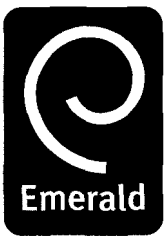
Fundamentals of Transportation and Traffic Operations

by

CARLOS F. DAGANZO
Department of Civil and Environmental Engineering and Institute of Transportation Studies, University of California, Berkeley



United Kingdom – North America – Japan – India – Malaysia – China – Australasia

JAI Press is an imprint of Emerald Group Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2007

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-08-042785-0

Preface

This book is an attempt to present in a self-contained way those basic concepts in the transportation and traffic operations field that should be well understood by every transportation professional. This includes graduate students planning to pursue more advanced studies, as well as newcomers to the field who may be readying themselves for an in depth review of the literature. It is also hoped that academics will find parts of this book suitable for teaching material and/or reading assignments.

The book has evolved from a set of course notes that were prepared for an introductory graduate course in transportation operations currently taught in the transportation engineering division at U.C. Berkeley. The goal of this course is to introduce the basics of transportation operations to a wide crossection of graduate students entering our interdisciplinary program, with backgrounds in civil engineering, city planning, operations research, economics, etc.

The structure and level of the book, as that of the course, is dictated by the necessity to reach such a wide audience in a pedagogically sensible manner. For example, probabilistic concepts are avoided to the extent possible until chapter 6 in order to allow some students to take a concurrent course on probability theory. Elementary calculus concepts, however, are used from the beginning. It is also assumed that the reader has the basic modeling skills that one would develop in an introductory physics course. An effort has been made to represent different things by different symbols within each chapter, and to use a unique symbol for the most important variables used throughout the book. Notational inconsistencies across chapters could not be totally avoided, however, due to the variety of subjects.

The book has chapters on tools (1, 2, 3, and the first part of 6) and others on applications (4, 5, 2nd part of 6, and 7). Very brief introductions to graphical methods, optimization, probability, stochastic processes, statistics and simulation are provided as part of the "tool" chapters. Somewhat unorthodox, these discussions have been made as self-contained as possible, emphasizing the most useful aspects of each tool. This is not the emphasis one usually finds in more specialized books. Readers already familiar with these subjects may skip chapter 3 and the first two sections of chapter 6, although they may find some portions of the discussion entertaining. Chapters 1 and 2 should not be skipped, however.

The book covers some of the application topics in more depth than

would be necessary for an introduction in order to fill gaps in the existing literature. Most notably, "Fundamentals" includes a fairly detailed treatment of "traffic flow theory" in Chaps. 1, 2 and 4. The second half of Chap. 4, covering "traffic dynamics", is more demanding than the rest of the book, but this was necessary for the sake of completeness. A more detailed treatment of this subtle topic is included because certain aspects of it are repeatedly missinterpreted in the published literature. The presentation of this topic stresses the simple traffic theories introduced in the fifties, whose successes and drawbacks are well understood, and ignores modern refinements which have not stood serious scrutiny.

The remaining application topics, "control" (Chap. 5), "observation" (2nd part of Chap. 6) and "scheduled modes" (Chap. 7), use a "building block" approach. Basic ideas involving simple systems (e.g., the timing of a simple traffic signal, the estimation of a bottleneck's "capacity", and the evaluation of passenger delay at a bus stop) are presented in detail and more complicated ones (e.g., networks, estimation of an origin-destination table, and coordination of transit schedules) more qualitatively. An objective was to present the issues clearly, more than a list of specific techniques. As with the material on traffic flow theory, an effort has been made to point out various pitfalls so that they can be avoided. Here too, only that material which is definitely known and correct has been presented in the hope that a newcomer to the transportation field will find in this book a useful source of basic culture.

The application subjects included do not represent a complete survey of those topics one could characterize as "transportation operations" because the book deemphasizes the description of facts (which change as technology changes) in favor of logic. Furthermore, only logical ideas which in my opinion have a solid grounding in physical reality have been included because those are the ones that have the best chance of standing the test of time. This seems appropriate for an introductory book (course) that attempts to prepare the reader for a critical understanding of the field. Of course, many excluded topics deserve treatment in journals and in more specialized books/courses. The reader should turn to these for proper coverage of the current literature.

"Fundamentals of transportation and traffic operations" may be used as a textbook if complemented with problems. A set of solved problems jointly developed with U.C. Berkeley colleagues will be available in the near future and can be ordered by writing to "Institute of Transportation Studies, Publications Office, 109 McLaughlin Hall, University of California, Berkeley, CA 94720" or by sending e-mail to "its@its.berke-

ley.edu". The book can also be used as background reading in graduate and undergraduate courses on transportation and traffic operations. "Fundamentals" also describes a number of computer spreadsheets that can be used for various purposes, including class demonstrations. These can be downloaded from the INTERNET by looking up the book title at "www.ce.berkeley.edu/~daganzo" and following instructions. The problems, but not the solutions can also be downloaded from the INTERNET.

I would be interested in learning of any errors, and plan to issue an errata sheet in conjunction with the set of problems when/if significant ones are found; the errata will also be posted on the INTERNET. Comments may be sent by e-mail to "daganzo@ce.berkeley.edu".

I wish to thank my mentor and colleague Gordon F. Newell for his valued comments, both on the course and on the book. Professor Rod Troutbeck of Queensland University of Technology, Brisbane, Australia, hosted me graciously during a sabbatical leave which made possible preparation of a first draft. His comments and encouragement are also deeply appreciated; the title of the book was suggested by him. Thanks are also due to Prof. Mike Cassidy of U.C. Berkeley for furnishing valuable feedback on those portions of the book he has used in the classroom, to Mrs. Ping Hale for patiently putting up with me while preparing a first draft of the manuscript and to Ms. Esther Kerkmann for doing the graphics. A grant from the University of California Transportation Center made it all possible. Most of all,however, I want to thank my wife, friend and companion, Valery, for her support and understanding.

Carlos F. Daganzo
Berkeley, California
December 29, 1996

Contents

CHAPTER ONE

The time-space diagram

Because the transportation field has not been developed to the point where many of the existing problems can be addressed with well established recipes, transportation professionals are often required to use basic modeling skills and think on their own. This, of course, can only be done effectively if one 'owns' a complete set of thinking tools. Since a basic set of tools is also necessary for a meaningful discussion of transportation operations, this book starts with brief introductions to those 'deterministic' tools that are not always covered in an undergraduate engineering curriculum: first the time-space diagram (Chapter 1), then cumulative plots (Chapter 2) and finally optimization (Chapter 3).

The material in chapters 1 and 2 is necessary to describe and think about the collective motion of items over guideways. The diagrams presented in these chapters are 'fundamental' in that they shed much light on time and motion problems that one may be trying to understand. Chapters 4 (traffic flow theory) and 7 (scheduled systems) rely on these diagrams extensively. The material in chapter 3 is often useful for the design and/or control of systems that are already well understood.

The present chapter introduces the time-space diagram and its application to the study of vehicular motion. It is organized in three sections. Section 1.1 discusses the motion of a single item, Sec. 1.2 that of many items sharing a guideway, and Sec. 1.3 some applications to more specific scheduled and unscheduled transportation problems.

1.1 Trajectories for a single vehicle

Very often in the analysis of a particular transportation operation one has to track the position of a vehicle over time along a 1-dimensional guideway as a function of time, and then summarize the relevant information in an understandable way. This can be done by means of mathematics if one uses a variable x to denote the distance traveled along the guideway from some arbitrary reference point, and another variable t to denote the time elapsed from an arbitrary instant. Then, the desired information can be provided by a function x(t) that returns an x for every t in the relevant range for our application.[1]

A graphical representation of x(t) in the (t,x) plane is a curve which we call a *trajectory*. As illustrated by two of the curves in Fig. 1.1, trajectories provide an intuitive, clear and complete summary of vehicular motion in one dimension. Curve 'a', for example, represents a vehicle that is proceeding in the positive direction, slows down and finally reverses direction. Curve 'b' represents a vehicle that resumes travel in the positive direction after nearly stopping. Curve 'c', however, is not a representation of a trajectory because there is more than one position given for certain t's (e.g. t_0); such a curve is not the representation of a (single-valued) function x(t). Valid vehicle trajectories must exhibit one and only one x for every t. For problems requiring a level of resolution comparable or finer than the vehicular length, e.g., when tracking the position of a mile-long train over a vertical curve, the curve x(t) should refer to a particular point of the vehicle such as the vehicle's front, rear or center of gravity. Any point is valid, provided the diagram is interpreted in accordance with the choice.

In some practical applications a vehicle's trajectory must be developed analytically from knowledge about the operating characteristics of the vehicle and the guideway such as: the vehicle mass, resistive forces, engine horsepower, guideway elevation profile, etc. Examples of such applications are: (1) determination of the minimum travel time between stations for a transit train when one is given the maximum operating speed, the maximum acceleration and the maximum allowable 'jerk'[2], as well as the distance between stations; (2) determination of a vehicle's

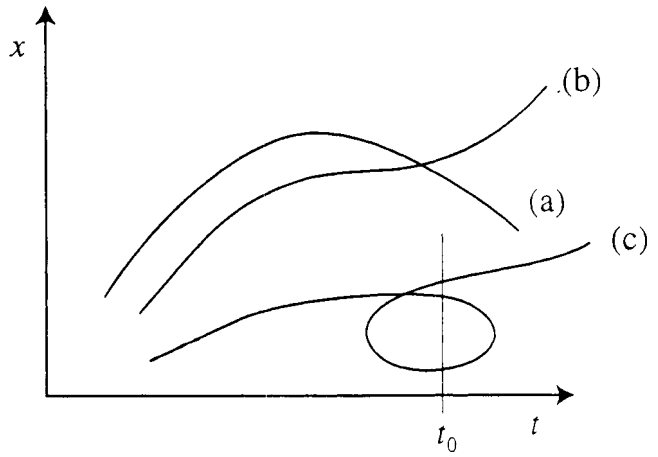


Figure 1.1 Time-space curves: (a) and (b) are vehicle trajectories; (c) is not.

initial speed from its skid marks and its estimated collision speed, given the vehicle's coefficient of friction and the road geometry; (3) studies of runway length and taxiway exit location, which use as inputs airplane deceleration characteristics and an initial speed to predict the distance traveled to achieve a target speed; (4) similar studies for the length of acceleration and deceleration lanes of freeway on-ramps and off-ramps, (5) calculation of high-speed rail travel times over rugged terrain as a function of the engine power and the vertical profile of a proposed alignment.

In other applications the trajectory of a vehicle can be recorded, e.g. with a video-camera, and the objective is to convert the raw data into a curve x(t) that can then be studied mathematically. Sometimes, as may happen for transit systems using an automated vehicle monitoring system, no conversion may be necessary at all. In other cases, however, data may only be available in the form of observed vehicle positions at discrete times, as happens for example in elevator and public transportation studies when only the times at which each vehicle arrives and leaves each stop are recorded. Then the full set of vehicle trajectories may be approximated by interpolation.

Later in this chapter we will see how representing multiple vehicle trajectories on the same time-space diagram, whether analytically or experimentally obtained, can help solve many problems. Before this can be done a more detailed look at single vehicle trajectories is in order, although it is with multiple vehicles that the technique really shines.

We recall from elementary physics that the first and second time-derivatives of a vehicle trajectory (e.g. curve 'a' of Fig. 1.1) represent the velocity, v, and acceleration, a, of the vehicle; i.e., that $v(t) = dx(t)/dt$ and $a(t) = d^2x(t)/dt^2$, or in abbreviated form:

$$v = dx/dt \text{ and } a = d^2x/dt^2. \tag{1.1}$$

Although Eq. (1.1) is widely known and its qualitative graphical consequences are rather obvious, it is worth emphasizing that steeply increasing (decreasing) sections of x(t) denote a rapidly advancing (receding) vehicle; horizontal portions of x(t) denote a stopped vehicle and shallow sections a slow-moving vehicle. Straight line segments depict constant speed motion (with no acceleration) and curving sections denote accelerated motion; here, the higher the curvature, the higher the absolute value of the acceleration. Concave downwards curves (such as curve 'a') denote deceleration and concave upwards (convex) sections denote accelerated motion.

The reader is encouraged not just to understand rationally these properties, but also to draw and study a large number of examples in an

effort to have the above qualitative properties become second nature. The ability to look at an x(t) curve and tell *immediately* what is happening will turn out to be rather useful later on.

The five problems mentioned earlier all have in common that from a knowledge of the forces acting on a vehicle and some initial conditions at time $t = 0$, it is possible to write an ordinary differential equation for the vehicle trajectory with x as the unknown, involving the acceleration. This can be done after calculating the resultant force, F, acting on the vehicle in the direction of travel and dividing it by the vehicle mass, m; i.e., invoking Newton's second law of motion: $F = ma$. Subsections 1.1.6 and 1.1.7 will show how this is done. Earlier subsections give a brief qualitative description of the various components of F and how they depend on key attributes of the vehicle and the guideway.[3]

Because for most problems it makes sense to assume that m is constant, we will express all the force components, F_i, in terms of the acceleration that they would induce, $a_i = F_i/m$. A further simplification is achieved by restricting our attention to longitudinal forces and accelerations; i.e., those acting in the direction of the guideway. This obviates the need for vector notation, since a + or − sign suffices to define the direction of each longitudinal component. For problems where vehicles don't reverse directions, we will take positive magnitudes to denote forces acting in the direction of travel (imparting acceleration) and negative magnitudes, forces in the opposite direction.

1.1.1 Propulsive force, F_p

This is a force that the guideway exerts on the vehicle (and vice versa). It usually varies with time as per the 'driver' input, but is always limited by the engine power and (for land vehicles) by the coefficient of friction in the following way:

$$\frac{F_p}{m} = a_p \leq g \min\left\{f, \frac{\kappa}{v}\right\} \tag{1.2a}$$

where g is the acceleration of gravity (about 32.2 ft/sec^2), f is a dimensionless coefficient of friction (about 0.3 for cars and buses[4], smaller for rail, and of no consequence for air and water vehicles), and κ is the power to weight ratio of the vehicle, which has units of speed.

The constant κ can be calculated as the ratio of the engine power and the weight of the vehicle assembly in some consistent system of units. The vehicle assembly must include any independent units pulled by the engine, and their loads; e.g. railroad cars pulled by a railroad

engine. Thus, care must be exercised in adopting a 'rated' power to weight ratio of a given power plant.

Formula (1.2a) recognizes that more acceleration can be developed at lower speeds and that the maximum is limited by the adhesion between the vehicle and the ground. The formula is only a coarse approximation to the actual acceleration that can be elicited from real engines in that (1.2a) assumes that the engine can develop maximum power at every speed. (This approximation is better for electric motors than for internal combustion engines, and realism of (1.2a) for the latter depends on the capability of the gearbox to keep the engine's rpm's near the optimum level at various speeds.)

Counteracting the propulsive (positive) force of the engine there usually are various forms of resistive (negative) forces; i.e. fluid resistance, rolling resistance, braking resistance and guideway resistance. These are reviewed next. The discussion below assumes initially that the vehicle in question is traveling on a horizontal path. Corrections for climbing, descending and curving paths are introduced in Sec. 1.1.5.

1.1.2 Fluid resistance, F_f

This is the force that air and/or water exerts against the vehicle. Although more accurate expressions can be developed for specific vehicles and specific media, a good all-purpose approximation for this force is:

$$\frac{F_f}{m} = -\alpha v_r^2 \tag{1.2b}$$

where v_r is the vehicle speed relative to the fluid, and α is a coefficient of drag. The coefficient of drag, which depends on the vehicle's cross-section and its aero/hydro-dynamic shape, has units of $(\text{distance})^{-1}$.

Equation (1.2b) is based on Bernoulli's principle that fluid resistance past a solid varies approximately with the square of the fluid speed. Although a more precise description of this law would involve the vector velocity and a vector force resultant, for most transportation applications (airplanes excepted) our scalar equation suffices if the F_f is interpreted to act in the opposite direction as the vehicle motion in a coordinate system moving with the fluid.

For the motion of a ground vehicle in still air, v_r is the vehicle speed $v_r = v$, and Eq. (1.2b) gives the force acting against the vehicle in the direction of travel. If the air is not still but cross-winds can be neglected, then $v_r = v - v_a$, where v_a is the air-speed relative to the ground in the

direction of travel. It is important to note that if $v_r < 0$ the sign of (1.2b) should be reversed.

For water and air vehicles the analysis is easier if one works in a coordinate system that moves with the fluid (by definition then $v = v_r$) and one later obtains the trajectory of the vehicle relative to the ground after composing the motions of the fluid and the vehicle (relative to the fluid) in the usual Galilean way.

1.1.3 Rolling resistance, F_r

For ground transportation vehicles we must also consider rolling resistance, and this is usually approximated by a linear function of speed. At high speeds rolling resistance is not as important a factor as fluid resistance.

1.1.4 Braking resistance, F_b

For ground vehicles, this force depends on the force with which the brakes are applied, up to a maximum that depends on the friction coefficient between the wheels and the guideway. Thus, we can write:

$$F_b/m \geq -gf. \tag{1.2c}$$

Note that (1.2c) is a bound; in practical applications $F_b = 0$ if $F_p > 0$ (and vice versa), since the brakes and the throttle are rarely applied simultaneously.

For air and water vehicles braking is achieved by increasing the coefficient of drag and/or reversing the engines so that the propulsive force applies in the negative direction; i.e., changing α and κ. In all cases, the braking force actually applied may be a function of time.

1.1.5 Guideway resistance, F_g: Motion along a profile

We consider here the contribution toward acceleration of the earth's gravity when the vehicle travels on a slanted path, and also the effect of path curvature. When the vehicle path is linear (e.g. a truck on a fixed grade) the force in question is negative (positive) if the vehicle is climbing (descending). If we let y denote the elevation of the vehicle at position x ($y = y(x)$), and we assume that the slope of the vehicle's path (expressed as a dimensionless fraction, $\beta(x) = dy(x)/dx$) is small compared with 1 we can write from a simple force diagram:

$$F_g/m = -g\beta. \tag{1.2d}$$

This formula is only an approximation because the dimensionless factor multiplying '$-g$' is actually the sine of the angle made by the path and the horizontal and not the tangent, β. Of course, for small angles (1.2d) is very close to the exact expression, which is $-g\beta/(1+\beta^2)^{1/2}$. For ground transportation, β is the slope of the guideway which is known a priori as a function of x. For air travel, β depends on 'driver' decisions, but these are usually prespecified in most practical applications.

As all the previously introduced forces, F_g acts in the direction of travel. For ground transportation problems, however, where $\beta \ll 1$, it is customary to measure distances, x, in the horizontal direction and then study the motion of the horizontal projection of the vehicle. Although the horizontal projections of forces (1.1) and (1.2) would then have to be used in the equation of motion of the horizontal projection, this step is usually (and justifiably) omitted when $\beta \ll 1$.

The use of a constant friction factor in expressions (1.2a) and (1.2c) assumes that the full weight of the vehicle rests on the road at all times. This is not accurate when a vehicle is traveling fast on a sharp vertical curve since it ignores the vehicle's centrifugal tendency, which has a levitating effect on crest vertical curves. The error can be corrected by replacing f by the following:

$$f_n \approx f[1 + v^2 c/g] \tag{1.3}$$

where c is shorthand for $c(x) = d^2y(x)/dx^2$. The amount in brackets represents the factor by which the normal force between the vehicle and the guideway is increased due to the latter's curvature. The expression can be easily derived from a force diagram.[5]

1.1.6 Analytical derivation of a trajectory

Summation of Eqs. (1.1) and (1.2) results in an expression for the instantaneous acceleration of a vehicle. If the actions of the 'driver', and the characteristics of the driver and the guideway are given, then the instantaneous acceleration will be a function of v, x and t: F(v, x, t).[6]

If the path followed by the vehicle is linear and homogeneous, and the driver inputs only depend on v (or are constant), then the instantaneous acceleration will only depend on v and we can write:

$$\frac{dv}{dt} = F(v), \tag{1.4}$$

after eliminating the irrelevant arguments from 'F'. This is a differential equation that can be solved by separating variables; i.e., by integrating

the separated version of the equation, $dv/F(v) = dt$, between some known initial conditions $(v, t) = (v_0, 0)$ and some unknown final (or intermediate conditions) (v_1, t_1):

$$\int_{v_0}^{v_1} \frac{dv}{F(v)} = \int_0^{t_1} dt.$$

This result defines a relation between v_1 and t_1, $G(v_1) = t_1$, which indicates the times at which various speeds are reached. Solving for v_1, we may also express the result as follows:

$$v_1 = G^{-1}(t_1), \tag{1.5}$$

where G^{-1} is the inverse of G. This equation can in turn be rewritten as:

$$\frac{dx}{dt} = G^{-1}(t) \tag{1.6}$$

after eliminating the 'dummy' subscript '1' from (1.5) and then substituting dx/dt for v on the left side.

Equation (1.6) can now be treated like Eq. (1.4), so that after separation of variables and another integration between initial and final limits $(x_0, 0)$ and (x_1, t_1) one obtains the vehicle trajectory in a form: $(x_1 - x_0) = H(t_1)$.

If the effects of speed and time can be neglected but not the effects of position, e.g. in the braking of a car at low speeds on a vertical curve, we can express the instantaneous acceleration as a function of x alone, F(x), and the ordinary differential equation of motion becomes:

$$\frac{dv}{dt} = F(x). \tag{1.7}$$

In this case the time variable can be eliminated from the equation by writing dv/dt as $(dv/dx)(dx/dt) = vdv/dx$. The result is an equation for v as a function of x that can be written in the separated form:

$$vdv = F(x)dx.$$

Integration between initial and final values $(v_0, 0)$ and (v_1, x_1) now yields:

$$1/2(v_1^2 - v_0^2) = I(x_1)$$

where 'I' is the result of integrating 'F' between 0 and x_1. This expression can be solved for v_1. Then, on eliminating the subscript '1' and replacing v by dx/dt we have:

$$\frac{dx}{dt} = (v_0^2 + 2I(x))^{1/2}, \tag{1.8}$$

which, again, can be integrated between initial and final limits after one more separation of variables. The result,

$$\int_0^{x_1} \frac{dx}{\sqrt{v_0^2 + 2I(x)}} = \int_0^{t_1} dt \tag{1.9}$$

is a relation between x_1 and t_1 which is the sought vehicle trajectory.

If (1.9) cannot be integrated, or the equation $dv/dt = F(v, x, t)$ is not of a form that can be easily solved, the solution can be found numerically.

1.1.7 Numerical derivation of a trajectory

Computer spreadsheets such as 'LOTUS', 'QUATTRO' and 'EXCEL' can be used to solve differential equations of the type introduced here quite easily. The idea is to step through time (or distance) in steps of small and constant size, Δt (or Δx), recording at each step all the information that is necessary to calculate the information at the next step; i.e.: t, x, v, a, y, $\beta = dy/dx$ and $c = d^2y/dx^2$.

To do this, one should reserve a separate column (properly labeled for future reference) for each of these seven pieces of information, and then should enter as data any columns that are known and the known information at $t = 0$, $x = 0$. Fig. 1.2 shows as an example the organization of a spreadsheet that calculates the travel times of a vehicle for different distances. The guideway profile is given and therefore the columns for x, y, $\beta(x)$ and c(x) have been entered as data. Also, as part of the initial data we have $t = 0$ and $v = v_0$. The remaining cells of the first and second rows of the spreadsheet contain recursive formulae that are the discrete equivalent of the equation: $dv = F(v, x, t)dt$. Once copied across all rows the table is automatically filled with the approximate solution to the problem. The formulas for this particular problem are displayed in Fig. 1.2; the arrows pointing to each of the three formulas indicate the source of the data for the arguments of the formulas. (The formula for v is the well-known expression for the final speed of a uniformly accelerated object over a fixed distance Δx.) The formula for F may use as inputs data (e.g. α, m, κ...) that may be stored in cells outside the range shown in the figure.

To make sure that the approximation is sufficiently accurate it is good practice to reduce Δx (or Δt, in other cases) and check that the solution doesn't vary appreciably.

