN (1996) A microscopic traffic simulator for evaluation of dynamic traffic management systems. Transp Res Part C 4:113–129
21. Fritzsche H (1994) A model for traffic simulation. Traffic Eng Control 35(5):317–321

Problems

1. Identify and briefly describe a specific traffic operational problem in your area. For example, consider overflowing left/right turning bays, spillback from a downstream signal to an upstream intersection, and other such congestion-related problems. Provide the exact location, along with a sketch of the pertinent geometric characteristics of the area. Would you recommend the use of simulation for studying this problem and recommending alternative solutions?
2. Conduct a literature review and identify three different applications of traffic simulation programs. For each of them describe the model type, the model capabilities, and any shortcomings identified during the model application. Briefly describe the application of the model and the results obtained.
3. Develop a GPSS simulation program to replicate the operations of an all-way stop-controlled intersection with the following characteristics:

   NB approach: Demand = 160 veh/h, 10 % left, 80 % through, 10 % right
   SB approach: Demand = 300 veh/h, 5 % left, 80 % through, 15 % right
   EB approach: Demand = 280 veh/h, 10 % left, 40 % through, 50 % right
   WB approach: Demand = 330 veh/h, 40 % left, 40 % through, 20 % right

   All approaches have one lane per direction. What is the average queue, the maximum queue, and the average delay per approach?
4. There is an entrance to a parking lot along a two-lane, two-directional highway. The driveway has a queuing capacity of two vehicles, in addition to the ticket pickup spot (see sketch below).



5. Vehicles arrive during the peak hour (8–9 a.m.) from the eastbound direction at a rate of 280 veh/h (arrivals are Poisson distributed). Of these, 40 % enter the parking lot. During the same time, the arrival rate from the westbound direction is 250 veh/h (Poisson distributed). Of these 15 % enter the parking lot. Each driver can pick up their ticket within 30–60 s, uniformly distributed. Model the entire system in GPSS, and provide:

   (a) A flowchart of the simulation model. Clearly state any assumptions you need to make.
   (b) The GPSS program that describes it.

(c) Run the program 15 times and provide a summary of the average queue lengths for each direction, and the average time it takes for each vehicle to travel through the system.
(d) Determine when and where there are queues developing. Do you have any suggestions for improving the system?

Part IV
Highway Facilities and Principles for Their Analysis

This part focuses on the characteristics of specific types of highway facilities and applies the methods and tools described in previous chapters to selected facilities. Each of the chapters in this part focuses on a specific facility and its specific design and traffic operational characteristics. It also applies the tools discussed in the previous three parts of this book to that specific facility and examines the advantages and disadvantages of using a specific tool to assess the operational performance of the facility.

Chapter 8 presents freeway systems, including merge and diverge junctions and weaving segments. Chapter 9 discusses the operation of signalized intersections and signalized networks. Chapter 10 presents unsignalized intersections, including two-way stop control, all-way stop control, and roundabouts. Chapter 11 focuses on two-lane highways.

Chapter 8
Freeways

Freeways are defined as those facilities that afford uninterrupted flow of traffic, i.e., there is full access control. Control of access refers to public access rights from properties along the freeway; access to freeway facilities is allowed only through selected public roads, typically on- and off-ramps [1]. Thus, freeways typically operate at higher speeds and higher capacities than urban arterial streets or local roadways.

Figure 8.1 provides an image of a freeway facility (I-4 in Orlando). In such facilities, the mainline has at least two lanes of traffic per direction, and design speeds are in the order of 50–70 mph depending on the roadway design characteristics [1]. Operating speeds are typically high, in the order of 60–70 mph, as they tend to have low grades and flat horizontal curves. Freeways have medians that separate the two directions of traffic. The Green Book [1] recommends lane widths equal to 12 ft and continuous paved shoulders on both sides of the travel way. Rural freeways tend to have greater right-of-way availability and thus higher design speeds and more generous design elements. Such facilities accommodate longer-distance intercity travel. Urban freeways on the other hand have more restrictive geometry, and they typically carry high traffic volumes as they serve urban commuter traffic which has a morning and evening peak hour. Reference [1] (chapter 8) provides detailed information regarding the design of freeways and various freeway configurations.

In the USA, freeways represent a relatively small portion of the highway network in terms of lane miles; however, they carry the majority of vehicle miles of travel. As of 2010, freeways represented 2.21 % (135,508 out of 6,117,785) of rural lane miles and 5.90 % (145,289 out of 2,463,373) of urban lane miles [2]. However, freeways carry 26.95 % (265,250 out of 984,148) of rural vehicle miles of travel and 35.24 % (698,554 out of 1,982,358) of urban vehicle miles of travel (in millions of vehicle miles per year).

Freight movement is a significant portion of the vehicle miles traveled along freeways. Truck trips represent 24.60 % (60,435 out of 245,647) of vehicle miles traveled (in millions of vehicle miles per year) on rural interstates and 10.53 % (50,297 out of 477,692) of vehicle miles traveled on urban interstates.

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_8,

Fig. 8.1 A Freeway facility (photo by Vipul Modi)

Operationally, freeways were designed to allow uninterrupted mainline movement. Vehicles enter the freeway mainline through on-ramps, which have yield signs at the entrance to the freeway; thus, vehicles can maintain or even increase their speed as they approach the freeway. Similarly, vehicles exit the freeway through off-ramps to join the surface (arterial) network.

This chapter first defines the types of freeway segments and describes their operation. The second part presents freeway management methods for optimizing freeway operational performance, and the third part summarizes methods of analysis for freeway segments and for the freeway system.

Freeway Segments and Systems: Configurations and Operations

From the perspective of the driver, freeways are seamless facilities. Drivers enter the freeway through an on-ramp junction; travel along the mainline freeway until they reach their desired exit, where they take the off-ramp; and join the surface network.

From the perspective of the traffic analyst however, freeways have been traditionally [3] divided into the following types of segments (Fig. 8.2):

- Merge junctions: They are the segments where an on-ramp joins the mainline freeway. An acceleration lane (or auxiliary lane) allows vehicles entering

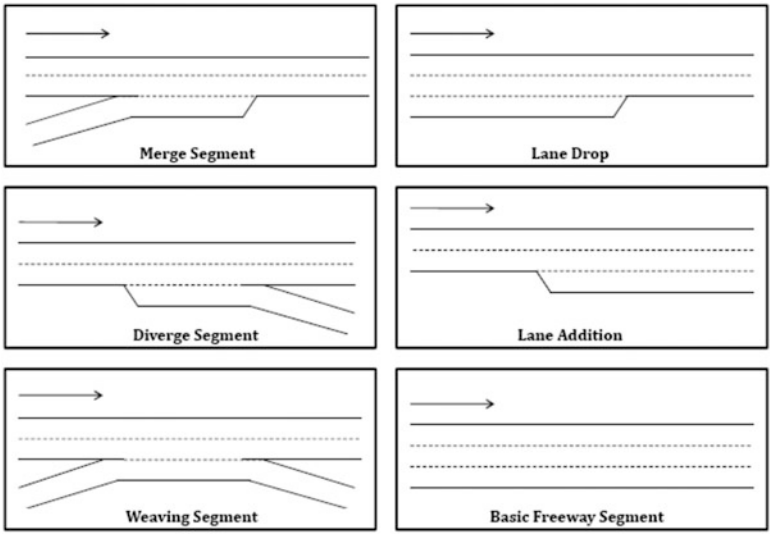


Fig. 8.2 Types of freeway segments

the mainline to reach a suitable speed and to identify a suitable gap through which to enter the traffic stream. These often function as bottlenecks when the total number of lanes approaching the area is higher than the number of lanes departing.

- Diverge junctions: They are the segments where an off-ramp connects the mainline freeway to the adjacent surface network. A deceleration lane (or auxiliary lane) is provided to allow exiting vehicles to decelerate without impacting the freeway mainline. Freeway off-ramps can create bottlenecks when the exiting demand exceeds the capacity of the diverge ramp. Spillback from the off-ramp may impact the adjacent freeway lane, or even the inside lanes of the freeway facility.
- Weaving segments: Those are freeway segments with an on-ramp followed by an off-ramp where the two are connected with an auxiliary lane. There are many different weaving configurations with different operating characteristics. Weaving areas are characterized by a high proportion of lane changes as vehicles position themselves in advance of the weave.
- Lane drops: Those are sections where a mainline freeway lane ends. This might occur downstream from a major merge junction, or as the freeway exits a major metropolitan area. Lane drops can become bottlenecks when the demand to the downstream segment exceeds its capacity.
- Lane additions: Those are sections where a mainline freeway lane is added. This might occur as a freeway approaches an urbanized area, or as part of a major diverge segment.

- Basic freeway segments: Those are segments with no on- or off-ramps, where the mainline freeway cross section is essentially the same throughout their length. Those segments can become bottlenecks due to their geometry (e.g., by the presence of steep upgrades or steep horizontal curves).

This section describes first the operations of merge, diverge, and weaving segments. Operations at a lane drop, a lane addition, and basic freeway segments are discussed next. The last subsection discusses the analysis of a freeway as a system, considering the interactions between each its components.

Merge Junctions

From a traffic operations standpoint, merge junctions are key elements in the freeway system, as they are likely to function as bottlenecks. Let us first examine the operations of merge junctions from a microscopic perspective. Figure 8.3 illustrates a merge junction with a vehicle (in the green circle) arriving from the on-ramp seeking a suitable gap. At first glance, there are two possible gaps the vehicle may select. It can slow down so that it can take gap 1, or it can accelerate and take gap 2. Another possibility is for the vehicle in the adjacent lane (in the yellow circle) to choose to move to the middle lane and take gap 3, creating thus a new gap for our subject vehicle, which is the combination of gap 1 and gap 2.

These options assume that the vehicles along the freeway forming the gaps do not accelerate or decelerate. The drivers of these vehicles, however, may decide to either "cooperate" or "compete" with the on-ramp vehicle. A freeway vehicle may cooperate by decelerating to enlarge a gap; it may compete by accelerating to pass the on-ramp vehicle and close the gap. Finally, drivers may choose to position themselves toward the inside lanes many hundreds of feet upstream of the merge, in anticipation of heavy oncoming ramp traffic. This typically occurs with drivers familiar with the traffic patterns in the vicinity of the merge. The graph of Fig. 8.3 only shows the current vehicle positions and does not provide any indication of whether vehicles positioned themselves earlier in anticipation of the merge. Incidentally, it is interesting to note that earlier efforts to match the release of an on-ramp vehicle with an anticipated gap failed [4], because vehicles did not maintain their speed and lane from the time they were detected until the time they arrived at the merge. At the same time, moving the mainline detection zone closer to the merge does not allow for adequate time for the release of the on-ramp vehicle. Thus, the wealth of options and unpredictability of driver behavior around merges make them difficult to model from a microsimulation perspective. Recent research [5] examined the behavior of various types of drivers around merge junctions and recommended categorizing drivers into three

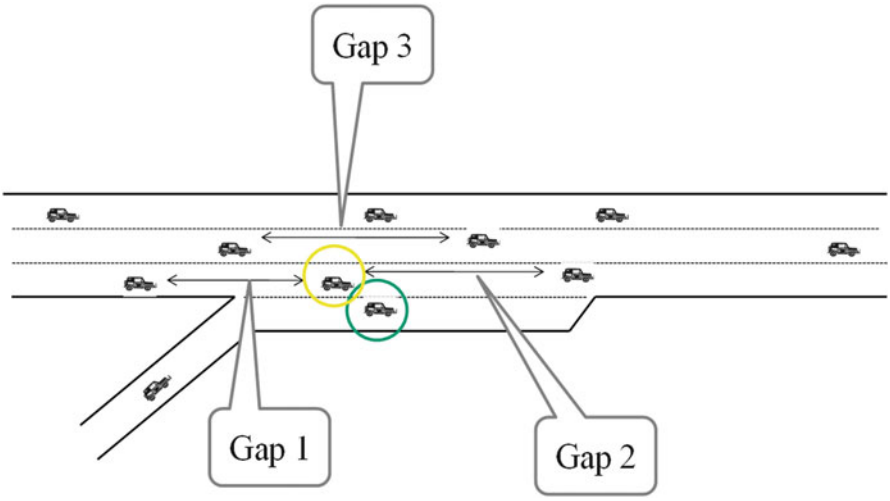


Fig. 8.3 Traffic operations at a merge junction

types (aggressive, average, and conservative) and examined three types of merging maneuvers (free, cooperative, and forced) based on the level of interaction between the merging vehicle and its surrounding vehicles.

Generally, driver actions around the merge area can affect the traffic patterns and capacity to a significant degree. Investigations related to breakdown occurrence at ramp junctions have found that when groups of several vehicles following each other enter the freeway from the ramp, they are likely to create turbulence resulting from lane changes and decelerations of vehicles on the mainline, forced merges by on-ramp vehicles. This turbulence can lead to breakdown when the mainline demand level is adequately high [6]. More recent research [7] has shown that the probability of breakdown increases more with the addition of a ramp vehicle than with the addition of a freeway vehicle. In other words, the impact of a ramp vehicle on capacity is higher than the impact of a mainline vehicle. Another research effort [8] indicated that "...the vehicle will simultaneously occupy two lanes during the process of changing lanes, momentarily decreasing the capacity of the link. This feature becomes particularly important near bottlenecks, where it might reduce the already limited throughput."

From a traffic operations perspective, the operations at a merge junction are affected by the gap availability (size and number of gaps) on the mainline and the gap usability (the manner in which drivers accept or reject each gap) for vehicles arriving from the on-ramp. Mathematically, the gap availability can be expressed through the interarrival distribution of the mainline traffic. The gap usability can be expressed through gap acceptance analysis (see Chaps. 2 and 10 for additional information on gap acceptance).

Generally, the most important traffic and design elements affecting the operations of a ramp merge junction are:

- *The mainline and ramp demands*: The number and size of gaps on the mainline are determined by the demand and the arriving pattern on the mainline, while the usability of those gaps depends on the ramp demand, as well as the their arrival pattern (e.g., whether there is ramp metering present or whether there is an upstream signal).
- *The presence of upstream and downstream ramps, as well as their distance to the subject on-ramp, and their respective demands*: The presence of an upstream on-ramp in close proximity would result in a relatively high number of vehicles along the shoulder lane, with reduced opportunities for merging for the subject on-ramp. The closer that on-ramp is, the higher the probability that vehicles do not have an opportunity to move to the inside freeway lanes before arriving at the subject on-ramp. Similarly, when there is a downstream off-ramp in close proximity, vehicles may be positioning themselves toward the shoulder lane in anticipation for exiting, creating thus fewer gap opportunities for the subject on-ramp.
- *The type and length of the acceleration lane*: The longer the acceleration lane, the more opportunities on-ramp merging vehicles have to enter the freeway. The literature has indicated that longer acceleration lanes generally result in higher capacities through the merge [5, 9].
- *The free-flow speed of the ramp and the mainline freeway, as well as the speed differential between the two*: These affect the usability of gaps, as the higher the speed differential the more difficult it is for entering vehicles to adjust to the mainline traffic speed and find a suitable gap.
- *Sight distances*: Sight distance from the on-ramp to the mainline freeway affects the gap acceptance characteristics as ramp vehicles approach the freeway. Also, sight distance from the mainline to the ramp vehicles may affect the positioning of mainline vehicles as they approach the merge.
- *Geometric elements*: The angle of approach of the on-ramp to the mainline freeway, the grades, cross section, and horizontal curvature in the vicinity of the merge all have the potential to affect traffic operations. More restricting geometry generally affects free-flow speeds and may create a bottleneck at relatively lower demands.

In addition to the typical merge configurations where there is a single lane ramp approaching the mainline freeway which clearly is the predominant movement, there are major merge configurations which bring together traffic streams from two major highways. The merge may have an equal number of lanes approaching and departing, or it may have fewer lanes approaching than departing.

Diverge Junctions

The literature has not devoted quite as much effort on diverge junction operations as it has on those for merge junctions. At diverge junctions, typically the total number of lanes arriving is equal or higher than the total number of lanes departing, and

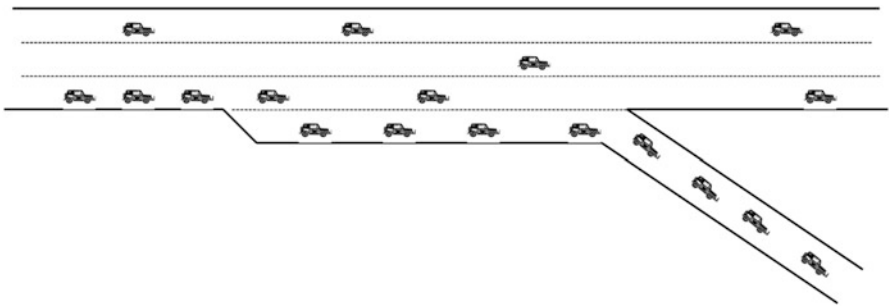


Fig. 8.4 Traffic operations at a diverge junction

thus, they may not seem prone to breakdowns. However, when the exiting demand exceeds the capacity of the off-ramp proper, or when there is a signal at the end of it limiting throughput, spillback is created on the mainline and can severely disrupt operations in the freeway system. Figure 8.4 illustrates a congested diverge junction which spills back into the mainline freeway. As shown, there is a queue created along the rightmost lane of the freeway, which can severely disrupt operations; furthermore, it can pose a safety hazard as vehicles approach the junction at high speeds and may not expect stopped or slow-moving vehicles ahead. The hazard increases when the visibility is low, either because of reduced sight distance or during low visibility conditions (e.g., foggy or rainy conditions).

Diverge areas are characterized by relatively high lane change activity, and driver actions can affect operations to a significant degree. Operations are mostly affected by the number of lane changes around the diverge, as well as the demand-to-capacity at each of the departing directions. The most important traffic and design elements affecting the operations of a diverge junction are:

- *The mainline and off-ramp demands*: The demand-to-capacity for the departing off-ramp number is a key variable which affects the operations of the diverge segment. If it exceeds capacity, spillback may affect upstream freeway sections. The higher the demand on the mainline, the greater the disruption to traffic. Also, higher demands along the mainline restrict lane changing for exiting vehicles.
- *The presence of upstream and downstream ramps, as well as their distance to the subject off-ramp, and their respective demands*: The presence of an upstream on-ramp in close proximity would result in a relatively high number of vehicles along the shoulder lane, with reduced opportunities for vehicles desiring to exit at the next off-ramp. Also, the closer that on-ramp is, the higher the percent of vehicles that remain along the shoulder lane before arriving at the subject off-ramp. Similarly, when there is a downstream off-ramp in close proximity, vehicles may be positioning themselves toward the shoulder lane in anticipation for exiting at both off-ramps, increasing the demand for the rightmost lane.

- *The type and length of the deceleration lane*: The longer the deceleration lane, the more storage would be available for exiting vehicles and the smaller the impact on the mainline freeway traffic. In cases when demand for the off-ramp is high, the ramp may be designed to have two parallel lanes to increase storage and minimize spillback.

Similarly to major merge junctions, there are major diverge junctions where the mainline traffic stream is split into two facilities. Operationally, it is lane changes around this area which tend to limit its capacity as vehicles seek to position themselves to the appropriate set of lanes.

Weaving Sections

Weaving is defined as the crossing of two or more traffic streams traveling in the same general direction along a significant length of highway without the aid of traffic control devices [3]. Weaving areas are formed when a merge area is closely followed by a diverge area or when an on-ramp is closely followed by an off-ramp and the two are joined by an auxiliary lane. Figure 8.5 illustrates four different types of weaving sections (there are several other types). As shown, depending on the configuration, the two crossing traffic streams may require none, one, or more lane changes. For ramp weaves (Fig. 8.5a), each of the two weaving streams has to make at least one lane change in order to reach their destination. For two-sided weaves however, the on-ramp to off-ramp demand (Fig. 8.5d) has to make two lane changes in order to do so.

The number of lane changes required and the demand of the respective movements affect capacity. Also, the capacity of one movement depends on the demand of the other movements, since that demand defines the number and size of gaps available. At each of those weaving types, one can define four different origin–destination pairs (Fig. 8.6).

The most important traffic and design elements affecting the operations of a weave are [3, 10]:

- *The weaving configuration*: The configuration of the weave affects the number of lane changes required by each of the weaving traffic streams and thus the amount of turbulence around the weave.
- *The demands of the four origin–destinations within the weave*: The demand of each movement relates to the number and size of gaps available to other movements and the ability of vehicles to complete the necessary lane changes. Generally, higher demand in the weaving movements results in higher turbulence and lower capacity.
- *The length of the weave*: Longer distances between the on-ramp and the off-ramp create increased opportunities for weaving vehicles to complete any necessary lane changes and thus improve its operations. As the length of the weaving segment increases, so does its ability to accommodate lane changes.

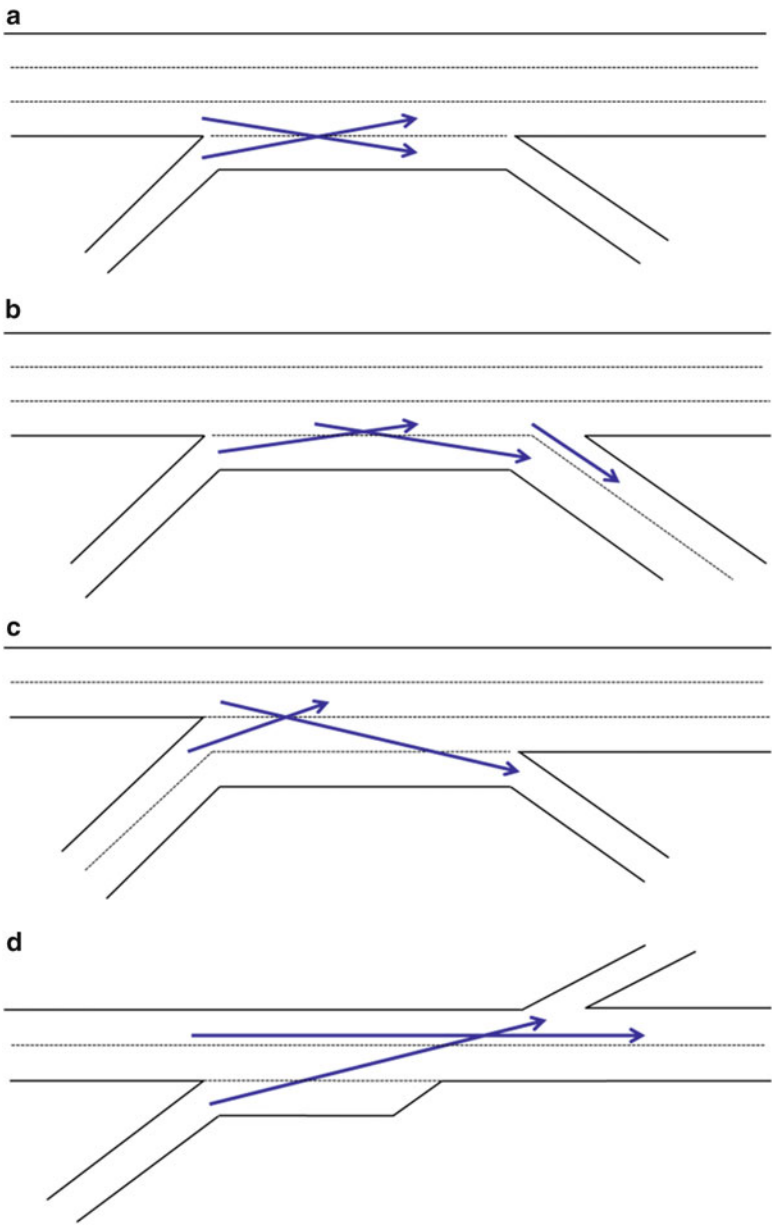


Fig. 8.5 Weaving section configurations. (**a**) Ramp weave, (**b**) major weave with zero or one lane changes, (**c**) major weave with zero or two lane changes, (**d**) two sided weave

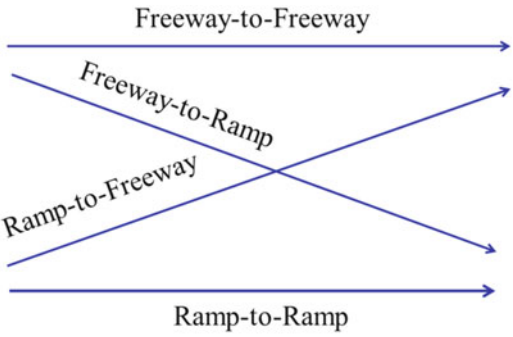


Fig. 8.6 Origin destination demands at a weave

Similarly to merge and diverge junctions, weaves can function as bottlenecks, particularly when the number of lanes departing is lower than the number of lanes arriving or when the demand for one of the departing movements exceeds its capacity.

Lane Additions and Lane Drops

Lane additions are not as critical to the operations of the system, as they provide increased capacity for the same demand. Figure 8.7a illustrates a lane addition configuration, where the total number of lanes departing is higher than those arriving.

Lane drops, illustrated in Fig. 8.7b, have less lanes departing than those arriving and can act as bottlenecks if the demand is sufficiently high. Their operation is be similar to that of merge junctions; however, the distance over which vehicles may merge is much longer, and thus, their operation is much smoother, with lane changes spread out over a longer distance.

Basic Freeway Segments

Basic freeway segments are those that are not influenced by merging, diverging, or weaving operations, and thus, the number of lane changes is relatively low [3]. At those segments, the total number of lanes remains constant, and there are no conflicting traffic streams. Bottlenecks may form only due to geometry (e.g., steep upgrades or sharp horizontal curves). Therefore, such segments are typically non-congested unless there is restrictive geometry or unless there is a downstream bottleneck which results in spillback into the basic freeway segment.

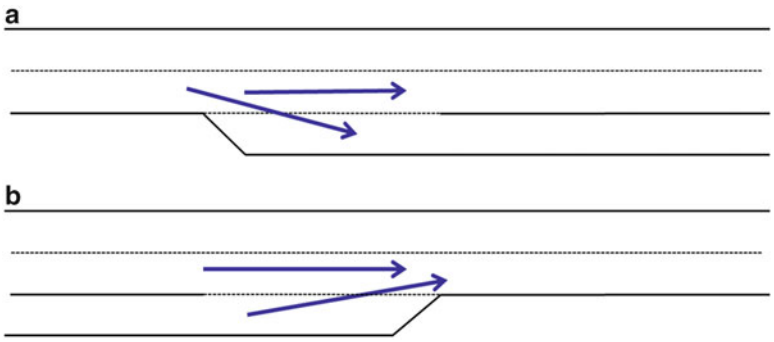


Fig. 8.7 (**a**) Lane addition and (**b**) lane drop

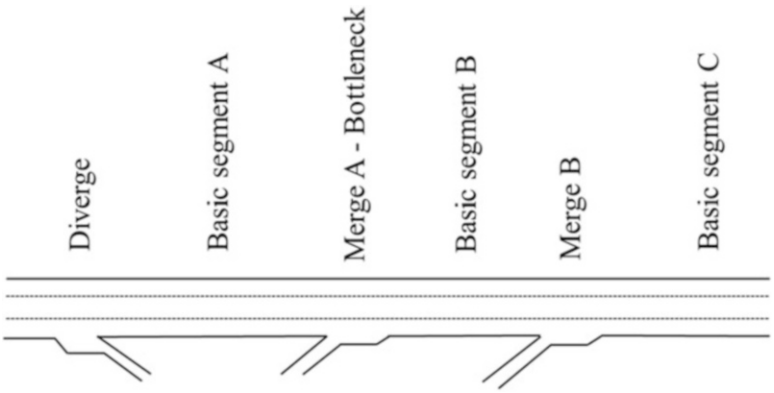


Fig. 8.8 Sketch of a freeway facility

Freeway Systems

When demand is below capacity, we can evaluate each of the segments described above in isolation and determine its operational performance (e.g., its expected density and speed). However, when one or more of the freeway segments become congested, surrounding segments cannot be individually evaluated, and we need to evaluate the freeway facility as a whole, considering simultaneously all operationally interrelated segments in the system. To illustrate this interdependency between freeway segments, let us examine Fig. 8.8, which illustrates a freeway facility with multiple segments. For this facility, the bottleneck is located at Merge A. When the bottleneck is activated, a queue forms and affects Basic segment A as well as the upstream Diverge. Thus, these upstream segments become congested, but only because of the impacts of Merge A. If analyzed as individual segments they will likely not appear congested, as spillback from a downstream segment cannot be detected.

Time Period		Diverge	Basic Segment A	Merge A – Bottleneck	Basic Segment B	Merge B	Basic Segment C
1	3:30 – 3:45 pm	60	59	56	57	58	60
2	3:45 – 4:00 pm	58	57	54	56	49	61
3	4:00 – 4:15 pm	45	42	41	50	52	59
4	4:15 -4:30 pm	40	37	35	50	53	61
5	4:30 – 4:45 pm	41	35	34	51	54	60
6	4:45 - 5:00 pm	39	34	35	52	53	59
7	5:00 – 5:15 pm	35	32	30	51	53	61
8	5:15 – 5:30 pm	36	28	29	50	55	60
9	5:30 – 5:45 pm	36	38	38	51	54	58
10	5:45 – 5:30 pm	38	40	47	52	56	61
11	5:30 – 5:45 pm	51	52	55	57	59	60
12	5:45 – 6:00 pm	60	57	57	57	60	61

Fig. 8.9 Speeds (in mph) along a freeway facility by time interval (Key: 26–35 mph, *red*; 36–45 mph, *orange*; 46–55 mph, *yellow*; 56–65 mph, *green*)

At the same time, the downstream segments may experience demand starvation when the demand is held up at Merge A. If Merge B is analyzed in isolation considering the total freeway and on-ramp demand from Merge A, that segment may appear congested. However, in the field it may not be congested, as a portion of the demand is held up in queues associated with the bottleneck at Merge A. Therefore, when congested conditions are expected along one or more segments, it is important that the facility be analyzed in its entirety, considering all possible interactions and interdependencies between individual segments. The analysis of the freeway as a system, particularly when congestion is present, requires an evaluation in both time and space.

Figure 8.9 provides an illustrative analysis of the freeway system shown in Fig. 8.8, considering the average operating speed of each segment for each 15-min period throughout the evening peak period (3:30–6:00 p.m.). Each row indicates the average operating speed for the particular time interval at each one of the segments, while each column provides the speed at each segment of the freeway. During the first

time period, there is no congestion along the freeway. During the second time period, the speeds begin to drop at the two merge segments, while starting with the fourth period Merge A has transitioned into congested conditions. The breakdown at Merge A results in speed drop at the upstream freeway segments. At the downstream segments, speed drops slightly as a result of the breakdown, and conditions are free flowing at Basic segment C.

Evaluation of the facility for the entire peak period is critical in taking into consideration the queue buildup and dissipation. Consideration of a single analysis period and the respective demand during that time (e.g., time period 6) would not provide a complete and accurate picture of the queue length and overall freeway conditions. The type of analysis shown in Fig. 8.9 provides a more comprehensive evaluation of the freeway system, as it recognizes the location of the bottleneck as well as its effects in time and space.

Freeway analysis should consider the development and dissipation of queues starting from the bottleneck(s) and considering their effects on adjacent segments. Consideration of the entire freeway system affected is especially important when considering the alleviation of a particular bottleneck, as the increasing capacity of that bottleneck may result in an increase in downstream demand and may reveal new bottlenecks downstream. For example, alleviation of the Merge A bottleneck may result in increasing demand at Merge B and may cause that location to become a bottleneck.

Advanced Traffic Management Methods for Freeway Facilities

As congestion continues to increase, agencies have considered alternative methods and advanced technology to make better use of existing capacity. These methods and technology are often integrated into freeway management programs that seek to manage, operate, and maintain freeway facilities in an efficient and cost-effective manner [11]. Most recently, these techniques are referred to as Advanced Traffic Management Strategies (ATMS). This section presents a few of the most widely used freeway management techniques: ramp metering, high-occupancy toll/high-occupancy vehicle (HOT/HOV) lanes, variable speed limits (VSL), use of shoulder, and incident management.

Ramp Metering

Ramp management is defined as "the application of control devices, such as traffic signals, signing, and gates to regulate the number of vehicles entering or leaving the freeway, in order to achieve operational objectives" [12]. Ramp management strategies may be used to control access to selected ramps or to control the manner in which vehicles enter a freeway. Ramp management strategies include ramp metering, ramp closures, and various ramp terminal treatments to provide increased storage.

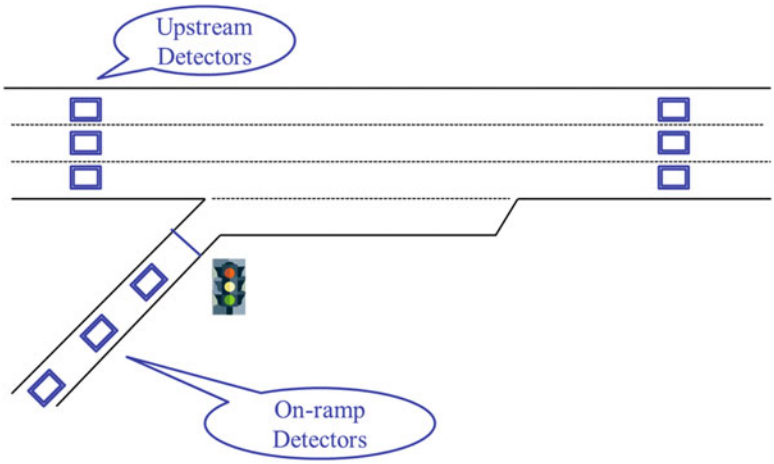


Fig. 8.10 Illustration of ramp metering

Ramp metering (also referred to as ramp signaling) is the use of a traffic signal at a ramp to control the rate at which vehicles enter the freeway. Ramp metering algorithms have been developed to limit the amount of on-ramp flow as a function of the freeway flow in order to avert or postpone breakdown. Figure 8.10 provides an illustration of ramp metering at a single ramp. Ramp metering algorithms use field data from upstream and/or downstream detectors along the freeway to determine what would be the maximum amount of flow that could be allowed from the on-ramp. In addition to reducing on-ramp demand, ramp metering also breaks up the platoons that would otherwise enter the freeway and create increased turbulence and expedite breakdown.

Such algorithms typically monitor the queue length on the ramp, to evaluate whether there is potential for spillback onto the adjacent arterial and local street network. Of course, when demand is exceedingly high, congestion cannot be avoided, and ramp metering algorithms have to consider the tradeoffs between congestion on the freeway versus congestion on the arterial network (i.e., a less restrictive on-ramp flow versus a longer queue and spillback from the on-ramp).

Reference [12] reports that "when ramp meters were turned off for a 6 week study in Minneapolis, Minnesota, a before and after evaluation concluded that meters were responsible for a 21 % reduction in crashes and a 9 % increase in mainline volumes" (pp. 5–4).

Most of the systems installed allow one vehicle per green (4–4.5 s of green indication), but the discharge rate can be increased by permitting two vehicles per green. In this case, the cycle length should be increased to 6–6.5 s of green, and the maximum discharge rate then increases to 1,100 or 1,200 vehicles per hour. Another technique employed at high-volume ramps is to widen the ramp to two or more lanes at the meter and permit one or two vehicles per lane per green.

Ramp metering has been used in the USA since the early 1960s, and since then, numerous ramp meters have been deployed in metropolitan areas around the world. Based on the experience gained so far, ramp metering systems have the potential to offer the following benefits:

- Smoother flow of traffic
- Increase in vehicle throughput
- Increase in average speed along the freeway
- Reduction in the total number of crashes and crash rate (especially rear-end and sideswipe collisions)
- Reduction of vehicle emissions
- Reduction of fuel consumption

The following are potential disadvantages in the use of ramp metering:

- Traffic diversion: Motorists may use parallel facilities, such as arterials, to avoid ramp meters.
- Equity: Many argue that ramp metering favors suburban motorists who make longer trips versus those who make shorter trips and live within the metered zones.
- Socioeconomic considerations: Ramp metering may shift traffic congestion and associated impacts from one location to another.
- Application: If not applied correctly, they may result in worsening of traffic conditions.

Based on their level of demand responsiveness and scope of operation, ramp metering systems can be characterized as [9]:

Pre-timed control: The metering rate is fixed and usually is determined using historic volume data. Meters are activated on preset schedules, and thus, those systems are best for addressing recurring congestion, but not special events and incidents. Such systems may involve a series of on-ramps.

Local traffic responsive control: The metering rate is based on detection near the ramp (immediately upstream or downstream of the ramp or at the merge point, as shown in Fig. 8.10). In the case of a system with multiple ramps, each ramp operates based on the local demand.

Area-wide (or system-wide) traffic responsive control: The metering rate is obtained considering the entire system in order to optimize traffic flow along a stretch of roadway, rather than at a specific location. Therefore, the metering rates at different ramps are interrelated. Those types of systems are the most complex, and also the most effective, as they can coordinate metering rates within a system of ramps to limit demand from the mainline to the critical ramp(s).

To successfully implement ramp metering, the analyst should consider the following:

1. *Where are the bottlenecks in the system, and is ramp metering the most effective strategy*? The analyst should identify the breakdown locations along the freeway facility, identify the critical ramps (see Chap. 4 for a definition), and evaluate

whether ramp metering is a suitable strategy. Ramp metering can improve traffic operations when the cause of congestion is the demand from an on-ramp or a series of on-ramps. Another consideration is the available acceleration lane length and whether it allows vehicles that have stopped at the ramp signal to reach speeds that would allow them to merge safely.

2. *What is the extent of the area to be included in the ramp metering system*? The freeway facility should be evaluated as a system considering the most downstream bottleneck and the entire length of the facility affected by the bottleneck.
3. *What is the most suitable ramp metering strategy?* Depending on the extent of the freeway facility to be considered, the analyst should select a local or an area-wide system. Also, depending on the demand patterns and the significance of infrequent or unplanned events, the analyst should select a pre-timed versus a demand responsive system.
4. *What is the most suitable ramp metering algorithm?* There are several different types of ramp metering algorithms that have been implemented around the world. Each type of algorithm uses different measures (speed, occupancy, or volume) and slightly different assumptions on what is the corresponding threshold to be used. To some degree, the selection of the algorithm depends on the types and locations of sensors already available along the freeway facility, as well as the overall objectives of the system. To date there are no specific guidelines developed to assist transportation professionals with the selection of a suitable ramp metering algorithm.
5. *What is the maximum queue that can be tolerated at each metered ramp?* In order for the ramp metering system to be successful, it should be able to weigh the relative importance of uncongested freeway operations versus vehicle delays and queues at the on-ramps. Suitable thresholds should be set by the analyst to achieve the desirable balance.

Variable Speed Limit Systems

Variable speed limits may be implemented either for safety reasons (i.e., during adverse weather conditions) or to alleviate congestion. VSLs have been implemented in numerous areas throughout the USA and are widespread throughout Europe. Most of the VSL systems in the USA have been implemented to address adverse weather conditions; however, several of the European systems have been implemented to smooth flow and reduce congestion-related crashes [13, 14]. Several studies have shown that mean speeds decrease when VSL is implemented, indicating that the VSLs do affect the speed at which motorists drive. Several studies have shown the speed standard deviation to decrease as well and that decrease has been associated with safety benefits. There has been little evidence to suggest that implementing VSLs has the potential to increase capacity. Congestion-related benefits have been shown mostly using simulation. However, safety benefits have been documented for several of the systems.

Table 8.1 Orlando I-4 VSL system thresholds

	Occupancy for decreasing speed limit (%)	Occupancy for increasing speed limit (%)	Speed limit (mph)
Free flow	<16	<12	50
Light congestion	16–28	12–25	40
Heavy congestion	>28	>25	30

Different algorithms have been developed based on the purpose of the VSL. For example, the VSL algorithm implemented along I-4 in Orlando, Florida, was implemented to alleviate congestion, and it is based on occupancy. The system uses in-ground inductive loops to measure traffic speed, volume, and occupancy at selected locations along I-4 [14]. The speed displayed on the VSL signs depends upon the traffic occupancy level observed by these inductive loops. Each sign is linked to selected downstream detectors. The algorithm selects the appropriate speed limit based on three categories of traffic: free, light, and heavy. Table 8.1 provides the thresholds used by the I-4 system to set variable speed limits. The system recommends an increase or decrease in speed based on the current occupancy level. Two different sets of thresholds are used (one for decreasing occupancy and one for increasing occupancy) in order to avoid too frequent fluctuations between speed limits. In order for the software to recommend a change between categories, the occupancy level must be sustained and observed for at least 120 consecutive seconds.

Other VSL algorithms have been developed based on flows, speeds, and combinations of these. For algorithms developed for congestion mitigation, VSL signs are almost always associated with downstream detectors to decrease flow entering a congested area. Algorithms based on weather or road condition parameters usually deal with VSLs associated with adjacent detectors. In both cases, it is common to gradually lower the speed limit in increments of 5 or 10 mph. Most algorithms also use a safety measure that prevents adjacent signs from having more than a 10 mph difference between them. In addition, nearly all systems use a mechanism to prevent hysteresis, or rapid fluctuation in the displayed speeds. Some systems use minimum time durations, and others use reverse thresholds to avoid this.

There are several different types of variable speed limit signs utilized. The signs can be categorized into two groups: overhead signs and roadside signs. Either of these technologies can be accompanied by changeable message signs or flashing beacons displaying a “Reduced Speed When Flashing” message. Figure 8.11 illustrates an overhead VSL installation in Seattle, Washington. Each lane has a display of the reduced speed limit.

One of the important issues in implementing VSLs which is crucial to their success is whether drivers will obey the speed limit signs. Research has shown [13, 14] that drivers tend to travel at their desired speed whenever there is no enforcement. Automated enforcement, highway patrol enforcement, and signs that display drivers’ speeds have all shown to be effective enforcement strategies. In Europe, most systems have had a positive response from drivers, and previous

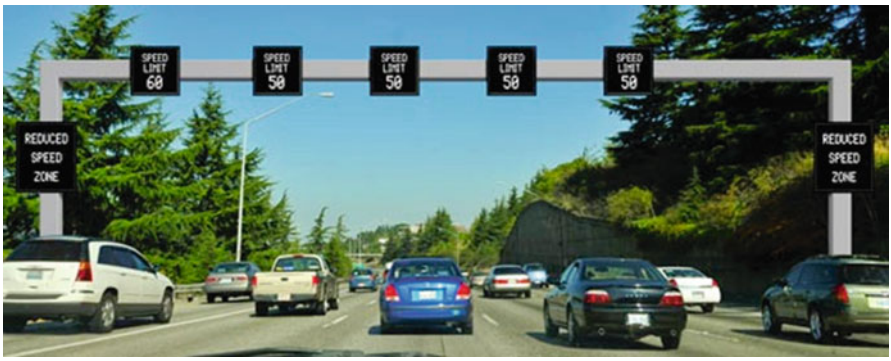


Fig. 8.11 VSL signs in Seattle, Washington (Photo by Washington State Department of Transportation, http://www.wsdot.wa.gov/smarterhighways/vsl.htm; Reproduced with permission of the Washington State Department of Transportation) [15]

studies have concluded that drivers are more accepting of these systems if they know why they are implemented. The effectiveness of a VSL system is dependent on the driver's acceptance of the system. Gaining increased compliance of variable speed limits can be accomplished through some method of enforcement, or by making drivers aware to the specific strategies of VSL implementation. Research also suggests that gradual speed limit reduction is more effective than sudden speed reduction.