A spreadsheet 'TRAJTRY.WK1' has been made available to the public. (See preface for information on how to obtain this and other

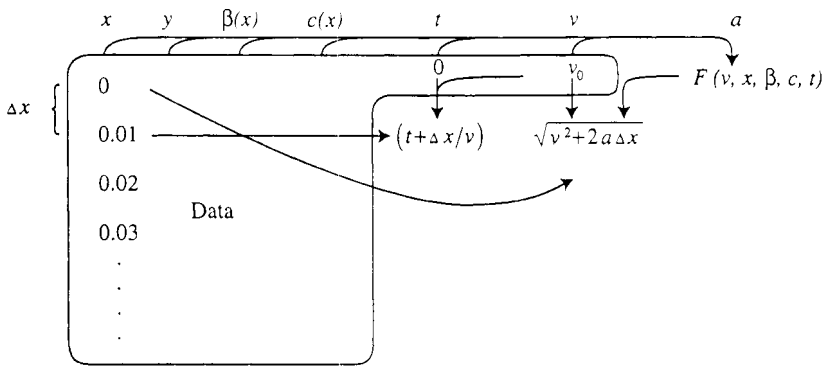


Figure 1.2 Organization of a vehicle-trajectory spreadsheet with a distance increment of 0.01 units.

spreadsheets mentioned in this book.) It has the organization of Fig. 1.2, but the terrain profile has been entered as the formula:

$$y = 100 + \bar{\beta}x + 10\sin(x/200)$$

which relates x and y (in meters). In this expression $\bar{\beta}$ is the user-specified average slope of the road. The term 10sin(x/200) has been added to mimic the profile of an undulating road with a slope that fluctuates in a $\pm 5\%$ range about the average every 1200 meters or so. Of course, a real vertical profile would be piece-wise parabolic, but if it undulated in the manner described it would be close to the one we have adopted and would generate similar vehicle trajectories. The advantage of defining the profile with a simple formula, for illustration purposes, is that the frequency and severity of the undulations may then be easily changed in the spreadsheet by substituting other values for the coefficients '10' and '200' that appear in the formula. With this goal in mind, the spreadsheet also includes a block of data with the parameters '$\bar{\beta}$', 'α' (in meters^{-1}), 'κ' (in m/s), and 'fg' (in m/s^2). It is then a simple matter to perform sensitivity analysis. The graphical capabilities of spreadsheets allow the engineer to see at the push of a button the resulting vehicle trajectory for a given set of conditions. For our particular illustration you might want to explore how vehicles with various power to weight ratios are affected by varying average slopes, '$\bar{\beta}$'. It is also possible to replace the 'y' column by data corresponding to an actual profile and in this way solve a real problem.

This example should serve to illustrate that spreadsheets can be used as powerful calculators that can solve integrals, and many kinds of equations if used properly.

1.1.8 Additional background

Chapters 4 and 5 of Hay's book (Hay, 1977), and other transportation texts, such as Mannering and Kilareski (1990) and Haefner (1986) contain extensive discussions of the performance characteristics of various types of vehicles. Texts on engineering mathematics, differential equations and/or elementary physics can be consulted for additional information and examples on ordinary differential equations of the type presented here. A concise description of numerical methods for differential equations is given in Chapter 15 of Press *et al.* (1986). Spreadsheet implementation is discussed in Oris (1987).

1.2 Trajectories for many vehicles

The discussion on single vehicle trajectories focused on their derivation from available information, and it was understood that once derived one could look at the picture of the trajectory to 'see' what was happening. The true power of the time-space (t,x) diagram, however, cannot be appreciated until one has trained oneself to see what is happening by looking at a plot *without having to rationalize why*; and then using this skill to interpret plots including many vehicle trajectories. It is for the analysis of problems where many vehicles interact while proceeding along a common right-of-way that the time-space diagram becomes an invaluable tool. Three applications come to mind: (i) airplanes with various gliding speeds sharing a landing approach path subject to minimum spacing requirements, (ii) scheduling of freight (slow) and passenger (fast) trains along a single-track railroad line with passing allowed at predetermined sidings, and (iii) estimation of safe passing sight distances on two-lane bidirectional highways from the acceleration and speed characteristics of the passing, passed and opposing vehicles. Similar applications exist in all modes of transportation; e.g. public transit scheduling, elevator system design, etc. While most analyses can be done without the (t, x) diagram, it helps to identify and correct errors that may have been committed in the formulation. This is dramatically illustrated by the following puzzle.

Three friends take a long trip using a tandem bicycle for 2 persons. Because the bike riders travel at 20 km/hr, independent of the number of riders, and all three persons walk at 4 km/hr, they proceed as follows: To start the journey, friends 'A, B' ride the bicycle and friend 'C' walks; after a while, friend 'A' drops off friend 'B' who starts walking, and 'A' rides the bicycle alone in the reverse direction. When

'A' and 'C' meet, they turn the bicycle around and ride forward until they catch up with 'B'. At that moment the 3 friends have completed a basic cycle of their strategy, which they then repeat a number of times until they reach their destination. What is their average travel speed?

Although possible, very few people can solve this problem in less than 5 to 10 minutes (about 5% in my experience) and this happens only because they haven't identified the easy way to look at it... the (t, x) diagram! If you plot the trajectories of the 4 moving objects (the bicycle and friends 'A, B, C') on a (t, x) diagram you will find by inspection that the average speed is 10 km/hr. To obtain this result, of course, you must make sure that the trajectories of the objects overlap with one another in a way that is consistent with our word description of the problem. You can arrive at the solution either by measuring the relevant slope in a (t, x) diagram drawn to scale, or analytically; in the latter case, the (t, x) diagram will help you set up the equations correctly.

The above exercise illustrates the usefulness of the (t, x) diagram for problem solving and design. The diagram, however, is also an excellent tool for diagnosing problems in existing systems because it includes all relevant information regarding the progress of the vehicles on the system during a study period and displays such information in a way that can be readily interpreted by the trained professional.

1.2.1 Construction of the (t, x) diagram from data: completeness

The form of the data based on which trajectories can be built will depend on the application. For example, if one has some means of tracking the positions of the buses on a transit line then the trajectory of each individual vehicle can be constructed (e.g. from dead-reckoning navigation data if such technology is used), and it is then a simple matter to superimpose all the trajectories on the same diagram. A similar result would be obtained from logs kept by drivers.

For other systems (e.g. elevators, pedestrians, autos, etc...) it is more convenient to record the times at which individual objects pass by stationary observers. When the data are displayed by means of tick marks on horizontal lines at the locations of the observers as shown in Fig. 1.3a, vehicle trajectories are retrieved by connecting the ticks for specific vehicles. For this to be possible each observer must identify each passage time with a vehicle 'signature', which in the case of an automobile could be its license plate or the electronic pattern it leaves

on an automatic detector. In most applications this is more easily said than done and often we have to work from the time-series alone. This would be fine if vehicles don't pass because then (barring detector errors) as soon as one of the trajectories is identified the rest follow. It would also be fine if the observation stations were to be so closely spaced that only small vehicular speed changes over the detector spacing could arise because then it would also be obvious which vehicles are which.

Another form of data (used in connection with freeway studies) arises from time-lapse aerial photographs. Because each photograph is taken at a specific time, t, it is associated with a 'vertical' line on the time-space diagram, as shown in Fig. 1.3b. One can then display by means of dots on the line the location of the 'noses' (or 'tails') of every vehicle at each sampled instant. The photographs automatically display vehicle 'signatures', thanks to their pictorial detail, and this makes it possible (although very tedious and impractical) to connect the appropriate points with smooth lines to develop the vehicle trajectories. This method of construction, however, illustrates that the time-space diagram is a complete summary of the 1-dimensional progress of our vehicles. We note that Fig. 1.3b could have been obtained by actually laying the strips of film side by side and that if these were viewed across a vertical slit that was moved from left to right at an appropriate rate, one would be replaying a movie of the system's evolution! In other words the (t, x) diagram gives a complete description of the history of our vehicles' longitudinal motion.

Besides displaying field data in a complete way, the recipes for constructing Fig. 1.3a and b are also important because they indicate a reverse way in which the (t, x) diagram can be 'read'. In particular note that a horizontal line through the diagram (e.g. at position x_3 in Fig. 1.3a) identifies the times at which successive vehicles pass a stationary observer, and that a vertical line at a given abscissa (e.g. time t_4 in Fig. 1.3b) identifies the vehicle positions at the given time. The truth of this statement does not depend on how the (t, x) diagram was developed. The times between consecutive vehicle observations at a fixed location, h_i, are usually called *headways*, and the distance separations between consecutive vehicles at a given instant, s_j, *spacings*.

1.2.2 Definitions of traffic stream features

The number of vehicles observed by a stationary observer during a given time interval, m, divided by the length of the time interval, T, is the

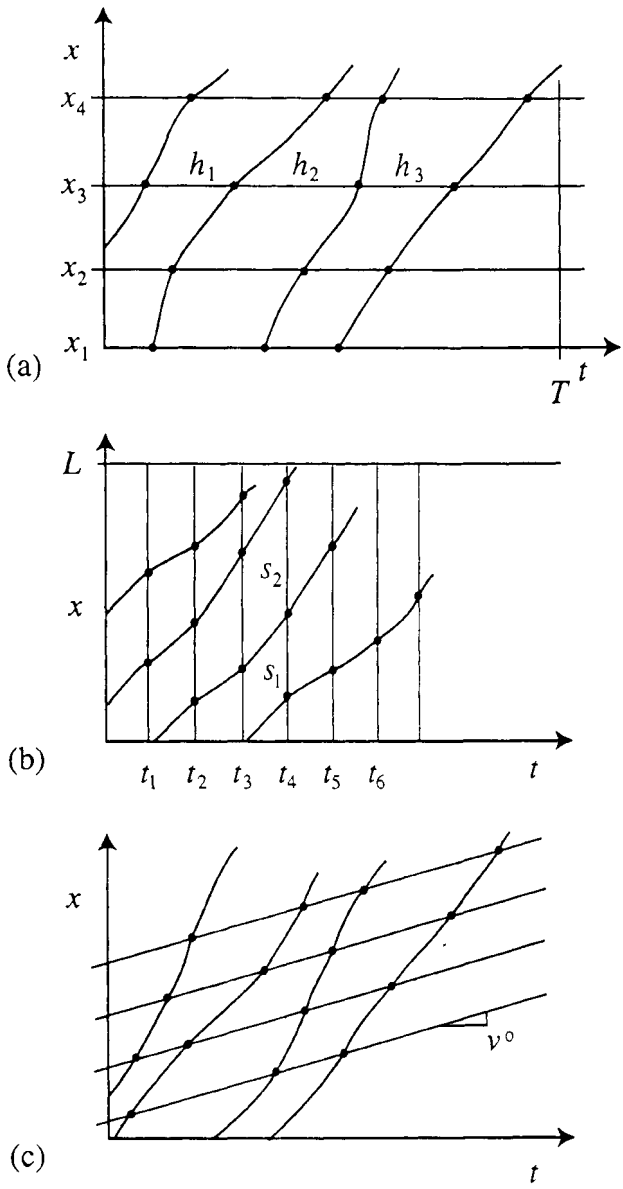


Figure 1.3 Three ways of gathering (t, x) trajectory data: (a) roadside observers at various locations; (b) aerial photographs at different instants; (c) moving observers.

flow, $q = m/T$, for the interval; e.g. the observer at $x = x_3$ in Fig. 1.3a observes a flow, $q = 4/T$, during $0 \leq t \leq T$. It should be clear from Fig. 1.3a that for long observation periods including many vehicles ($m, T \rightarrow \infty$) with comparable headways,

$$T \approx \sum_{i=1}^{m} h_i,$$

and therefore, on dividing both sides of this expression by m, we obtain the important relation:

$$q^{-1} = \frac{T}{m} \approx \frac{1}{m} \sum_{i=1}^{m} h_i = \bar{h}\,; \tag{1.10}$$

i.e., under the conditions stated, the flow over an interval is approximately equal to the reciprocal of the average headway seen by a stationary observer during the interval. We note that this relationship is exact if the observation period starts and ends immediately before the arrival of a vehicle. The concept of flow is equivalent to the terms '*volume*', used in certain traffic engineering circles, and '*frequency*', used in connection with scheduled transportation modes.

A similar treatment of the number of vehicles seen on a photograph, n, over a stretch of road of a given length, L, leads to the concept of *density*, $k = n/L$, over the stretch and a parallel relationship of the density with the average spacing:

$$k^{-1} = \frac{L}{n} \approx \frac{1}{n} \sum_{j=1}^{n} s_j = \bar{s}. \tag{1.11}$$

As with headways, the quality of the approximation improves for $L \rightarrow \infty$, and the relationship becomes exact when both ends of the interval are immediately ahead of a vehicle.

It should be noted here that other vehicle characteristics (besides spacings and headways) can be averaged across space or time as well; e.g. vehicle occupancies, speeds, etc... and that there is no a-priori reason to expect averages taken across space or time to be the same. Averages taken at a specific location (with time-varying over an interval) are called 'time-means', whereas those taken at an instant over a space interval are termed 'space-means'; e.g. space-mean speed, $\bar{v}_s$, and time-mean speed, $\bar{v}_t$, are the terms used to denote the speed averages obtained in the aforementioned way.

Fig. 1.3c describes one more way in which trajectory data can be recorded (and in which the t, x diagram can be interpreted). It involves observers traveling at a constant speed v° that record the times at which

vehicles pass them. The observer trajectories are then plotted and used to locate points in the (t, x) plane through which the vehicle trajectories must pass. In Fig. 1.3c traffic passes the slow-moving observer, but similar figures could be drawn for observers moving faster than traffic and moving against traffic; e.g. if observers on a fast-moving car pool (or contra-flow) lane record the times at which they pass individual vehicles on the general use lanes. Note that the interpretation of Fig. 1.3c generalizes the prior two interpretations because $v^\circ = 0$ leads to Fig. 1.3a and $v^\circ \to \infty$ to Fig. 1.3b.

1.3 Applications of the (t, x) diagram

Here we present two applications of the time-space diagram. The first application is a preview of traffic flow theory for an idealized case that, despite its simplicity, clearly reveals some interesting relationships between traffic flow variables; in this application, the (t, x) diagram helps in the mathematical development, but most importantly it shows physically why the derived expressions are true. The second application is a scheduling problem where vehicles compete for a common right-of-way; there the (t, x) diagram is also used as an aid for thinking that helps eliminate mistakes, and just as importantly, it can be used as an elegant way of displaying the solution that could be used in a professional report.

1.3.1 Traffic flow theory with straight trajectories

We consider a section of road length L that is observed for time T and assume that vehicles travel over the section (approximately) at constant speed without interacting with one another. This scenario could arise in lightly traveled multi-lane freeways with fast and slow vehicles, and in airport corridors with mechanical transportation devices that only a fraction of the people use.

We will also assume that there are only a finite number of speeds v_l that vehicles adopt and that the trajectories of each vehicle family are evenly spaced straight lines. This means that all the vehicles of family 'l' have the same headway *within the family*, h_l. Here, h_l denotes the time separation between two consecutive vehicles of family l; the headway between consecutive vehicles will in general be smaller and will not be constant. This can be seen clearly from the diagram of Fig. 1.4 for the special case where there are two vehicle classes, $l = 1, 2$.

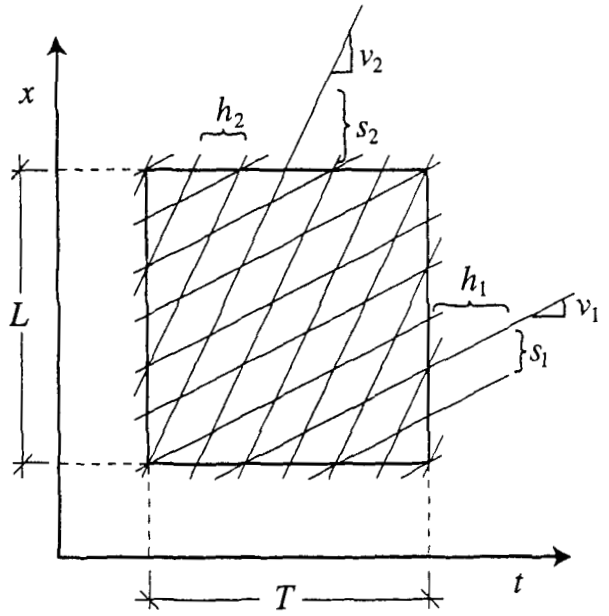


Figure 1.4 Time-space trajectories of two vehicle families.

It can be seen from the geometry of the figure that $h_l v_l = s_l$ for each vehicle class, l, where s_l is the spacing for the class. If the class flows, q_l, and densities, k_l, are defined over intervals containing many vehicles, we can accurately rewrite this relation as:

$$q_l \approx v_l k_l \tag{1.12}$$

by virtue of (1.10) and (1.11).[7] If (1.12) is now added across l, and we recognize that the total flow and density are:

$$q = \sum_l q_l \quad \text{and} \quad k = \sum_l k_l,$$

we find that:

$$q \approx k \sum_l v_l (k_l/k) = \bar{v}_s k. \tag{1.13}$$

The second equality is justified on noting that the summation in the middle member of (1.13) defines a weighted average of the vehicle speeds where the weighting factors are the fraction of vehicles by type *seen in an aerial photograph.* (This statement follows from the definition of density, given earlier.)

The emphasis is given because the fractions of vehicle types seen in an aerial photograph are usually different from those that would be counted by a stationary observer (q_i/q). To see this intuitively you should refer to Fig. 1.4 and note that a stationary observer sees approximately two fast vehicles for every slow one, but a photograph would show two slow ones for every fast one! You should ponder why, and realize that the stationary observer will invariably see higher fractions of fast vehicles than shown in the aerial photos, independent of the specific speeds, flows and densities of the vehicle families. (Can you imagine what would happen if one of the families had $v_i = 0$; e.g. it corresponded to parked cars?). A related result which should come as no surprise is that the inequality $\bar{v}_t \geq \bar{v}_s$ is generally true[8] for long observation intervals when traffic behaves as described in this section.

A similar disparity should also be expected between time and space averages of other quantities that vary across families, but remain constant within a family. For example, if the fast vehicles of Fig. 1.4 are car-pools (with a 2 person vehicle occupancy) and the slow vehicles are driven without passengers, it should be clear that the average vehicle occupancy will be different depending on the method of observation. Can you figure out what the average vehicle occupancy measured by a moving observer with speed v° (for $v^\circ = 0$, v_1, v_2 and ∞) would be in the example of Fig. 1.4?

The same multiplicity of averages would be obtained for other measures such as energy consumption and pollution generation that vary across vehicle classes with different speeds (e.g. buses and cars; commercial jets and private airplanes, etc...). Which is the 'real average' then? The answer to this question cannot be given absolutely. It depends on the practical problem that motivates your particular analysis and this is why it is important to understand the fundamentals.

1.3.2 Closed loops

It should also be noted that the (t, x) diagram can be used to describe closed loop systems. If we use x to denote the position of a vehicle within the loop ($0 \leq x \leq L$, where L is the length of the loop) then the vehicle's trajectory will 'disappear' upon reaching the coordinate $x = L$, and will simultaneously reappear at $x = 0$. The trajectory of a vehicle that travels at a constant speed along the loop then adopts a 'saw-tooth' shape as shown in Fig. 1.5.

The figure depicts the (parallel) trajectories of 4 vehicles equally spaced on the loop. Such a diagram could represent the behavior of

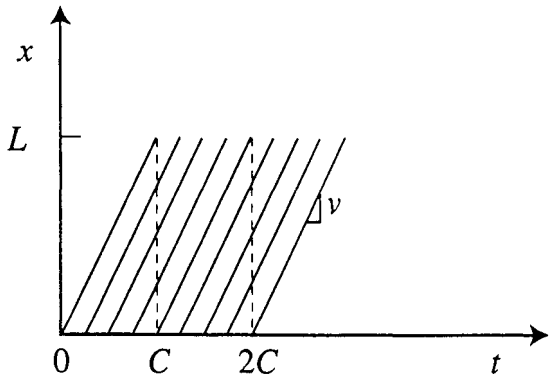


Figure 1.5 Vehicle trajectories on a loop.

four buses serving a specific route, if the wiggles in the individual trajectories due to (random) fluctuations in speed and brief stops had been smoothed out. For such a system the average speed of a vehicle is (by definition) $v = L/C$, where C is the vehicle's cycle time. If there are n vehicles on the route at all times, then, $q = n/C$ is the vehicular flow during any given cycle (see figure) and $k = n/L$ is the density on the route. This means that the definition $v = L/C$ is equivalent to the relation $v = q/k$.

A similar commentary could be made for two-way systems such as elevators, where on the way back the vehicles retrace the positions traveled on the way out (e.g. from point O to point P and then from point P to point O.) Then, if the variable x is defined as the distance traveled within a cycle, with point P located at $x = L/2$, Fig. 1.5 still applies. For two-way systems, however, it is also correct to define x as the distance from O (with $0 \leq x \leq L/2$) and work with a slightly different (t, x) diagram. Can you draw such a diagram for the system of Fig. 1.5?

The relationships $q = k\bar{v}_s$ and $\bar{v}_t \geq \bar{v}_s$ also hold for closed loop systems with straight vehicle trajectories, since the mathematical arguments given about the superimposition of various vehicle classes also extend to this case. As an illustration, the reader may verify that if one has three vehicles on a 2 mile racetrack ($k = 1.5$) traveling at $v_k = 100$, 120 and 140 miles/hr (so that $\bar{v}_s = 120$), then the vehicular flow seen by a spectator is $q = (120)(1.5) = 180$ veh/hr (Hint: add the hourly frequencies of each car driver.) That $\bar{v}_t$ is greater than $\bar{v}_s$ should be intuitive without any calculation since faster vehicles go by the spectator more often.

1.3.3 Applications to scheduling

The time space diagram is not just useful to describe what is happening on a path, but also to coordinate the schedules of *various vehicles* that *interact* while *traveling on the same path*, so as to operate the system as efficiently as possible. The emphasized words summarize the real-life conditions under which one should suspect that some use of a (t, x)-diagram would be useful. Besides the examples mentioned earlier, the (t, x) diagram is also useful for traffic signal coordination on two-way arterial streets and for the determination of the maximum safe service frequency at a rapid transit station. The following problem and its solution deal with the evaluation/design of a waterway; a related problem arises in connection with temporary lane closures on two-lane bi-directional roads and in two-way railroad scheduling on a single track line.

Problem: This problem illustrates use of the time-space diagram to analyze the interaction of vehicles in a narrow way. The waterway depicted in Fig. 1.6 is wide enough for one ship only, except in the central siding which is wide enough for two ships. Ships can travel at an average speed of six miles/hour; they must be spaced at least one-half mile apart while moving in the waterway and 0.25 miles apart while stopped in the siding. Westbound ships travel full of cargo and are thus given high priority by the canal authority over the eastbound ships which travel empty. Westbound ships travel in four ship convoys which are regularly scheduled every 3-1/2 hours and do not stop at the siding. Then, we should find:

1. The maximum daily traffic of eastbound ships, and
2. The maximum daily traffic of eastbound ships if the siding is expanded to one mile in length on both sides to a total of three miles.

Note: We assume that eastbound ships wait exactly five minutes to enter either one of the one way sections after a westbound convoy has cleared it. We do this to take into account that ships do not accelerate instantaneously. ■

Solution: We first draw the time-space diagram for the problem at an adequate scale. Next, we plot the trajectories of the high-priority

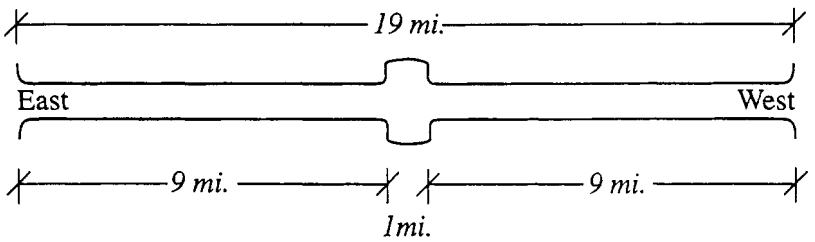


Figure 1.6 Sketch of a waterway with an intermediate siding for ship crossings.