Research has not yet shown conclusively what the impacts of the VSL are, nor are there guidelines regarding the installation of VSL. We don't know under what conditions and for what bottleneck configurations they are effective. Several VSL algorithms are reported in the literature, but it is not clear what are their specific advantages and disadvantages, nor how the respective thresholds should be set. Some researchers have claimed that VSL reduce demand to the bottleneck and thus postpone breakdown. However, the reduction in speeds may adversely affect traffic operations and may result in overall higher travel times, especially if the system is implemented prematurely. Thus, there are still many questions regarding the operational effects and the implementation of VSL. Finally, VSL effectiveness hinges upon driver compliance—if the system is not correctly implemented and it is not enforced, drivers won't comply.

HOV/HOT Lanes

High-occupancy vehicle (HOV) lanes are those designated only for use by vehicles with more than one traveler (some are designated for 2+ people, while others for 3+). They were originally implemented to encourage carpooling and alleviate congestion. However, carpooling rates in the USA have steadily declined during the past 20 years [16], and thus, HOV lanes have often been underutilized. More recently,

agencies have begun converting HOV lanes into high-occupancy toll (HOT) lanes. Those lanes, in addition to allowing high-occupancy vehicles and transit, charge other users for traveling in them while guaranteeing a minimum level of service. The first such facility was installed in California in 1995. As of this writing, there are approximately ten HOT lane installations around the USA. Some of these facilities operate by setting tolls using historical data, while others are dynamic and they set tolls as a function of prevailing traffic conditions. Several algorithms have been reported in the literature for setting the HOT lane tolls dynamically [17].

Use of Shoulder

The use of shoulders during peak periods has been tested in the USA and is fairly common in Europe [18]. In the USA, several states (Florida, Massachusetts, Minnesota, Virginia, Washington) have used the right- or the left-side shoulder to increase capacity. In some cases (Florida, Minnesota), the use of shoulder was implemented as part of a managed lanes project, as an HOV/HOT. In other cases (Washington), it is a permanent installation, while in others (Massachusetts), it is a first step toward expanding the facility. In some installations, use of shoulder is restricted to buses. Based on data from the existing installations in the USA, use of shoulders does not decrease safety, while it has the potential to significantly increase capacity. Concerns relate to the availability of access for emergency vehicles in case of an incident, as well as maintenance costs.

European installations combine use of shoulder with VSL and restrict speeds upstream of the shoulder opening to smooth flow and reduce the probability of incidents. Such systems have reported capacity increases in the order of 7–22 % depending on the configuration [18]. Great Britain, Germany, and the Netherlands have implemented systems where the temporary shoulder lane use is deployed only in conjunction with variable speed limit strategies, and it is activated in the field only after the speed limit has been reduced [18]. Overhead gantries and dynamic message signs provide travelers with information on reduced speed limits and the availability of the shoulder lane for travel. Emergency refuge areas are provided at regular intervals. In Germany, shoulder use may be discontinued when approaching an on-ramp, to allow on-ramp approaching vehicles to enter the freeway unencumbered (see Fig. 8.12). This strategy may be very effective when there is heavy demand from the on-ramp but the mainline demand is not as high.

Incident Management

Incidents reportedly account for 25–30 % of total congestion delay [11]. Incident management is “the systematic, planned and coordinated use of human, institutional, mechanical and technical resources to reduce the duration and impact of traffic

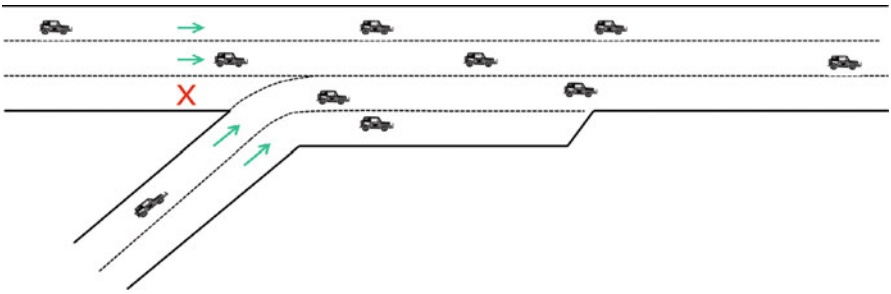


Fig. 8.12 Termination of shoulder use upstream of an on-ramp [18]

incidents, and improve the safety of motorists, crash victims, and traffic incident responders" [11].

Traffic incident management is often the most important element of a freeway management program. Incident management requires the cooperation of many different agencies including law enforcement, emergency medical services, and various transportation agencies. The objectives of incident management are to detect incidents as quickly as possible, respond in a timely and appropriate manner by notifying the appropriate parties, and clear the incident and restore traffic as quickly as possible. Reference [19] provides a thorough discussion of issues and challenges related to incident management and an overview of best practices in the USA.

Freeway Analysis Methods

In freeway analysis, the types of questions we attempt to answer are:

What should the number of lanes be so that the freeway operates at a particular LOS?
What is the best strategy for alleviating a freeway bottleneck?
How will a given design change affect operations?

This section summarizes the two most frequently used techniques for analyzing freeway segments and freeway systems: the Highway Capacity Manual methods and microsimulation.

The HCM Analysis Methods for Freeways

The HCM 2010 [3] contains analysis methods for evaluating basic freeway segments, freeway weaving segments, merge and diverge segments, as well as freeway facilities (or freeway systems). The methodologies for evaluating each segment type are most appropriate for use when conditions are non-congested. For those, analysis is completed for one analysis period. For congested conditions,

Table 8.2 Level of service boundaries for basic freeway segments and freeway facilities

Level of service	Density (pc/mi/ln)
A	≤11
B	>11–18
C	>18–26
D	>26–35
E	>35–45
F	>45 or $v/c > 1.0$

it is recommended using the freeway facilities method which takes into account the impacts of congestion in time and space. Freeway facilities are analyzed for several time periods such that the entire duration of congestion is included.

The methodology for analyzing basic freeway segments is based on a series of speed–flow curves (Fig. 3.7); each curve represents a particular free-flow speed level. The analyst first determines the free-flow speed (either through field data or using an equation provided in the HCM). Based on the selected free-flow speed, the analyst selects an appropriate speed–flow curve. Based on this curve and given a specified demand, we can obtain the operating speed and then calculate the resulting density (see Example 3.5). The LOS can be obtained using Table 8.2.

In order to use the speed–flow curves, the demand needs to be converted to passenger cars per hour. The HCM 2010 provides Passenger Car Equivalents (PCEs) for the "typical" truck, which establish quantitative equivalencies between trucks and passenger cars. For example, for level terrain, the impact of a typical truck is estimated to be equal to that of 1.5 passenger vehicles; thus, the truck PCE in those conditions is 1.5. In order to convert vehicles into passenger cars, the HCM provides PCE values as a function of terrain (grade and length of grade) and percent of trucks. It also provides formulas for converting flow in vehicles per hour into flow in passenger cars per hour.

For weaving and merge/diverge segments, the HCM 2010 estimates the speed and density within their influence areas. For weaving segments, operating speeds are obtained as a function of the lane changes estimated to occur. For merge and diverge segments, operating speeds are estimated considering the demand and capacity of the two rightmost lanes of the segment.

As indicated earlier, analysis for freeway systems should include the entire time period of congestion. The HCM 2010 recommends that the first and last time periods of analysis are non-congested; it also recommends that the first upstream and last downstream segment of the facility analyzed do not get congested. The methodology is based on the estimation of the capacities of each segment along the facility, considering lane closures and other such conditions which are not analyzed with the segment methods. The HCM 2010 defines the LOS of the freeway facility based on the weighted average density of all its segments. The densities are weighted on the basis of segment length and number of lanes. The criteria of Table 8.2 apply for defining LOS at freeway facilities as well.

Simulation for Freeway Systems

Freeway facilities are often analyzed using microsimulation. There are several tools that are capable of analyzing freeway networks. The strength of simulation is that it can analyze various types of configurations (e.g., toll facilities) which the HCM 2010 cannot address. Some tools can even evaluate freeway and arterial systems in combination, considering demand adjustments as a function of operating conditions. Stochastic microsimulators can consider the randomness inherent in traffic operations due to the differences in driver behavior as well as the wide variability in vehicle capabilities.

However, simulation tools are not as rigorous as the HCM in considering geometric effects such as lane and shoulder widths. Many of these tools allow the analyst to adjust the free-flow speed in order to account for such geometric effects. In those cases, the HCM is an excellent source as it provides an equation which estimates free-flow speed as a function of geometry. The user should be cautious when using such an equation, to ensure that none of the variables are already considered within the simulation package. For example, in the HCM 2010, free-flow speed is estimated using the number of ramps per mile; however, in the simulator, the ramps are directly replicated in the network and their effect is already considered in this manner.

References

1. American Association of State Highway Transportation Officials (AASHTO) (2011) A Policy on Geometric Design of Highways and Streets, 6th edn
2. FHWA, Highway Statistics 2010. http://www.fhwa.dot.gov/policyinformation/statistics/2010/hm20.cfm
3. Transportation Research Board, National Academies of Science (2010) Highway capacity manual 2010. Transportation Research Board, National Academies of Science, Washington, DC
4. Courage K. Implementation and evaluation of a moving merge control system in Tampa. FHWA-RD-78- 29 Final Report, Apr 1979, p 99
5. Kondyli A, Elefteriadou L (2011) Modeling driver behavior at freeway-ramp merges. Transportation Research Record: Journal of the Transportation Research Board of the National Academies, No 2249, Washington, DC, pp 29–37, Jan 2011
6. Elefteriadou L, Roess RP, McShane WR (1995) The probabilistic nature of breakdown at freeway—merge junctions. Transportation Research Record 1484. National Academy Press, pp 80–89
7. Brilon W, Elefteriadou L, Kondyli A, Hall F, Persaud B, Washburn S (2011) Proactive ramp metering based on breakdown probabilities. In: 6th international symposium for highway capacity, Stockholm, June 2011
8. Coifman B, Mishalani RG, Wang C, Krishnamurthy S (2006) Impacts of lane change maneuvers on congested freeway segment delays, pilot study. Transportation Research Record No. 1965, pp 152–159
9. Ci Y, Wu L, Pei Y, Ling X (2009) Gap acceptance capacity model for on-ramp junction of urban freeway. J Transp Syst Eng Inf Technol 9(4):116–119
10. Lertworawanich P, Elefteriadou L (2003) A methodology for estimating capacity at ramp weaves based on gap acceptance and linear optimization. Transp Res Part B Methodol 37B (5):459–483

11. Freeway management and operations handbook. FHWA-OP-04-003, 2003. Federal Highway Administration (FHWA), Washington DC, 564 pages
12. Ramp management and control handbook. FHWA-HOP-06-001, Jan 2006. Federal Highway Administration (FHWA), Washington DC, 342 pages
13. CTC and Associates LLC (2003) Variable speed limit signs for winter weather. transportation synthesis report. Bureau of Highway Operations, Division of Transportation Infrastructure Development, Wisconsin Department of Transportation (CTC and Associates LLC, http://wisdotresearch.wi.gov/wp-content/uploads/tsrwintervariablespeedlimitsigns1.pdf)
14. Elefteriadou L, Washburn S, Yin Y, Modi V, Letter C (2012) Variable speed limits: best management practice. Final Report BDK77-TWO977-11. Florida Department of Transportation, Aug 2012
15. Robinson MD (2000) Examples of variable speed application. Prepared for the speed management workshop at Transportation Research Board, 79th annual meeting, 2000. http://safety.fhwa.dot.gov/speedmgt/vslimits/docs/vslexamples.ppt. Accessed 26 Mar 2010
16. Latta J (2011) HOV vs HOT: tweaking old fashioned busy lanes could smooth flow. Better Roads 81(4):43–45
17. Michalaka D (2012) Enhancement and evaluation of dynamic pricing strategies of managed toll lanes. Ph.D. Dissertation, University of Florida, Aug 2012
18. Kuhn B (2010) Efficient use of highway capacity summary. Report No. FHWA-HOP-10-023. Texas Transportation Institute, Texas A&M University, College Station, TX
19. Carson J (2010) Best practices in traffic incident management. Report No. FHWA-HOP-10-050. Texas Transportation Institute, Texas A&M University, College Station, TX

Problems

1. Conduct a literature review regarding advanced traffic management for freeway systems. What types of strategies are available internationally? Are there specific guidelines for implementing a specific strategy under specific conditions?
2. Conduct a literature review to document the interactions between VSL, ramp metering, and HOV/HOT lanes.
3. Conduct a literature review to identify simulation packages that evaluate freeway systems. Document the manner in which each simulator replicates the network, merging operations, and upgrade sections. What are the advantages and disadvantages of each approach?
4. A freeway merge section has three lanes of traffic along the mainline with demand equal to 5,200 vph. The ramp has one lane and demand equal to 750 vph. There are 10 % trucks in the traffic stream, and the free-flow speed is 65 mph. The terrain is level. Conduct an HCM 2010 analysis to determine the LOS at this section. Also, use your favorite microsimulator package to replicate the same facility. What is the LOS estimated by this microsimulator?
5. Estimate the operating speed of a freeway with the following characteristics: three lanes per direction, free-flow speed 68 mph, peak hour demand 6,350 vph, and 5 % trucks with PCE 2.0. What is the LOS of the facility, and what is its capacity? Is capacity exceeded during the peak hour? Clearly state any assumptions necessary for solving the problem. (To solve this problem, use Chap. 11 of the HCM 2010.)

Chapter 9
Signalized Intersections and Networks

Signalized intersections operate with the assistance of a traffic signal, and they cyclically assign the right-of-way to a movement or combination of movements. Ideally, the amount of time assigned to each combination of movements minimizes travel time and delay and/or the number of stops in the network and allocates capacity optimally. An arterial is a highway facility with a series of signalized intersections along its length, while an urban network or surface street network consists of a group of interconnected arterials and side streets.

This chapter focuses on the fundamental relationships between traffic and signal control and on the traffic operational aspects of signalized intersections, arterials, and urban networks. It does not discuss the details of signal control design or signal control equipment and operation. The interested reader can consult [1] for detailed information on signal design strategies and implementation.

The chapter first presents signal control principles, including the definition of key terms used in signal control, the fundamentals of signalization, and the relationship between traffic demand, signalization, capacity, and delay. The second section discusses the design of signalization patterns and their optimization for an isolated intersection. The third section discusses the operation of signalized arterials and networks, while the fourth section provides a brief overview of signal control analysis and optimization methods. The last section summarizes current development efforts related to advanced technologies for signal control optimization.

Signalization Principles and Traffic Operations

At signalized intersections, conflicting traffic streams compete for the common intersection area. The total amount of time available within an hour should be allocated wisely to maximize capacity and minimize delays through the intersection. The operational quality of each movement depends on the green time allocation and thus indirectly on the demand of the conflicting movements.

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_9,

This section first defines some of the most important terms used in signal control and then presents capacity and delay estimation methods at a signalized intersection approach.

Key Terms and Their Definitions

The following are definitions of some of the most important terms used in signal control:

- *Cycle length* (C): It is the length of time required to go through all the signal indications at the signalized intersection and serve all movements in all approaches. Its length is typically 60–120 s, although much longer cycles are provided at more complex and heavily used intersections.
- *Phase*: It is the part of the cycle serving a movement or group of movements, and it includes the green, yellow, and all-red intervals.
- *Interval*: It is the part of the cycle where all signal indications remain the same. There is a green, yellow, and all-red interval for most phases through the cycle.
- *Change and clearance intervals*: Those are the yellow (change) and all-red (clearance) intervals which are used to transition the assignment of the right-of-way to the next phase. The yellow interval (Y) is provided to warn approaching drivers that the signal is about to change to red, and its duration is a function of the approach speed (typically 2–4 s). The all-red interval (AR) is of relatively short duration (1–3 s), and it is provided to ensure the intersection has cleared from all vehicles from the subject phase before the right-of-way is given to the next phase. The length of those intervals is set to ensure the safe operation of the intersection, and it is estimated as a function of the design of the intersection. The yellow and all-red intervals are estimated independently of the signal control optimization process. Additional information regarding yellow and all-red duration is provided in [1].
- *Green time* (G): It is the part of the phase when its respective movements are given the right-of-way. It should be set to minimize travel time and delay throughout the network, considering the demands of all conflicting movements of the intersection.
- *Lost time* (L_P): It is the part of the phase that is essentially not used by any movement. It consists of the start-up lost time and the clearance lost time. The start-up lost time (l_1) is the lost time experienced by vehicles at the beginning of the green, and it includes the driver reaction time before vehicles start to clear the intersection. The clearance lost time (l_2) is the lost time experienced when vehicles do not use portions of the yellow and all-red intervals.
- *Effective green time* (g): It is the part of the phase effectively used by the respective movements, considering the lost time experienced at the beginning

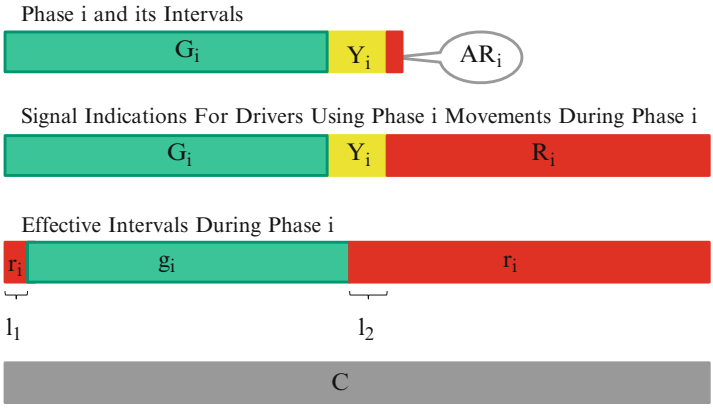


Fig. 9.1 Intervals, effective intervals, and lost time for phase *i*

of the cycle as well as the potential utilization of portions of the yellow and all-red interval at the end of the phase. It is calculated as follows:

$$\boxed{g_i = G_i + Y_i - (l_{1i} + l_{2i}) = G_i + Y_i - L_P} \tag{9.1}$$

where

g_i is the effective green time for phase i
G_i is the green time for phase i
Y_i is the yellow interval time for phase i
$l_{1i} + l_{2i}$ is the sum of the lost time for phase i
L_P is the total lost time for phase i

- *Effective red time* (r): It is the part of the phase that is not effectively used by the respective movement(s). It can be estimated as:

$$r_i = C - g_i$$

Figure 9.1 provides a graphic representation of the terms defined above. The top bar of the graph shows the actual signal indications for phase i. The length of the entire bar represents the phase duration. The second bar provides the actual signal indications shown to drivers in movements served during phase i. The entire length of the bar is the cycle length of the intersection. The third bar shows the effective green as well as the effective red for phase i. The start-up lost time l_1 and the clearance lost time l_2 are also shown on this bar. The effective red includes the lost time for the phase, as the signal is not effectively used by any movement during that time. The last bar shows the cycle length for the intersection.

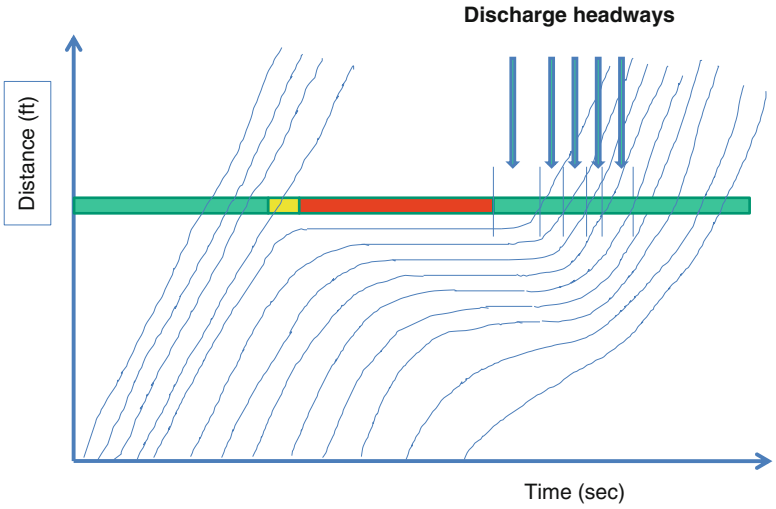


Fig. 9.2 Vehicle trajectories and discharge headways

Capacity of a Signalized Intersection Movement

The amount of traffic that can be processed through a signalized approach (i.e., capacity) is a function of the duration of the green time for the respective movement. To understand the movement of traffic through a signal, let us start by examining the trajectories of vehicles traveling through a signalized intersection for a particular movement. Figure 9.2 shows the time–space diagram of vehicle trajectories as they approach a signalized intersection. As shown, when the signal turns red, vehicles queue behind the stop bar. As soon as it turns green, they accelerate and depart the intersection. The time headways of vehicles departing the stop bar from a queue are called discharge headways.