(westbound) convoys. These have been plotted in Fig. 1.7. The dashed band (width - 1 mile) represents the siding, where eastbound and westbound trajectories may cross. These dashed lines will help us draw the eastbound trajectories.

Part 1:

We start by drawing the trajectory of a ship entering the western end of the canal at 3:30 p.m. (the earliest possible time for that particular gap in between convoys). Note how it must stop at the eastern end of the siding to yield the right of way to the last ship for the westbound convoy; note also how it makes it within the 5 min. allowance to the eastern end of the canal. The same process is followed successfully with the second trajectory. In that case we must also watch for safe spacings while moving and stopped. The third ship, however, would not be able to arrive to the western end of the siding within the 5 min. allowance and it cannot be dispatched. Therefore, we find:

$$\text{Capacity} = 2(\text{ships per } 3\tfrac{1}{2}\text{hours}) \times \frac{24(\text{hours})}{3\frac{1}{2}(\text{hours})} = 13.71 \text{ ships/day}$$

Part 2:

Enlarging the siding diminishes the length of the one way sections and, thus, more ships can make it in time (see Fig. 1.8). The result is:

$$\text{Capacity} = 4 \times \frac{24}{3\frac{1}{2}} = 27.42 \text{ ships/day}$$

The reader may now test his scheduling skills, by looking for a way of scheduling six ships per $3\frac{1}{2}$ hour period instead of the 4 depicted in the figure. (Hint: one may have to stop some of the eastbound ships in the siding). ■

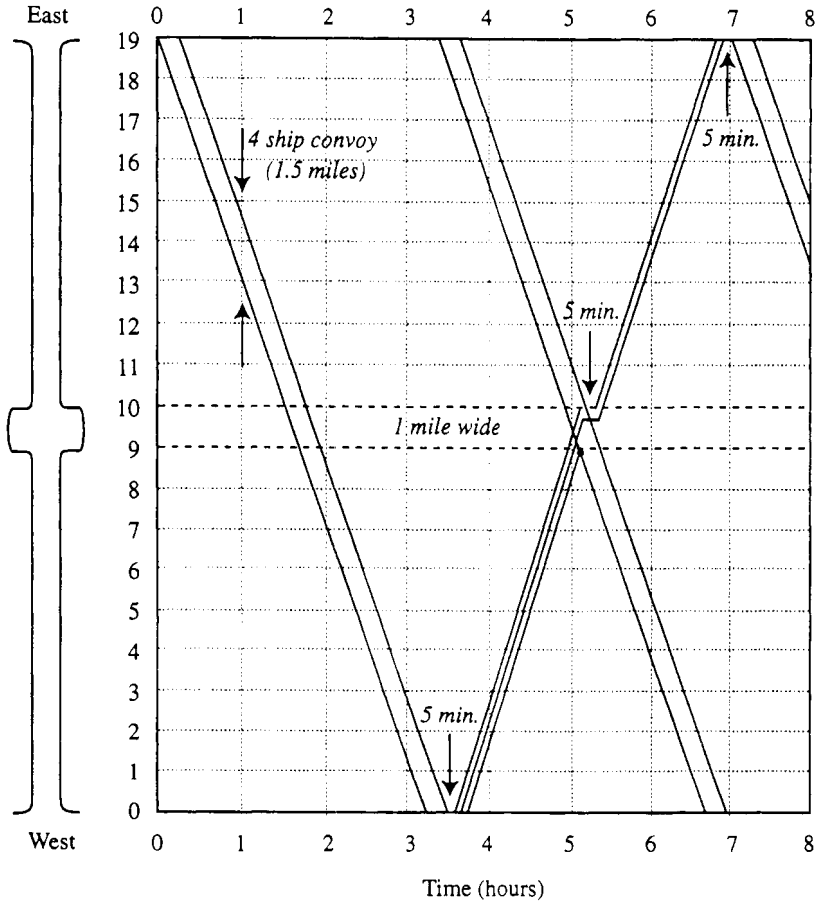


Figure 1.7 Time-space diagram; a short siding.

Notes

1. As is conventional in some branches of mathematics, we use the same symbol 'x' to denote the output of the function and the function itself.
2. The jerk is defined as the absolute value of the derivative with respect to time of acceleration. If the time-derivative of acceleration is sufficiently high, standing passengers can't adjust their position quickly enough to avoid falling. The motion of the transit vehicle is perceived as 'jerky'. A maximum jerk of 0.1g per second (1 second to reach 1/10 of the acceleration of gravity) is reasonable for design purposes.

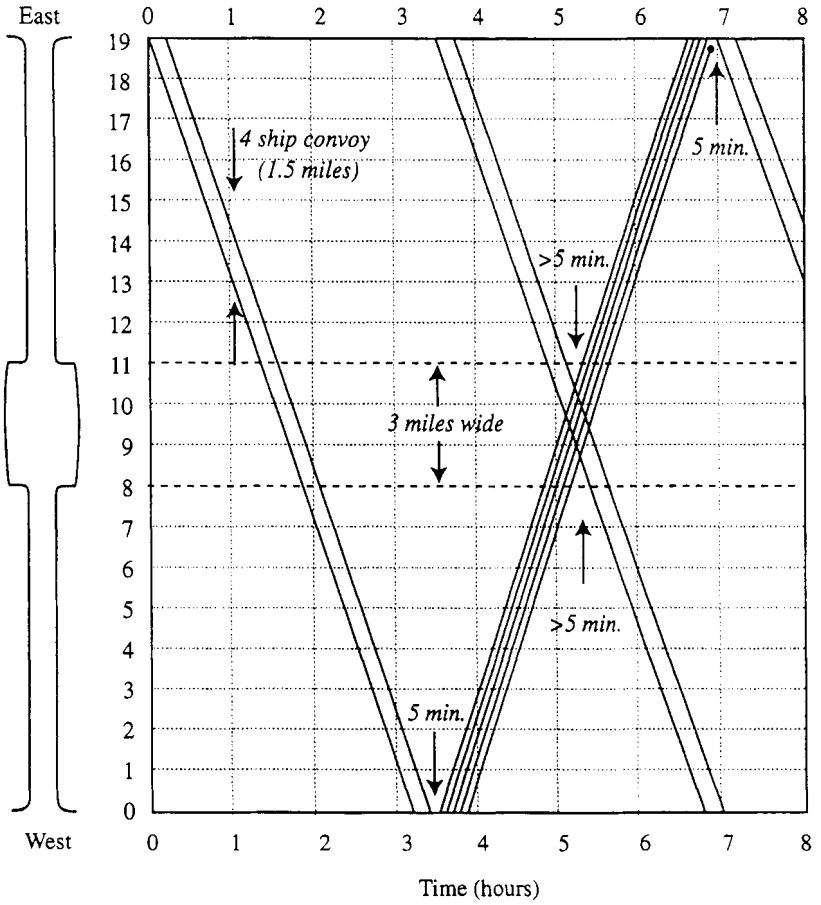


Figure 1.8 Time-space diagram; an enlarged siding.

3. Additional details can be found in most introductory books on 'Transportation Engineering'. Section 1.1.8 includes some references.
4. A smaller f is often used for design purposes to recognize that the 'friction' between passengers and vehicles is even smaller. A value of $f = 0.2$ for seated passengers and $f = 0.1$ for standing passengers is reasonable.
5. Examination of such a diagram reveals that the magnitude of the revised normal force is simply the sum of the centrifugal pseudo force experienced by the vehicle (which is approximately mv^2c and is taken to be positive if $c > 0$) and the magnitude of the normal component of the vehicle weight. If the latter is approximated by the actual vehicle weight, which is quite

reasonable if $\beta \ll 1$, then Eq. (1.3) is obtained.

6. This assumes that driver decisions only depend on v, x, and t. If this is not the case, the function 'F' would have a different set of arguments, and the analysis we are about to present needs modification.
7. Note that (1.12) becomes exact in the limit of infinite sampling intervals, when q_l and k_l no longer depend on interval length.
8. A simple proof is obtained by expanding the product $\bar{v}_s\bar{v}_t$ and seeing that it is greater than $\bar{v}_s^2$:

$$\bar{v}_s\bar{v}_t = \sum_l \bar{v}_s v_l (q_l/q) = \sum_l v_l^2 (k_l/k) \geq \left[\sum_l v_l (k_l/k)\right]^2 = \bar{v}_s^2$$

The second equality is based on the equivalence of $\bar{v}_s/q$ and $1/k$, and that of $v_l q_l$ and $v_l^2 k_l$, for long observation intervals. The inequality is based on the fundamental principle stating that the average of the square is never less than the square of the average.

CHAPTER TWO

Cumulative plots

A cumulative plot is the graph of a function N(t) that gives the cumulative number of vehicles (or other moving objects) to have passed an observer by time t starting from an arbitrary initial count, e.g., at t = 0. Cumulative plots are useful because the count in any interval (t_1, t_2) is given by the change in N(t) across the interval, and this information can be seen at a glance from the graph of N(t).

Cumulative plots are the tool of choice when one must analyze the flow of items past one or several restrictions. Their usefulness in hydrologic synthesis has been recognized for over a century; in that field they form the basis of a technique known as 'mass curve analysis' for determining the capacity of reservoirs (Linsley and Franzini, 1955). Cumulative plots appear to have been introduced to transportation by Caltrans engineer Karl Moskowitz (see Moskowitz, 1954) and then again by Gazis and Potts (1965), but it was Gordon Newell who demonstrated their full potential as an analysis and thinking tool in connection with his queueing theory work (Newell, 1971 and 1982) and later in his reexamination of traffic flow theory (Newell, 1993). Our brief description in these notes cannot do justice to the subject but it should serve as an introduction. Pages 1-24 of Newell (1982) are a good complementary reading.

Just like the time-space diagram is most useful in situations when more than one vehicle trajectory must be depicted, so the time-count diagram is most useful when it displays the cumulative curves upstream and downstream of a series of bottlenecks, and in particular the 'arrival' and 'departure' curves at a single bottleneck. Another way of thinking about the proper application context for cumulative plots is that while the (t, x) diagram is suited to describe how items compete for space over a transportation 'link', the cumulative count diagram is suited to describe the competition for service time through a 'node'.

Our introduction to cumulative plots starts with some definitions (Sec. 2.1), continues with an application of these basic ideas to the study of bottlenecks (Sec. 2.2) and is completed with a discussion of stochastic fluctuations (Sec. 2.3). The chapter also includes a brief section (Sec. 2.4) where the relationship between cumulative plots and time-space diagrams is examined.

2.1 Definitions

We start by establishing the relationship between cumulative count curves and flow. To this end, note that N(t) is related to our earlier definition of average flow in an interval (0, T) by:

$$q = [N(T) - N(0)]/T. \tag{2.1}$$

In most transportation applications cumulative count is made up of indivisible objects (passengers, buses, cars, etc...) and, thus, N(t) is a step function as in Fig. 2.1. However, in applications dealing with many moving objects, where predicting the exact number to the nearest integer is not important, it is preferable to work with a smooth approximation of N(t) that can be differentiated, e.g., an interpolation passing through the crests of every step, such as the smooth curve shown in Fig. 2.1. As in the current case, where $\tilde{N}(t)$ is the approximation of N(t), a tilde will be used to denote the smooth approximation of a given curve whenever the discussion refers to both versions of a particular set of counts. The tilde is not used in other cases.

The main advantage of working with smooth functions is that one can then use differential calculus as a powerful method of analysis. For example, it allows us to define the instantaneous flow at time $t = 0$ as the limit of (2.1) for $T \to 0$, assuming that $\tilde{N}(t)$ has been substituted for N(t) in (2.1). This limit, of course, can be taken for any other time, $t \neq 0$. The result is the definition of *instantaneous flow* at time t, q(t):

$$q(t) = d\tilde{N}(t)/dt. \tag{2.2}$$

Note that this limit does not exist for the discrete data. From now on, the word 'flow' will be taken to mean 'instantaneous flow'. It is important to make this distinction because in practice the word flow is used in the senses of both (2.1) and (2.2), and this can be a source of confusion. Let us now turn our attention to the relationship between the curves recorded by two different observers.

We shall denote by A(t) and D(t) the exact (discrete) arrival and departure curves that might have been recorded by two observers upstream and downstream of some bottleneck–or any other location for that matter. Figure 2.2 shows those curves for a hypothetical example involving a long observation period and so many observations that the individual steps of the curves cannot be seen. (This is a typical application context of cumulative plots.) Then, if we define Q(t) to be the number of items in between the observers at time t, and choose the starting counts so that $A(0) - D(0) = Q(0)$, we see that the vertical separation between the two curves will continue to represent the

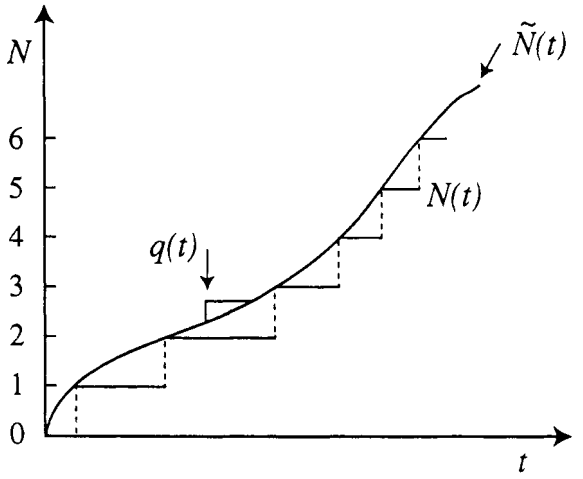


Figure 2.1 A curve of cumulative counts and its continuous approximations.

number of items between the observers provided that vehicles do not enter or leave the intervening space. (The letter 'Q' is used because this accumulation would be called a 'queue' in many cases.) The statement is justified by noting that the input to the system during (0, t) is A(t) – A(0) and, similarly, that the output is D(t) – D(0); then on noting that the difference between input and output is the change in accumulation, we can write:

$$Q(t) = Q(0) + [A(t) - A(0)] - [D(t) - D(0)] = A(t) - D(t). \quad (2.3)$$

We also note from Fig. 2.1 that the time at which the N^{th} item is observed is obtained by finding the 't' where a horizontal line across ordinate N meets the crest of a step. We denote the function that returns t for a given N by means of the superscript '-1'; e.g., $N^{-1}(N)$.[1] Thus, $A^{-1}(N)$ and $D^{-1}(N)$ describe the arrival and departure times of the N^{th} observation recorded at both locations. If items pass through our system in a first-in-first-out (FIFO) order, then these N^{th} observations correspond to the same individual, and the difference

$$w(N) = D^{-1}(N) - A^{-1}(N) \quad (2.4)$$

represents the trip time through the system for the N^{th} individual. Graphically, this is the horizontal separation between the crests of curves A and D at ordinate N. Relation (2.4) is also true for properly smoothed curves. The letter 'w' is used in (2.4) because the trip time can be interpreted as waiting or delay in applications where the observers are close. The terms 'wait' and 'trip time' (between the observers) will

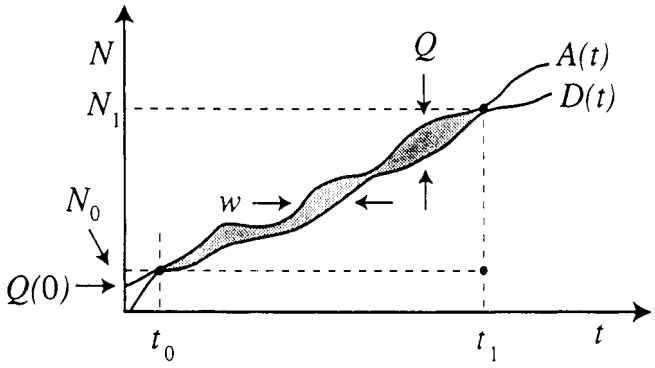


Figure 2.2 Hypothetical arrival and departure curves observed at two locations.

be used interchangeably until the next subsection. Finally, note that Equation (2.4) is not true if there is passing within the system.

Since horizontal separations in the (t, N) diagram represent time and vertical separations represent accumulation, it should not come as a surprise that the area of the region enclosed by A(t), D(t) and any two vertical lines, $t = t_0$ and $t = t_1$ ($t_0 < t_1$), should be the total wait done in the system in the time interval (t_0, t_1). As an example, the shaded area in Fig. 2.2 is the wait done between $t = t_0$ and $t = t_1$. In this figure we have deliberately chosen t_0 and t_1 to be instants when the system is empty, but the foregoing statement is always true. It is justified by noting that the total wait (e.g. in vehicle-hours) done in a time interval where the accumulation is constant is the product of said accumulation and the length of the interval. In particular, the wait done in (t, t + dt) is Q(t)dt, and the (Riemann) summation of this elemental wait from t_0 to t_1 is the aforementioned area:

$$\text{Area} = \int_{t_0}^{t_1} Q(t)\,dt = \int_{t_0}^{t_1} [A(t) - D(t)]\,dt. \tag{2.5}$$

A similar argument (using horizontal elemental rectangles of area w(N)dN, and the exact discrete versions of A(t) and D(t)) reveals that the wait done by items $N_0 + 1$ through N_1, ($N_0 < N_1$) in a FIFO system is given by the area of the region enclosed by curves A(t) and D(t) and by the two horizontal lines, $N = N_0$ and $N = N_1$. Although this result is also intuitive it should be noted that it does not hold generally for systems with passing.

An exception arises if, as shown in Fig. 2.2, the arrival and departure curves 'touch' when $N = N_0$ (at time t_0) and also when $N = N_1$ (at time t_1). In this special case, the area between the A(t) and D(t) curves enclosed by lines $N = N_0$ and $N = N_1$, which is shaded in Fig. 2.2, coincides with the area enclosed by lines $t = t_0$ and $t = t_1$; i.e. it is the wait done in the interval (t_0, t_1). However, we also know that all the items doing some waiting in the interval (t_0, t_1) must have arrived and departed in that interval because otherwise the system could not be empty at times t_0 and t_1. This means that the set of arrivals from $N_0 + 1$ to N_1 matches the set of departures from $N_0 + 1$ to N_1, even if they do not occur in the same order, and that the shaded area in the figure is the wait done by this set of items.

The result is also approximately true if one can be sure that the combined wait of those items that only arrived or only departed in the observation period is a small fraction of the total wait. In particular this is true of any system where the waits are either bounded or stationary (e.g. don't grow with time) as $t_1 - t_0 \to \infty$. We can now state as obvious an important result of queueing theory.

For the conditions where (2.5) represents (approximately) the wait done by the items arriving between t_0 and t_1 we can write the average wait as:

$$\bar{w} = \frac{\text{Area}}{A(t_1) - A(t_0)} = \left(\frac{\text{Area}}{(t_1 - t_0)}\right)\left(\frac{(t_1 - t_0)}{A(t_1) - A(t_0)}\right). \qquad (2.6)$$

where the overbar is used to denote the average over time of a quantity. As per (2.5), we see that the first term of the product on the right side of (2.6) is the mean value of Q(t) in the interval (t_0, t_1); i.e. the average number of items in the system, $\bar{Q}$. The second term is the reciprocal of the slope of the line connecting the two thick dots in Fig. 2.2; i.e. the time averaged arrival rate, $\bar{\lambda}$. (The Greek letter, λ, is often used in queueing theory to denote an arrival rate). Thus, (2.6) is equivalent to the well known queueing formula among time averages:

$$\bar{Q} = \bar{\lambda}\bar{w}. \qquad (2.7)$$

Just like the (t, x) diagram simplified the derivation of the relationship between flow, density and space mean speed, cumulative plots have made the derivation of (2.7) rather trivial; they have also helped us identify some rather general conditions under which it holds true.[2]

As with time-space trajectory plots, it will again be useful if the reader makes an effort to develop a 'feel' for the (t, N) plots, which at a glance show the time variation of accumulations and trip times.

2.2 Applications

So far we have seen cumulative plots as an efficient way in which input-output data can be presented, but in practical applications the complete data are not always available. On the contrary, in any attempt to answer a practical question, one usually has to construct the plot from limited information; e.g. when, from knowledge of A(t) and the operating characteristics of the restriction, one tries to infer the maximum accumulation in the system. This is the central question in the hydrologic synthesis problem mentioned earlier, and the question also arises in the design of 'transportation reservoirs' such as left turn pockets at traffic signals and transit station platforms. Of course, given the same data, we may be interested in other measures of (transportation) performance such as average and maximum trip times. Restrictions to which the technique can be applied appear in all transportation modes. Examples are: airport ticketing counters (people), signalized pedestrian crossings (pedestrians), bus stops (passengers), railroad yards (rail-cars), container cranes (containers), and 4-way highway stops (cars); the words in parentheses denote the counted items in each example.

2.2.1 Restrictions with constant service rates: virtual arrivals and 'delay'

In many cases knowledge of A(t) and D(t) up to the present time ($t = t_0$) allows us to predict (at least approximately) the evolution of the system in the very short term, to time $t_0 + \Delta t$. If this is the case, one can then move the 'present' to time $t_0 + \Delta t$ and repeat the procedure to obtain A(t) and D(t) up to time $t_0 + 2\Delta t$. It is then possible to step through time, repeating this process with small Δt, to predict our system's evolution in the study period (t_0, t_1).

In the common case where A(t) is known for all times in the study period, but D(t) is unknown and a single restriction exists between the observers, the feasibility of the foregoing prediction method hinges on an understanding of the 'service mechanism'; i.e. the rules under which items pass through the restriction. In applications where the observers are so close that the unrestricted trip time between them can be neglected, many service processes have the property that items (or customers in the queuing theory jargon) flow at a constant maximum rate μ through the restriction whenever there is a 'queue', and pass through the bottleneck undisturbed (at a rate that cannot exceed μ) otherwise. Using an overdot to denote the derivative with respect to

time, we can express these conditions in terms of the unknown D(t) as follows:

$$\dot{D}(t) = \mu \quad \text{if} \quad A(t) > D(t) \tag{2.8a}$$

and

$$\dot{D}(t) = \min\left(\mu, \dot{A}(t)\right) \quad \text{if} \quad A(t) = D(t). \tag{2.8b}$$

Equations (2.8) assume that the upstream and downstream observers of our system were very close to each other. When items take considerable physical space, however, the physical queue may back-up past the upstream observer and this would prevent us from claiming a-priori knowledge of the A(t) curve. To restore this a-priori knowledge we must place the upstream observer beyond the reach of any effect of the restriction. This poses a problem, however, because then A(t) can exceed D(t) when there is not a queue directly upstream of the restriction, and this invalidates (2.8a).

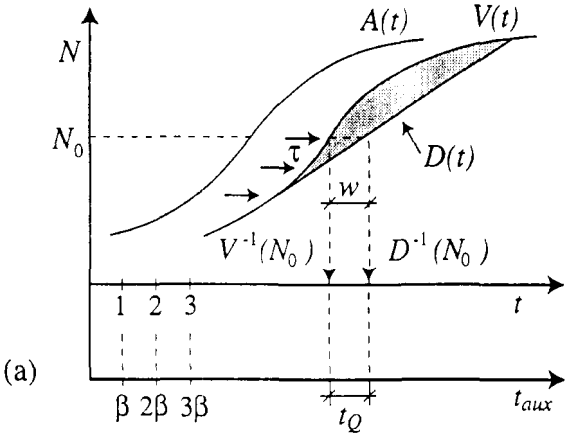
Fortunately, the difficulty can be resolved if we replace (2.8a) by the requirement that any delayed vehicles should discharge at the maximum rate. This new rule, however, is not completely general. An exception arises on 'flared' approaches to traffic signals when the queue is allowed to back up into the narrow part of the highway. In this case, the vehicles delayed in the narrow part cannot be discharged during the green at the maximum rate allowed by the flares because the constriction starves the signal for flow. The techniques presented in Chapter 4 can be used to address this kind of complication, e.g. for the analysis of vehicular flow on an inhomogeneous highway through a series of bottlenecks. Here we explore the (common) case where the highway is homogeneous and the proposed replacement for (2.8a) is reasonable.