When examined microscopically, the discharge headways follow a distinctive pattern. Figure 9.3 provides a typical sketch of discharge headways. The horizontal axis in the figure represents consecutive vehicles discharged, while the vertical axis represents the time headway between the subject vehicle and the previous vehicle. The first discharge headway in the graph is the time between the beginning of the green and the discharge of the first vehicle. This first discharge headway is typically the largest among the first several vehicles, as it includes a large portion of the start-uplost time (i.e., the extra time it takes the first vehicle to react to the change in the signal indication). The next few discharge headways include smaller portions of the start-up lost time, as vehicles accelerate to cross the stop bar at the intersection. Discharge headways decrease gradually until they stabilize around a nearly constant number. This nearly constant discharge headway is called the saturation headway, s_h, and it is defined as the time headway observed under conditions of continuous queuing after the first few vehicles in the queue have departed. It is

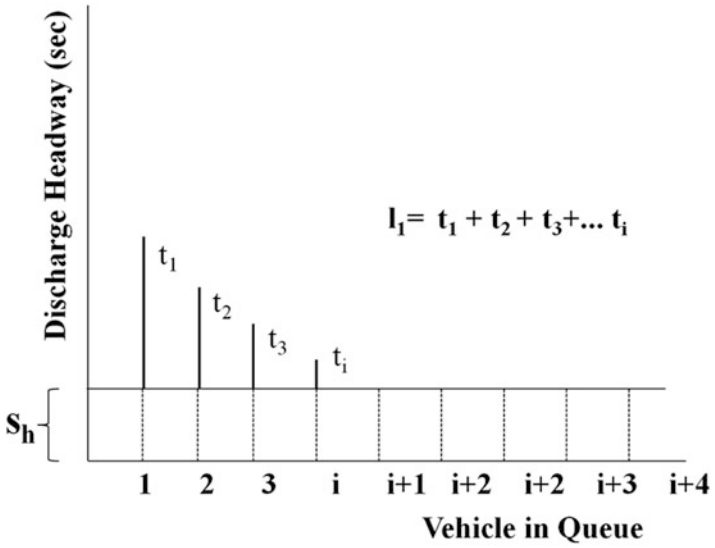


Fig. 9.3 Discharge headways from a signalized approach

estimated as the nearly constant average headway between vehicles occurring after the *N*th vehicle in the queue and continuing until the last vehicle in the initial queue has cleared the intersection. Its length is a function of the type of movement (left, through, or right), as well as the design of the intersection and the traffic stream characteristics. Subtracting the saturation headway from each of the first few time headways provides the start-up lost time (l_1), which is the sum of the lost times experienced by each of these first few vehicles in the queue.

Based on the saturation headways, which represent the minimum headway for the movement, one can obtain the saturation flow (S) which is defined as the maximum throughput for the signalized intersection approach if the approach were given the green for a full hour. The saturation flow for a single lane is calculated as:

$$\boxed{S = 3,600/s_h} \tag{9.2}$$

where

S is the saturation flow (veh/h of green/ln)
s_h is the saturation headway (s/veh)

Because traffic flow approaching a signalized intersection is regularly interrupted to serve conflicting traffic, this saturation flow does not represent capacity, but rather it can be used to estimate capacity. The capacity (c) of the approach is a function of

the amount of the effective green (g) given to the respective movements within a given time interval. Mathematically:

$$\boxed{c = S\frac{g}{C}} \tag{9.3}$$

where

S is the saturation flow (veh/h of green/ln)
g is the duration of effective green for the approach (s)
C is the cycle length for the intersection (s)

The g/C is a very important parameter for a signalized intersection approach, as it indicates the percent of time a movement or group of movements can effectively use the intersection. The sum of all the g/C ratios for all phases at a signalized intersection should be less than 1.0 (considering also the lost time throughout the cycle).

Equation (9.3) is an important relationship in signalized intersection analysis, and we will use it extensively in the remainder of this chapter; thus, its terms and function should be thoroughly understood before proceeding to subsequent sections of the chapter.

Example 9.1 A series of vehicles are discharged from a signalized intersection approach as shown in Fig. 9.4, where vehicle departures are represented by the vertical lines along the time axis (horizontal line). The cycle length at this intersection is 100 s. At the beginning of the green, there were ten vehicles queued at the approach. As shown in Fig. 9.4, the volume during this green interval is 16 vehicles. Determine the saturation headway, the saturation flow rate, the start-up lost time, and the capacity of the approach.

Solution to Example 9.1
As shown in Fig. 9.4, the discharge headways stabilize after the third vehicle, and the discharge headways for the third through the tenth vehicle are 2 s/veh. Thus, the saturation headway for the approach is 2 s/veh.

The saturation flow is:

$$\begin{aligned} S &= 3,600/s_{\mathrm{h}} \\ &= 3,600/2 \\ &= 1,800\,\mathrm{veh/h\ \ of\ \ green/ln} \end{aligned}$$

The start-up lost time is the extra time experienced by each vehicle in addition to the saturation headway. Since the first ten vehicles took 25 s to clear the stop bar, and the saturation headway is 2 s/veh, the start-up lost time is:

$$25s - (10\,\mathrm{veh} \times 2s/\mathrm{veh}) = 5\,\mathrm{s}$$

The capacity of the approach is estimated as follows:

$$c = Sg/C$$

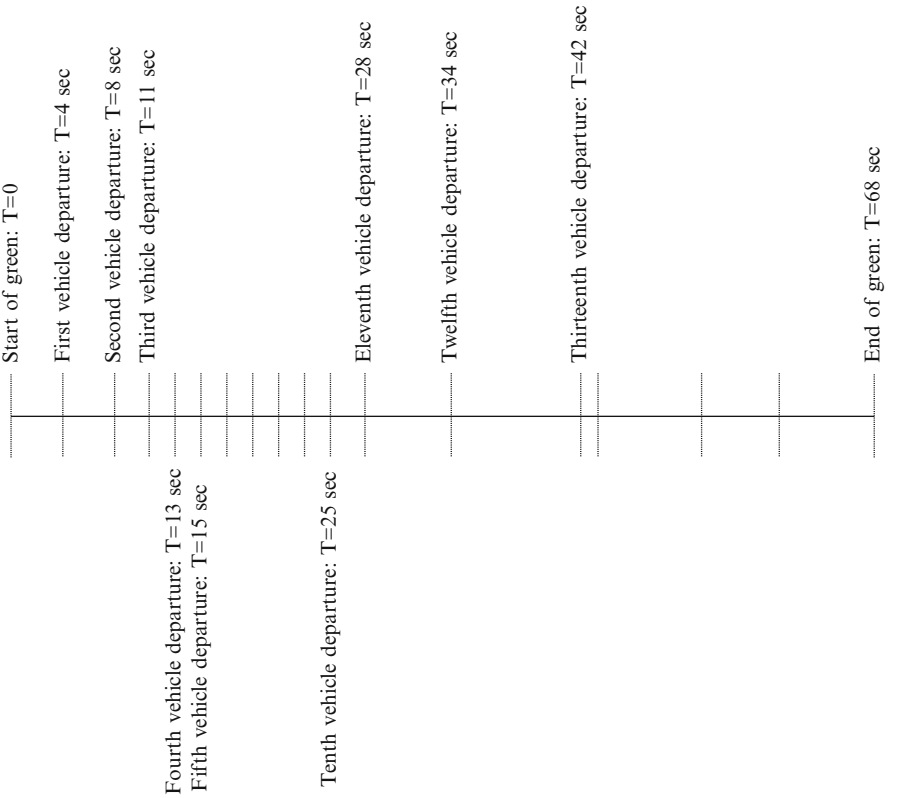


Fig. 9.4 Time headways for Example 9.1

Assuming that no portion of the yellow and all-red is used by this movement ($l_2 = 0$), the effective green time is estimated as:

$$\begin{aligned}\text{Effective green time} &= \text{actual green time} - \text{lost time}\\ &= 68\,\text{s} - (25\,\text{s} - 10\,\text{veh} \times 2\,\text{s})\\ &= 63\,\text{s}\end{aligned}$$

Using this number, the capacity of the approach is:

$$c = 1,800 \times 63/100 = 1,134\,\text{veh/h}$$

An alternative method to estimate capacity within the hour is to estimate the capacity per cycle and multiply by the number of cycles within the hour. In this case, the capacity per cycle is:

$$\begin{aligned}\text{Capacity per cycle} &= \text{effective green/saturation headway}\\ &= 63\,\text{s}/2\,\text{s}\\ &= 31.5\,\text{vehicles per cycle}\end{aligned}$$

The total number of cycles in the hour is:

$$\text{Number of cycles} = 3,600/100\,\text{s} = 36\,\text{cycles per hour}$$

Thus, the capacity within an hour is:

$$\begin{aligned}\text{Capacity} &= \text{capacity per cycle} \times \text{number of cycles} \\ &= 31.5\,\text{vehicles per cycle} \times 36\,\text{cycles} \\ &= 1,134\,\text{veh/h}\end{aligned}$$

Delay at a Signalized Intersection Approach

As discussed in Chap. 5, delay is a key measure of performance, particularly for signalized intersections. The HCM 2010 [2] bases level of service (LOS) designations for signalized intersections on control delay. Figure 10.5 illustrates the arrivals and departures at a signalized approach. The horizontal axis represents time and illustrates the effective green and red durations for the approach. The vertical axis represents the cumulative number of vehicles. The graph assumes uniform arrivals and departures. When the signal is green, each vehicle that arrives can depart the intersection. When the signal turns red, the rate of arrivals remains unchanged, but the number of departures becomes zero. When the signal turns green again, the initial rate of departures is equal to the saturation flow rate until the queue clears; at that time, departures become again equal to the arrivals.

If we draw a horizontal line across the graph of Fig. 9.5, for each vehicle n, the time between its arrival and its departure represents the delay incurred due to the signal. If we draw a vertical line, we can obtain the queue length at a particular time. The area of the triangle formed between the arrivals and the departures during the red represents the total delay for all vehicles delayed. If we know the rate of arrivals, the saturation flow rate, and the effective green and red durations, we can estimate the total delay at the signalized approach, as well as the average delay per vehicle. The derivation is provided below.

Using geometry, the delay for all vehicles during a given cycle (i.e., the area within the triangle) is:

$$\boxed{\text{DL} = 0.5RN_\text{d}} \tag{9.4}$$

where

R = duration of red interval (s)
N_d = total number of vehicles encountering delay during the cycle

Next we convert the terms R and N so that they can be expressed as a function of signalization parameters such as cycle length and g/C, as well as hourly demand.

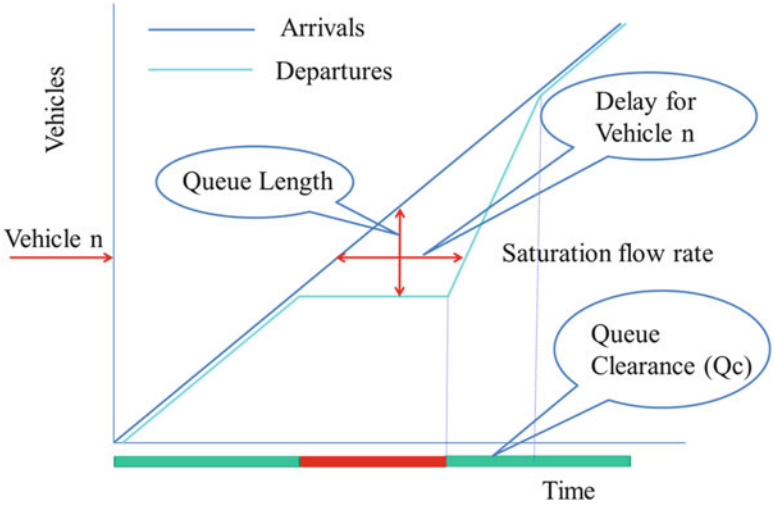


Fig. 9.5 Arrivals, departures, and delay at a signalized approach

The duration of the red interval can also be written as a function of the cycle length and the effective green time:

$$\boxed{R = C - g = C[1 - g/C]} \tag{9.5}$$

The number of vehicles encountering delay during the cycle is:

$$\boxed{N_{\mathrm{d}} = V(R + Q_{\mathrm{c}})} \tag{9.6}$$

where

Q_{c} is the time it takes for the queue to clear the approach, in seconds (illustrated in Fig. 9.5)
V is the arrival rate

To estimate Q_{c}, we use the observation that during the time $R + Q_{\mathrm{c}}$, the total number of vehicles arriving is equal to the total number of vehicles departing:

$$\begin{aligned} V(R + Q_{\mathrm{c}}) &= SQ_{\mathrm{c}} \Rightarrow \\ R + Q_{\mathrm{c}} &= (S/V)Q_{\mathrm{c}} \Rightarrow \\ R &= (S/V)Q_{\mathrm{c}} - Q_{\mathrm{c}} \Rightarrow \\ R &= Q_{\mathrm{c}}(S/V - 1) \Rightarrow \\ \boxed{Q_{\mathrm{c}} &= R/[S/V - 1]} \end{aligned} \tag{9.7}$$

Substituting Eq. (9.7) into Eq. (9.6):

$$N_d = V[R + R/(S/V - 1)] \Rightarrow$$
$$N_d = VR[1 + V/(S - V)] \Rightarrow$$
$$N_d = R[SV/(S - V)]$$

Substituting Eq. (9.5) into this last equation, we have:

$$\boxed{N_d = C[1 - g/C][SV/(S - V)]} \tag{9.8}$$

Finally, substituting Eq. (9.5) and Eq. (9.8) into Eq. (9.4), we can estimate the total delay as follows:

$$\boxed{\mathrm{DL} = 0.5C^2[1 - g/C]^2[SV/(S - V)]} \tag{9.9}$$

This value represents the total delay for all vehicles that were delayed (in seconds or hours). To estimate the delay per vehicle, we divide the result of Eq. (9.9) by the total number of vehicles that arrive during the cycle, $N = VC$:

$$\mathrm{DL}_{\mathrm{avg}} = 0.5C[1 - g/C]^2[S/(S - V)] \Rightarrow$$
$$\boxed{\mathrm{DL}_{\mathrm{avg}} = \frac{1}{2}C\frac{\left[1 - \frac{g}{C}\right]^2}{\left[1 - \frac{V}{S}\right]}} \tag{9.10}$$

If we substitute $S[=c/(g/C)]$, we obtain Webster's delay equation, which is also the equation provided in the HCM [2] for estimating the first term of delay, called uniform delay. This term, labeled d_1 in the HCM, estimates the delay that can be expected to occur if we assume that arrivals are uniform. Note that delay increases with increasing cycle length; this is an important observation we need to remember when discussing signal control optimization.

The delay Eq. (9.10) is not valid if demand exceeds capacity or if there is a residual queue at the end of the cycle. If there is congestion and vehicles stop at the signal for more than one cycle, the shape of the area is not a closed triangle anymore, and the delay cannot be estimated by this equation alone. Figure 9.6 illustrates the arrivals, departures, delay, and residual queue at the end of green for the case of oversaturated conditions. The area between the arrivals and the red line represents the additional delay incurred when demand exceeds the capacity of the approach, and it can be estimated as:

$$\text{Oversaturation delay} = T \times (\text{arrivals} - \text{departures})$$

Example 9.2 The arrival rate at a signalized intersection approach is 800 veh/h, the saturation headway is 1,800 veh/hg, the cycle length is 60 s, and the g/C is 0.50. Estimate the capacity and the uniform delay at this signalized approach.

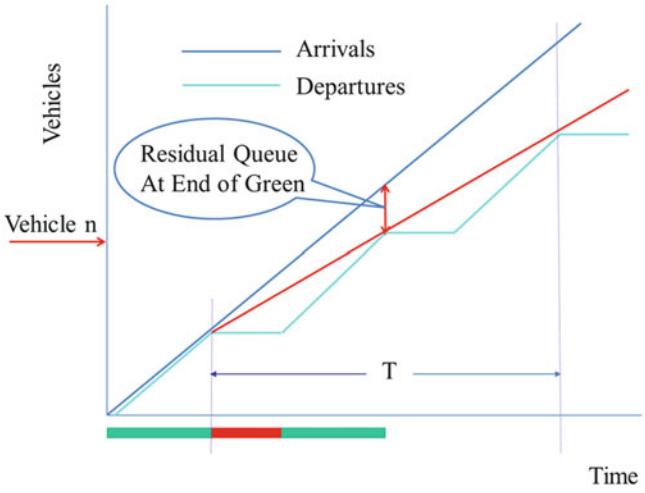


Fig. 9.6 Delay considering random arrivals and oversaturated conditions

Solution to Example 9.2
The capacity is: $c = Sg/C = 1,800 \times 0.50 = 900$ veh/h
Since the $v/c = 800/900 = 0.89 < 1$, the uniform delay can be estimated using Eq. (9.10):

$$\mathrm{DL}_{\mathrm{avg}} = \left(0.5 \times 60[(1-0.50)]^2\right)/(1-800/1,800) = 13.5\,\mathrm{s/veh}$$

The Operation of Signalized Intersections

This section focuses on the signal control optimization of single signalized intersections. The first section examines the entire signalized intersection and formulates relationships between signalization patterns at the intersection and traffic operational quality. The second section explains the functionality of actuated control, while the last section discusses a few additional elements of interest in the operation of isolated signalized intersections.

Signalized Intersection Phasing Plans and Optimal Cycle Length

Signalized intersections operate with a variety of signal phasing plans, which are developed as a function of the approach demands and the geometry of the intersection. Figure 9.7 illustrates a signalized intersection along with four different possible phasing plans. The first plan is the simplest, with two phases. In this case,

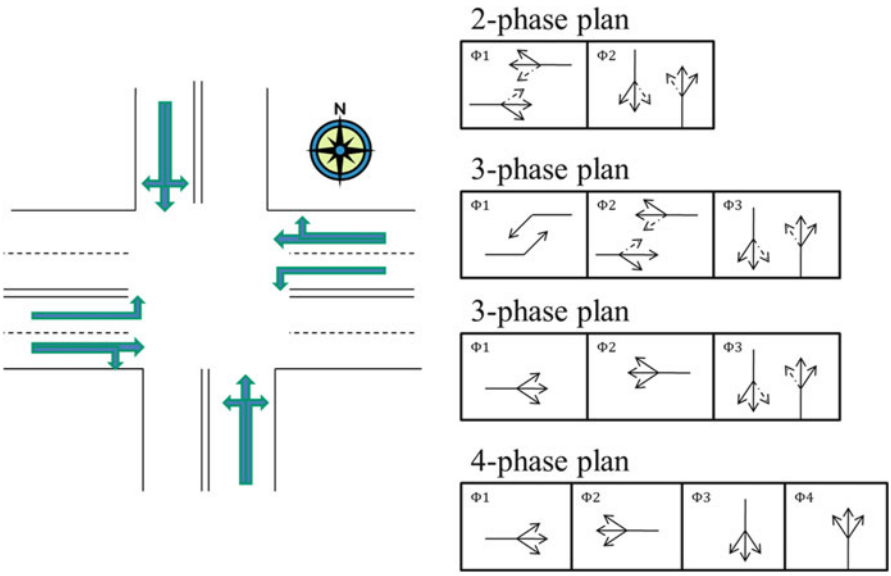


Fig. 9.7 Phasing plan options for a signalized intersection

the left turns for all movements are permitted, i.e., they can be performed with caution, using the gaps in the opposing through movement. The second plan has three phases, and in this case, the eastbound and westbound left turns are protected (i.e., they have the right-of-way) during the first phase. This type of a phase should be used when the opposing through movement does not have an adequate number of gaps to accommodate the left-turn demand. Another option for a three-phase plan is to split the eastbound and westbound movements into two completely protected phases. This option is typically preferable when the left turns from both directions are heavy, but unbalanced, i.e., one approach has heavier demand than the other, and thus requires significantly more green time than the other approach. The four-phase plan splits both the eastbound/westbound and northbound/southbound movements into their own phases. This is a less typical type of plan, reserved for major intersections where demand is heavy from most or all its approaches. An important question is, which plan is operationally more efficient? To answer this question, let us use the principles of the previous section to analyze a signalized intersection considering all its conflicting approaches.

Let us first examine a very simple signalized intersection with two one-lane approaches (four legs) and two phases, shown in Fig. 9.8. Since the two movements of the intersection are in conflict, only one movement can use the intersection at a time. The maximum amount of traffic that can use the intersection is the sum of the conflicting lane capacities. If we assume that the traffic of these two approaches behaves identically (i.e., lost time and saturation flow is the same for these two approaches), then it doesn't matter whether vehicles are coming from the northbound or the eastbound approach; their impact on intersection capacity will be the same. Using the equations of the previous section, let us estimate

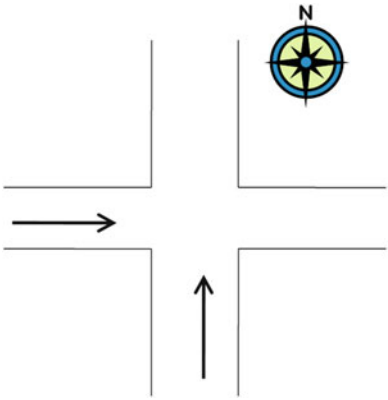


Fig. 9.8 Signalized intersection operations

the maximum throughput this intersection can handle and establish the interrelationships between that maximum throughput and intersection signalization.

Based on the terms and equations presented in the previous section, one can estimate the total lost time for the cycle, considering the two conflicting approaches and their respective phases. Assuming that the lost time for each phase is equal ($L_{P1} = L_{P2} = L_P$), the total lost time for the cycle is:

$$L_C = 2 \times L_P$$

where

L_C is the total lost time for the cycle (s)
L_P is the lost time per phase (s)

The equation can be generalized to consider any number of conflicting approaching lanes and their respective phases:

$$L_C = N \times L_P$$

where N is the total number of phases at the intersection.

We can estimate the total lost time during the hour, by calculating the total number of cycles in the hour (=3, 600/C) and multiplying that by the lost time per cycle. Then the total lost time within the hour is:

$$\text{Total lost time in the hour} = (N \times L_P)\left(\frac{3,600}{C}\right)$$

This amount of time represents the total time the intersection is not effectively used by any movement. Then, the total usable time within the hour is:

$$\text{Total usable time in the hour} = 3,600 - (N \times L_P)\left(\frac{3,600}{C}\right)$$

If we assume that the saturation headway for all approaches is the same (s_h), then the maximum number of vehicles that can be accommodated from both approaches during the cycle is:

$$V_T = \frac{1}{s_h}\left[3,600 - NL_P\frac{3,600}{C}\right] \tag{9.11}$$

where V_T is the maximum number of vehicles that can be accommodated during the hour at this intersection.