The analysis method about to be presented yields the accumulation between the observers, as well as the trip time and the delay for any vehicle. It is important to note that the latter is *not the same* as the time spent in the queue, and that the time in queue is always greater. It will be shown at the end of this section how to derive the time in queue and the queue length (expressed in distance units) in a special case that can be studied easily. The prediction method for delay and accumulation is now described.

Let τ denote the trip time between the observers for an isolated vehicle (no queue), and assume that this time is approximately the same for all the vehicles. In this case we can introduce a *'virtual' arrival curve*, V(t), giving the number of items that would have been seen directly upstream of the restriction by time t if the physical extent of the queue

had been eliminated. Obviously, $V(t) = A(t - \tau)$, so that the virtual arrival curve is a translation to the right of A(t) by the amount, τ; see Fig. 2.3a.

If customer N_0 of the figure is delayed (i.e. it cannot depart the bottleneck at his/her desired time $V^{-1}(N_0)$) then the departure curve must be strictly to the right and beneath point $(V^{-1}(N_0), N_0)$ since according to our assumptions the service rate for delayed customers



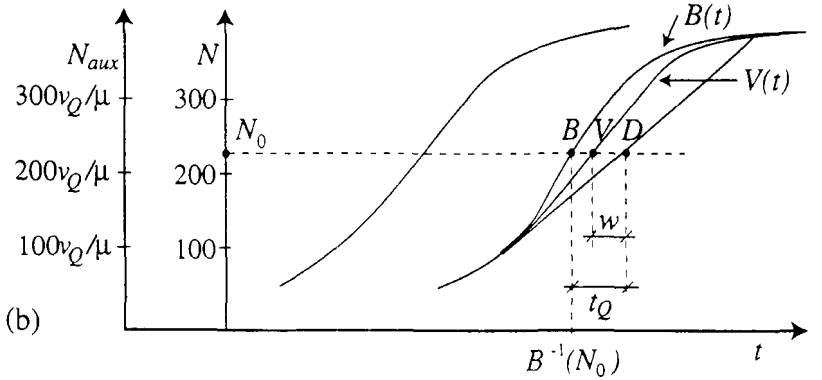


Figure 2.3 Solution of the single bottleneck problem with variable arrivals: (a) construction of the virtual arrival and departure curves; (b) construction of the curve of cumulative arrivals to the back of the physical queue.

must be μ.[3] For customers that are not delayed, the server works at the rate required by the *actual* arrival curve at the bottleneck which, under our assumption of constant trip time when there is no queueing delay, must coincide with the *virtual* arrival curve. This means that the departure curve must track the virtual arrival curve when there is no delay.

In summary, if A(t), τ and μ are known, a graphical construction recipe that is based on the above considerations is as follows:

(i) Translate A(t) to the right by τ to obtain V(t)
(ii) Draw D(t) as close to V(t) as possible, but never exceeding it and making sure that $\dot{D}(t) \leq \mu$; in other words, obtain D(t) from (2.8) as if A(t) was replaced by V(t).

Figure 2.3a displays the result for our example.

It should be noted that step (i) is not commonly taken by most practitioners, and thus the 'standard' approach is only valid when queues take little physical space. In cases where the extent of the queue cannot be neglected, step (i) allows the simple technique embodied in step (ii) to be used quite effectively and simplifies analyses that otherwise would be quite difficult. The interpretation of the results is still as in Sec. 2.1 in that the area of (t, N) regions between A(t) and D(t) represents time in the system. The term 'wait', however, is somewhat of a misnomer because it includes a significant free-flow travel time component that is positive even if there are no queues.

We will use the term '*delay*' in this book to denote that portion of the total 'trip time' that is due to the limited ability of the bottleneck to process vehicles; i.e. delay is the amount of time that can be saved by letting $\mu \to \infty$. For a FIFO restriction the delay for a specific vehicle, N_0, is $D^{-1}(N_0) - V^{-1}(N_0)$; i.e., the difference, w, between the vehicle's actual departure time $D^{-1}(N_0)$ and the time at which the vehicle would have liked to have departed, $V^{-1}(N_0)$, (See Fig. 2.3a). It follows that total delay is measured by the area of relevant regions between V(t) and D(t) in the same way as total time is measured by the area of region between A(t) and D(t). As before, this result is true for non-FIFO systems when the study period begins and ends with customers that find no congestion (no delays/no queues), as in the case shown by a shaded area in Fig. 2.3a.

The interpretation of vertical separation between the curves requires some care. While A(t) – D(t) still represents the number of vehicles between the two observers, the difference V(t) – D(t) does not correspond to anything real; it is the length of the vehicle (item) queue that

would have existed if vehicles (items) traveled in time τ to the bottleneck and formed a 'point' queue next to the bottleneck (as if they were stacked on top of one another) when delayed. As a result of their dimension, real queues contain more vehicles than point queues but the vehicles within them are not stationary; they move slowly. Later in this book (chapter 4) we will present more detailed methods of analysis for tracking queues. The brief subsection that follows examines a simple but important special case. What is important and rather remarkable about the simple procedure we have presented is that it allows us to predict delays and the time-varying vehicle accumulation between observers in a very simple way, and *independently of the detailed behavior of the system between the observers.*[4]

2.2.1.1 Time in queue and the distance taken up by the physical queue

In applications where varying arrivals cause a queue to grow and dissipate upstream of a constant-capacity bottleneck, it is possible to determine the length of the physical queue (in distance units), d_Q, as well as the queueing time, t_Q, of any vehicle. The method about to be presented assumes that: (i) the queue forms directly upstream of the bottleneck, (ii) vehicles travel at an average speed v_Q within the queue, and (iii) vehicles travel at a 'free" speed $v_f > v_Q$ when approaching the queue. In practice, both speeds can be measured quite easily.[5]

We assume that the V(t) and D(t) curves have already been constructed as explained earlier, so that the delay is known. Then, our first goal is obtaining formulae for the t_Q and d_Q of a given vehicle (N_0) in terms of its (known) delay, w, as well as v_f and v_Q. Our two unknowns can be determined from the relation $t_Q = d_Q/v_Q$ and the following equation:

$$w = t_Q - d_Q/v_f.$$

This relation expresses that the delay of vehicle N_0 is the difference between its travel time inside the queue and the time that it would take to travel the same distance without any hindrance.

On substituting d_Q/v_Q for t_Q in the above relation, and then solving for d_Q we find the desired information for vehicle N_0:

$$d_Q = w/(1/v_Q - 1/v_f),$$

and

$$t_Q = w/(1 - v_Q/v_f).$$

Since d_Q and t_Q are related to w by two factors that are the same for all vehicles, $\alpha = (1/v_Q - 1/v_f)^{-1}$ and $\beta = (1 - v_Q/v_f)^{-1} \geq 1$, we see

that the sum of all the vehicle-miles (or vehicle-minutes) spent in queue by any collection of vehicles is simply the product of α (or β) and the sum of the delays experienced by said vehicles. This means for example that if the shaded area shown in Fig. 2.3a is multiplied by β one obtains the total time in queue.

If desired, one can introduce two auxiliary, appropriately rescaled, horizontal axes in Fig. 2.3a, which would allow one to read the d_Q and t_Q for a particular vehicle by measuring the horizontal separation between the V(t) and D(t) curves on the relevant axis. The figure shows the measurement of t_Q for vehicle N_0 as an example.

If we want to predict the queue length or the time in queue at a particular time (as opposed to those encountered by a particular vehicle) additional calculations are needed. It is then convenient to construct a curve B(t), such as that shown in Fig. 2.3b, that gives the cumulative count of vehicles having joined the queue at time t. Read in reverse, the curve should give the times, $B^{-1}(N_0)$, at which specific vehicles join the back end of the queue, so that its horizontal separation from D(t) is t_Q. It should then be clear that the curve can be constructed from the known D(t) and V(t) simply by extending every segment VD of Fig. 2.3b toward the left by a factor β to create segment BD, and that the locus of all points 'B' identified in this manner is the desired curve.

Since the curve's vertical separation with D(t) at a given time gives the number of vehicles in the queue at that particular time, and the average vehicle spacing within the queue is v_Q/μ (see chapter 1, Eqs. (1.11) and (1.13)) the physical queue length is simply $[B(t) - D(t)]v_Q/\mu$. Thus, if we read the vertical separations between the curves using an auxiliary vertical axis with a scale v_Q/μ times smaller than the original, we obtain the physical queue lengths. This is shown on Fig. 2.3b.

Knowledge of the physical queue length is important in applications where the queue must be contained in order to avoid blocking upstream facilities. The method just described can then be used to evaluate the effectiveness of flow control strategies that would change A(t). The method can also be used to evaluate the effect of permanent changes in μ, but in that case one must recognize that v_Q will also change. Knowledge of the total number of vehicle-minutes and/or vehicle-miles traveling at an average speed v_Q is also important and necessary for the calculation of environmental impacts such as total vehicular emissions.

Generalization of the physical queue calculations to bottlenecks with time-dependent capacity such as traffic signals is possible but requires

subtle manipulations that are beyond the scope of this introductory chapter. Fortunately, delay calculations do not require significant modifications and this is explained next.

2.2.2 *On-off service: traffic signal example*

Some bottlenecks have an on-off service mechanism. During the 'on' periods customers are served at a rate μ and during the 'off' periods not at all. The on/off cycle can be periodic or variable, and either independent or dependent of the arriving traffic. Examples are: traffic signals (pre-timed or vehicle actuated), accesses to inland ports restricted by tides, highway railroad crossings and bus stops.

The method of analysis is a minor variation on the two step process described in Sec. 2.2.1 for constant-rate bottlenecks. Given the virtual arrival curve at the bottleneck (or the actual arrival curve if the effects of physical dimensions are deemed to be unimportant), we simply determine D(t) by finding the 'highest' curve that does not exceed V(t) while satisfying $\dot{D}(t) \leq \mu$ in 'on' periods and $\dot{D}(t) = 0$ in 'off' periods. If the on-off periods are fixed, they should be labeled properly (e.g. 'G' for green and 'R' for red) on the time axis before deriving D(t). If the on-off periods depend on the size of the queue, one must step through time in small increments deciding at each step if the server switches its on/off state and determining the $\dot{D}(t)$ that will be consistent with the resulting state.

As an example we consider a pretimed traffic signal that operates on a cycle of C secs, a red (off) phase of R secs and a green (on) phase of G secs. We assume that portions of the amber phase have been appropriately included in R and G so that $R + G = C$, and that the maximum service rate during the green is μ veh/sec. The reader can verify that if vehicles arrive at a constant rate λ (assuming that λ is so small that the arrivals in a cycle, λC, don't exceed the number that can be served, μG) then a periodic diagram such as that shown in Fig. 2.4 is obtained.

The shaded area in that figure represents the total delay during a cycle, which we denote by W. Also shown are the number of vehicles served in one cycle, n′, and the number delayed, n. It should be clear from the geometry of the picture that $R = (n/\lambda) - (n/\mu)$, and thus $n = R/(\lambda^{-1} - \mu^{-1})$. On requiring that $n < n' = \lambda C$ (to force the queue to dissipate at the end of each cycle) we recover the already-mentioned stability condition, $\lambda C < \mu G$. If this condition is not satisfied the signal is over-saturated and the queue would grow steadily cycle after cycle.

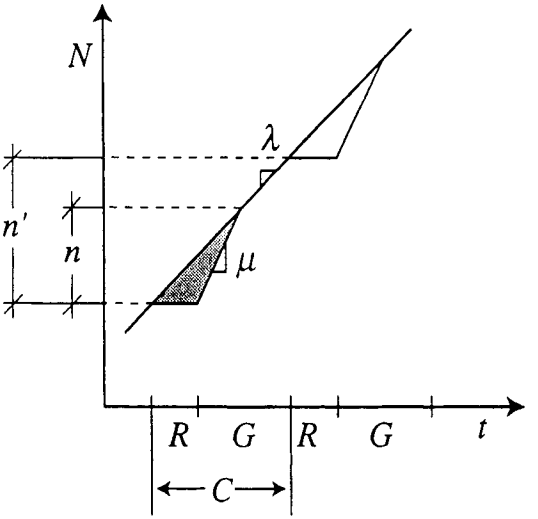


Figure 2.4 Delay at a pre-timed traffic signal.

Since the shaded area is $W = nR/2$, and $n = R/(\lambda^{-1} - \mu^{-1})$, we obtain $W = \frac{1}{2}\lambda\mu R^2/(\mu - \lambda)$. The long-run average delay per car is thus:

$$\bar{w} = \frac{W}{n'} = \frac{1}{2}\frac{\mu R^2}{(\mu - \lambda)C}. \tag{2.9}$$

It should be noted that (2.9) was obtained by prorating the delay in one cycle to all the cars arriving (and served) in one cycle, including those that were not delayed at all. Had W been divided by n, one would have obtained a larger value (R/2) that represents the average delay for delayed vehicles.

Expression (2.9) can be traced back to 1927 and the New Jersey Dept. of Highways (Sloan, 1927), although it is often attributed to Matson (1929) and Clayton (1940).[6] A more general discussion, encompassing time-dependent arrival rates, can be found in Sec. 2.6 of Newell (1982). Section 2.7(a) of that reference explains how (2.9) can be used to choose the phases of a signal to accommodate two competing traffic streams. We will return to this example in Chapter 3.

The scope of application of cumulative plots is much broader than implied by the foregoing types of bottlenecks. The technique can be applied to rather complex service mechanisms and to systems of serial restrictions, where the output of one becomes the input to the next.

This latter application is particularly useful because a single diagram depicts the interconnection between the status of all the bottlenecks. We will return this idea in Chapter 5.

2.3 Stochastic fluctuations

Generally one finds in analyzing queues for transportation systems, that the detailed behavior of a system is not reproducible. If one repeats the observation another day under what seem to be 'identical conditions,' the new curves A(t), V(t) and D(t) may, on a coarse scale, be nearly the same as the old curves, but the exact times of the steps will be different, and the curves will, in general, have wiggles in different places.

If the magnitude of the wiggles is small compared with the separation between V(t) and D(t), then the queues and delays should not change much from day to day. In that case it should be possible to apply the approach of Sec. 2.2 with an average 'μ' to the specific (virtual) arrival curve $V_j(t)$ on any given day, j, or to the average arrival curve, $V_{avg}(t)$, over J days

$$V_{avg}(t) = \frac{1}{J} \sum_{j=1}^{J} V_j(t),$$

and obtain a satisfactory approximation.[7] Favorable conditions for this to happen arise if the arrival curves exhibit a strong 'time-of-day' pattern - or 'seasonal effects' if working on a larger time scale.

If the queue length on a given day, $Q_j(t)$, is determined by the daily wiggles more than by the general shape of the arrival curve then $Q_j(t)$ will vary drastically from day to day and may be difficult to predict. In such a case, one would be interested in the average behavior of the queue over a number of days. In particular, one would like to make predictions that do not require knowledge of the specific $V_j(t)$'s. Care must be exercised, however, because if the deterministic procedure of Sec. 2.2 is applied to the $V_{avg}(t)$ curve when the daily variations in $V_j(t)$ are not small, then the effects of congestion will be underestimated, and perhaps severely. After this fact is demonstrated in Sec. 2.3.1 below, Sec. 2.3.2 will explain how the average delays and queues (across many days) can be correctly estimated for the special case when $V_{avg}(t)$ is linear in t; i.e., if the arrival process is 'stationary'.

2.3.1 Relationships among averages

If we define $D_{avg}(t)$ as the average of $D_j(t)$ over many days, and define in the same way the averages for the 'queue' at a specific time and for the

total delay in a specific time interval, $Q_{avg}(t)$ and W_{avg}, we find that the definitional linear relations:

$$Q_j(t) = V_j(t) - D_j(t) \tag{2.10a}$$

and

$$W_j = \int_{t_0}^{t_1} \left[V_j(t) - D_j(t) \right] dt, \tag{2.10b}$$

also hold for the averages across days; i.e.:

$$Q_{avg}(t) = V_{avg}(t) - D_{avg}(t), \tag{2.11a}$$

and

$$W_{avg} = \int_{t_0}^{t_1} \left[V_{avg}(t) - D_{avg}(t) \right] dt. \tag{2.11b}$$

This is true because an average is a 'linear transformation' and Eqs. (2.10) are linear in $V_j(t)$ and $D_j(t)$. A third relation cannot be written for the average delay of a specific FIFO customer $w_{avg}(N)$ because the averages of an inverse function, e.g., $V_j^{-1}(N)$ or $D_j^{-1}(N)$, is not in general equal to the inverse of the average; thus, the average of $w_j(N) = D_j^{-1}(N) - V_j^{-1}(N)$ is not in general equal to $D_{avg}^{-1}(N) - V_{avg}^{-1}(N)$. Equations (2.11) show that if we could predict $D_{avg}(t)$ from $V_{avg}(t)$ we could also estimate queues and *total* delay. Unfortunately, as is shown with the simple example below $D_{avg}(t)$ does not follow the same rules as $D(t)$ and one cannot in general determine it from $V_{avg}(t)$. Information regarding the arrival curves for every day is needed.

Example: Consider the two-day sequence with (virtual) arrival curves $V_1(t)$ and $V_2(t)$ depicted in Fig. 2.5, parts a and b. From these and from the service rate μ one can get $D_1(t)$ (part a), $D_2(t)$ (part b) and then, by averaging the results, $V_{avg}(t)$ and $D_{avg}(t)$ (part c). Note that the slope of $D_{avg}(t)$ on the parts where it is rising is $\mu/2 < \mu$. This means that the actual delay is greater than would be predicted if the deterministic approach of Sec. 2.2 had been applied to $V_{avg}(t)$; as has been done in part (d).

It can be shown mathematically that the result illustrated by the figure is general; i.e. both W_{avg} and $Q_{avg}(t)$ are larger when the arrival curves vary substantially from day to day.

Because situations giving rise to the same average arrival curve will, in general, generate different average departure curves and different queue lengths, it is difficult to estimate delays and queue lengths when only $V_{avg}(t)$ is known since one needs some additional information in order to obtain $D_{avg}(t)$. There are two instances, however, when obtaining average queue lengths is extremely easy:

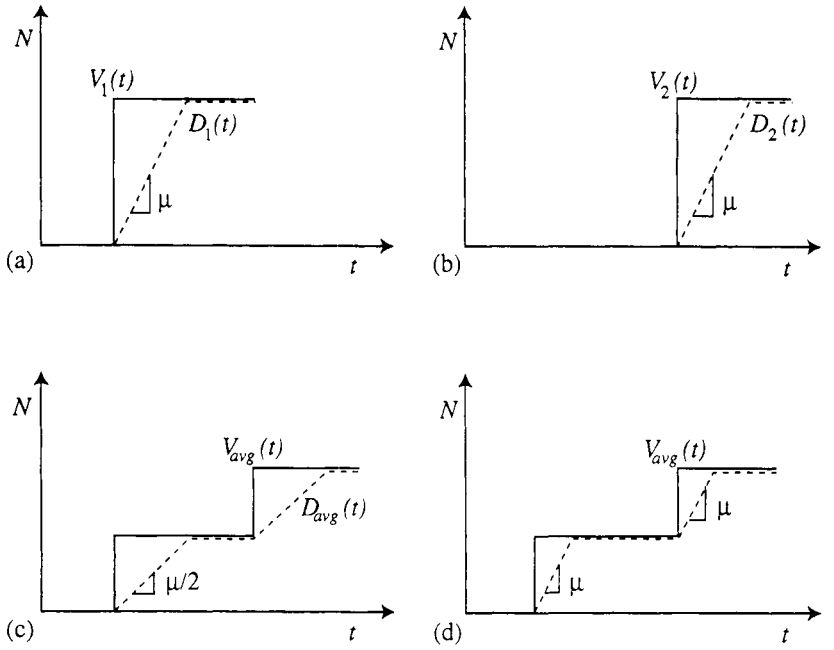


Figure 2.5 Effect of day-to-day variations in arrivals on customer delay: (a) and (b) cumulative curves on two different days; (c) averaged cumulative curves; (d) naive construction of the departure curve from the average arrival curve.

1. We have already stated that if the queue lengths that develop on a typical day are large compared with the wiggles on $V_j(t)$ from day to day, one can neglect the wiggles and approximately obtain $D_{avg}(t)$ from $V_{avg}(t)$ or from the arrival curves on an individual day. This is a valid technique since one can always draw the picture at a scale where the wiggles don't show but queue lengths and total delays do.
2. When the average arrival curve is a straight line for long periods of time (i.e., when the average arrival rate, $\lambda_{avg}(t) = dV_{avg}(t)/dt$, remains constant for long periods of time) and the system is (nearly) empty at the start of our observation period. This is done in the next subsection.

Other instances can also be analyzed, but the theory is more involved. Cox and Smith (1971) and Newell (1982) provide excellent descriptions of the subject from two different perspectives.

2.3.2 Equilibrium queues

In this subsection we explore the behavior of queues with time-independent average arrival rates, $\lambda_{avg}(t) = \text{constant}$, over long periods of observation. The value of the average arrival rate is denoted λ and the duration of the observation period T. It is assumed that the initial queue is of comparable magnitude to those that will develop, and that the latter are negligible when compared with the extra number of arrivals that could have been served in time T if the server had been busy all the time; i.e., compared with $(\mu - \lambda)T$.

Since queues are small the V(t) and D(t) graphs for one typical day should look roughly as two superimposed straight lines; see Fig. 2.6. As shown in the inset, however, short-lived queues may arise due to fluctuations in the arrivals, and these generate delay. Although the timing and magnitude of these congested episodes varies from day to day because of the random nature of transportation phenomena, theoretical analyses of queueing systems have shown that for most queueing systems with random arrivals and departures, the average of the queue length over a time period of long duration on any given day is given by:

$$\bar{Q} \approx \frac{\Delta/2}{1 - \lambda/\mu} \quad \text{if } \lambda < \mu \tag{2.12}$$

where Δ is a constant capturing the variability of the arrival and service processes (how random they are) that is comparable with 1 whenever items do not arrive and are not served in groups.[8]

The average waiting time per customer can be obtained as a byproduct of (2.12). We first note from (2.7) that on any given day $\bar{w} = \bar{Q}/\bar{\lambda} \approx \bar{Q}/\lambda$. (The approximate equality is justified because for long periods of observation, $\bar{\lambda} \approx \lambda$.) Thus, we can write:

$$\bar{w} \approx \bar{Q}/\lambda \approx \frac{\frac{1}{2}\Delta/\lambda}{1 - \lambda/\mu} \quad \text{if } \lambda < \mu. \tag{2.13}$$

When $\lambda \geq \mu$ a formula like (2.12) cannot be developed because it is not possible to find a value of T for which $V_{avg}(t)$ and $D_{avg}(t)$ look like two superimposed straight lines. Actually, if $\lambda > \mu$ and $T \to \infty$, $V_{avg}(t)$ and $D_{avg}(t)$ look like the two lines with slopes λ and μ shown in Fig. 2.7, and the average queue length depends on T in the following way: $\bar{Q} = (\text{shaded area})/T = \frac{1}{2}(\lambda - \mu)T$.

A rule of thumb to know whether T is so long that equation (2.12) can be applied (when $\lambda < \mu$) is:

$$T \gg T^* = \Delta\mu/(\mu - \lambda)^2, \tag{2.14}$$

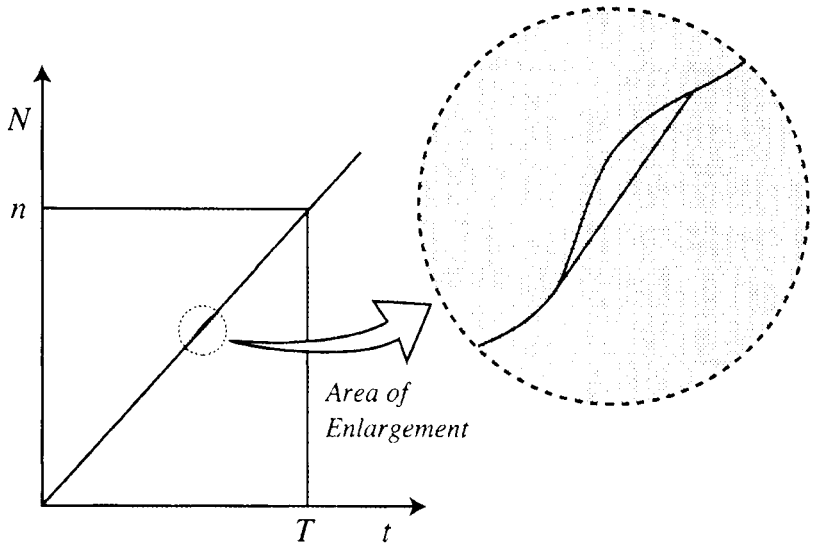


Figure 2.6 Stationary arrival and departure curves for an undersaturated server.