As shown in Eq. (9.11), this maximum number of vehicles is a function of the cycle length and the number of phases. When the cycle length increases, there are fewer cycles in the hour, and thus less lost time, and the throughput of the intersection increases. Conversely, increasing the number of phases at the intersection increases the total lost time, and thus, the throughput decreases. Generally plans with longer cycles and fewer phases provide the highest capacity. However, increasing the cycle length increases delay, as shown earlier in Eq. (9.10). Therefore, to optimize operations, we are interested in implementing the lowest cycle length that can handle the expected demand.

If we solve Eq. (9.11) for C, we can obtain the minimum cycle length that would accommodate this maximum volume for the intersection:

$$C_{min} = \frac{N \times L_P}{1 - \left(\frac{V_T}{3,600/s_h}\right)} \tag{9.12}$$

This equation can be used in the preliminary stages of signal design to determine the cycle length required so that the intersection can handle the total demand from all movements. The equation can be modified so that fluctuations within the hour as well as a desirable v/c can be considered. The desirable v/c should be considered so that the intersection does not operate too close to capacity, since this would likely lead to several oversaturated cycles within the hour. In this case, V_T can be replaced by $V_T/[\text{PHF} \times (v/c)]$, and the equation becomes:

$$C_{opt} = \frac{N \times L_P}{1 - \left(\frac{V_T/\left[\text{PHF}\left(\frac{v}{c}\right)\right]}{3,600/s_h}\right)} \tag{9.13}$$

where C_{opt} is the optimum cycle length for achieving the desired v/c.

Once the cycle length has been selected, the green time can be allocated proportionally to their respective demands. For the intersection of Fig. 9.8, the green times

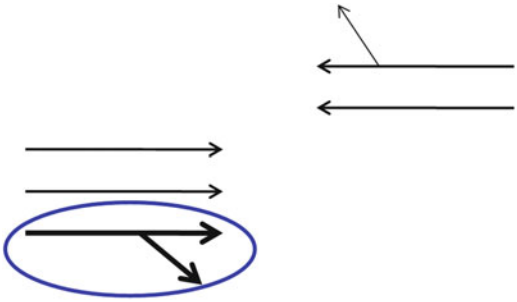


Fig. 9.9 The critical lane of a phase

for the two approaches can be estimated as follows, assuming that the yellow and all-red intervals for each phase are equal:

$$\text{Green for Approach } A = \left[C_{\text{opt}} - (Y + \text{AR}) \times 2\right] \times \frac{V_{\text{A}}}{V_{\text{A}} + V_{\text{B}}}$$

$$\text{Green for Approach } B = \left[C_{\text{opt}} - (Y + \text{AR}) \times 2\right] \times \frac{V_{\text{A}}}{V_{\text{A}} + V_{\text{B}}}$$

The above discussion assumed that there is one lane per approach and phase. How can we consider multilane approaches and multiple movements per phase? When we have multiple movements or multiple lanes per phase, we identify the most heavily used lane in the phase and use the demand of that lane as the basis of our calculations. For example, in the sketch of Fig. 9.9, the most heavily used lane is the eastbound through and right lane. We can perform the analysis outlined above using the demand of this lane to represent this particular phase. The lane is defined as the critical lane of the phase. Similarly, we can identify the critical lane for each phase and estimate the optimal cycle length and allocate green times based on the critical lanes.

The concepts presented here can also be used to determine the required number of lanes for accommodating the demand at each approach. The equations provided above can be used assuming a given number of lanes and a given per lane utilization to estimate or assume the worst lane demand. Trial-and-error analysis can then be performed to identify the required number of lanes that can accommodate a given demand, for a given cycle length and for a given number of phases.

Example 9.3 There is a signalized intersection that is to operate with three phases. The critical lane volumes for each phase are 100, 500, and 400 veh/h. The saturation headway for all lanes is 2 s/veh, while the lost time per phase is 2 s. Estimate (a) the minimum cycle length so that the intersection operates under capacity; (b) the optimal cycle length so that the demand to capacity ratio (v/c) is 0.7, assuming that the PHF is 0.85; and (c) the green times to be allocated to each of the three phases, assuming that the yellow plus all-red times for each phase is 4 s.

Solution to Example 9.3

The total demand at the intersection considering the three critical lanes is:

Total intersection demand = 100 + 500 + 400 = 1,000 vehicles per hour

The minimum cycle length is estimated using Eq. (9.12) as follows:

$$C_{\min} = \frac{3 \times 2}{1 - \left(\frac{1,000}{3,600/2}\right)} = 13.5s$$

To estimate the optimal cycle length, the volume is adjusted to consider the desirable v/c as well as the prevailing PHF. In this case the equation can be solved using the following volume:

$$\text{Adjusted volume} = \frac{Vc}{v/c\,\text{PHF}} = \frac{1,000}{0.7 \times 0.85} = 1,680\,\text{veh/h}$$

Thus, the optimal cycle length is:

$$C_{\text{opt}} = \frac{3 \times 2}{1 - \left(\frac{1,680}{3,600/2}\right)} = 90s$$

The total yellow plus all-red time in the cycle is 4×3 phases $= 12$. Thus, the total green time available in the cycle is $90 - 12 = 78$ s. Assuming that the green time is allocated proportionally to the critical lane demands of each of the three phases, the respective green time durations are:

For the phase with 100 veh/h critical lane volume: $78 \times [100/1,000] = 7.8$ s ≈ 8 s
For the phase with 500 veh/h critical lane volume: $78 \times [500/1,000] = 39$ s
For the phase with 400 veh/h critical lane volume: $78 \times [400/1,000] = 31.2$ s ≈ 31 s

The equations presented above for estimating the minimum and optimal cycle lengths assume that each movement and lane has the same saturation flow rate. However, different types of movements (left, through, and right) have different saturation headways. Furthermore, geometric design and traffic stream composition also affect the saturation flow rate. Thus, the calculations shown above should be adjusted to accommodate this difference in saturation headways.

Pre-timed and Actuated Control for Isolated Intersections

Pre-timed control is defined as that occurring when the phasing pattern and the duration of each phase are fixed. Pre-timed control is implemented based on time of day, i.e., one pattern is implemented for the am peak period, another for the pm peak, and a third one for off-peak periods. It works best when the demand patterns are relatively uniform within each of the assigned periods. However, when demands can vary significantly within a particular time period, it is best to provide additional flexibility in the duration of each time interval as well as in the phasing sequence.

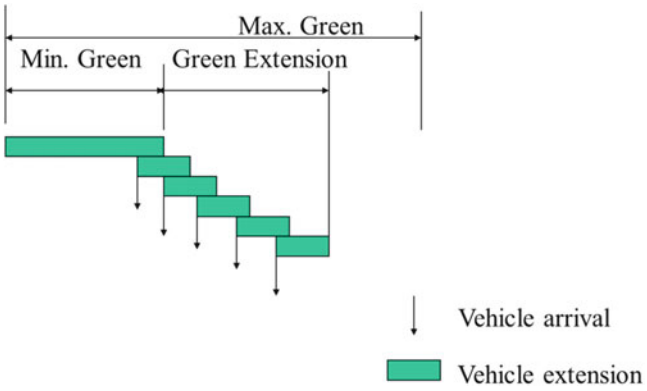


Fig. 9.10 An actuated phase

In actuated control the signal controller (i.e., the box which controls the intersection) has the capability to adjust the timings and the phasing sequence, based on a predetermined set of rules, so that time can be allocated to each phase based on the demand. In actuated control, movements that are actuated have detectors in the respective lanes, which indicate vehicle presence or passage, and can modify the respective phase. These detectors can detect the presence (or passage) of a vehicle and send a signal to the controller. This signal can extend the green according to the rules implemented in the controller.

Figure 9.10 illustrates conceptually the operation of an actuated phase. The traffic engineer sets the minimum green, the maximum green, and the vehicle extension. The minimum green is the minimum time given to the phase regardless of the demand. The maximum green is the maximum time that can be given to the phase. The vehicle extension is the time by which the controller extends the green with each vehicle actuation.

The minimum green is set considering the presence of pedestrian, as well as the time required to clear the queue between the detector and the stop bar at the intersection approach. Similarly, the vehicle extension is set to ensure the vehicle can safely clear the intersection before the green is terminated. Additional information on actuated control is provided in [1], while details on detector functions are provided in [3].

Additional Elements of Interest in Signalized Intersection Operations

When optimizing isolated signalized intersection operations, it is important to consider the following:

Left turns and left-turn bays: These are one of the key elements in the determination of the optimal phasing plan. High left-turn demand dictates exclusive left-turn

phases or splitting of the N–S or E–W phase to give the right-of-way to one direction, then the other. The presence of turn bays, or the possibility to include them, is also an important consideration; if these are not available, we cannot accommodate protected left-turn movements. When turn bays are designed, it is important to consider the anticipated maximum queue for the respective movements in order to provide adequate length and avoid spillback to the adjacent lane.

Balance of v/c and delay: As indicated above, increasing the cycle length at an intersection increases its capacity, but it also increases the delay. In signalizing an isolated intersection, the analyst should provide the minimum possible cycle length so that the demand can be comfortably met without unnecessarily increasing the delay.

Demand fluctuation: Demands may fluctuate significantly within the hour, within the day and the week, and even for the same time period on a weekly basis. The analyst should consider all types of conditions prevailing at the intersection (peak periods, special events) and develop plans that can accommodate these fluctuations, either through well-selected pre-timed plans or through the use of actuated control.

Pedestrians and bicycles: Pedestrians and bicycle demands are steadily increasing, and thus, it is important to consider and implement signal timing plans that can accommodate these users as well. Pedestrian movements are typically served concurrently with the adjacent through movement of the intersection (e.g., crossing the side street when the main street has the green). These may conflict with the right-turning or permissive left-turning movements, and the analyst should consider such conflicts and determine whether they can be mitigated through the use of exclusive pedestrian phases and other such considerations [1].

Signalized Arterials and Networks

The previous section focused on isolated signalized intersection operations. This section discusses various types of facilities with two or more signalized intersections. When there is more than one signal, we need to consider coordination in the phasing and timings between adjacent intersections, so that we can minimize the number of stops and the delay encountered by vehicles as they travel along an arterial. This section presents the principles of coordination and briefly discusses coordination at signalized interchanges, which are a special case of signal coordination.

Principles of Coordination for Signalized Arterials

Signal coordination aims to minimize stops and delay in a network, maintain emissions, and minimize storage required between successive intersections. In Fig. 9.11a, the vehicle traveling westbound along the arterial will not be stopped

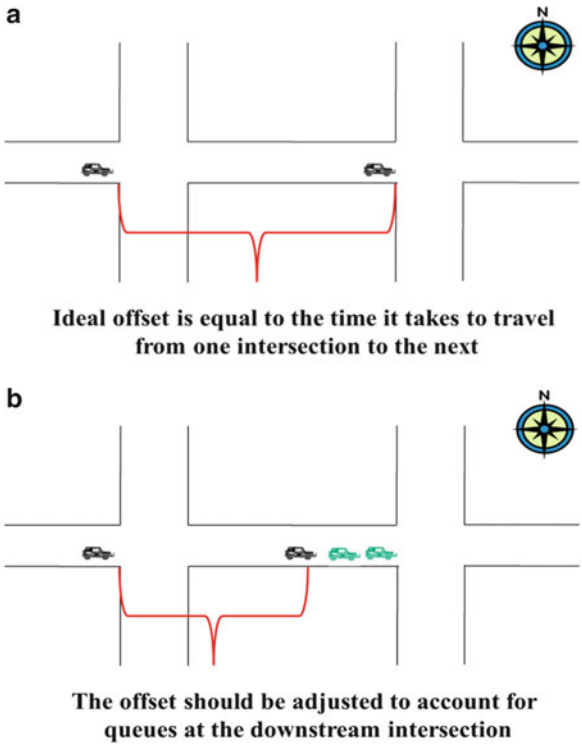


Fig. 9.11 Signal coordination and offset. (**a**) Ideal offset and (**b**) adjusted offset

at the second intersection if the green for the westbound approach starts as the vehicle arrives at the intersection. The time between the start of green at the upstream intersection and the start of green at the downstream intersection is called the offset.

We can easily estimate the ideal offset (i.e., assuming there are no other vehicles in the network) as follows:

$$\boxed{O_{\text{ideal}} = d/v} \tag{9.14}$$

where

O_{ideal} is the ideal offset, s
d is the distance between the two intersections, ft
v is the speed of the vehicle between the two intersections, ft/s

When there are other vehicles in the network, this ideal offset needs to be adjusted to account for the time it takes to clear the queue at the downstream signal. Those vehicles may be arriving from the side streets of the first

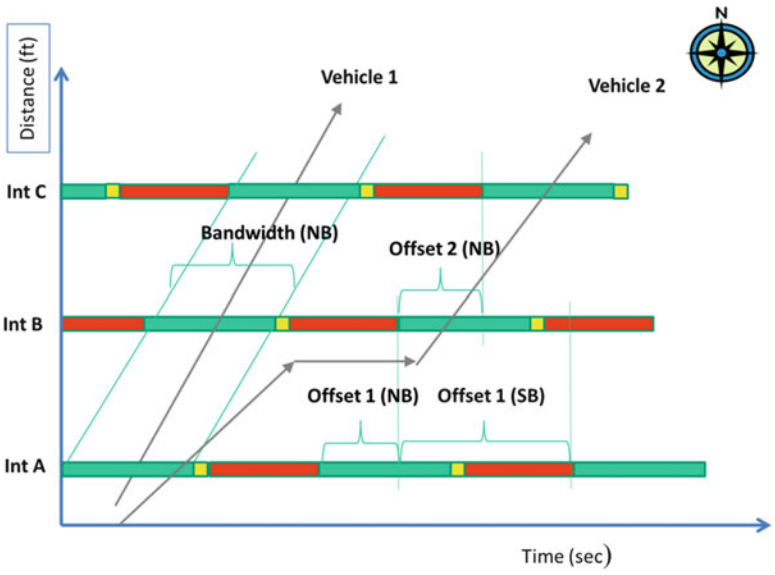


Fig. 9.12 Offsets and bandwidth in a time–space diagram

intersection or may be entering the arterial midblock. The adjusted offset is then estimated as follows:

$$\boxed{O_{\text{adjusted}} = \frac{d}{v} - (Ns_{\text{h}} + L_{\text{s}})} \tag{9.15}$$

where

O_{adjusted} is the adjusted offset, s
d is the distance between the two intersections, ft
v is the speed of the vehicle between the two intersections, ft/s
N is the number of vehicles queued at the downstream intersection
S_{h} is the saturation headway, s/veh
L_{s} is the start-up lost time, s

Figure 9.12 provides a time space diagram for three signalized intersections along an arterial street. Offset 1 NB is the offset between intersections A and B in the northbound direction, while Offset 2 NB is the offset between intersections B and C in the same direction. Offset 1 SB is the offset between intersections B and A in the southbound direction. By definition:

$$\boxed{\text{Offset 1 NB} + \text{Offset 1 SB} = \text{Cycle length}} \tag{9.16}$$

This is shown graphically in Fig. 9.12. In other words, when we set the offset in one direction, it is also set in the opposing direction.

Bandwidth, illustrated in Fig. 9.12, is defined as the time between when the first and the last vehicle can travel through the arterial at the same speed without stopping. Maximizing the bandwidth used to be the objective of signal control optimization, as it is a seemingly reasonable and easy to devise visual indicator of coordination. However, this approach does not consider overall network delay, and particularly delays to the side streets, nor does it consider the amount of traffic that enters the arterial from these side streets and results in queues that makes the estimation of an offset unpredictable (see Eq. 9.15).

Coordination between a series of signals requires that all signals have the same cycle length; otherwise, the offset is not constant from one cycle to the next. In a coordinated system, the system cycle length is set considering the maximum of the minimum cycle lengths for the intersections along the arterial (C_{maxmin}) to ensure all intersection movements operate with adequate capacity. When demand exceeds capacity for one or more arterial movements, coordination cannot help improve congestion.

In setting offsets, the analyst should consider signalized intersection spacing, prevailing speeds along the arterial and traffic volumes. Coordination is less effective when there are heavy side street turning movements, since there are fewer vehicles taking advantage of a "green wave" through the arterial. Complicated intersections with multiple phases also reduce the benefits of coordination, as they restrict the amount of green that is available for the mainline arterial movement. Substantial side friction (multiple midblock access points, parking) can adversely affect coordination, as it creates fluctuations in the queues and make the estimation of an offset unpredictable (see Eq. 9.15).

For additional information on signal control, [1] provides more detailed information geared toward practitioners, while [4] provides an excellent overview for novices and students.

A Special Case of Signalized Arterials: Two-Intersection Interchanges

An interchange is a special type of an arterial segment which connects the freeway to the surface network. There are several different types of interchanges, including various types of diamond and partial cloverleaf interchanges. Figure 9.13 provides an example of a diamond interchange, which is one of the most common types. At the surface street, the diamond interchange consists of two signalized intersections in relatively close proximity. The intersection on the left allows access to the southbound direction of the freeway, while the intersection on the right provides access to the northbound direction.

Interchange ramp terminals have some unique differences from other arterial segments [2]. First, they have a large percentage of turning movements, as they provide access to and from the intersecting freeway. Second, the two intersections are relatively closely spaced, and thus, there is limited storage space for queues

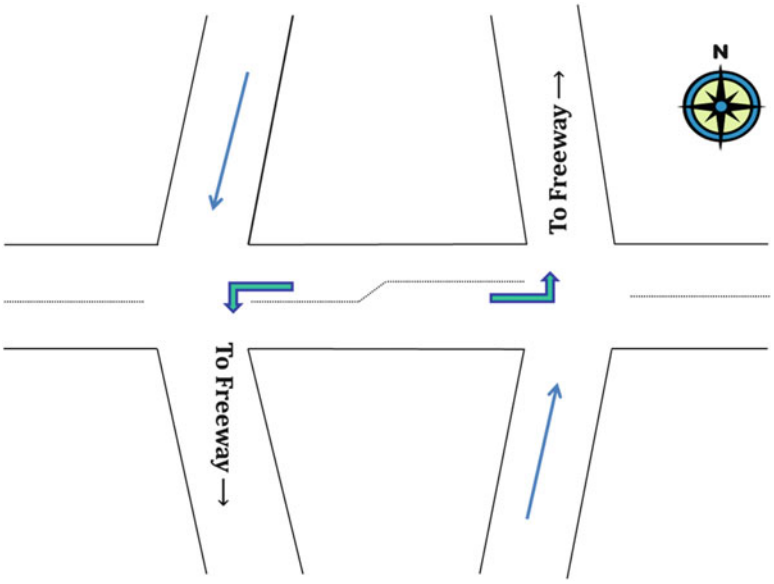


Fig. 9.13 Example of a diamond interchange

that develop along the link between the two intersections. This is particularly true for urban interchanges where space is at a premium. The presence of a downstream queue may reduce or completely block the traffic discharging from the upstream intersection. These characteristics create some unique challenges in their signalization.

Often the two signals are operated by one controller with complete coordination between the two intersections. Figure 9.14 illustrates an example of a phasing plan for a diamond interchange. This type of signalization allows for traffic to move through the arterial first for the eastbound direction, then for the westbound direction, and then the freeway off- ramp traffic from both ramps. The offset between the two intersections can be adjusted based on the distance between the two intersections and the demands. Additional information about the signalization of interchanges and their operation is provided in [2].

More recently, a new type of interchange has emerged: the diverging diamond interchange (DDI) or double crossover diamond interchange (DCD). Figure 9.15 provides a sketch of this type of interchange. Operationally, the major advantage of this configuration is that the crossover of the two traffic streams eliminates the conflict between the left turns and the opposing through traffic, and the two intersections only need to accommodate the two crossing through movements. The two signalized intersections only need a two-phase signal and can operate with shorter cycle lengths, reduced lost times, and increased capacity. This design is particularly appropriate when the demands of the left turns into the freeway and those of the freeway off-ramps are high. The first such interchanges were built in France, while the first one in the USA was built at Springfield, MO, and completed

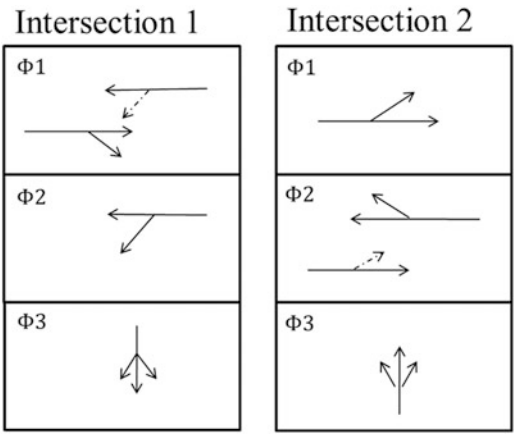


Fig. 9.14 Signalization plan at a diamond interchange

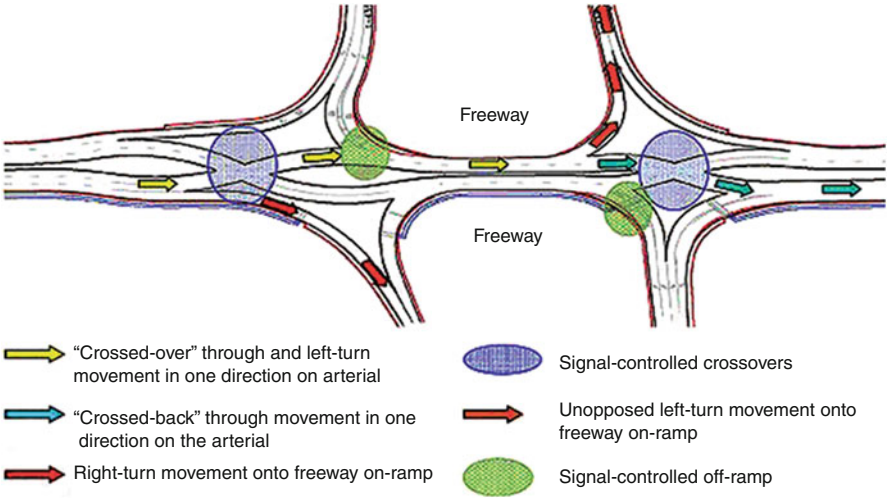


Fig. 9.15 Diverging diamond interchange (Federal Highway Administration; http://www.fhwa.dot.gov/publications/research/safety/09054/)

in 2009 (http://www.fhwa.dot.gov/publications/research/safety/09054/). Several such interchanges have been built since, and several are already in the design stage.

Operational Analysis Methods for Signalized Intersections and Networks

This section discusses existing analysis methods for signalized intersections and networks. There are three main types of tools for analyzing such facilities: the HCM, signal optimization software, and traffic simulation software.

Table 9.1 LOS Criteria for urban streets [2]

	LOS by critical volume-to-capacity ratio	
Travel speed as a percentage of base free-flow speed (%)	≤1.0	>1.0
>85	A	F
>67–85	B	F
>50–67	C	F
>40–50	D	F
>30–40	E	F
30	F	F

Overview of the HCM 2010 Procedures for Signalized Intersections and Networks

The HCM 2010 provides methods for evaluating urban street facilities, urban street segments, signalized intersections, and unsignalized intersections. An urban street segment is defined as the link between two intersections, while an urban street facility consists of multiple such segments. Intersection types considered in the analysis include unsignalized intersections and roundabouts (these are discussed in Chap. 10).

Chapter 16 of the HCM 2010 provides the methodology for analyzing urban street facilities, and it considers multiple travel modes: automobile, pedestrians, bicycles, and transit. The method provides the LOS of each mode separately, using different performance measures, and there is no combined LOS for the facility. Chapter 17 of the HCM 2010 provides methods for evaluating urban street segments. Table 9.1 provides the LOS criteria for the auto mode for urban streets.

LOS for pedestrians is evaluated considering traveler perception through the use of a score, as well as the average pedestrian space on the sidewalk. Similarly, LOS for bicycle and transit is estimated using an LOS score.

Chapter 18 of the HCM 2010 provides the methodology for evaluating signalized intersections. The chapter considers automobiles, pedestrians, and bicycles. The automobile methodology replicates fully actuated controller operation. This is achieved through an iterative process by which the analyst determines first the proportion of arrivals during the green and then the respective signal phase duration and then evaluates whether the two converge. If not, the estimates are revised, and the process continues until the two converge. Thus, the method is computationally intensive and cannot be solved by hand. The methodology ultimately estimates control delay for each lane, lane group, and intersection approach.

To address oversaturated conditions, in addition to d_1, the HCM [2] estimates two more delay components. The second component of delay, d_2, called incremental delay accounts for the randomness in arrivals and occasional cycle failures,

as well as the sustained oversaturation during the analysis period. It is estimated as follows:

$$d_2 = 900\,T\left[(X_A - 1) + \sqrt{(X_A - 1)^2 + \frac{8kIX_A}{c_A T}}\right] \tag{9.17}$$

where

T is the duration of the analysis period
$X_A = v/c_A$ is the average volume-to-capacity ratio
k is the incremental delay factor and it accounts for the effect of controller type on delay; it varies from 0.04 to 0.50; the value 0.50 is recommended for pre-timed phases
I is the upstream filtering adjustment factor, which accounts for the effect of an upstream signal on vehicle arrivals to the subject movement group; for isolated intersections its value is 1.0
c_A is the available capacity for a lane group served by an actuated phase (veh/h)

The third component of delay, d_3, called initial queue delay, accounts for any additional delay due to the presence of an initial queue, i.e., unmet demand during the previous time interval. It is estimated as follows:

$$d_3 = \frac{3,600}{FT}\left(t_A\frac{Q_b + Q_e - Q_{eo}}{2} + \frac{Q_e^2 - Q_{eo}^2}{2c_A} - \frac{Q_b^2}{2c_A}\right) \tag{9.18}$$

where

F is the demand flow rate during the analysis period, T
T is the duration of the entire analysis period
t_A is the duration of unmet demand during the analysis period
Q_b is the initial queue at the beginning of the analysis period, T
$Q_e = Q_b + t_A(v - c)$ is the queue at the end of the analysis period
$Q_{eo} = T(v - c)$ is the queue at the end of the analysis period when the volume exceeds capacity and the initial queue Q_b is zero

There is a significant body of literature devoted to the development of delay equations, and the interested reader may want to consult [4] for an overview of delay considerations as well as [5, 6] for the development of the equation that currently provides d_2. Additional information on the entire HCM 2010 procedure and details and background on the development and application of these equations is provided in Chap. 18 of the HCM 2010 [2] as well as in [4].

The LOS criteria for signalized intersections are shown in Table 9.2. A simpler, planning level, signalized intersection analysis method is also provided in Chap. 31 of the HCM 2010. That method is considerably simpler and can be solved by hand.

Table 9.2 LOS Criteria for signalized intersections [2]

Control delay (s/veh)	LOS by volume-to-capacity ratio	
	≤1.0	>1.0
≤10	A	F
>10–20	B	F
>20–35	C	F
>35–55	D	F
>55–80	E	F
>80	F	F

Traffic Signal Optimization Software

There are several signal optimization software which can be used to obtain optimal signal timings for isolated intersections, arterials, and networks. Some of the most popular packages include TRANSYT-7F, Synchro, and PASSER. Each of these packages has its own internal algorithm for optimization.

Generally, an important first step in signal control optimization is to estimate the minimum cycle length at each intersection (as if it is isolated), and the maximum of these would be able to accommodate the demand at all intersections (C_{maxmin}). The determination of offsets can be accomplished through several methods. One of the simplest methods estimates the performance measures under all or most possible scenarios and then identifies the set of offsets which result in the optimal traffic conditions. The PASSER family of models [7, 8] takes such an approach. Another method, used by TRANSYT-7F, uses the hill-climbing optimization method, which searches for the set of offsets that minimize a particular performance measure or index [9, 10]. Another method, also used by TRANSYT 7F, is based on genetic algorithms.

These optimization tools provide equivalent pre-timed signal timings. The analyst is still required to determine the minimum and maximum green, the detector design, and the vehicle extension for actuated control configurations [1].

Simulation of Signalized Intersections and Networks

There are several simulation tools that can be used to replicate the operation of signalized intersections and networks (as discussed in Chap. 7). These simulation tools can replicate traffic operations and evaluate a variety of signalization plans; however, these cannot directly provide a set of optimal signal timings. When these are used, they can provide optimal timings through trial and error. For example, TRANSYT 7F can generate optimal signal timings using CORSIM as its estimation tool [9, 10]. Also, microsimulators can be used in conjunction with signal optimization tools as an additional validation step before these are implemented in the field.

Advanced Technologies in Signal Control

Recent developments in signal control seek to automate the signal optimization function by developing or implementing optimal plans real time. There are various types of systems that have been developed and are implemented in a few locations, which can provide increasing levels of demand responsiveness. The Federal Highway Administration recently highlighted four adaptive signal control systems that use real-time traffic information to optimize signal control (http://www.fhwa.dot.gov/everydaycounts/technology/adsc/).

SCOOT (http://www.scoot-utc.com/), which stands for Split Cycle Offset Optimization Technique, can be used to optimize traffic signal control in the field using real-time information. It requires the installation of detectors at the upstream end of each link, and it operates based on principles similar to those used in TRANSYT 7F. The system models the progression of traffic from the detectors to the stop bar, considering prevailing traffic conditions and queues downstream. Based on this information, SCOOT continually adapts the splits, offsets, and cycle times for the intersections in the network.

SCATS (Sydney Coordinated Adaptive Traffic System) is another such system for coordinating signalized intersection operations (http://www.scats.com.au/). SCATS can integrate freeway and arterial systems and can interface with several microsimulation packages.

OPAC (Optimized Policies for Adaptive Control), initially developed in the 1980s, is a distributed real-time signal control system that continuously modifies signal timings to minimize a performance function based on delay and stops. It is based on dynamic programming, and it can operate in distributed individual intersection control or in a coordinated network [11].

RHODES (Real-Time Hierarchical Optimized Distributed Effective System) uses upstream detector information for each intersection to predict the arrivals at each approach. Similarly to OPAC, its optimization process is based on dynamic programming. The system architecture decomposes the traffic control problem into several subproblems that are hierarchically interconnected [12].

References

1. ITE (2009) Traffic signal timing manual. Institute of Transportation Engineers (ITE), Washington, DC
2. Transportation Research Board, National Academies of Science (2010) Highway Capacity Manual, Transportation Research Board, National Research Council, Washington, DC
3. Traffic detector handbook, 3rd edn (2006) Federal Highway Administration, McLean, VA, Publication Number: FHWA-HRT-06-108, October 2006
4. Roess RP, Prassas ES, McShane WR (2011) Traffic engineering, 4th edn. Pearson/Prentice Hall, Upper Saddle River, NJ
5. Akcelik R (1980) Time-dependent expressions for delay, stop rate, and queue lengths at traffic signals, Report No. AIR367-1. Australian Road Research Board, Vermont, VIC

6. Akcelik R. Traffic signals: capacity and timing analysis. ARRB Report 123, Australian Road Research Board, VIC, Australia, Mar 1981
7. PASSER™ II-02. http://ttisoftware.tamu.edu/fraPasserII_02.htm
8. Federal Highway Administration (1991) PASSER II-90, User's guide. In: Methodology for optimizing signal timing: MOST, vol 3. Federal Highway Administration, Washington, DC
9. TRANSYT 7F. http://mctrans.ce.ufl.edu/featured/TRANSYT-7F/
10. Wallace CE, Courage KG, Hadi MA, Gan AC (1998) Transyt-7F user guide. In: Methodology for optimizing signal timing (MOST), vol 4. Transportation Research Center, University of Florida, Gainesville, FL
11. Gartner NH (1983) OPAC: a demand-responsive strategy for traffic signal control, Transportation Research Record No 906. Transportation Research Board, Washington, DC, pp 75–81
12. Mirchandani PB, Head L (2001) A real-time traffic signal control system: architecture, algorithms, and analysis. Transp Res Part C Emerg Technol 9(6):415–432

Problems

1. Identify in your area a signalized intersection approach that regularly experiences congestion. Observe and record lost time and saturation flow rate for at least ten cycles. What is your recommended value for lost time and saturation flow rate for this approach? What is the capacity of the approach?
2. For the data provided in Example 9.2, plot uniform delay as a function of demand, assuming the following demand values: 200, 400, 600, 800, 1,000, 1,200, 1,400, 1,600, 1,800, 2,000. What do you observe? Calculate incremental delay for the same values. How do the two types of delay vary by demand?
3. For the data provided in Example 9.2, plot uniform delay as a function of the cycle length and as a function of the g/C for reasonable ranges of these values. What do you observe? What are the implications of your observations for signal control optimization?
4. Review the HCM 2010 signalized intersection analysis method, and provide a step-by-step overview of the iterative process for estimating signal phase durations. Develop a simple example problem for illustrating its application.
5. A four-approach isolated signalized intersection has a two-phase signal. The critical lane demand for the first phase is 1,050 vphpl, while the critical lane demand for the second phase is 725 vphpl.

   (a) What cycle length would you recommend for this signal, assuming:

   PHF = 0.95, Desirable $v/c = 0.90$
   Phase 1: Lost time = 2 s/phase, critical lane saturation flow = 1,700 vphgpl
   Phase 2: Lost time = 3 s/phase, critical lane saturation flow = 1,650 vphgpl

   (b) A new development is planned which will increase the demand of the critical lane for the second phase to 1,350 vphpl. What cycle length would you recommend in this case? Discuss the results and provide your professional opinion on traffic operations under this scenario.

6. The arrival rate at a two-lane signalized intersection approach is 350 vehicles per hour per lane. The service flow rate at saturation is 1,900 vehicles per hour per lane. The intersection operates at a cycle length of 120 s, while the red interval for the study approach is 45 s. Construct the cumulative vehicles versus time diagram for both arrivals and departures, and calculate the following: percent time queue is present, number of vehicles per cycle, average queue length, average individual delay, average individual delay while queue is present, and maximum queue length.
7. Conduct a literature review search to document the optimization process used in TRANSYT 7F and SCOOT.
8. Document the use of saturation headways and lost time in your favorite traffic simulation software package. How are these two defined? How does the software account for different geometric features (such as lane width and grade) and different traffic movements (left, through, and right) in these values?

Chapter 10
Unsignalized Intersections

Unsignalized intersections are those where at least one of the movements is controlled by a STOP or a YIELD sign. Operations of such facilities require the drivers on the controlled movements (usually referred to as minor movements) to judge the size of the gaps along the major (or uncontrolled) street and select a suitable one to cross or to merge into.

Unsignalized intersections include two-way stop-controlled (TWSC) and all-way stop-controlled (AWSC) intersections, as well as roundabouts. Figure 10.1 provides sketches of these three types of facilities.

At the TWSC intersection shown in Fig. 10.1a, minor movements include left, through, and right turns from the two minor approaches (NB and SB), as well as left turns from the two major approaches (EB and WB). High demand along the major street limits the opportunities for minor street vehicles to enter or cross. The left-turn movements from the major street have priority over the minor street movements, while the through and right movements from the minor streets have priority over the respective left-turn movements. These priority rules, combined with the intersection geometry and the relative demands of each movement, affect the capacity of each minor movement.

At AWSC intersections (Fig. 10.1b), arriving vehicles are required to proceed in a priority sequence (first come, first served), and thus, operations are slightly different than those at TWSC. Drivers do not need to judge the size of gaps in order to proceed, but they do need to observe their priority position relative to other vehicles at the intersection when they arrive at the stop bar. The capacity of each approach is a function of the demand in the other approaches; the higher the demand in the other approaches, the lower the capacity of the subject approach.

Roundabouts are traffic circles where the circulating traffic has priority and the entering traffic has to yield to the circulating traffic. Thus, at roundabouts, drivers have to judge the gaps within the circulating stream. Roundabouts are designed to reduce the speed of approaching vehicles without requiring them to stop. The angle of the approach to the circulating traffic stream allows vehicles to enter the roundabout at a higher speed. Therefore, all things being equal, the capacity of a roundabout approach would be higher than the equivalent right-turn movement

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_10,

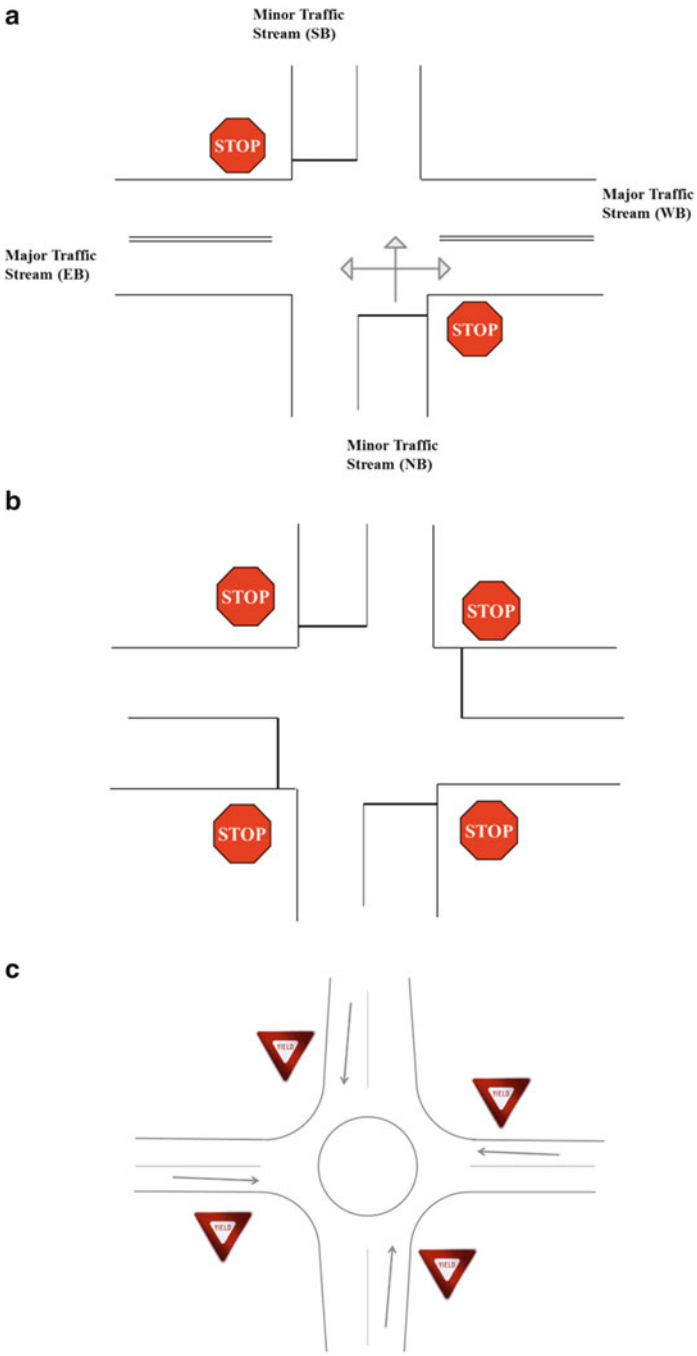


Fig. 10.1 Three types of unsignalized intersections

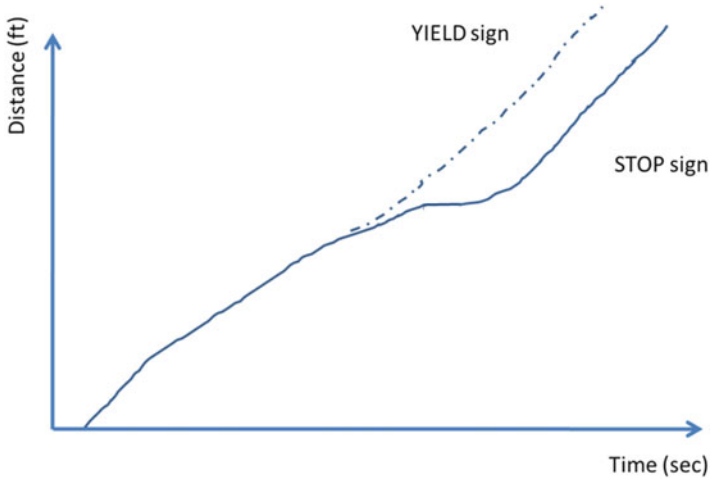


Fig. 10.2 Trajectory of vehicle approaching a stop sign (TWSC) vs. a yield sign (roundabout)

from a stop-controlled approach. Similarly, the delay to the roundabout approach vehicles would be lower than that of a stop-controlled intersection. Figure 10.2 illustrates the trajectory of a vehicle approaching a stop sign vs. that of a vehicle approaching a yield sign. As shown, the presence of a stop sign produces higher delay and longer travel times for minor street vehicles.

This chapter first reviews principles of gap acceptance as it applies to TWSC and roundabouts (Chap. 2 discusses gap acceptance principles more broadly and considering lane changing, merging, and stop/yield control). It then discusses operations of these three types of unsignalized intersections in more detail, along with analysis methods.

Principles of Gap Acceptance

The operation of TWSC intersections and roundabouts (yield controlled) is different from that of signalized intersections because drivers approaching a stop or yield sign use their own judgment proceeding through conflicting traffic movements. Each driver of a stop- or yield-controlled approach must evaluate the size of gaps in the conflicting traffic streams and judge whether he or she can enter the intersection or roundabout safely. The capacity of a stop- or yield-controlled movement is a function of the following parameters:

The availability of gaps in the major (uncontrolled) traffic stream: Gap is generally defined as the time elapsing between the crossing of the lead vehicle's *rear bumper* and the crossing of the following vehicle's *front bumper*.

The availability of gaps is a function of the arrival distribution of the major traffic streams. Some methods use headways rather than gaps in their analysis.

The gap acceptance characteristics and behavior of the drivers in the minor movements: The same gap may be accepted by some drivers and rejected by others. Also, when a driver has been waiting for a long time, he or she may accept a shorter gap while have rejected longer ones. The parameter used most often in gap acceptance is the *critical gap* (*or in some cases the critical headway*), defined as the minimum time between successive major street vehicles in which a minor street vehicle can complete a maneuver.

The follow-up time of the subject movement queued vehicles: The *follow-up time* is the time headway between consecutive vehicles using the same gap under conditions of continuous queuing, and it is a function of the perception–reaction time of each driver.

The use of gaps in the major traffic stream by movements of higher priority or for shared lanes: This results in reduced opportunities for lower-priority movements or for different movements sharing a lane (this is not applicable for roundabouts because there is only one minor movement).

The example of Fig. 10.3 is a simplified illustration of the capacity estimation process for a stop- or yield-controlled movement. The capacity of the northbound (NB) minor through movement of Fig. 10.3 is estimated based on field measurements at the intersection. There is only one major traffic stream at the intersection (EB) and one minor street movement (NB).

Example 10.1 The critical gap for the NB approach of the intersection of Fig. 10.3 was measured to be 5 s. It is assumed that any gap larger than 5 s will be accepted by every driver, whereas every gap smaller than 5 s will be rejected by every driver. The follow-up time was measured to be 3 s. Estimate the capacity of the approach assuming that a brief data collection effort recorded the following gaps for the EB approach: 5, 12, 15, 7, 9, 11, 3, 8, 3, 17, 4, 4, 10, 2, 17.