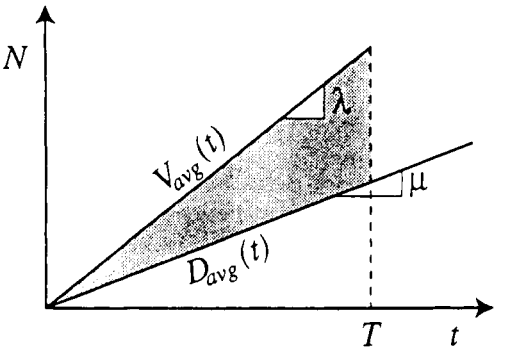


Figure 2.7 Stationary arrival and departure curves for an oversaturated server.

where it is assumed that the initial queue is of order $\overline{Q}$ or smaller. The right side of (2.14) will be called the 'relaxation time'. The greater the ratio T/T^* the lesser is the variation of $\overline{Q}$ across days and the more accurate (2.12) becomes; see Eq.(6.34) of Chapter 6 for more details.[9] Note that as the system approaches saturation ($\lambda \rightarrow \mu$) the necessary

period of observation increases rapidly. If T does not satisfy (2.14) then a precise description of the arrival process, together with some knowledge of the conditions a time $t = 0$, is necessary to make predictions. We could then apply (2.11) but to predict each $D_j(t)$ we would need the corresponding $V_j(t)$ and $Q_j(0)$.

Example: A toll booth can handle one vehicle every 6 seconds. Can you tell approximately what average delay per vehicle at the toll booth one should expect during an hour if the system is initially empty and then

a. 300 vehicles arrive randomly during that hour?
b. 580 vehicles arrive randomly during that hour?

Solution: Since the exact arrival curve is not given, we try to use the equilibrium formulas for an observation period of one hour.

Case a.

$\mu = 600$ vph
$\lambda = 300$ vph

Since 1 hour is much greater than $600/(600\text{-}300)^2 = 150^{-1}$ hours and the initial queue is small, we can use equation (2.12). The result is:

$$\bar{Q} \approx \frac{1/2}{1 - 0.5} = 1$$

and

$$\bar{w} \approx 1/\lambda = 1/300 \text{ hrs} = 12 \text{ sec.}$$

Case b.

In this case 1 hour is not much greater than $600/20^2$ hours and equilibrium theory cannot be used. A more detailed description of the arrival process would be needed. ■

2.4 Relationship between (t, x) and (t, N) plots

Suppose that vehicles (or items) move over a guideway without passing each other, as shown in the (t, x) diagram of Fig. 2.8a, and that they have been numbered in increasing order in the direction of increasing time starting with an arbitrary reference vehicle. We then let N denote the vehicle number and define a function N(t, x) that assigns to each point in the (t, x) plane the number of the last vehicle to have passed. This function will be discontinuous, exhibiting bands of constant N(t, x)

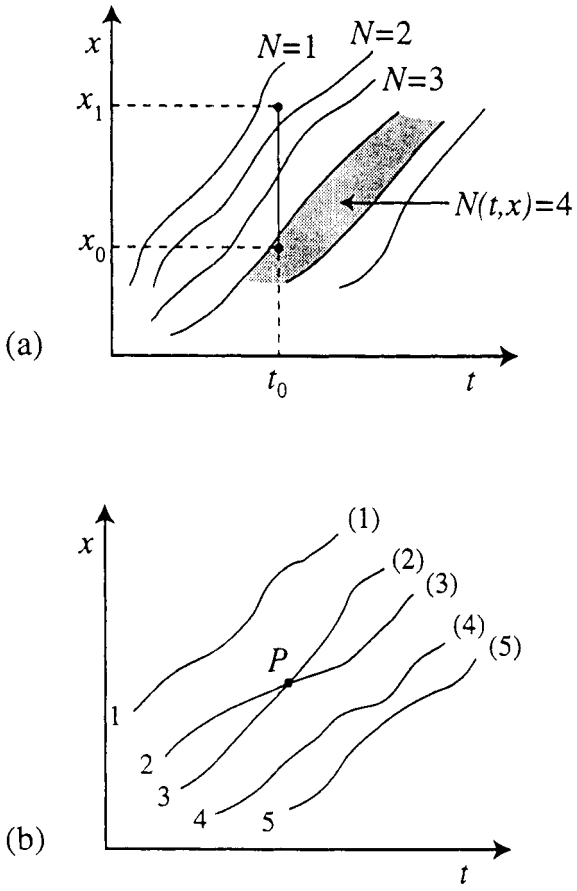


Figure 2.8 Some vehicular trajectories and the associated N(t,x) function: (a) no-passing traffic; (b) passing traffic.

separated by 'faults' at the vehicle trajectories. If desired (and this is convenient for theoretical purposes) one can also define a smooth approximation of N(t, x), $\tilde{N}(t, x)$, that coincides with N(t, x) on the vehicle trajectories; i.e. if $x_j(t)$ is the trajectory of the N = j vehicle, then $N(t, x_j(t)) = \tilde{N}(t, x_j(t))$.

It should be clear that for any fixed x_0, $N(t, x_0)$ is the raw cumulative counting function N(t) that would be recorded by the observer at x_0, as in Fig. 2.1, and that $\tilde{N}(t, x_0)$ would be a smooth approximation passing through the crests of N(t), also as in Fig. 2.1. The same could be said for other locations, x_1. We also recognize that the difference in the values

of either of the functions (N or $\tilde{N}$) at the two locations at any given time, t_0, is (approximately in the case of $\tilde{N}$) the number of vehicles between the two locations (see Fig. 2.8a). Thus, a plot of $\tilde{N}(t, x_0)$ and $\tilde{N}(t, x_1)$ as t varies is nothing but the input-output plot of Sec. 2.1 and Fig. 2.2.

It should also be clear that the partial derivatives of $\tilde{N}(t, x)$ can be interpreted as the instantaneous flow and (the negative of) local density:

$$\frac{\partial \tilde{N}(t,x)}{\partial t} = q(t,x) \tag{2.15a}$$

and

$$\frac{\partial \tilde{N}(t,x)}{\partial x} = -k(t,x) \tag{2.15b}$$

prevailing at point (t, x). The negative sign in (2.15b) arises because $\tilde{N}$ decreases in the direction of increasing x (see Fig. 2.8a).

All the observations made in this section can be extended to passing traffic if one modifies what is meant by a vehicle number or label. To see this consider Fig. 2.8b, which displays the trajectories of 5 consecutively numbered vehicles as per the labels on the bottom left part of the picture. Trajectories 2 and 3 intersect at a point P where a vehicle '3' overtakes another '2'. Examination of the figure shows that the vehicle number does not increase monotonically with time for locations downstream of 'P'. Therefore, $N(t, x_1)$ cannot be interpreted as a cumulative count curve for any x_1 downstream of point P.

On the other hand, we also see from the figure that this difficulty is removed if we imagine that the two interacting vehicles exchange their labels 'N' after the passing maneuver at point P, with the parenthetical results displayed. It should perhaps be intuitive that if similar label-exchanges are used with every passing maneuver, then the difficulty is removed in general.

This can be verified more formally by imagining that physical tags with a label are actually exchanged among vehicles and then defining our function N(t,x), or $\tilde{N}(t,x)$, as that which describes the flow of *tags*. This eliminates the difficulty because tags never pass each other and there is at all times one and only one tag attached to every car. Thus, a tag count must always match the corresponding vehicle count. This change, of course, means that N now represents, not individual vehicles, but positions in the traffic stream; e.g., vehicle 3 advances to position (2) and 2 recedes to (3). Thus, only if there is no passing can the N and $\tilde{N}$ functions be used to track individual vehicles.

The functions N and $\tilde{N}$ were proposed by Moskowitz (1965) and later examined in more detail by Makigami *et al.* (1971) as a convenient way of tying together (t, x) and (t, N) plots in a single 3-dimensional representation. They will be used in Chapter 4.

Notes

1. The inverse function of $\tilde{N}(t)$ also returns the observation time of item N if $\tilde{N}(t)$ passes through the crests of N(t). Otherwise, the returned value is a good approximation.
2. If we use L to denote the separation between the observers and divide both sides of (2.7) by L, then the left side of the equality becomes the time average of the density and the right side the product of the average flow $\overline{\lambda}$ and $(\overline{w}/L)$, which is the time average of the reciprocal speed. Thus, the new expression is a generalized version of the '(flow) ≈ (density) × (speed)' relationship introduced in chapter 1. Equation (2.7) can also be shown to be a special case of the more general relation (4.12) presented in Chapter 4.
3. As an aside we note that the server must work at rate μ as long as the customer is being delayed. This can be seen from the geometry of Fig. 2.3a on realizing that customers prior to N_0 cannot define a V(t) curve that will touch the segment of D(t) lying between the two vertical dotted lines.
4. An application and further elaboration on these issues can be found in Daganzo (1983).
5. Experienced freeway drivers know that the queued speed depends quite heavily both on μ, increasing with an increasing μ, and on the geometry of the road upstream of the bottleneck, decreasing with increasing road width for a given μ. These ideas will be revisited in chapter 4.
6. A detailed historical review can be found in Stohr (1966).
7. We are using the subscript 'avg' instead of an overbar to denote an average across days. This will allow us to highlight the difference between averages within a day and averages across days for variables, such as the queue length, that can vary both ways.
8. We will see in Sec. 6.1.3 that Δ is related to the size of the batches in the arrival and service processes, and will discuss in Sec. 6.3.1 how it can be measured experimentally.
9. If T is only a few times greater than T^* the variations in $\overline{Q}$ across days may be larger than acceptable, but Eqs. (2.12) and (2.13) can still be used. In this case, one should interpret their predictions as being the averages across days of the mean queue and the mean delay.

CHAPTER THREE

Optimization

This chapter reviews elementary optimization concepts, which sometimes arise in connection with transportation system control problems. Unlike chapters 1 and 2, however, this is a purely 'tools' chapter with little emphasis on transportation. Thus, readers already familiar with optimization or systems analysis at the level of an undergraduate course may skip the chapter and refer to it as if it was an appendix. This recommendation does not apply, however, to Sec. 3.2.2 on dimensional analysis because conventional books on systems analysis and optimization do not cover this material.

The chapter introduces some of the terminology used in the optimization field and then shows how to formulate, analyze and solve simple problems. Optimization algorithms are not presented at all because the main goal of the chapter is helping the uninitiated reader to recognize various classes of optimization problems and their possible treatment. This coverage level is deemed sufficient because optimization concepts are not used extensively in this book. The chapter also introduces dimensional analysis, which is a technique whose modeling usefulness transcends the optimization field.

The chapter is divided into three main sections that introduce: problem formulation and classification concepts (Sec.3.1), analytical solution of simple problems and interpretation of results (Sec. 3.2), and the numerical treatment of simple problems with spreadsheets (Sec. 3.3).

3.1 Definitions and basic concepts

The tools presented in chapters 1 and 2 of these notes answered 'what if?' questions; i.e. given a hypothetical scenario, those tools allowed us to predict what would happen to a system. By applying the techniques to a variety of scenarios one could try to find a scenario that was 'best' in some sense. Such a process is called an *optimization.*[1]

For example, Eq. (2.9) can be used to predict the average delay per vehicle at a traffic signal as a function of the 'red' and 'green' phases offered to the approaching traffic. It should be intuitive that if such a formula is now applied to all the approaches of an intersection one can

obtain an expression for the average delay experienced by all the vehicles passing through the intersection during a long observation period. One could then predict what effect changes in the signal phases (i.e. changes in the basic scenario) would have on the overall delay so that, conceivably, one could identify the scenario that minimized delay. Section 3.1.1 below uses this simple problem as the venue to introduce the steps one should normally follow in formulating an optimization problem. Sections 3.1.2 and 3.1.3 then discuss the post-formulation analyses that should precede a solution attempt.

3.1.1 Formulation, terminology and example

As a specific illustration that will be used later we develop below the average delay formula for a simple intersection of two 1-way streets controlled by a traffic signal with cycle time C (in secs) that alternatively displays green phases G_1 and G_2 (in secs.) to approaches 1 and 2. As in Chapter 2, we assume that during these phases queues 1 and 2 discharge at rates μ_1 and μ_2. Each cycle includes a positive lost time, $L = C - G_1 - G_2 > 0$, during which no vehicles discharge from either approach; effectively then, the red periods for approaches 1 and 2 are $R_1 = L + G_2$ and $R_2 = L + G_1$. If we associate subscripts 1 and 2 with the remaining variables of Eq. (2.9) that are not common to both approaches (e.g. λ and $\bar{w}$) and we assume that arrivals are stationary we can write:

$$\bar{w}_i = \frac{1}{2}\frac{\mu_i R_i^2}{(\mu_i - \lambda_i)C} \quad (i = 1,2). \tag{3.1a}$$

provided that $\lambda_i C \leq \mu_i G$.

It should be clear that the total delay per unit time observed at the site is the following weighted average,

$$y = \sum_{i=1,2} \lambda_i \bar{w}_i, \tag{3.1b}$$

since λ_i vehicles of each type pass through the intersection per unit time. Thus, after replacing R_1 by $(L + G_2)$, R_2 by $(L + G_1)$ and C by $L + G_1 + G_2$ in (3.1), we find the following expression for the average total delay per unit time:

$$y = \frac{1}{2}\left[\frac{\lambda_1 \mu_1 (L + G_2)^2}{\mu_1 - \lambda_1} + \frac{\lambda_2 \mu_2 (L + G_1)^2}{\mu_2 - \lambda_2}\right]\left(\frac{1}{L + G_1 + G_2}\right). \tag{3.2}$$

If we work in a consistent system of units this expression will have units of veh-secs/secs; i.e., 'vehicles', which is logical since y also represents the sum of the average virtual queues on both approaches. Note as well that if (3.2) is divided by the number of vehicles passing through the system in a unit time $(\lambda_1 + \lambda_2)$ we would obtain the average delay for a typical vehicle. If λ_1 and λ_2 are independent of G_1 and G_2 then the multiplying factor $(\lambda_1 + \lambda_2)^{-1}$ is independent of G_1 and G_2, and a choice of G_1 and G_2 that minimizes y will also minimize $y/(\lambda_1 + \lambda_2)$. Thus, in comparing various signal settings it does not matter whether we use y or $y/(\lambda_1 + \lambda_2)$ as a criterion for selection. If such criteria was acceptable, one could then evaluate (3.2) for many possible combinations of G_1 and G_2 and choose the one with the smallest y.[2]

The field of optimization deals with the systematic search for 'best' scenarios, and the following terminology is standard: the levers on which the decision-maker can exercise control (such as variables G_1 and G_2 in our example) are called *decision variables*, the criterion which takes the form of a mathematical function with the decision variables as arguments is called the *objective function* (e.g. Eq. (3.2) in our case), and any restrictions placed on the levers are called *constraints*. These usually take the form of a finite number of mathematical inequalities and/or equalities among functions of the decision variables. We have not yet introduced any constraints for our example but this is done now.

We recall that Eq. (2.9) applied only if the arrivals in one cycle could be served in a green phase ($\lambda C \leq \mu G$) because otherwise the delay grows without bound. Because this inequality must be true for both approaches, we specify the following two constraints:

$$\mu_1 G_1 \geq \lambda_1 (G_1 + G_2 + L), \tag{3.3a}$$

and

$$\mu_2 G_2 \geq \lambda_2 (G_1 + G_2 + L). \tag{3.3b}$$

As part of the set of constraints one must also specify whether the decision variables should be non-negative or unrestricted, and integer or real. In our case the G_i are real and non-negative, but the non-negativity requirement is obviated by the stronger requirement of a minimum phase length for safe pedestrian crossings: i.e. by the inequalities:

$$G_1 \geq G_1^{\min}, \tag{3.3c}$$

and

$$G_2 \geq G_2^{\min} \tag{3.3d}$$

for some positive $G_1^{\min}$, $G_2^{\min}$.

It is also good engineering practice to limit the cycle length to avoid the appearance of a malfunctioning signal, and therefore we include one last constraint in our set:

$$L + G_1 + G_2 \leq C^{max}. \tag{3.3e}$$

This last inequality completes the formulation of our idealized signal timing *minimization* problem.

If this had been a practical problem, the *parameters* μ_i, λ_i, L, G_i^{min} and C^{max} would have been numerical data and the objective function and constraint set of our own problem would have only involved decision variables.

Our problem is a minimization problem because the objective function is to be minimized. One can also formulate maximization problems, but an alternative theory doesn't have to be developed; a maximization problem can always be converted into an equivalent minimization problem by changing the sign of the objective function.

The set of equations defining an optimization problem (e.g. (3.2) and (3.3)) is called a '*mathematical program*'. When all the equations in the program are linear in the decision variables and these are real we say that we have a '*linear program*' (or an LP). This is by far the most common application of optimization, and powerful software exists for solving problems with many (thousands) of decision variables in the blink of an eye. Contemporary spreadsheets can solve L.P.'s of moderate size when fed the coefficients of the linear equations that define the program.

In looking at our 'toy' example we recognize that both sides of each constraint are linear but that the objective function is not. This means that (2.2) and (2.3) define a *non-linear program* (NLP) with linear constraints. Of all possible types of NLP's those with linear constraints have received the most attention. General NLPs are more difficult to solve than LPS, and the degree of difficulty depends greatly on the properties of the functions in the formulation; i.e. whether they are convex, concave, linear, etc.... The subject is vast and Avriel (1976) is an excellent introduction. Press *et al.* (1986) contains 'canned' routines for both LP and NLP problems, and briefly describes their logic.

Another degree of difficulty is introduced when one or more of the decision variables is restricted to be integer and we then have to resort to *integer* and *mixed-integer* programming techniques. Brief introductions to these subjects can be found in specialized chapters of introductory texts to operations research such as Hillier and Lieberman (1995), and also in more specialized books. The next two subsections show how NLP problems can be classified according to the properties of the

objective function, examined in Sec. 3.1.2, and the constraints. The latter are discussed in Sec. 3.1.3.

3.1.2 Convex and concave functions

Given a trial solution to an NLP, consisting of a set of decision variables satisfying the constraints (a *feasible point*), one would like to have a procedure for finding a small change of the current feasible point that would improve the objective function. Such a procedure would be very useful and general because it could then be applied iteratively to obtain a sequence of improved solutions to the NLP, the last of which could be adopted for implementation. For most problems, the limit of such a sequence is a point, called a *local optimum*, that cannot be improved further by additional infinitesimal perturbations.

It turns out that efficient numerical perturbation schemes exist for finding local optima, so that the problem of finding a local optimum of an NLP can be considered 'easy'. Unfortunately, this is not enough to be practical because there is no guarantee that the objective value obtained in this convenient way is close to the *global optimum*, or perhaps even satisfactory.

Convex and concave functions are important in that, when they appear in certain ways in an NLP, they guarantee that a local optimum is also global.

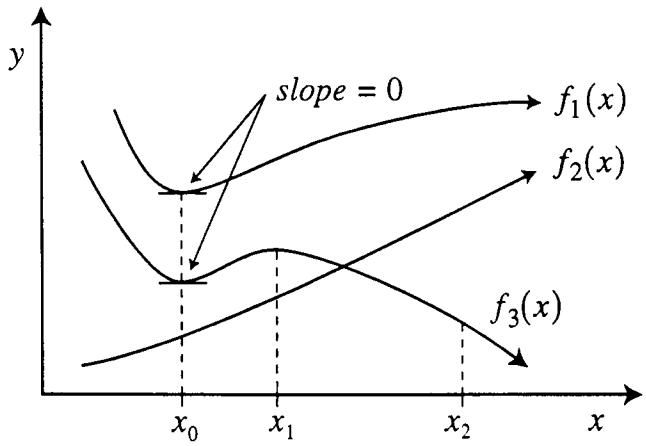


Figure 3.1 Three smooth curves with different optimization properties.

Figure 3.1 illustrates the ideas of local and global optimality for a problem with a single decision variable defined by: min $\{f_i(x)\}$, subject to the constraint $x \geq x_0$. The three objective functions displayed ($i = 1, 2, 3$) have a local optimum at $x = x_0$ because the only small perturbation allowed, $x = x_0 + dx$, increases the value of the objective function in any of the 3 cases. In case $i = 3$, however, the minimum is not global because a large perturbation, e.g. $x = x_2$, reduces the objective value. We also note that f_2 does not have an *unconstrained local optimum* at x_0 because (in the absence of constraint $x \geq x_0$) we could find a small feasible perturbation $x = x_0 - dx$ that would reduce y. The remainder of this section discusses unconstrained problems, and in particular, ways in which one can check if their local minima or maxima are also global. Constrained problems are discussed in Sec. 3.1.3.

As is well known from elementary calculus, a necessary condition for $x = x_0$ to be an unconstrained local minimum of f(x) is that:

$$df(x)/dx|_{x=x_0} = 0. \tag{3.4}$$

The condition becomes sufficient if in addition

$$d^2f(x)/dx^2|_{x=x_0} > 0, \tag{3.5}$$

as happens for f_1 and f_3 in Fig. 3.1.

Curve $f_3(x)$ of Fig. 3.1 illustrates that (3.4) and (3.5) are not sufficient conditions for global optimality. The mechanism for this failure is a change in the function's curvature, which allows it to start a sustained decline for large values of x. It should be clear in light of this observation that if the second derivative of $f_3(x)$ had been positive for all $x \geq x_0$ then the curve would have continued to increase to the right of x_0 at a faster and faster rate and that small values of f(x) could not be found to the right of x_0. Something similar can be said to the left of x_0. This justifies informally that the condition

$$d^2f(x)/dx^2 \geq 0 \quad \text{for all } x \tag{3.6}$$

guarantees that if $f(x_0)$ is an unconstrained local maximum it is also an unconstrained global minimum. Functions of 1 variable satisfying (3.6) are said to be convex.[3] It should be clear from this that establishing the convexity of the objective function of a minimization problem is highly desirable.[4]

A more general definition of a convex function that does not require f to be differentiable is that the segment connecting any two points on the curve f(x) should lie above or on the curve. We don't prove here the equivalence of the two definitions here, but it should be intuitive on graphical grounds that curves obeying the second definition must 'bend upward' and cannot have multiple minima.

This definition of convexity also applies to functions of two and more variables where the 'chord-above-the-curve' property implies that the function is 'bowl-shaped'. In this case the differential property equivalent to (3.6) pertains to the 'definiteness' of the matrix of second derivatives $\{\partial^2 f/\partial x_i \partial x_j\}$, which should be 'positive semi-definite' for all values of the x's. We don't explain what this means here because in many practical cases one can establish convexity in a far easier way from the following facts, which are given without proof:

(i) Linear functions are convex.
(ii) Sums of convex functions are convex.
(iii) A convex function multiplied by a positive constant remains convex.
(iv) A convex function of a linear function is convex.
(v) A product of two positive convex functions of single but different variables is convex if one of the functions is increasing and the other is decreasing.

Finally we note that the negative of a convex function is defined to be *concave*. Concave functions satisfy the reverse of (3.6) and share with convex functions properties (i) - (iv).