Solution to Example 10.1

Table 10.1 summarizes the field data and subsequent calculations and provides the respective capacity estimate. Column 1 of Table 10.1 provides the gaps measured in the field. Column 2 indicates whether the gap is usable, that is, whether it is larger than the critical gap. Column 3 estimates the number of vehicles that can use each of the usable gaps. For two vehicles to use a gap, it should be at least:

$$5\,\text{s}\ \ (\text{first vehicle}) + 3\,\text{s}\ \ (\text{second vehicle}) = 8\,\text{s}$$

The following equation can be used to estimate the maximum number of vehicles that can use a gap of size X:

$$\boxed{\text{Number of vehicles} = 1 + (\text{gap size } X - \text{critical gap})/\text{follow-up time}} \tag{10.1}$$

The total number of vehicles that can travel through the NB approach is the sum provided at the bottom of Column 3 (28 vehicles over 126 s). Assuming that the

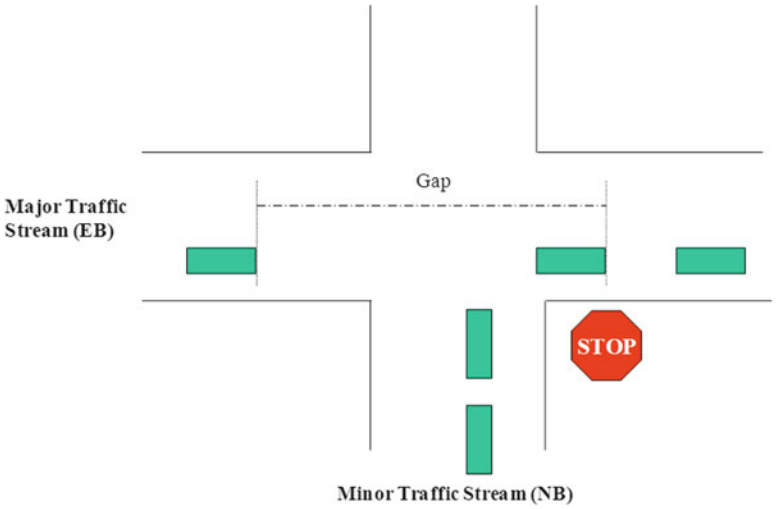


Fig. 10.3 Capacity of a stop-controlled approach

Table 10.1 Capacity estimation for a stop-controlled movement

Gap size (s) (1)	Is the gap usable (Y or N)? (2)	Vehicles in NB movement that can use the gap (veh) (3)
5	Y	1
12	Y	3
15	Y	4
7	Y	1
9	Y	2
11	Y	3
2	N	0
8	Y	2
3	N	0
17	Y	5
4	N	0
4	N	0
10	Y	2
2	N	0
17	Y	5
126 s	10 usable gaps	28 vehicles

gaps measured during this short interval are representative of the demand at the intersection, the capacity of the movement is:

$$(3,600/126) \times 28 = 800\,\text{veh/h}$$

The *HCM 2010* methodology for estimating the capacity of TWSC intersections is based on the principles just outlined using mathematical expressions of gap

distributions and probability theory for establishing the use of gaps by higher-priority movements. An overview of this method is presented in the next section.

Gap acceptance theory considers the availability of gaps in the major traffic stream and the acceptability of gaps by the minor traffic stream to determine the traffic operational performance of the facility. As indicated in Chap. 2, there is a wealth of research on critical gap estimation for various types of movements, facilities, and drivers. The principles discussed in this section apply also to permitted left turns at signalized intersections, as well as freeway merging and lane changing.

Operational Analysis Methods for Unsignalized Intersections

There are various methods and models used to analyze unsignalized intersections. Each method typically pertains to a specific type of intersection, and thus, comparison of the performance of various types of intersections (including signalized intersections) is very difficult. The HCM 2010 estimates control delay at all types of intersections to facilitate comparisons in their performance. However, the LOS boundaries differ at each facility. For example, a delay of 40 s/veh at a signalized intersection represents LOS D, while at a two-way stop-controlled intersection it represents LOS E. There is no research yet documenting driver perception around each of those intersection types; thus, LOS criteria may not necessarily reflect driver perceptions.

This section provides an overview of analysis techniques for TWSC, AWSC, and roundabouts. The two primary analysis methods for all three types of intersections are the HCM and simulation.

TWSC Intersections

The HCM 2010 [1] analysis method is provided in Chap. 19 of the HCM, and it is based on the principles of gap acceptance described earlier. The basis of the methodology is documented in [2]. The methodology steps are generally as follows (for detailed information on the methodology, consult [1, 2]):

Step 1. Determine the movement priorities at the intersection At a typical TWSC intersection, minor movement priorities are as follows:

- Left turns from the major street
- Right turns from the minor streets
- U-turns from the major street
- Through movements from the minor streets
- Left turns from the minor streets

Therefore, if there is a gap available along the major street in the NB direction, it will be the SB left-turn movement that will be able to use it first; movements of lower priority may not be able to use it. Thus, in determining the capacity of each movement at a TWSC, it is important to determine whether movements of higher priority will be able to use the gaps available.

Step 2. Determine conflicting flow rates for each minor movement During this step, the analyst identifies the movements that conflict with each minor movement, along with their flow rates. Conflicting movements may include pedestrians. For example, the left turns from the major street conflict with the opposing through and right-turn movement, as well as the pedestrians crossing the respective minor street.

Step 3. Determine the critical headways and follow-up headways for each movement The HCM 2010 method is based on headways rather than gaps. During this step, the analyst estimates the critical headway for each movement, as a function of the number of lanes along the major street, the presence of heavy vehicles, grade, and intersection geometry. The HCM provides an equation and adjustment factors for calculating the critical headway for each movement. The adjustment factors consider whether there is a median separating the two directions of major street traffic, as in this case vehicles from the minor street may complete a "two-stage" gap acceptance in order to cross the facility. The follow-up headway is similarly estimated for each movement as a function of the presence of heavy vehicles. The equations the HCM 2010 provides for determining the critical headways and follow-up headways were calibrated based on field data. The critical headway increases for movements with lower priority. For example, the base critical headway for the left turn from a major street with one lane per direction is 4.1 s, while for the left turn from a minor street and assuming a one-stage gap acceptance, it is 7.1 s.

Step 4. Compute potential capacities Potential capacity is the capacity the movement would have if it was unimpeded by higher-ranked movements or shared lanes. The potential capacity of each movement is estimated using the following equation:

$$c_{\mathrm{p},x} = F_{\mathrm{c},x} \frac{\mathrm{e}^{-F_{\mathrm{c},x} t_{\mathrm{c},x}/3,600}}{1 - \mathrm{e}^{-F_{\mathrm{c},x} t_{\mathrm{f},x}/3,600}} \tag{10.2}$$

where

$c_{\mathrm{p},x}$ is the potential capacity of movement x (veh/h)
$F_{\mathrm{c},x}$ is the conflicting flow for movement x (s)
$t_{\mathrm{c},x}$ is the critical headway for minor movement x (s)
$t_{\mathrm{f,x}}$ is the follow-up headway for minor movement x (s)

Mathematically, this equation is obtained as follows [2]. If the major street volume in vehicles per second (vps) is V_{m}, and the probability density function

(pdf) of major street headways is $f(t)$, then the expected number of gaps of size t per hour along the major street is:

$$\text{Number of size } t \text{ gaps per hour} = 3,600 \times V_{\text{m}} \times f(t)$$

If the number of minor movement vehicles that can enter into a gap of size t is $g(t)$, then the capacity available for all gaps of size t is:

$$\begin{aligned}&\text{Capacity available for all gaps of size } t \text{ per hour}\\ &\quad = \text{Number of size } t \text{ gaps per hour} = 3,600 \times V_{\text{m}} \times f(t) \times g(t)\end{aligned}$$

Thus, the capacity available for all gaps can be obtained by integrating the above function for all t:

$$c = 3,600 \; V_{\text{m}} \int_{(t=0)}^{n} f(t)g(t)\text{d}t$$

Let us assume that the major street headways follow the negative exponential distribution:

$$f(t) = \lambda \text{e}^{-\lambda t}$$

where λ is the arrival rate of the major street.

Let us further assume that $g(t)$ is a step function:

$$g(t) = \sum_{n=0}^{\infty} n \, p_n(t)$$

where $p_n(t)$ is the probability that n vehicles would accept a gap of size t:

$$p_n(t) = \begin{cases} 1 & \text{for } t_{\text{c}} + (n-1)t_{\text{f}} \le t < t_{\text{c}} + nt_{\text{f}} \\ 0 & \text{elsewhere} \end{cases}$$

These two assumptions, together with the assumption that drivers are both consistent and homogeneous (i.e., critical headways, t_{c}, and critical follow-up headways, t_{f}, are constant), result in Eq. (10.2). Reference [2] indicates that these assumptions produce satisfactory results for all practical applications.

Next, this potential capacity must be adjusted to consider higher-ranked movements as well as shared lanes.

Step 5. Compute movement capacities The capacity of each movement is estimated based on the geometric design of the intersection (i.e., shared movements and two-stage gap acceptance) as well as the specific rank of the movement. The method takes into account the probability that a gap would not be used by a higher-ranked movement (i.e., the probability that any higher-ranked movements would operate in a queue-free state), as well as impedance due to shared lane operations.

Table 10.2 LOS criteria for TWSC intersections [1]

	LOS by volume-to-capacity ratio	
Control delay (s/veh)	≤1.0	>1.0
≤10	A	F
>10–15	B	F
>15–25	C	F
>25–35	D	F
>35–50	E	F
>50	F	F

Step 6. Compute movement, approach, and intersection control delay Control delay is estimated for each movement and then can be used to obtain weighted averages for the approach and the intersection. Table 10.2 provides the LOS criteria for TWSC intersections. The control delay LOS boundaries are generally lower than those for signalized intersections (Table 9.2). This difference is to indicate that users generally expect higher delays at signalized intersections than at unsignalized ones.

The HCM 2010 method provides LOS for pedestrians crossing a traffic stream that is not controlled by a STOP sign and for midblock pedestrian crossings. Pedestrian LOS is based on average pedestrian delay, which is the delay pedestrians are expected to incur as a function of the critical headway for pedestrians, the availability of vehicular gaps, and the probability that motorists yield to pedestrians. Bicyclists are not explicitly considered in the methodology. Overall, the methodology considers constant demand and capacity and does not explicitly evaluate the impacts of residual queues from previous analysis periods. Those situations should be analyzed using simulation and other tools.

Many simulation models replicate unsignalized intersection operation and gap acceptance. However, the literature does not provide detailed information on the replication of TWSC intersections and the respective gap acceptance and follow-up time algorithms. This is likely for two reasons. First, unsignalized intersections tend to have lower demands, are often undersaturated, and they are seldom a key component of the congested urban networks most often analyzed. Second, there are significant difficulties with collecting field data to model gap acceptance and follow up time behavior, particularly considering the differences between different drivers, vehicles, and intersection characteristics.

Example 10.2 Calculate the capacity of the WB left-turn and the NB right-turn movements for the intersection shown in Fig. 10.4. The critical headway for the WB left turn is 5 s, and its follow-up headway is 3 s. The critical headway for the NB right turn is 7 s, and its follow-up headway is 4 s. The EB through movement flow is 800 vph. How would the intersection operate if the demand for the WB left-turn movement is 500 vph and the demand for the NB right turn is 450 vph?

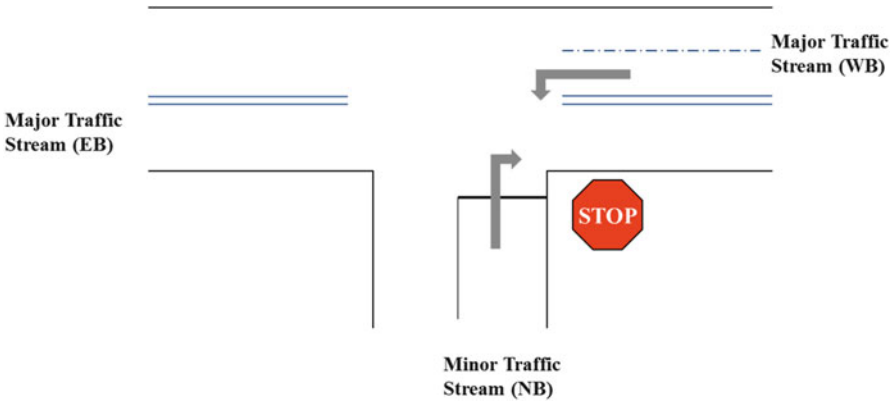


Fig. 10.4 Sketch for Example 10.2

Solution to Example 10.2
Using Eq. (10.2), the potential capacity of the WB left turn is calculated as follows:

$$c_{\mathrm{p},x} = F_{\mathrm{c},x}\frac{\mathrm{e}^{-F_{\mathrm{c},x}t_{\mathrm{c},x}/3,600}}{1-\mathrm{e}^{-F_{\mathrm{c},x}t_{\mathrm{f},x}/3,600}} = 800\frac{\mathrm{e}^{-(800\times5)/3,600}}{1-\mathrm{e}^{-(800\times3)/3,600}} = 541\,\mathrm{vph}$$

Similarly, the potential capacity of the NB right turn is calculated to be 287 vph.

Both of these two minor movements conflict with the EB through traffic; however, they can both utilize simultaneously the same gaps. Thus, their potential capacity is equal to their actual movement capacity.

The demand of the WB left-turn movement is below its capacity, and thus, no queues are expected. However, the right-turn movement would be significantly over capacity, and queues are expected to form under the existing configuration and set of demands.

AWSC Intersections

The HCM 2010 provides a methodology for analyzing AWSC intersections in Chap. 20 [1]. AWSC intersections do not operate based on gap acceptance principles. All vehicles approaching such an intersection must stop and proceed on a first come, first served basis, and thus, their operation is based on the arrivals and the origin–destinations of the traffic streams. For those types of intersections, the capacity of each approach is a function of the demand in the other approaches. The HCM methodology (which is based on [3]) is iterative and cannot be solved by hand. It estimates the capacity and LOS of each approach (subject approach) as a function of the demand in the opposing and conflicting approaches.

Depending on the intersection geometry, AWSC intersections work roughly either as a two-phase pattern or as a four-phase pattern. In the two-phase pattern, vehicles arriving from the NB and the SB approaches move simultaneously whenever they do not conflict, and similarly does EB and WB traffic. Multilane AWSC intersections tend to operate in a four-phase pattern, with a single approach moving in turn.

The capacity of an AWSC intersection subject approach depends on its departure headways. When there is no traffic in any of the other approaches, the departure headway is at its minimum (i.e., flow is maximized); this departure headway is generally termed the saturation headway. As demand in the other approaches increases, the departure headway of the subject approach increases, and thus flow is reduced.

The steps of the HCM 2010 methodology are generally as follows:

Step 1. Estimate the saturation headway of each movement through the intersection The analyst determines the minimum departure headway (i.e., saturation headway) through the intersection for each movement as a function of the proportion of turning movements and the presence of heavy vehicles.

Step 2. Estimate the initial degree of saturation, x *(or degree of utilization), for each approach* This is estimated as the arrival rate divided by the service rate using the following equation:

$$x = \frac{F\, h_{\mathrm{d}}}{3,600}$$

where

x is the degree of saturation or degree of utilization
F is the flow rate for the approach
h_{d} is the departure headway (s)

The first iteration of this calculation is initiated with a default value for h_{d} (=3.2 s).

Step 3. Estimate the actual departure headway This is based on the flow in the opposing and conflicting approaches and is computed considering the probability that each lane has a vehicle waiting at the stop bar.

Step 4. Check convergence If the actual departure headway estimated in Step 3 is different by less than 0.1 from the initial one, the process concludes. However, if the difference is larger, the process is repeated until convergence is achieved.

Step 5. Estimate capacity and control delay for each approach and determine LOS The capacity of an AWSC intersection approach is estimated by increasing the flow of the subject lane and holding the flows of the opposing and conflicting approaches constant until the degree of utilization becomes equal to 1. The control delay for each lane is estimated similarly to TWSC intersections. The LOS boundaries are also identical to those of TWSC intersections.

There are very few references addressing simulation of AWSC intersections. It is also not clear how they are treated within existing microsimulators, if at all. Since they are rarely part of a congested urban network, they have not been given as much attention as some of the other intersection types.

Roundabouts

Roundabouts in the USA have become very popular only recently. Thus, the HCM 2010 is the first edition to include a methodology for obtaining LOS for roundabouts. Chapter 21 of the HCM 2010 provides a methodology for the analysis of roundabouts, and it includes analysis methods for pedestrians and bicycles (the methods for pedestrians and bicycles are based on research in Europe and Australia, where roundabouts have been widely used for many years). The HCM 2010 methodology proceeds as follows (the method is developed based on [4], while specifics of the method are provided in [1]):

Step 1. Determine circulating and exiting flows The circulating flow conflicting with each entering flow is obtained here based on the origin–destinations through the roundabout and converted to passenger cars per hour. Similarly, the exit flows in passenger cars per hour are calculated for each leg of the roundabout; these flows conflict with right-turn bypass lanes.

Step 2. Determine entering flows by lane For single lane entries, this flow is simply the total entering flow. For multilane approaches, specific guidelines are provided for lane assignments as a function of the roundabout configuration and demands.

Step 3. Calculate the capacity of each entering lane including bypass lanes The capacity of each lane is estimated using calibrated equations that have been developed for various roundabout configurations. For example, the following equation is used to estimate the capacity of a single lane approach that is opposed by a single lane circulating traffic stream:

$$\boxed{c_{\text{pce}} = 1,130\, \text{e}^{\left(-1.0\times10^{-3}\right)F_{\text{c,pce}}}} \tag{10.3}$$

where

c_{pce} is the capacity of the single lane approach (pc/h)
$F_{\text{c,pce}}$ = is the conflicting flow (pc/h)

Similar equations are provided in the HCM 2010 for other roundabout configurations. These equations were developed based on field data around the USA [4].

Step 4. Determine pedestrian impedance This step estimates the pedestrian impedance and adjusts the capacity of each lane accordingly. The effect of pedestrians is higher when there is a significant number of pedestrians and when there are low

vehicular flows; at high flows, pedestrians tend to walk between vehicles, and thus, their effect on capacity is minimal.

Step 5. Compute the average control delay for each lane, as well as the approach and the entire roundabout The average control delay is estimated similarly to that for TWSC intersections with a slight adjustment to account for the presence of a YIELD sign instead of a STOP sign at entry. The average delay is a function of the *v*/*c*, and the model assumes there is no initial (residual) queue at the beginning of the analysis period. The LOS criteria are identical to those provided for signalized intersections (see Table 9.2). As discussed in Chap. 5, there is no strong basis for the establishment of these boundaries. The existing boundaries are based primarily on consensus from TRB's Highway Capacity and Quality of Service (HCQS) committee, and additional research is required to determine how users perceive the delay at various types of unsignalized intersections vs. that at signals.

In addition to the HCM 2010 procedures, there are two other popular analytical models for the analysis of roundabouts. RODEL (http://rodel-interactive.com) was developed in the early 1970s at the UK's Transportation Research Laboratory (TRL). It can estimate capacity for a variety of configurations, and it can consider mini-roundabouts, pedestrian impacts, as well as crashes and economic impacts. The newest version provides the HCM 2010 delay estimates as well. SIDRA (http://www.sidrasolutions.com) was developed in Australia, and in addition to roundabout analysis, it can evaluate a variety of intersection types. It can provide HCM 2010 outputs as well as RODEL and ARCADY outputs.

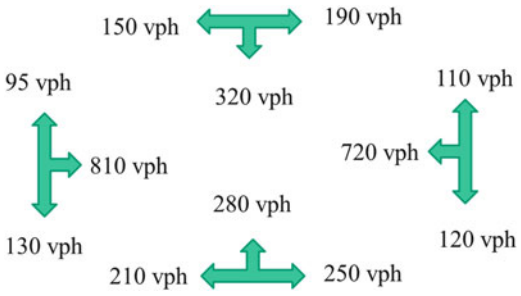
Most traffic microsimulators can analyze roundabouts. In addition, ARCADY is a popular microsimulator developed in the UK by the TRL for the analysis of roundabouts (https://www.trlsoftware.co.uk/products/junction_signal_design/arcady).

References

1. Transportation Research Board, National Academies of Science (2010) Highway Capacity Manual, Transportation Research Board, National Research Council, Washington, DC
2. Kyte M, Tian Z, Mir Z, Hameedmansoor Z, Kittelson W, Vandehey M, Robinson B, Brilon W, Bondzio L, Wu N, Troutbeck R (1996) NCHRP Web Doc 5 capacity and level of service at unsignalized intersections: final report vol 1—two-way-stop-controlled intersections. Transportation Research Board, National Academies of Science, Washington, DC
3. Kyte M, Tian Z, Mir Z, Hameedmansoor Z, Kittelson W, VandeheyM, Robinson B, Brilon W, Bondzio L, Wu N, Troutbeck R (1996) NCHRP Web Doc 5 capacity and level of service at unsignalized intersections: final report vol 2—All-way-stop-controlled intersections. Transportation Research Board, National Academies of Science, Washington, DC
4. Rodegerdts L, Blogg M, Wemple E, Myers E, Kyte M, Dixon M, List G, Flannery A, Troutbeck R, Brilon W, Wu N, Persaud B, Lyon C, Harkey D, Carter D (2007) NCHRP Report 572: roundabouts in the United States. TRB, National Academies of Science, Washington, DC

Problems

1. Conduct a literature review documenting methods for obtaining the critical gap of unsignalized intersection approaches. Which one would you recommend?
2. Examine your favorite microsimulator and document its algorithms for analyzing roundabouts. How does the model handle differences between the presences of a STOP sign used at TWSC intersections vs. a YIELD sign? How does the model handle critical gap determination for various driver populations and various vehicle types?
3. For the intersection of Example 10.2, use the HCM 2010 to calculate the capacity of the NB left-turn movement if it is provided its own lane, and the demand of the WB through movement is 670 vph. How does the capacity of this movement change if the NB approach has a single lane? Assume that the demand of the NB left is 150 vph.
4. Conduct a literature review to identify methods for the analysis of AWSC intersections. Document the process for estimating the probability of the presence of vehicles in the opposing and conflicting approaches.
5. Consider the demands shown in the figure below. Use your favorite microsimulator to evaluate the performance of the intersection if it is assumed it is a TWSC, an AWSC, or a roundabout. Which of the three designs is most effective for this particular set of demands? How would your response change as a function of the relative demands?



6. Conduct a literature review to identify guidelines when roundabouts are operationally more effective than TWSC or AWSC intersections.

Chapter 11
Two-Lane Highways

Two-lane highways, which have one lane per direction, are unique operationally, since they may allow passing through the use of the opposing traffic stream. According to the US Federal Highway Administration [1], two-lane highway facilities represent about 97 % of the total highway system and for more than 65 % of the total nonurban vehicular travel in the USA. Hence, two-lane highways provide most of the primary interurban highway network as well as being the basis of the secondary highway and collector networks. Figure 11.1 provides a sketch of a two-lane highway.

When passing zones are provided, they are marked using dashed lines between the two opposing traffic streams, and vehicles following a slower vehicle must evaluate the gaps in the opposing direction before deciding whether to pass. Thus, around passing zones, the capacity and operations of one direction depend on the demand in the other direction. For segments with no-passing zones, vehicles must follow the lead vehicle in a platoon. Facilities with no-passing opportunities increase the probability of car-following (formation of platoons) and reduce the ability of drivers to travel at their desired speeds. Therefore, one important performance measure that has been used to evaluate two-lane highways is the percent time spent following (PTSF).

Two-lane highways are often designed with passing lanes, which are additional lanes provided at selected locations throughout the length of the facility to provide passing opportunities. A climbing lane is a passing lane added along an upgrade section to allow traffic to pass heavier and slower vehicles. In Europe (e.g., Finland and Sweden), there are two-lane highways designed with wider cross sections at selected locations intended to provide passing opportunities without providing an additional lane; at those facilities, vehicles use part of the opposing lane to pass. Other features used along two-lane highways are turnouts and two-way left-turn lanes. Turnouts are shoulder areas designed to allow vehicles to pull off in order to allow faster vehicles to pass. Two-way left-turn lanes are additional lanes provided between the two directions to allow left-turning vehicles to access driveways and intersections along the facility without obstructing through traffic.

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6_11,

Fig. 11.1 Sketch of a two-lane highway

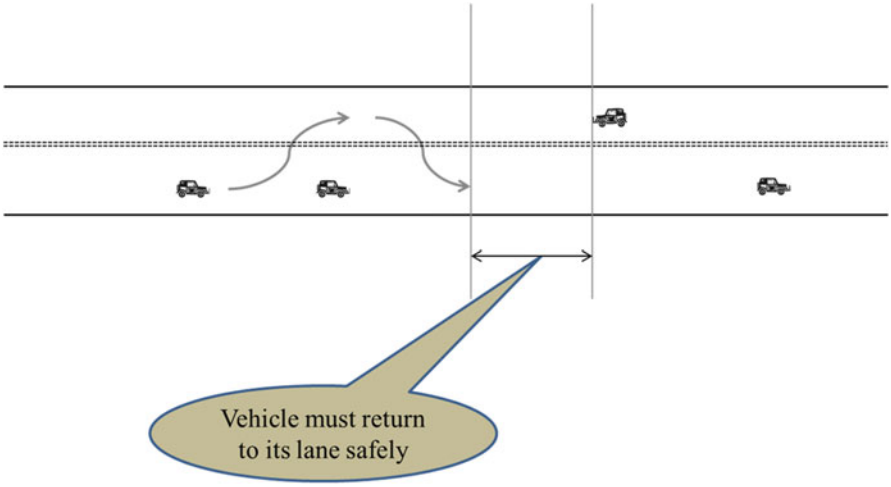


Fig. 11.2 Passing maneuver along a two-lane highway

This chapter first discusses the operation of two-lane highways and then provides an overview of capacity estimation for these facilities. The third section discusses the HCM 2010 analysis procedures for two-lane highways, while the last section provides an overview of microsimulation tools available to evaluate their operation.

Principles of Two-Lane Highways Operations

To understand traffic operations along two-lane highways, let us first examine passing. Figure 11.2 illustrates a passing maneuver. The passing vehicle has to judge the gap to the next vehicle in the opposing direction as well as its speed and decide whether it can safely return to its lane without colliding.

The passing maneuver has significant similarities with other gap acceptance-related maneuvers, such as those during lane changes and at TWSC intersections. Important elements during a passing maneuver are sight distance to the opposing vehicle, speed of the opposing vehicle, speed of the passing vehicle, and speed of

the vehicle being passed. A passing maneuver generally involves the following three steps from the perspective of the individual driver:

Desire to pass: The decision to pass is based on a driver's desired speed versus the desired speed of the lead vehicle. It is also a function of the availability and length of a passing zone.

Decision to pass: The decision to pass is a function of the availability of gaps in the opposing direction, i.e., the demand levels. It is also a function of the number of vehicles in the platoon in front of subject vehicle. The driver's decision to pass would depend on the relative size of the gap available versus the time required to pass all slower vehicles within a platoon ahead.

Execution of passing maneuver: The execution of a passing maneuver is a function of the subject vehicle's acceleration capabilities as well as the terrain and the vehicle's overall performance capabilities including the maximum speed it can attain given the prevailing grade and length of grade.

Highway designers consider the passing sight distance in order to allow for safe passing at passing zones. Traffic simulation modelers also consider the passing sight distance in order to replicate passing maneuvers at two-lane highways. According to [2], the vehicle desiring to pass must travel an additional distance relative to the slower vehicle, d_r. If the speed of the slower vehicle is v_s and the speed of the faster vehicle is v_f, the faster vehicle can travel this additional distance d_r in time t_p:

$$\boxed{t_p = \frac{d_r}{v_f - v_s}} \tag{11.1}$$

where t_p is the passing time, i.e., the time required for the faster vehicle to pass the slower one.

The actual distance the faster vehicle travels while passing is d_p and is estimated as follows:

$$\boxed{d_p = v_f t_p = \frac{v_f d_r}{v_f - v_s}} \tag{11.2}$$

The equations above assume that the faster vehicle can pass "on the fly", i.e., it is able to pass as it is approaching the slower vehicle and does not need to decelerate. If the vehicle needs to decelerate and wait for a suitable opportunity to pass, the time it requires to pass, t_p, would be longer since it would need extra time to accelerate to v_f.

The most current edition of the Green Book [3] provides minimum passing sight distance values, but it does not provide specific equations for estimating them. An older version of the Green Book [4] indicates that the minimum passing sight distance consists of the following four distance components (Fig. 11.3):

$d1$: distance traversed during perception and reaction time and during the initial acceleration to the point of encroachment on the left lane

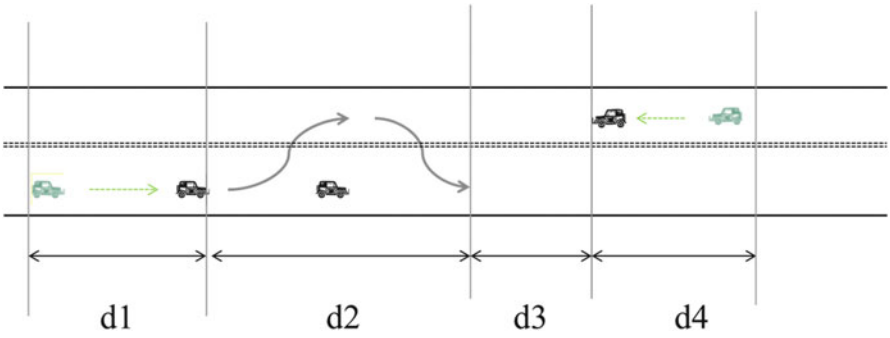


Fig. 11.3 The elements of passing sight distance [3]

$d2$: distance traveled while the passing vehicle occupies the left lane [this can be estimated using Eq. (11.2)]
$d3$: distance between the passing vehicle at the end of its maneuver and the opposing vehicle
$d4$: distance traversed by an opposing vehicle for two-thirds of the time the passing vehicle occupies the left lane, or 2/3 of $d2$ above

Example 11.1 Calculate the passing sight distance for a two-lane highway passing zone where heavy trucks are expected to travel at speeds of 38 mph, and faster traffic is expected to travel at 50 mph. Assume that the acceleration of the faster vehicles is 3 ft/s^2, that the length of the slower vehicle is 30 ft, that the length of the faster vehicle is 18 ft, and that the clearance distance $d3$ is 200 ft.

Solution to Example 11.1
The passing sight distance is estimated as the sum $d1 + d2 + d3 + d4$, as described above. Using Eqs. (1.4) and (1.7), we can calculate $d1$, the distance traveled during perception–reaction time, and acceleration to the final speed, assuming constant acceleration:

$$d1 = v_s t_r + \left(\frac{v_f^2 - v_s^2}{2a}\right)$$

where the first component represents the distance traveled during perception–reaction time while the last two components represent the distance traveled during acceleration.

Based on the information provided:

$$d1 = v_s t_r + \left(\frac{v_f^2 - v_s^2}{2a}\right) = 139.3 + 378.6 = 517.9\,\text{ft}$$

Next, we calculate $d2$ using Eq. (11.2), with d_r as the sum of the respective vehicle lengths plus the spacing between the two vehicles at the beginning and

at the end of the maneuver. If we assume that the spacing between the two vehicles is 30 ft:

$$d_r = 30 + 18 + 30 + 30 = 108\,\text{ft}$$

Then $d2$ is calculated as follows:

$$d2 = \frac{v_f d_r}{v_f - v_s} = 450\,\text{ft}$$

$d3$ is given as 200 ft, while $d4$ is 2/3 of $d2 = (2/3) \times 450 = 300$ ft.
Therefore, the entire passing sight distance based on the assumptions provided is: $d1 + d2 + d3 + d4 = 517.9 + 450 + 200 + 300 = 1{,}467.9$ ft.

Capacity of Two-Lane Highways

Although capacity is regularly observed in other types of facilities such as freeways, there are additional issues associated with the definition of capacity, as well as where and when it should be measured (see Chap. 4 for a detailed discussion of these issues). These issues have not been explored yet to the same degree for two-lane highway facilities.

HCM's two-lane highway capacity estimate has evolved over time. The 1950 HCM [5] specified the basic capacity of two-lane highways (possible capacity under ideal conditions) as being 2,000 pcph total for both directions, regardless of directional split. The practical capacity under prevailing conditions was obtained by using adjustments reflecting effects of lane width, road alignment, trucks, and grades.

The 1965 HCM [6] indicated that the total capacity under ideal conditions is 2,000 pcph for both directions, regardless of the directional split of traffic, which was equal to the basic capacity in the 1950 HCM. The 1965 Manual gives service volume reduction factors for each level of service considering lane width, lateral obstruction, alignment, and trucks. Unlike the 1950 HCM, the 1965 Manual presented a precise procedure for the computation of two-lane road capacities and service volumes. The capacity under prevailing conditions was determined as:

$$C = 2{,}000\, W_c T_c \tag{11.3}$$

where

W_c = width adjustment factor at capacity
T_c = truck adjustment factor at capacity

In the 1985 HCM [7], the capacity of a two-lane highway under ideal conditions was assumed to vary from 2,000 pcph (directional split, 100/1) to 2,800 pcph (directional split, 50/50), total in both directions as a function of the directional distribution of traffic. This capacity reflects the impact of opposing vehicles on passing opportunities and therefore on the ability to efficiently fill gaps in the traffic stream.

Compared with the 1985 HCM, the capacity in the 2000 HCM increased from 2,000 pcph to 3,400 pcph for both directions (or 1,700 pcph for each direction). According to the 2000 HCM [8], for extended lengths of two-lane highway, the capacity will not exceed 3,200 pcph for both directions of travel combined. Reference [9], which was the basis for the HCM 2000 method, indicated that the revised capacity values of 3,200 pcph for two-way flow and 1,700 pcph for directional flow are much less influenced by directional split than was suggested in the 1985 HCM. They indicate that the recommended two-way capacity value is less than twice the directional capacity value, and this 200 pcph difference between 3,200 and 3,400 pcph represents the influence of directional split on capacity.

Similarly, in the 2010 HCM [10], the capacity of two-lane two-way highways is given as 1,700 pcph (passenger car per hour) for each direction of travel (3,200 pcph for both directions), which is independent of the directional distribution of traffic on the facility and of site geometry. A single capacity value however does not reflect the effect by various driver, geometric, and traffic characteristics such as opposing flow, the presence of driveways, horizontal curves, grade, and trucks.

It has been stated in various references (e.g., [9, 10]) that capacity conditions on two-lane roads are very difficult to observe because very few two-lane highways operate at or near capacity. Those references indicate that typically a two-lane highway with high volume is widened to four lanes long before the demand approaches capacity. Reference [10] indicates that higher capacity volumes than ones suggested in the HCM have been observed in the field. Also [11, 12] suggest that the capacity of two-lane two-way highways may reach 3 600 pcph or 4,000 pcph for both directions.

Reference [13] used simulation to estimate the capacity of two-lane highways. It suggested a capacity under base conditions as follows: 1,800 pcphpl at 40 mph average free-flow speed, 2,050 pcphpl at 50 mph, and 2,100 pcphpl at 60–70 mph. The presence of passing zones was not found to have an effect on capacity. The authors suggested that when a driveway or horizontal curve or upgrade is present, the capacity reduction ranges between 3 and 30 % when there are no trucks and may reach up to 40 % when trucks are present in the traffic stream.

Overview of the HCM Procedures

The HCM 2010 [10] procedures for analyzing two-lane highways are based on a combination of traffic flow theory, simulation, and field data collection. The procedures are primarily based on [9]. The LOS is based on two performance measures: percent time spent following and average travel speed. Table 11.1 provides the specific LOS criteria for Class I two-lane highways.

The first step in the methodology is to compare the demand to the capacity; if the ratio is over 1 the LOS is F. If not, the analyst proceeds to estimate the percent time spent following and the average travel speed.

Table 11.1 LOS criteria for two-lane highways: Class I [10]

LOS	Percent time spent following	Average travel speed
A	≤35	>55
B	>35–50	>50–55
C	>50–65	>45–50
D	>65–80	>40–45
E	>80	≤40
F	Flow > Capacity	

The percent time spent following is estimated as a function of the demand, the directional distribution of traffic, and the percent no-passing zones. The average travel speed is estimated by initially obtaining the free-flow speed of the facility and adjusting it to reflect the prevailing demand. Figure 3.8 provides the speed–flow relationship for two-lane highways and for various free-flow speeds. The demand should be estimated in pcph.

One of the complications of the two-lane highways method is that the PCE values are given as a function of flow, which is not known before the PCE value is obtained. Thus, the process to establish the demand values in pcph is iterative.

The HCM 2010 also provides a procedure for evaluating the operational effects of passing lanes.

Microsimulators for Two-Lane Highways

For the analysis of two-lane highways, there are two existing microscopic simulators: TWOPAS (TWO-lane PASsing) and TRARR (TRAffic on Rural Roads). TWOPAS has been developed in the USA, and it considers vehicle characteristics as well as highway design parameters in estimating the performance of two-lane highways. It provides output such as average travel speed, percent time spent following, trip time, delay, number of passes, vehicle miles traveled, and travel time. This simulator however cannot provide capacity estimates, since its algorithms assume that capacity is a fixed value, equal to 1,700 vph. It does not consider the impacts of any geometric or traffic conditions on capacity and cannot handle oversaturated conditions. Determination of capacity cannot be accomplished by a simple change in the software, as its estimation involves many different model components. TRARR also does not account for intersections and varying traffic flow along the simulated road, and it cannot determine capacity. Its output includes derived macroscopic traffic measures such as travel times, journey speeds, percent of time spent following, and overtaking rates.

More recently, CORSIM has included the capability of simulating two-lane highways [14]. The main component added to previous versions of CORSIM is the simulation of passing maneuvers.

References

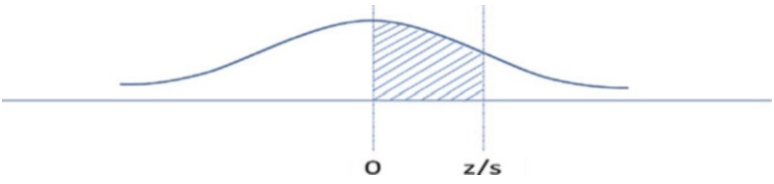
1. FHWA, Highway Statistics 2010. http://www.fhwa.dot.gov/policyinformation/statistics/2010/hm20.cfm
2. Tapio Luttinen R (2001) Traffic flow on two-lane highways: an overview. TL Consulting Engineers Ltd, Lahti
3. American Association of State Highway and Transportation Officials (2011) A policy on geometric design of highways and streets, 6th edn. American Association of State Highway and Transportation Officials, Washington, DC
4. American Association of State Highway and Transportation Officials (2004) A policy on geometric design of highways and streets, 4th edn. American Association of State Highway and Transportation Officials, Washington, DC
5. Transportation Research Board, National Academies (1950) Highway capacity manual 1950. Transportation Research Board, National Academies, Washington, DC
6. Transportation Research Board, National Academies (1965) Highway capacity manual 1965. Transportation Research Board, National Academies, Washington, DC
7. Transportation Research Board, National Academies (1985) Highway capacity manual 1985 Special Report 209. Transportation Research Board, National Academies, Washington, DC
8. Transportation Research Board, National Academies (2000) Highway capacity manual 2000. Transportation Research Board National Academies, Washington, DC
9. Harwood DW, May AD, Anderson IB, Leiman L, Archilla, AR (1999) Capacity and quality of service of two-lane highways. NCHRP Final Report 3-55(3), Midwest Research Institute
10. Transportation Research Board, National Academies of Science (2010) Highway capacity manual 2010. Transportation Research Board National Academies of Science, Washington, DC
11. Yagar S (1983) Capacities for two-lane highways. Aust Road Res 13(1):3
12. Rozic P (1992) Capacity of two-lane, two-way rural highways: the new approach. Transp Res Rec 1365:19–29
13. Kim J, Elefteriadou L (2010) Estimation of capacity of two-lane two-way highways using simulation model. ASCE J Transp Eng 136(1):61–66
14. Li J, Washburn S (2011) Implementing two lane highway simulation into CORSIM. In: 6th International symposium on highway capacity, vol 16. Stockholm, pp 293–305

Problems

1. Conduct a literature review and document two-lane highway designs and traffic operational analysis methods around the world.
2. Using your favorite traffic simulator, analyze operations along a two-lane highway facility with EB demand equal to 1,540 vph and WB demand equal to 1,420 vph. Assume the facility is 2 miles long and is located on level terrain with no horizontal curves. There are 5 % heavy vehicles, and passing is allowed along the entire section. What is the estimated capacity of the facility according to the HCM 2010, and what is its capacity based on the simulation results?
3. Estimate the minimum passing distance for a vehicle with initial speed 50 mph and length 18 ft, if it is to pass a heavy truck with initial speed 25 mph and length 35 ft. Assume that the acceleration of the passing vehicle is 2.5 ft/s^2 and that the clearance distance $d3$ is 200 ft.

4. Conduct a literature review to assess capacity estimation at two-lane highways. How do these estimates differ across different parts of the world and for different designs?
5. Develop a passing algorithm flowchart and program it in your favorite micro-simulation platform. State any assumptions you need to make in order to replicate passing maneuvers. How does your algorithm compare to the AASHTO Green Book-recommended passing sight distance values?
6. Conduct a literature review to assess the simulation of passing maneuvers and the associated gap acceptance models. Develop a recommended process for collecting relevant data to simulate passing maneuvers for various driver and vehicle types as a function of the geometric design characteristics of the two-lane highway.

Appendix A: Standard Normal Table



	Second decimal place for z/s									
z/s	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0800	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.3790	0.3810	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

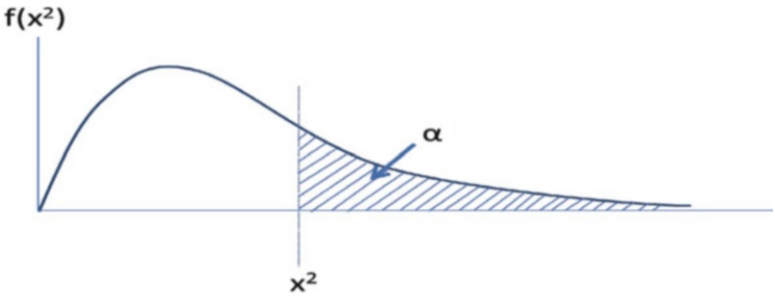
(continued)

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6,

(continued)

	Second decimal place for z/s									
z/s	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998

Appendix B: Chi-Square Table



Degrees of freedom	Level of significance α										
	0.995	0.990	0.975	0.950	0.900	0..500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80

(continued)

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6,

(continued)

Degrees of freedom	Level of significance α										
	0.995	0.990	0.975	0.950	0.900	0..500	0.100	0.050	0.025	0.010	0.005
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.38	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.50	13.24	20.38	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.4	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.43	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

References

Appendix A

1. Blank L (1980) Statistical procedures for engineering, management, and science. McGraw Hill Book, London
2. May AD (1990) Traffic flow fundamentals. Prentice Hall, Engle-wood Cliffs, NJ

Appendix B

3. Blank L (1980) Statistical procedures for engineering, management, and science. McGraw Hill Book, London
4. May AD (1990) Traffic flow fundamentals. Prentice Hall, Engle-wood Cliffs, NJ
5. Washington SP, Karlaftis MG, Mannering FL (2011) Statistical and econometric methods for transportation data analysis, 2nd edn. CRC, Boca Raton, FL

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6,

Index

L. Elefteriadou, *An Introduction to Traffic Flow Theory*, Springer Optimization and Its Applications 84, DOI 10.1007/978-1-4614-8435-6,