Example: Consider the functions: (1) $y = x + x^{-1}$, (2) $y = x(2 - x)$, (3) $y = 1/x_1 + 8/(x_1 + x_2)$ and (4) $y = 1/x_1 + 8/x_2 + x_1x_2$. Determine if they are concave, convex or neither for positive values of their arguments. ■

Solution: Function (1) is convex for positive x's because $d^2y/dx^2 = 2x^{-3} > 0$ for $x > 0$. This can also be seen directly if one remembers that the curve x^{-1} bends upward for positive x (i.e., is convex) and that the other term in the expression, x, is also convex (it is linear and thus convex); therefore their sum (i.e. function '1') must itself be convex.

Function (2) can be written as $2x - x^2$. We know that the curve x^2 is convex for all x and therefore $-x^2$ is concave. Function (2) is therefore concave because it is the sum of two concave terms. This result is confirmed by the inequality $d^2y/dx^2 = -2 < 0$, which is true for all x.

Function (3) is also the sum of two terms. The first term is convex for $x_1 > 0$. The second term can be written as $8(z)^{-1}$, where $z = x_1 + x_2$, and we have already seen that z^{-1} is convex for $z > 0$. Property (iii) ensures that $8z^{-1}$ is a convex function of z for $z > 0$, and since $z = x_1 + x_2$ is linear (and $z > 0$ if $x_1, x_2 > 0$), property (iv) ensures that

the second term is also convex if $x_1, x_2 > 0$. Therefore (by property ii) we conclude that function (3) is convex in the positive quadrant.

Function (4) cannot be proven to be concave nor convex using any 'tricks' because it isn't. The easiest way to verify this is by observing how y varies along the segment $x_1 + x_2 = 10$ and noting that a graph of the relation between y and x_1 (which is $y = x_1^{-1} + 8(10 - x_1)^{-1} + x_1(10 - x_1)$) changes curvature in the interval $0 < x_1 < 10$ and therefore cannot satisfy the 'chord-above-the-curve' property. The interested reader may want to plot the graph in order to verify this statement pictorially. Function (4) is an example of a non-convex function that is convex in either one of its arguments when the other one is held constant. ■

Function (4) of the example also illustrates that non-convex functions can be'unimodal' in the sense that any unconstrained local minimum is also global. This can be shown in this case by changing the (positive) decision variables (x_1, x_2) to (z_1, z_2) by the smooth 1:1 transformation $x_1 = e^{z_1}$, $x_2 = e^{z_2}$ where $-\infty < z_1, z_2 < \infty$. It should be clear that such a transformation does not corrupt the optimization problem because every possible scenario defined by a combination of x_1 and x_2 has one and only one alternative representation in the new space. Thus, the original minimization problem $y = F(x_1, x_2)$, is equivalent to minimizing:

$$y = H(z_1, z_2) = e^{-z_1} + 8e^{-z_2} + e^{-z_1}e^{-z_2}$$
$$= e^{-z_1} + 8e^{-z_2} + e^{-(z_1 - z_2)}$$

which is a convex function. (See if you can prove this to yourself using properties (i) to (v) after recognizing that e^{-z} is a convex function of z.) Clearly, a local minimum of H corresponds to a local minimum of F and any such minimum must be global (since H is convex). Therefore convexity of H implies unimodality for F.

The above illustrates that the desirable 'unimodality' property may be unveiled by a suitable change of the decision variables. Sometimes unimodality can be established after a monotonic-increasing change of variable for the dependent variable, $u = g(y)$. The idea is to find a $g(\cdot)$ such that the composed function $g(F(\cdot))$ is convex or concave. The change of variable is chosen to be monotonic-increasing to ensure that $y_1 > y_2$ if and only if the transformed values satisfy $u_1 > u_2$. This in turn implies that the original and transformed objective functions have local and global minima (maxima) for the same decision variables. In that case convexity or concavity of the transformed function ensures unimodality for the original. An example is the bell-shaped function $y = e^{-x^2}$, which becomes concave after the logarithmic transformation $u = \ln(y) = -x^2$.

Example: Show that (3.2) is convex in the decision variables, G_1 and G_2. ■

Solution: We see that (3.2) can be expressed in terms of new variables $z_1 = L + G_1$, $z_2 = L + G_2$ and $z = L + G_1 + G_2$ as: $y = az_1^2 z^{-1} + bz_2^2 z^{-1}$ where a and b are positive constants. The new expression is convex (for $z > 0$) because it is a sum of two terms of form $z_i^2 z^{-1}$ that are convex for $z > 0$. This can be seen from property (v) because z_i^2 and z^{-1} satisfy the required condition; i.e. z_i^2 is positive, *increasing* and convex and z^{-1} is positive, *decreasing* and convex (for $z > 0$). Since y is convex in the z's and the z's are linear in G_1 and G_2 we can say that y is convex in G_1 and G_2 (for $z > 0$). Because condition $z > 0$ is guaranteed by constraints (3.3c) and (3.3d) we can say that our objective function is convex over the *feasible region* defined by the constraints. The reader familiar with positive semi-definite matrices may also verify this statement directly, although in a considerably more tedious manner. ■

3.1.3 Convex sets. Convex programming

A related concept to that of a convex function is that of a *convex set*. A region on the 2-dimensional plane is said to be a convex set if the chord connecting any two points within the region lies entirely within the region. Fig. 3.2 displays an example of two sets in the (x_1, x_2)-plane: set A is not convex because the shown segment is not entirely within 'A', but set B is convex. The 'chord-within-the-region' definition also applies to regions in n-dimensional space. For 1-dimension, in particular, we find that intervals are convex regions. In higher dimensions convexity of sets is more difficult to establish, but the following rules are useful for regions defined by sets of (in)equality constraints such as Eqs. (3.3):

(a) the intersection of convex sets is convex
(b) a linear (in)equality constraint defines a convex set
(c) the inequality

$$f(x_1, x_2 \ldots x_n) \leq \text{constant} \tag{3.7a}$$

defines a convex set if f is a convex function.
(d) Similarly,

$$f(x_1, x_2 \ldots x_n) \geq \text{constant} \tag{3.7b}$$

defines a convex set if f is a concavefunction.

As an example consider the inequality $x_1^2 + x_2^2 \leq 9$ which defines the circle of radius 3 centered at the origin, i.e., a convex region. The

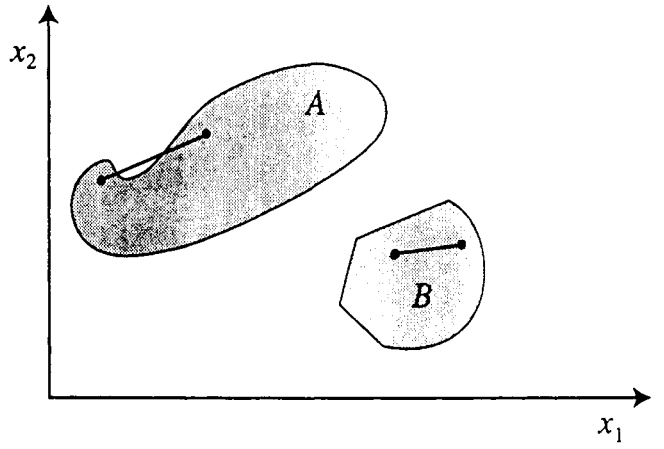


Figure 3.2 Examples of convex and non-convex sets.

convexity of the circle is confirmed by item (c) since the function $x_1^2 + x_2^2$ is convex. An important corollary of (a) is that a mathematical program in which each constraint satisfies either (b), (c) or (d) has a *convex feasible region.*

Minimization problems that involve a convex objective function and a convex feasible region are termed *convex programming* problems. Note in particular that linear programs fall in this category. Convex programs exhibit the desirable unimodality property of unconstrained unimodal functions; i.e. that all local minima are global. This is important because on establishing that a problem falls in this category one can use standard perturbation searching schemes that have been 'canned' to solve the problem numerically; otherwise solutions are more difficult to find.

As an exercise, the reader may want to show that the mathematical programing problem defined by (3.2) and (3.3) is a convex programming problem and therefore computationally 'easy'. Let us now turn our attention to solution methods.

3.2 Analytical solution methods

Analytical solution methods are most likely to succeed with problems involving few decision variables. Our discussion starts with 1-variable problems because an understanding of this basic case will help us deal with more general problems.

3.2.1 One decision variable problems

Unconstrained minimization of a smooth function f(x) is achieved by finding the roots x^* of the equation $df(x)/dx = 0$, choosing the one with the smallest $f(x^*)$ and comparing the result with the values of f for $x \to \pm\infty$.[5] If the function is known to be unimodal then we are only required to find *one* root, which will be a local minimum. The reader can verify using this method that function (1) of the first example in Sec. 3.1.2 is minimized for $x^* = 1$ in the range of positive x.

If the problem involves a feasible region that is a closed interval, i.e., a restriction of the form $x_i \leq x \leq x_j$ (for given $x_i < x_j$) and the objective function is convex we are in the 'nice' realm of convex programming and it suffices to identify a local optimum. Inspection of Fig. 3.3 reveals that if the unconstrained minimum of f(x), x^*, is in the feasible region (e.g., $x_i = x_2$ and $x_j = x_3$) then x^* is the sought global minimum. However, if x^* is to the left of the feasible region (e.g. $x_i = x_3$, $x_j = x_4$) then the local minimum is the left end of the interval (e.g. $x_i = x_3$). The converse happens for the right side so that the location of the global constrained minimum, $\tilde{x}$, is the middle value of x_i, x_j and x^*:

$$\tilde{x} = \text{middle}\{x_i, x_j, x^*\}. \tag{3.8}$$

This simple formula also applies to the maximization of a concave function, but does not apply if the feasible region is not convex. If it is a collection of closed intervals, then one should find the $\tilde{x}$ for each one of the intervals and select the one yielding the lowest (highest) $f(\tilde{x})$.

A simple recipe for minimizing concave (maximizing convex) functions also exists. As is illustrated by the behavior of $f_3(x)$ in Fig. 3.1 over the interval $[\, x_1, x_2 \,]$, the minimum of a concave function over a closed interval, $\tilde{x}$, is at one of its ends; and, specifically, the one yielding the lowest f(x). Again, if we are interested in the minimum over a collection of intervals, the various $\tilde{x}$ should be compared. Note that this recipe doesn't require any differentiation.

The main advantage of an analytical approach is that if the objective function includes some parameters then the final result ($\tilde{x}$ and $\tilde{y}$) will be given by closed form formulae with the parameters as arguments, and by inspection of such formulae one may make statements about the dependence of the solution on the parameters that would be impossible to make from the results of a purely numerical approach.

As an example suppose that we have determined the daily cost of providing shuttle bus service between two points to be: βC/h, where β is the cost of operating 1 bus for 1 day (\$/day), and C/h is the number

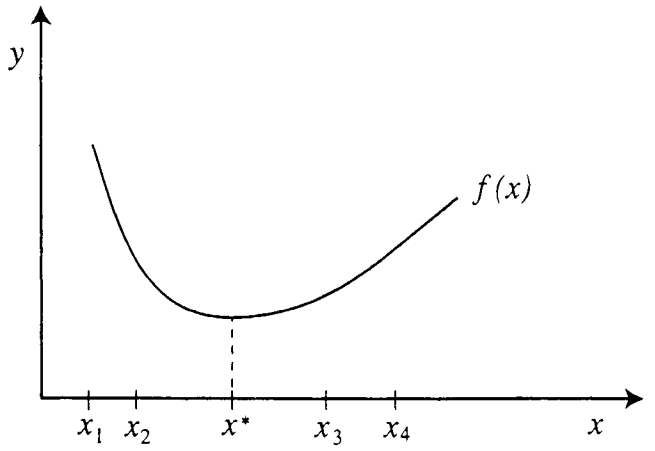


Figure 3.3 Minimization of a convex function of one variable in an interval.

of buses needed to operate the system with a headway of h (days) and a bus cycle of C (days). (The expression for the number of buses follows from the discussion on closed loops seen in Sec. 1.3.) Suppose as well that the average waiting time per passenger is h/2, that the passenger demand rate is λ (pax/day) and that each passenger values waiting at a fixed rate of α ($/pax-day). If we then decide to choose an operating headway by minimizing the sum of the daily cost of providing service and the monetized waiting cost to passengers,[6] the following objective function ($/day) is obtained with h as the decision variable:

$$y = (\tfrac{1}{2}\lambda\alpha)h + (\beta C)h^{-1} \tag{3.9a}$$

This expression should be minimized, subject to:

$$h \geq 0. \tag{3.9b}$$

We note that convex function (1) of the example in Sec. 3.1.2 is a special case of (3.9a). Similar logic reveals that (3.9a) is also convex and that (3.9a and b) define a 1-dimensional convex program. Its global solution can therefore be obtained from (3.8), and the result is:

$$h^* = \tilde{h} = (2\beta C/(\lambda\alpha))^{1/2} \tag{3.10a}$$

and

$$y^* = \tilde{y} = (2\lambda\alpha\beta C)^{1/2}. \tag{3.10b}$$

These expressions indicate at a glance how the optimum cost and the optimum headway depend on the bus cycle, the bus operating cost and the passengers' value of time.

3.2.2 Dimensional analysis and interpretation of results

It is important in a problem like this to verify the units of the solution so as to make sure that no mistakes have been made. For example the right side of (3.10a) has units of:

$$\left[(\$/\text{day})(\text{days})(\text{pax/day})^{-1}(\$/\text{day-pax})^{-1}\right]^{1/2} = (\text{days}).$$

A dimensional argument can also be applied advantageously *before optimization.* To do this one should redefine the decision variables in dimensionless form and do the same for the objective function value and any constraints. One starts by looking for combinations of the parameters appearing in the formulation of the problem that have the same units as the objective function, the decision variables, etc. In our case, for example, we may recognize that $(\lambda\alpha\beta C)^{1/2}$ has the same units as the objective function ($/day). We can then divide y by this constant to define an equivalent new objective, $u = y(\lambda\alpha\beta C)^{-1/2}$. The new objective function is obtained after dividing both sides of (3.9a) by this constant. The result can then be expressed as follows:

$$u = u(z) = \tfrac{1}{2}z + z^{-1} \tag{3.11}$$

where $z = (\lambda\alpha/(\beta C))^{1/2}$ h is the new (dimensionless) decision variable.

Equation (3.11) is not the only dimensionless formulation possible. Instead, we could have used the constants C (days) and β ($/day) to define a new dimensionless decision variable $w = h/C$ and a new objective variable $v = y/\beta$. The revised dimensionless objective function would then become: $v = 1/2sw + w^{-1}$ which now includes the dimensionless constant $s = (\lambda\alpha C/\beta)$. As we are about to see, one should usually try to eliminate as many constants as possible from a problem because this facilitates interpretation of the results. In our present case, s can be eliminated by multiplying both sides of the dimensionless formula by $s^{-1/2}$ and introducing $ws^{1/2}$ as the new decision variable. The result is: $(vs^{-1/2}) = 1/2(ws^{1/2}) + (ws^{1/2})^{-1}$; i.e., Eq. (3.11) again.

The advantage of reducing a problem to dimensionless form with as few parametric constants as possible such as (3.11) cannot be overemphasized. Once this has been done one can make statements about the nature of the solution even before a solution is obtained. If all parameters have been eliminated as in (3.11) we can exploit the fact that z^* and u^* (or $\tilde{z}$ and $\tilde{u}$ if the problem was constrained) are *constants* independent of any data. For example, because $z^* = (\lambda\alpha/\beta C)^{1/2}h^*$ must remain constant for our bus scheduling example, we can say that if the demand (λ) were to quadruple it would be best to reduce headways (h)

by a factor of 2 (which will keep z^* constant); and that this should be true *independent of all other parameters of the problem.* For example, we can make this assertion even if we don't know people's 'value of time', α. This is a very general observation that could not be made from purely numerical analyses, and therein lies the significance of dimensional analysis. Of course, the above is not the only statement possible; additional insights can be derived from further inspection of the formulas for u and z.

It should be noted that similar insights could be derived from inspection of the analytical solution (3.10). What is attractive about the elimination of constants by suitable changes of variable prior to solution is that it allows one to make broad statements, *even before the solution to the problem is known*! In our specific case the minimum of the dimensionless objective function u(z) could be found analytically ($z^* = (2)^{1/2}$) but the statements that were made didn't require the constant z^* to be known. They would have remained valid even if the equation du(z)/dz was transcendental and had to be approximated numerically. The idea of eliminating constants by suitable change of variables is particularly useful for problems with multiple decision variables, because dimensional analysis can extend the generality of numerical solutions.

3.2.3 Multiple decision variables

For unconstrained minimization of a continuously differentiable function $f(x_1,...,x_n)$, the usual first order condition:[7]

$$\frac{\partial f(x_1 x_2 \ldots x_n)}{\partial x_i} = 0 \quad (i = 1, 2, \ldots n) \tag{3.12}$$

identifies a global minimum, $(x_1^*, x_2^*,...,x_n^*)$ if the function is convex, or if it can be made convex by a change of variable. As an example, consider the 4^{th} objective function of the example in Sec. 3.1.2., $y = x_1^{-1} + 8x_2^{-1} + x_1x_2$, which is to be minimized over the set of positive x's. This dimensionless function could be the final result of the parameter elimination efforts described in Sec. 3.2.2, applied to a transportation optimization problem with 2 decision variables. We saw in Sec. 3.1.2 that this function could be transformed into a convex function (for x_1, $x_2 > 0$) by a change of (decision) variables. Thus, the solution of (3.12) for x_1, $x_2 > 0$, i.e., of the system $-x_1^{-2} + x_2 = 0$ and $-8x_2^{-2} + x_1 = 0$, yields the global optimum. Elimination of x_2 in these equations yields: $x_1 = 8x_1^4$. Thus, the solution for the global minimum is $x_1^* = 1/2$, $x_2^* = 4$, and $y^* = 6$.

Unfortunately, the ease with which the solution was identified in this example is the exception rather than the rule and one then has to treat problems numerically. In some cases it may not even be possible to establish unimodality. For problems with only 2 or 3 decision variables a *conditional decomposition* approach can sometimes be useful. It consists in 'freezing' all the variables except one, e.g. x_1, and then obtaining an analytical solution for x_1^* and y^* with the other variables as parameters. The resulting function $y^* = f_1(x_2, x_3,...)$ can then be treated likewise to eliminate x_2. Successive steps can then (theoretically) lead to an analytical solution. The approach is general in that it does not require the functions f, f_1, f_2, etc... to be well-behaved in any particular sense, provided that the 1-dimensional global optimum can be identified at each step.

The approach can be applied to the foregoing example. If we 'freeze' x_2 first, the optimum x_1^* is the global minimum of $x_1^{-1} + 8x_2^{-1} + x_1\ x_2$ (treating x_2 as a positive parameter). Because this expression is convex in x_1 the global minimum is easy to find; it is $x_1^* = x_2^{-1/2}$, which yields $y^* = 2x_2^{1/2} + 8x_2^{-1}$ as the new objective. The expression for y^* can now be 'defrosted' and minimized with respect to x_2. Although it is not convex, it is easy to see by inspection that it is unimodal and that its global minimum is the root of $x_2^{-1/2} - 8\ x_2^{-2} = 0$, which is $x_2^* = 4$, as we had previously found.

The conditional decomposition approach can be applied even if the problem is constrained and may be applied to groups of variables. For example, in mixed-integer programming problems it is sometimes useful to 'freeze' all the integer variables, in order to treat the conditional optimization with a powerful (continuous) optimization tool. One can then repeat this process for different combinations of the integer variables in the search for an ideal combination of the latter. The following section describes some basic ideas pertaining to numerical searches over continuous feasible regions.[8]

3.3 Numerical approaches

As you can imagine, an extensive literature exists on numerical solution methods. Entire courses at many universities are devoted to *subsets* of all possible methods. Libraries of optimization routines are also readily available. Even computer spreadsheets have significant capabilities in this respect. Here we can only introduce searching methods for 1-dimensional problems. It is also shown how virtually all 2-dimensional problems can be solved quickly by numerical enumeration with spreadsheets.

3.3.1 One decision variable

When the minimum of a function has to be found numerically just once and you have access to a computer, the best solution method is simply inspection of a graph produced with an adequate resolution.

When the human eye cannot be used (e.g. because the objective function is just one of many that must be systematically tried) automatic identification of the optimum is necessary. Interval reduction methods are sometimes useful in this context. They work by iteratively shortening an interval that straddles the optimum. The bisection method is the simplest of these methods. It can be used when the derivative of the function is known to change sign once in an initial interval. One simply calculates the derivative at the mid-point and retains for further consideration the side of the interval over which the derivative still changes sign. With this method, an additional digit of accuracy is obtained every 3 or 4 iterations. Faster and slower automatic methods exist depending on whether additional (e.g. 2^{nd} derivative) or less information (e.g. no 1^{st} derivatives) is available for use in the search.[9]

As an aside, the remainder of this subsection describes a simple trick that can be used with both analytical and numerical approaches when one wants to optimize a function for all values of a parameter, α, that appears in the objective function, f(x, α). The trick is useful if the first order condition $\partial f(x, \alpha)/\partial x = 0$ can be solved easily for α but not for x. This occurs for example if we seek the minimum of $y = \alpha x^2 + x - x\ln x$ for all positive α's over the range of positive x, because the first order condition is:[10]

$$2\alpha x - \ln x = 0, \tag{3.13a}$$

which cannot be solved for x in closed form but can be solved for α to establish the following relationship between α and x^*:

$$\alpha = (\ln x^*)/(2x^*). \tag{3.13b}$$

Now, instead of finding x^* for every α [with (3.13a) and a series of searches], which might have been our first instinct, we can use the fact that α is known for every x^* to eliminate all the searches. To do this, one generates a table of three columns with a suitable range of x^* as the independent column and then fills the two other columns with the corresponding values of α and y^*. These are obtained with (3.13b) and the objective function formula. The results can then be plotted, putting α on the horizontal axis and x^* and y^* on the vertical axis, to display the sought result for all possible α's. A computer spreadsheet can be used to do this rapidly. (One simply needs to specify an 'xy-plot' with the column for 'α' identified as the 'x-range' for the graph.)

As an exercise, see if you can develop a general solution of the unconstrained convex minimization problem $\min\{x^{10} + x^2 + \alpha x\}$ (for all α) after verifying the convexity of the objective function.

3.3.2 Two decision variables

Independent of its complexity, any problem that includes only two decision variables (x_1 and x_2) can also be solved by evaluating its objective function in tabular form over the range of values allowed by the set of constraints (the feasible region). A computer spreadsheet can be used to fill a 2-way table with appropriate values of x_1 and x_2 assigned to each row and column of the table. Each entry of the table should contain the objective function formula using as arguments the headings of the corresponding row and column (x_1 and x_2) in a form suitable for copying. Constraints can be incorporated by including in the

CE 251 EXAMPLE: SIGNAL TIMING-DETERMINISTIC

Arrival rate #1 =	0.40
Arrival rate #2 =	0.20
Service rate #1 =	1.00
Service rate #2 =	0.50
Lost time =	10.00 secs
Min Green #1 =	10.00 secs
Min Green #2 =	10.00 secs
Max cycle =	90.00 secs

TOTAL DELAY PER UNIT TIME = AVG. # CARS IN A QUEUE

GREEN #2 \ GREEN #1 (SECS.)	10.00	12.00	14.00	16.00	18.00	20.00	22.00	24.00	26.00	28.00	30.00	32.00	34.00	36.00	38.00	40.00	42.00	44.00
10.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
12.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
14.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
16.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
18.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
20.00	ERR	ERR	ERR	ERR	ERR	9.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
22.00	ERR	ERR	ERR	ERR	ERR	ERR	9.48	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
24.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	9.97	10.02	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
26.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	10.41	10.45	10.51	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
28.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	10.90	10.94	11.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR
30.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	11.38	11.43	11.50	ERR	ERR	ERR	ERR	ERR	ERR
32.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	11.87	11.93	12.00	12.08	ERR	ERR	ERR	ERR
34.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	12.36	12.42	12.50	12.58	ERR	ERR	ERR
36.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	12.91	12.99	13.07	13.16	ERR	ERR
38.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	13.49	13.57	13.66	13.76	ERR
40.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	14.08	14.17	14.27	ERR
42.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	14.72	ERR	ERR
44.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
46.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
48.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
50.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
52.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
54.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
56.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
58.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR
60.00	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR	ERR

Figure 3.4 Spreadsheet solution of a 2-variable optimization problem.

formula a conditional statement which returns an error 'ERR' if any of the constraints is violated. [A convenient way of incorporating an inequality constraint of the form $g(x_1, x_2) \geq 0$ is adding the term '$0 \cdot (g(x_1, x_2)^{1/2})$' to the objective function. Because the square root is multiplied by zero, the objective function is not corrupted by the modification. Other types of constraints can be treated similarly.] Then, when the spreadsheet is recalculated the cells including numerical values will form a pattern on the screen in the shape of the feasible region; e.g. a region that may look like those displayed in Fig. 3.2. One can then visually scan the result to identify the optimum. Alternatively, one can use the @MIN (or @MAX) functions to search for the minimum (or maximum) automatically. In this case, it is better to add (or subtract) a very large constant, M, to the objective function for each constraint that is violated to prevent the error messages from interfering with the automatic search; the term '@IF($g(x_1, x_2) \geq 0$, 0, M)' can be used instead of '$0 \cdot (g(x_1, x_2)^{1/2})$'.

Figure 3.4 displays the result of applying this procedure (with the square root option) to the traffic signal optimization problem defined by Eqs. (3.2) and (3.3). The formulas in the spreadsheet (named 'SIGNAL.WK1") use the block of data on the top as arguments so that sensitivity analysis can be readily performed. The displayed solution, which exhibits a minimum delay of 9 car-secs per second for $G_1 = G_2 = 20$, is typical in that (3.2) is usually minimized for the smallest possible values of G_1 and G_2 for typical input data.

Notes

1. We should caution at this point that the results of an optimization are most useful when the problem at hand has been formulated in terms of an unambiguous objective, and that quite often this is not possible. Transportation problems usually exhibit multiple objectives (e.g. reduction of travel time, cost, air pollution and noise) and a group of actors with different "tastes" for these objectives. In such situations there appears to be no proper way of weighing the various impacts on different actors so as to obtain an objective measure of the group's well-being, as will be explained in Sec. 7.4.2. In these cases optimization techniques can only play the role of search tools for *reasonable* solutions.
2. The example we are using rarely arises in practice because signals are usually part of a system and their arrivals are cyclic (non-stationary). Furthermore, one can question our delay-minimization objective (or any other objective for that matter in the case of traffic signals) since it is not clear that delay avoidance is a good measure of society's welfare for city travel.

3. If property (3.6) is only satisfied in an interval of x, then the function is said to be convex over that range; e.g. for $x \geq 0$.
4. Some non-convex functions (such as f_1 in Fig. 3.1) can be shown to share with convex functions the local/global equivalence property for their minima. This will be illustrated later in this section by means of two examples.
5. For more general functions one should also check the functions' behavior near singular (non-smooth) points; graphical examination of the curve is recommended.
6. Although this approach may be pragmatic, it cannot be formally justified; see Sec. 7.4.2.
7. Recall as well that local minima (or maxima) of equality constrained problems can be found by the method of Lagrange, covered in most calculus books, in which the constraints are introduced into the objective function by introducing auxiliary (multiplier) variables.
8. Although they exist, systematic search methods over discrete state spaces are beyond the scope of our discussion; as a rule, they are much slower than their continuous counterparts.
9. Newton's method of elementary calculus uses 2^{nd} derivative information and converges much more rapidly. Although it sometimes fails to converge, this becomes quickly apparent.
10. See if you can show that the objective function is unimodal

CHAPTER FOUR

Traffic flow theory

An objective of traffic and transportation engineering is to control the traffic streams on a set of roads (a network) so as to reduce delay or improve flow without inducing undesirable side effects to society at large. This is attempted for example when engineers change the signal-timing plan on a network in order to reduce both congestion and vehicular emissions. Some other times the objective is to achieve a societal goal, such as preventing traffic from flowing through neighborhoods, while inconveniencing as little as possible those who have to travel. In these two cases, and others as well, transportation engineering objectives are usually pursued by means of system control and redesign measures intended to affect traffic on the network in a desirable way.

Clearly, in order to be capable of developing effective system designs and control strategies, engineers must understand thoroughly how the system in question might respond to possible engineering changes. In particular, they should be able to predict (e.g., by means of mathematical models) any figures of merit that are relevant to the affected public, and should also have an intuitive 'feel' for the likely response of the system to a control or redesign. The latter skill is precious when the time comes to develop a short list of potential improvements for further evaluation.

In an attempt to help the reader develop such skills, the present chapter discusses some elements of 'traffic flow theory'. Highly relevant, this is the subject whose objective is predicting what would happen to a (set of) traffic stream(s) if it (they) had to flow on a given (set of) road(s) under conditions not yet observed. The chapter examines in some detail the case of a single traffic stream flowing on a facility with a single entrance and a single exit. More complicated networks are discussed more qualitatively, together with control issues, in Chapter 5. The discussion in both chapters focuses on fundamental issues and is not complete.

A deeper coverage of traffic flow theory, including theories now in vogue, is more appropriate for an advanced book and it is therefore not given here. As a result of this limited scope, Chapter 4 includes few references. Justice has not been done to the fundamental advances made in the 50's and 60's, but the reader can find a comprehensive

annotated bibliography, circa 1963, in Newell (1995). Since 1963 the field has continued to grow but publications have tended to emphasize computer modeling over experimental work and theoretical understanding; i.e., they have tended to be less 'fundamental'. Thus, although some modern ideas have been included here, our coverage of this literature is even more sparse.

On the other hand, the subtle nature of traffic flow calls for a more detailed and precise treatment of this subject than is used elsewhere in this book, especially since there exist a number of published errors on the subject. As a result, the present chapter is somewhat mathematical (Secs. 4.3, 4.4 and 4.5 in particular) and should be read slowly. Passages that could be skipped on a first reading have been noted.

The chapter is organized as follows. Section 4.1 describes basic concepts, including generalized definitions of flow and density and the concept of stationarity. Section 4.2 presents tools for treating stationary traffic. These include several diagrams that are customarily used to depict the relationship between flow, density and speed under stationary conditions, as is done in manuals such as the Highway Capacity Manual (1994). Stationary analyses have some merit for design purposes but more detailed studies (of operations) often require an understanding of time-and-space varying phenomena; i.e., of traffic dynamics. This material is covered in the last three sections of the chapter: Section 4.3 discusses the conservation equation and its applications, Sec. 4.4 some elementary ideas in traffic dynamics where traffic is treated continuously like a fluid, and Sec. 4.5 some aspects of discrete theories.

4.1. Basic concepts

This section introduces two main ideas: first some generalized definitions of flow, density and speed that describe the average behavior of a traffic stream over different locations and observation periods, and then the application of these ideas to homogeneous (stationary) traffic streams.

The generalized definitions can be viewed as ways in which the measurements from various, many, or even an infinite number of locations should be averaged so as to maintain exactly the basic relationship among flow, density and space-mean speed that was introduced in Chapter 1 as Eq. (1.13). A clear understanding of these ideas is useful to interpret traffic data.

We will also see that if traffic is stationary then it is possible to predict the traffic stream features observed in a photograph from road-side observation, and viceversa.

4.1.1 Generalized definitions

Figure 1 depicts a time-space diagram including the trajectories of nine vehicles that have been labeled by consecutive integer numbers, with the convention used in Fig. 2.8b of Chapter 2. Recall that labels represent positions in the traffic stream, and that they increase consecutively across the vehicles seen by a stationary observer or decline consecutively across those seen on a photograph.

Chapter 1 had introduced the concept of density over a section of road at a specific time as the number of vehicles observed in a photograph of the section at the given time divided by the length of the section. For the road section of length L shown in Figure 4.1, the density at time t_0 is: $k = n/L$ (where $n = 5$, since vehicles 1 through 5 would appear on the photograph). On multiplying the numerator and denominator of this expression by a small differential of time, dt, the formula for density becomes:

$$k = \frac{n}{L} = \frac{ndt}{Ldt}. \tag{4.1}$$

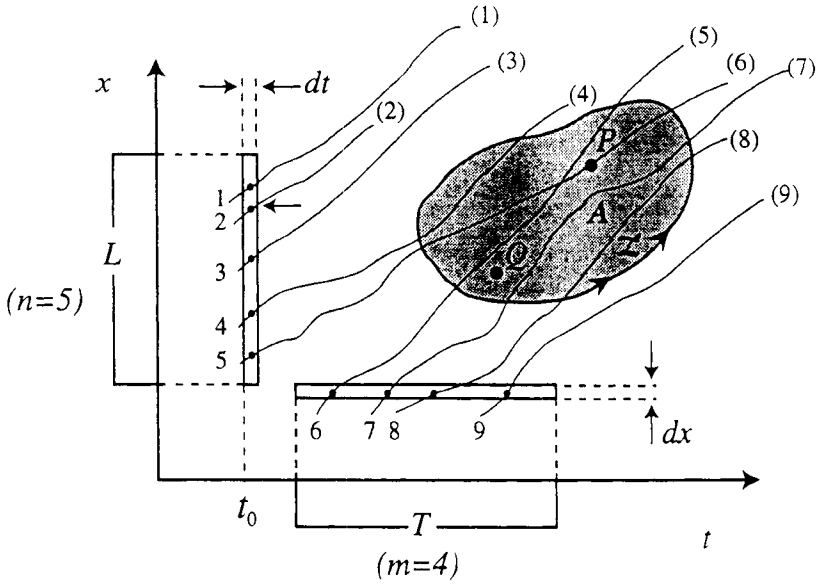


Figure 4.1 Vehicle trajectories and their intersection with a region of time-space.

If we now refer back to Figure 4.1, we note that the denominator of (4.1) is the area of the thin vertical rectangle shown. If we ignore the possibility that one of the vehicle trajectories in the photograph may leave or enter the rectangle through its top or bottom side, which is logical for $dt \to 0$, we can say that every trajectory observed spends dt time units inside the rectangle. Because n trajectories are observed, the numerator of (4.1) represents the total vehicular time that is spent inside the rectangle (e.g., in vehicle-hours). Thus, we can write for our thin, rectangular time-space region:

$$k = \frac{\text{total time in time-space region (veh-hrs.)}}{\text{'area' of time-space region (mile-hrs.)}}. \tag{4.2}$$

The word area is enclosed in quotation marks because its units of measurement are distance $\times$ time and not distance $\times$ distance.

The advantage of Eq. (4.2) over (4.1) is that the former can be applied to arbitrary (not infinitessimally thin and not even rectangular) time-space regions such as region A of Fig. 4.1. This capability will turn out to be important from an experimental point of view because it allows one to compare the behavior of two traffic streams even if they have been observed in different ways, and it also allows one to characterize the status of a single stream with the least expensive measurement techniques. We will return to these ideas in Sec. 4.1.2, below, and in Chapter 6.

The result, k(A), is a generalized definition of density because, as wehave just seen, the original definition is recovered for suitably defined regions (thin rectangles). If we let $|A|$ denote the area of region A and t(A) the total time spent in A by all vehicles, (4.2) can be written as:

$$k(A) = t(A)/|A|. \tag{4.3}$$

This generalized definition was proposed by Edie (1963).

A useful interpretation of (4.3) in terms of averages can be obtained with the help of Fig. 4.2a. The total amount of time spent in the lightly shaded rectangle is equal to the product of number of vehicles in it, n(t), and dt (as in the numerator of Eq. (4.1)). Since the total time spent in A is the sum of the time in the component rectangles we can write:

$$t(A) = \int_{t_l}^{t_u} n(t)\,dt. \tag{4.4}$$

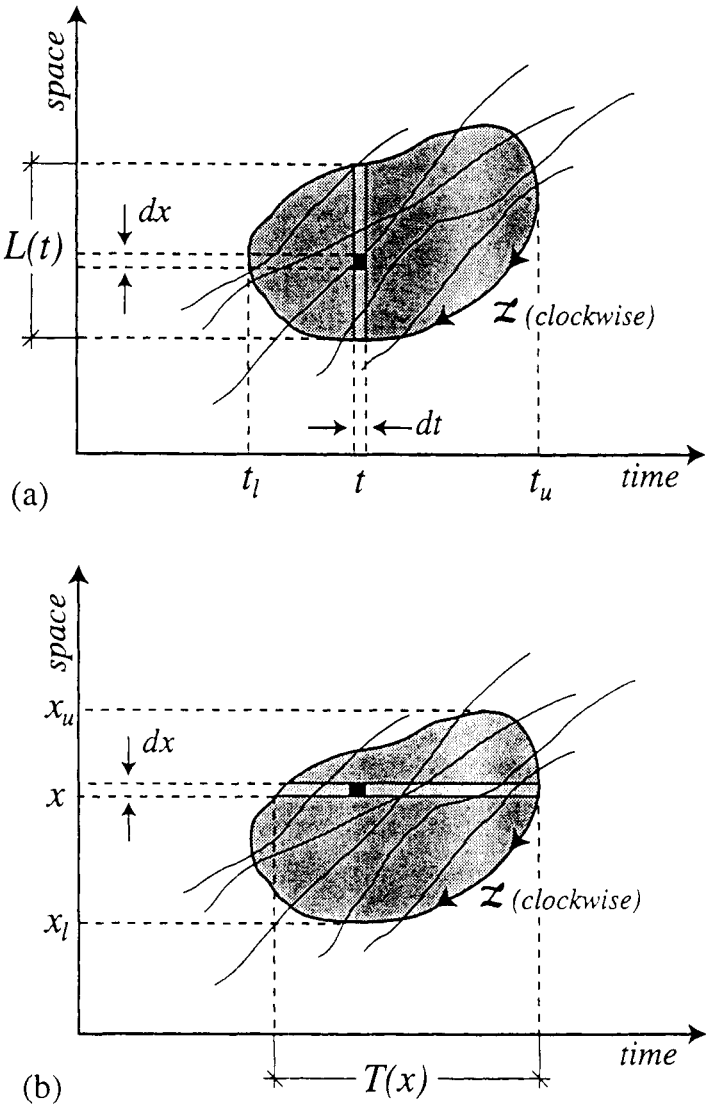


Figure 4.2 Representation of generalized traffic stream variables as averages of conventional measurements: (a) density; (b) flow.

If we now use L(t) to denote the length of the rectangle found at time t (see Fig. 4.2) and k(t) to denote the conventional value of density at

time t, i.e., k(t) = n(t)/L(t), then (4.4) can be rewritten as follows:

$$t(A) = \int_{t_l}^{t_u} k(t)L(t)dt.$$

The denominator of (4.3) can also be written as:

$$|A| = \int_{t_l}^{t_u} L(t)dt$$

so that (4.3) becomes:

$$k(A) = \int_{t_l}^{t_u} k(t)L(t)dt \Big/ \int_{t_l}^{t_u} L(t)dt. \tag{4.5}$$

This seemingly more complicated expression is interesting because it indicates that the generalized density is simply the average across time of all the (conventional) densities encountered in region A, weighted by length.

In the important case where a road section of given length, L, is observed for a fixed time the region of interest is a rectangle. Then L(t) = L, and we see from (4.5) that the generalized density is simply the unweighted time-averaged density observed in the road section.

We also note, as a curiosity more than anything else, that it is possible to express the three quantities appearing in (4.3) in terms of the vehicle counting function, N(t, x), defined in Sec. 2.4. The advanced reader may want to verify that

$$t(A) = -\int_{\mathscr{L}} N dt,$$

$$|A| = \int_{\mathscr{L}} x dt$$

and

$$k(A) = -\int_{\mathscr{L}} N dt \Big/ \int_{\mathscr{L}} x dt \tag{4.6}$$

where '$\mathscr{L}$' is the closed curve defining the boundary of A. It is assumed in these expressions that the line integrals are calculated in the clockwise direction; i.e., leaving the region A to the right.

There seems to be some confusion in the literature by what is meant by 'density'. Note that in all our definitions we have always defined

density as the ratio of two unambiguously defined quantities in a specific region of the time-space plane; and this is a precise definition. The problem arises when, in an analogy to models of continuous phenomena in the physical world, we attempt to define a point density by taking the limit of (4.2), (4.5) or (4.6) for a region **A** that shrinks around a specific point in time and space (e.g., point Q in Fig. 4.1). Such a definition would have practical appeal because for example we could then make simple statements to describe traffic at specific points in time-space such as 'traffic density at the Hegenberger off ramp of freeway I-880 was 100 veh/mile at 5:00 pm on Tuesday' without any further qualification. Unfortunately, the result of such a limit would either be zero (as for point Q in Fig. 4.1) or infinity (as for point P) and we must rethink more carefully what it is meant by a point density.

Some practitioners define density at a point (t, x) as the density in a road segment of definite length with x at its center, but this requires mentioning the length of the segment in any statement; the method has the further disadvantage that the resulting function k(t, x) varies discontinuously with x and t, so that some smoothing is required if differential calculus methods are to be applied to the data. An alternative approach consists in smoothing the data, replacing N(t, x) by a smooth monotonic function $\tilde{N}(t, x)$ as in Chapter 2, and then defining the density at a point, k(t, x), as the limit of (4.2), (4.5) or (4.6), where **A** is a rectangle of vanishing dimensions enclosing point (t, x). The advanced reader is encouraged to verify that this is equivalent to the equality:

$$k(t, x) = -\frac{\partial \tilde{N}(t, x)}{\partial x}, \tag{4.7}$$

in agreement with the definition given in Sec. 2.4.

The reader may complain that the results of (4.7) depend somewhat on the recipe that one uses for smoothing, i.e., on the chosen $\tilde{N}(t, x)$, and that therefore the definition of a point density is ambiguous. This is true, but fortunately not too important for most practical questions. Practical questions can usually be answered with any of the possible $\tilde{N}(t, x)$ just as satisfactorily as with the exact N(t, x) because one rarely requires answers to an accuracy of a single vehicle.

The exact same discussion, from (4.1) to (4.7), can be repeated for flow instead of density if the roles of time and distance are reversed in all the steps. Then, the generalized definition of flow in a region **A** becomes:

$$q(A) = d(A)/|A| \tag{4.8}$$

where d(A) is the total distance traveled by vehicles in A. This is the analog of Eq. (4.3). Instead of (4.5) we now find from a summation of the total distance traveled in *horizontal* rectangles partitioning A (see Fig. 4.2b):

$$q(A) = \int_{x_l}^{x_u} q(x)T(x)dx \Big/ \int_{x_l}^{x_u} T(x)dx \tag{4.9}$$

where T(x) is the length of the rectangle found at position x and q(x) is the flow corresponding to said rectangle. Likewise, one finds that

$$q(A) = -\int_{\mathscr{L}} N dx \Big/ \int_{\mathscr{L}} x dt \tag{4.10}$$

and that

$$q(t, x) = \frac{\partial \tilde{N}(t, x)}{\partial t}, \tag{4.11}$$

again in agreement with the definition of 'instanteneous' flow given in Sec. 2.4.

As in the case of density, the generalized definition of flow can be interpreted as a weighted average of its conventional definition. We see from Eq.(4.9) that q(A) is the average across space of all the (conventional) flows encountered in region A, weighted by time of observation. And again, when a road section of given length is observed for a fixed time, T, so that the region A is a rectangle, then the generalized flow becomes the unweighted average of the flows measured at all the locations on the road section. In the special case where the road section is empty at the beginning and end of our study, so that the flow is the same at all locations, then q(A) is the conventional flow in the observation interval. This quantity may be denoted $\bar{q}$ if one wishes to stress that the conventional flow is the time-average of the instantaneous flow.

Let us now consider a generalized definition of average 'speed'. The ratio of (4.8) and (4.3) reduces to the ratio of the total distance traveled in A to the total time spent A, which is a measure of average speed in A:

$$v(A) = \frac{q(A)}{k(A)} = \frac{d(A)}{t(A)}. \tag{4.12}$$

When A is a vertical rectangle such as the one in Fig. 4.1, t(A) = ndt and

$$d(A) = \sum_{i=1}^{n} v_i dt.$$

Thus, for instantaneous photos

$$v(A) = \frac{1}{n} \sum_{i=1}^{n} v_i; \tag{4.13}$$

i.e., v(**A**) is the average of the speeds seen in the photograph represented by the thin rectangle, which matches our prior definition of space-mean speed. Thus, we can view (4.12) as a generalization of the space-mean speed concept to arbitrary regions of the time-space plane.

For arbitrary regions, v(**A**) can also be interpreted as the average across vehicles of their (mean) speed in the region, weighted by the time spent in the region. To see this, let t_i and d_i denote the time and distance that vehicle i spends in region **A**, and then note that the last member of (4.12) can be expressed as follows:

$$d(A)/t(A) = [\Sigma_i d_i]/[\Sigma_i t_i] = [\Sigma_i (d_i/t_i) t_i]/[\Sigma_i t_i].$$

The last expression is the weighted average of the observed mean speeds (d_i/t_i) as claimed. Equation (4.13) is a special case of this formula because in a thin vertical rectangle all the vehicles spend the same time in the region and the average does not have to be weighted.

In the special case where we observe a road section of length L that is initially empty, i.e., $d_i = L$ for all i, the above expression reduces to:

$$v(A) = Ln/\Sigma_i t_i = L/\bar{t},$$

where $\bar{t}$ is the average across vehicles of the trip time in the 'system'. We have also seen that in this case q(**A**) = $\bar{q}$. Therefore (4.12) reduces to $\bar{q}\bar{t}$ = Lk(**A**). Since k(**A**) is the average of the conventional density, as per (4.5), we see that the second member of the equality is the average number of vehicles in the section during the observation period, $\bar{n}$, and therefore we can write $\bar{q}\bar{t} = \bar{n}$, which is just Eq.(2.7) in new notation. In other words, Eq.(4.12) is also a generalization of an important queuing relation. This alternative derivation of Eq.(2.7) illustrates that both the (t, x) and the (t, N) diagrams can often be used to describe the same phenomena and/or solve the same problem.

The above remarks also illustrate that the approximate relation, 'flow' ≈ "density' × 'speed', given by Eq.(1.13) of Chapter 1, becomes exact when applied to properly weighted averages of the three basic variables[1].

It is finally noted that it is also possible to characterize the generalized average speed as the weighted time-average of the conventional

space-mean speeds measured over time, using as weights the number of vehicles appearing in each 'photograph'. The justification of this statement is left as an exercise for the reader.

Some special cases. It is of particular interest to write the formulae for the flow in an instantaneous photograph of a road section of length L and for the density at a particular location for a specified time, T, in terms of the numbers of vehicles observed in each scenario (n or m) and their speeds (v_i or v_j).

Since the total distance traveled in the photograph is

$$\mathrm{d}(A) = \sum_{i=1}^{n} v_i dt$$

the flow for that scenario is:

$$q(A) = d(A)/(Ldt) = \frac{1}{L}\sum_{i=1}^{n} v_i. \tag{4.14}$$

For a horizontal rectangle (corresponding to a stationary observer) the time spent in it by car j is dx/v_j. Thus,

$$\mathrm{t}(A) = dx \sum_{j=1}^{m} \frac{1}{v_j}$$

and

$$k(A) = \frac{1}{T}\sum_{j=1}^{m} \frac{1}{v_j}. \tag{4.15}$$

For this scenario, $\mathrm{d}(\mathbf{A}) = m dx$ and the (generalized) space-mean speed becomes:

$$v(A) = \frac{d(A)}{t(A)} = \frac{1}{\frac{1}{m}\sum_{1}^{m} \frac{1}{v_j}} = 1/[\overline{1/v}] \tag{4.16}$$

Equation (4.16) states that the generalized space-mean speed at a fixed point in space is the reciprocal of the average of the reciprocal speeds, which is called the 'harmonic mean'. Since the harmonic mean never exceeds the arithmetic mean, this proves in a different way the statement made in Chapter 2 that the space-mean speed never exceeds the time-mean speed for situations, such as those that will be introduced in Sec. 4.1.2 (of which Chapter 1 is a special case), where the conventional and generalized space-mean speeds take on the same

values. We don't pursue this any further here, because the time-mean speed is of little practical importance.

Alternatively, we can think of $1/v$ as being a slowness or '*pace*' (in the sense used by long distance runners who like to gauge their performance by a 'pace', p, measured in minutes per mile, i.e., $p = 1/v$) and we could thus have expressed $v(\mathbf{A})$ for observation at a fixed point in space as the reciprocal of the average pace.

If we define the generalized mean pace by the reciprocal of (4.12), then we could also state for the same scenario that $p(\mathbf{A})$ is the arithmetic mean of the observed paces. Conversely, the formula for $p(\mathbf{A})$, if $\mathbf{A}$ is a thin vertical rectangle, would then be the harmonic mean of the paces in the corresponding photograph.

Table 4.1 summarizes the results of this subsection. Each row corresponds to a different generalized variable, whose label is given in column 1. Column 2 gives the formulas that apply to a vertical slice, $\mathbf{A}$, of the time-space diagram when the data come from an instantaneous photograph; the formulas assume that $i = 1, \ldots, n$ vehicles with speeds v_i appear on the photograph. Column 3 gives the formulas for a horizontal slice of the time-space diagram when the data come from observation at a fixed location; it is assumed that $j = 1, \ldots, m$ vehicles with paces p_j are observed. The boxed quantities are the conventional definitions of Chapter 1. The formulas are simpler if, as we have done, one uses 'pace' for the expressions corresponding to observation of a fixed location during a time interval and speed for the expressions pertaining to instantaneous observation of long road sections. Of course, the expressions remain valid if one replaces v_i by p_i^{-1} and p_j by v_j^{-1}.

We note that what Table 4.1 does not do is give formulas for the generalized definitions of a vertical slice when observation is at a fixed point in space, or for a horizontal slice when observation is at a fixed point in time. If this was possible, then we would be able to estimate the (conventional) density and space-mean speed that would be seen on a photograph without the expense of aerial surveillance, simply by taking measurements from a fixed location. The next section shows when and how this can be done.

4.1.2 Stationary traffic

We say that traffic on a long stretch of road is stationary during a period of observation if you cannot get any clues as to what time it is or where you are by inspecting the time-space diagram through a small window in a template. Traffic is not stationary if conditions change over

Table 4.1.. Generalized formulas for various traffic characteristics using two observation methods. Boxed expressions correspond to the original definitions introduced in Chapter 1:

	Method of Observation	
	Instantaneous photograph (section length, L)	**Observation from a fixed location (duration, T)**
Density, k(**A**)	$\boxed{n/L}$	$\frac{1}{T}\sum_{j=1}^{m} t_j$
Flow, q(**A**)	$\frac{1}{L}\sum_{i=1}^{n} v_i$	$\boxed{m/T}$
Space-mean speed, v(**A**)	$\boxed{\frac{1}{n}\sum_{i=1}^{n} v_i}$	$[\frac{1}{m}\sum_{j=1}^{m} p_j]^{-1}$
Average pace, p(**A**)	$[\frac{1}{n}\sum_{i=1}^{n} v_i]^{-1}$	$\frac{1}{m}\sum_{j=1}^{m} p_j$
t(**A**)	ndt	$dx\sum_{j=1}^{m} p_j$
d(**A**)	$dt\sum_{i=1}^{n} v_i$	mdx

time as in Fig. 4.3a, where at a given time (e.g., the end of a hail-storm) everyone increases their speed; you can see that the pattern seen through the template changes before and after the critical time.

Figure 4.3b and c include other examples where clues about your location in the time-space plane would be derived from the above form of observation. Case b is not stationary because all vehicles decelerate at a given location (perhaps because of a change in grade). Case c is not stationary either because flow and density increase behind the middle vehicle and this also yields clues as to when and where you are looking.

Traffic is stationary, however, if all the vehicle trajectories are parallel and equidistant. And it is also stationary if it is a superposition of families of trajectories with these properties (e.g., of fast and slow drivers). Of course, by using a very small hole in the template one could sometimes view an empty region of the diagram and other times not, so

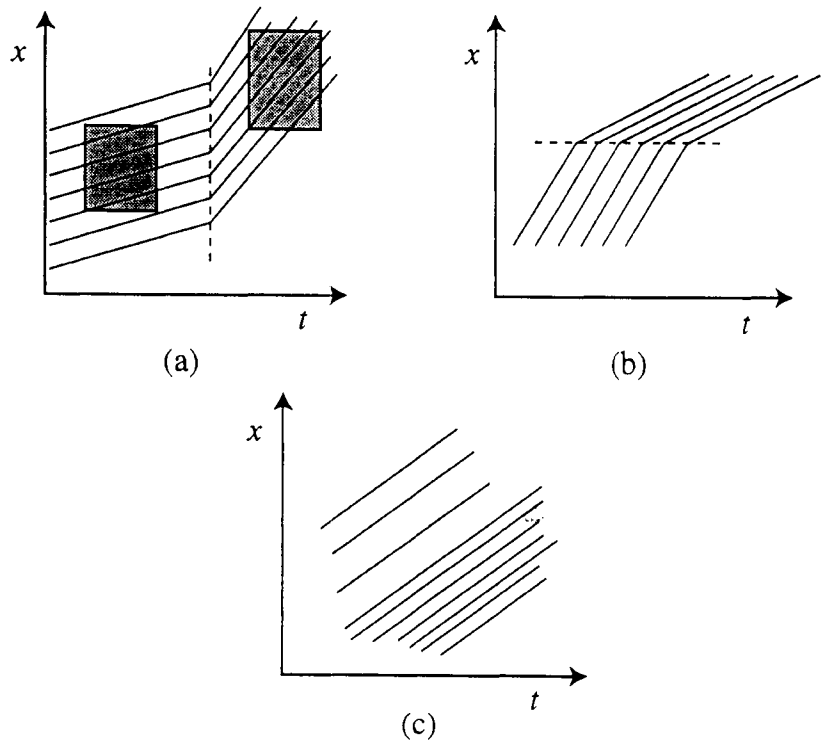


Figure 4.3 Three examples of non-stationary traffic.

that even in these cases, one could say that traffic was not stationary. Clearly, for such fine level of observation, stationary traffic does not exist. Obviously, we must exclude such a microscopic level of observation from our definition and must be satisfied if traffic 'looks the same' through larger windows. In fact, we relax the definition even further by only requiring that the quantities t(A) and d(A) be approximately the same regardless of where the 'large' window (A) is placed. It is in this sense that we will talk about stationarity in this chapter.

A direct consequence of this definition is that t(A) and d(A) should only depend on A through the area |A|; i.e.:

$$t(A) = \tau_0|A| \quad \text{and} \quad d(A) = \delta_0|A|, \tag{4.17}$$

where τ_0 and δ_0 are constants, representing the total time and distance traveled in a unit size rectangle of the (t, x) plane. That (4.17) is true

should be clear since A is made up of a large collection of identical-looking elementary rectangles of smaller size and the total time (distance) in A is the sum of the time (distance) in the elementary rectangles. Since these are identical (by stationarity) only their number influences t(**A**) (or d(**A**)), and since the number is proportional to |**A**|, Eq. (4.17) follows. Eq. (4.17) could in fact have been used as our definition of stationarity.

Aside from t(**A**) and d(**A**), all the remaining quantities in Table 4.1 are ratios of two of the following: t(**A**), d(**A**) and |**A**|. As such, they don't involve A in any way[2]. This is noteworthy because it entitles us to say that the limiting definitions (4.7) and (4.11), of q(t, x) and k(t, x), are independent of (t, x). It also allows us to state that our generalized definitions of traffic characteristics for horizontal and vertical rectangles are equivalent; i.e., that *the two columns of Table 4.1 yield the same values*. We can therefore say, for example, that density (in the conventional sense) is given by the ratio of the sum of the paces and the period of observation recorded by a stationary observer; i.e., that the density can be estimated from easily obtained data. We will see later in this book how these concepts are useful in the analysis of highway traffic detector data. The concepts, of course, are totally general and apply to the movement of any type of objects.

An equivalent definition of stationarity can be given in terms of the smoothed function $\tilde{N}(t, x)$. Namely, we should not be able to tell where in the (t, x) plane we are by looking at $\tilde{N}$ through any template—even an infinitesimal one. This means that the limiting definitions of density and flow, k(t, x) and q(t, x), given by (4.7) and (4.11) must be independent of t and x, as was stated earlier. This, of course, can only happen if $\tilde{N}(t, x)$ is a plane, i.e., the following linear function of x and t:

$$\tilde{N}(t, x) = \tilde{N}(0, 0) - xk + tq, \tag{4.18}$$

where k and q are constants[3]. Thus, we can also say that traffic is stationary in a (t, x) region if a plane is a good approximation for N(t, x) in said region. Since stationary regions are usually odd-shaped, and the N(t, x) not perfectly planar, the generalized definitions can be used to characterize an observed stationary traffic regime in a way that averages properly all the information in the (t, x) region of interest[4]. The next section concerns itself with the behavior of a traffic stream when conditions do not change with time.

4.2 Time-independent models

It is often observed empirically that if traffic flow does not change with

time significantly at a given location for an extended period of time (e.g., 15 minutes), then the (space-mean) speed observed for the period tends to be reproduced whenever the same (stationary) flow is observed. Subsection 4.2.1, below, explores the consequences of such a belief. It introduces diagrams that can be used to predict two of the three basic traffic variables at a point in space given the third, and then shows how the diagrams are related. The section also discusses time-independent traffic over inhomogeneous road sections, and in particular the stable states that can exist in a section including a bottleneck. Section 4.2.2 shows by means of an example how the diagrams appearing in manuals can be used for design, and discusses some issues pertaining to their use. Finally, Sec. 4.2.3 introduces a simple causal theory that explains why the relations observed for light traffic should be of a certain shape. Causal theories are important because, when a particular performance measure needs improvement, they suggest which causal factors should be changed.

4.2.1 Diagrams

It seems reasonable to postulate that if traffic conditions on a given road are stationary, there should be a relationship between flow and speed that will be a property of the road (e.g., number of lanes, grade, etc.), the environment (e.g., whether it is icy, sunny, raining, etc...) and the population of travelers (e.g., percentage of heavy vehicles, commuters in a hurry, etc...). This assumption is plausible since one can reasonably expect drivers to do the same *on average* under the same average conditions. Note, however, that we have not said whether flow influences speed or the reverse is true. This was done because both possibilities can arise.

For example, if traffic enters the upstream end of a highway at rate q until a stationary state develops downstream, then we would expect the space-mean speed that develops (downstream) to be a consequence of the input flow and the behavior of drivers as they interact with one another while passing. As a result of the definitional relation ($q = k\bar{v}_s$) for stationary traffic, we can also say that the downstream density would then be a consequence of the input flow. In this case, causality comes from upstream.

On the other hand, if a stationary state develops behind a slow-moving obstruction to traffic (e.g., a snow-plow), then through the reproducible behavior of drivers we would expect the average spacing inside the queue *upstream of the obstruction* to be a function of the travel

speed; and therefore for k and q to be a consequence of the obstruction's speed (with $q = 0$ if the speed is zero.) In this case, we see that causality comes from downstream.

Be that as it may, one would expect the basic features of a given traffic state, q, k, and $\bar{v}_s$, for a given road and environment to vary with a single parameter, α; i.e., for the possible traffic states to define a parametric curve in the $(q, k, \bar{v}_s)$ space: $q(\alpha)$, $k(\alpha)$, $\bar{v}_s(\alpha)$. We also note that q cannot be the parameter because in practice a given flow is observed for more than one speed (e.g., with low flows arising for very low and very high speeds). On the other hand, because speed can be expected to be a declining function of density (and vice versa), either k or $\bar{v}_s$ can serve as the parameter.

The justification for this statement is usually given in terms of a 'car-following' argument in which it is said that drivers keep wider spacings when traveling faster and therefore one would expect k to decline with speed. But one should not be surprised if experiments also show that a unique relation between v and k exists in situations where cars do not 'follow the leader'; e.g., when traffic is very light and there is almost free passing or when there is active lane changing[5]

It should be stressed at this point that these relationships are only postulated to be true 'on average'; i.e., for large stationary (t,x) regions containing many vehicles. If one measures flows (or densities) on small scales, then one should expect substantial deviations from our curve, e.g., because of driver differences.

Because three-dimensional curves cannot be plotted on a sheet of paper, various two-dimensional representations of the postulated relationship are often used. Some of those are shown in Fig. 4.4. Part (a) displays a diagram of speed vs. density, including a curve that represents the possible traffic states for a given highway. (We have dropped the over-bar and the subscript of $\bar{v}_s$ for simplicity of notation and will continue to do so from now on.) Because $q = kv$, flow is represented on this diagram by the area of the rectangle with corners at the origin and at the point in question, and sides parallel to the axes of coordinates. Thus, the three state variables can be conveniently read from the diagram; e.g., for point '1' of the figure. As in the figure, one would expect the curve to intercept both axes at coordinates that we label, v_f and k_j. The former represents the speed that arises when there is no (or very little) traffic, and is called the 'free flow speed'. The latter represents the maximum density observed, when traffic is bottled-up and at a stand-still; and it is normally called the 'jam density'.

In one of the earliest (if not the earliest) studies on this subject, Greenshields *et al.* (1947) proposed a linear relationship between v

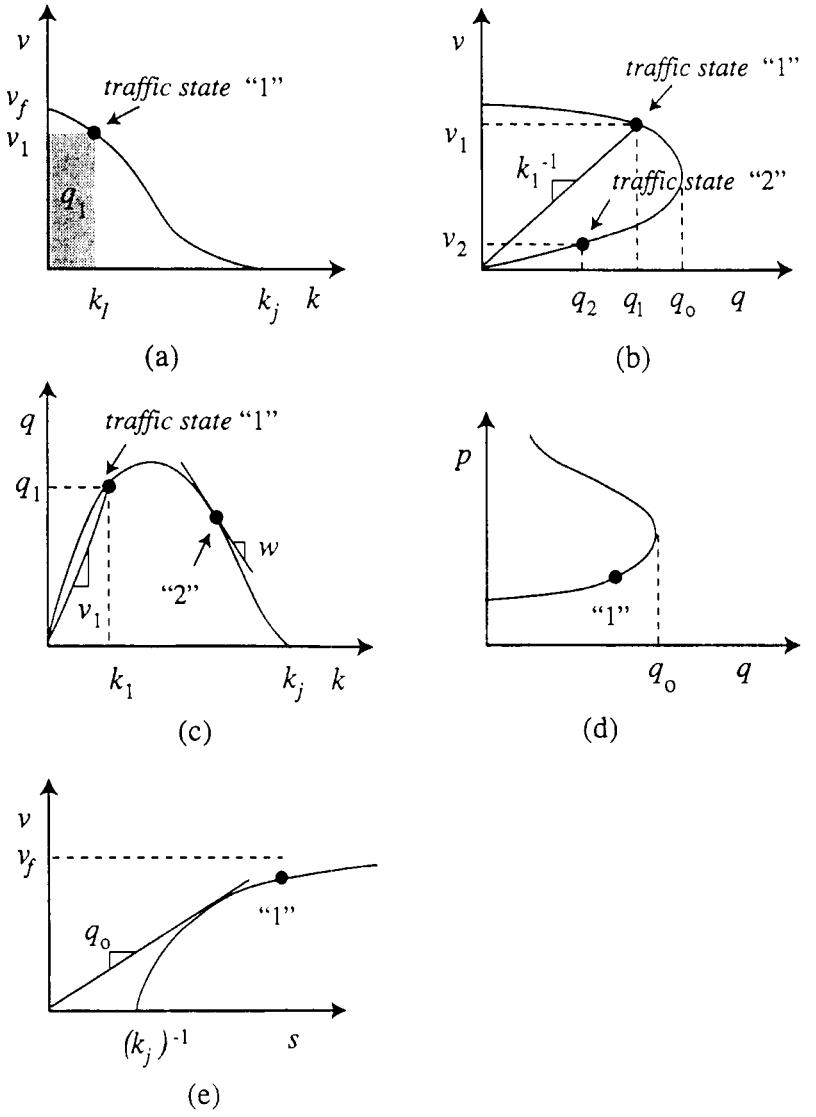


Figure 4.4 Equivalent representations of possible stationary traffic states, using diagrams for: (a) speed vs. density; (b) speed vs. flow; (c) flow vs. density; (d) pace vs. flow; and (e) speed vs. spacing

and k:

$$\frac{v}{v_f} = 1 - \frac{k}{k_j} \tag{4.19}$$

from limited field observations. We now know that a linear relationship is quite inaccurate, but (4.19) is of historical interest. It will be used here for the purposes of *illustration only* since its simplicity is appealing.

It can be seen at a glance from Fig. 4.4a that q varies smoothly as the point denoting the state is moved along the curve and that its value must reach a maximum, denoted q_o, at some point. Note as well that $q = 0$ for the extreme points.

Alternatively, one could plot q versus v as the point moves along the curve; and one would obtain a graph such as part (b) of the figure. On it, the axes have been reversed as is customary, so that the maximum stationary flow (or *capacity*) is shown on the abscissa. In this (v vs. q) representation of the possible stationary states, density is the reciprocal of the slope of the radius connecting a point with the origin (since $k^{-1} = v/q$).

The equation corresponding to Fig. 4.4b can be obtained by substituting q/v for k in the expression for v as a function of k and solving for q; the result for (4.19) is:

$$\left(\frac{q}{k_j v_f}\right) = \left(\frac{v}{v_f}\right)\left(1 - \frac{v}{v_f}\right) \tag{4.20}$$

which is the equation of a parabola. The maximum flow is obtained when $v/v_f = 1/2$ so that the capacity for the Greenshields et.al. (1947) model is $q_o = (1/4)k_j v_f$.

Part (b) of the figure also illustrates that although each v identifies a unique flow, the reverse is not true. For the relations typically postulated, in which the curve has an upper branch where v decreases with q and a lower branch where v increases with q, there are two possible speeds for every flow.[6] If the stationary state arises from traffic entering the upstream end of an empty homogenous highway that is free of downstream obstructions, where traffic reaches an equilibrium (q is given), then we would expect states on the upper branch of the curve to arise. This seems reasonable, since we assume that there are no downstream restrictions to reduce traffic's speed, but cannot be formally justified without an accurate model of how people drive and an analysis of the problem; a theory of traffic dynamics—the subject of our next

section—is needed for this. We can say, however, that every reasonable theory proposed to date leads to equilibrium states on the upper portion of the curve if the highway is initially uncongested, and that this is consistent with observation.

The bottom part of the curve describes states that can only persist if downstream conditions prevent traffic from moving faster (queued traffic). This can occur for example if a slow vehicle with speed v_2 (see Fig. 4.4b) blocks the road and traffic is forced to follow. The state values q_2, v_2, and $k_2 = q_2/v_2$ then describe the conditions prevailing in the (t, x) region containing the queue behind the slow vehicle. A similar effect is obtained if a fixed partial obstruction restricts the passage of flow. Now too, models and experience predict that queued slow-moving traffic will be created in the region upstream of the restriction. The resulting (stationary) conditions will correspond to a point on the lower branch of the curve[7]. Because of these considerations, we will refer to the upper branch as the 'unrestricted' branch and to the lower branch as the 'restricted' or 'queued' branch. This terminology is not universal but is evocative of the underlying causes determining the branch that arises in a specific case. Theories of traffic dynamics (Sec. 4.4) are needed to describe how traffic moves from a stationary state on one branch to the other as conditions change.

It is important to develop a 'feel' for how the diagrams in Fig. 4.4 (a) and (b) depend on road width. To a first level of approximation, we can say that traffic on neighboring freeway lanes interacts weakly and that therefore the flows that result for a given (average) speed should be roughly proportional to the number of lanes. And so should the densities. This is recognized in traffic manuals (such as the Highway Capacity Manual, 1994) which often use flow per lane and density per lane as the basic inputs for determining operating conditions; in this manner, the same diagram can be used independent of the number of lanes. When engineers see diagrams such as those of Fig. 4.4 they often think of 'per-lane' measurements. This convention is *not* adopted in this book so as to preserve maximum generality. Thus, we would expect k_j and q_o to be proportional to the number of lanes. You should try to imagine how the curves in the figure would change as you increase the number of lanes while holding everything else constant.

The diagrams, of course, will also depend (although to a lesser extent) on other features of the highway and its environment, and this has been the subject of much research. Traffic manuals from various countries summarize it. It is important to note that most of the information determining the particular form of the curve comes from characteristics of the highway (e.g., the number of lanes) although some also comes

from the 'quality' of traffic that moves over it (e.g., the percentage of trucks). What does not enter in the determination of the curve is the 'quantity' of traffic (i.e., q or k). It is thus useful to view the curve and its parameters, v_f, q_o, k_j, etc... as inherent properties of the highway that exist even in the absence of traffic. A particular amount of (stationary) traffic simply determines a point on the curve.

Engineers like to think of Fig. 4.4b as a diagram that is useful for design but they prefer to work with a different representation, Fig. 4.4c, for more detailed analyses of traffic operations. In this diagram flow is plotted versus density, and speed is given by the slope of the radius connecting a particular point with the origin. The equation of the curve in this diagram is obtained by substituting q/k for v in the relation of v vs. k and solving for q. The result of these manipulations for (4.19) is again a parabola:

$$\left(\frac{q}{v_j k_j}\right) = \frac{k}{k_j}\left(1 - \frac{k}{k_j}\right) \tag{4.21}$$

which exhibits a maximum for an 'optimum' density k_o that satisfies $k_o/k_j = 1/2$. The capacity, of course, is as before: $q_o = 1/4 v_f k_j$. In reality, values of k_o/k_j for freeways, are close to 1/6 with q increasing rapidly toward q_o as k increases toward k_o and then declining more gradually for $k > k_o$. The parabola (4.21), symmetric about $k = k_j/2$, does not exhibit these qualitative features.

While (4.21) or its equivalent for a general v(k) equation defines a unique q for every k, i.e., a function q = Q(k), the reverse is not true. Fig. 4.4 shows that this curve also exhibits two branches, i.e., two k's for every q, corresponding the unrestricted and queued regimes[8]. It is reasonable to expect q to increase with k up to a point (k_o) and then to decrease toward zero again. Although there is some controversy in the field today as to whether the q vs. k curve is discontinuous (and even multi-valued) near capacity, it should be noted that the data supporting such a conjecture may correspond to the inadvertent measurement of non-stationary traffic conditions (Cassidy, 1995). Some of these questions will be addressed in Chapter 6.

Figure 4.4c also displays a second point '2' and denotes by 'w' the slope of the q vs. k curve at that point. Like any slope in the (k,q) plane, w has units of speed; this variable plays a significant role in the theory of traffic dynamics.

Finally, we display two additional representations of the possible stationary states: a curve of 'pace' vs. 'flow' (d), and another curve of 'speed' vs. 'spacing' (e). The former is the cousin of the increasing travel time vs. flow curves used in transportation planning and large scale

network models, since pace is simply the average travel time normalized by length. Curve (d) is obtained after replacing v by $1/p$ in case (b). In network models one normally assumes that p increases with q and ignores the top part of the curve; i.e., the part of the curve that corresponds to obstructed flow. This can lead to difficulties when modeling congested networks.

Curve (e) is obtained from case (a) after replacing k by s^{-1}. In this representation, flow is given by the slope of the radius from a point to the origin, $q_1 = v_1/s_1$; thus, the capacity, q_0, is given by the slope of the tangent to the curve passing through the origin. Because they relate speed to spacing such curves have been used in studies of drivers' car-following; see Sec. 4.5.

Heterogeneous highways. The concepts in the present section can be extended to a heterogeneous highway (e.g., one whose width changes along its length) that is stationary in time. By time-stationarity we mean that flows and densities can only depend on location, or equivalently that the partial derivatives of $\tilde{N}$ are independent of t. The most general form of $\tilde{N}(t, x)$ satisfying this condition is $\tilde{N}(t, x) = \tilde{N}(x) + qt$, where q is a constant. Thus, we see from (4.11) that $q(t, x) = q$; i.e., that the flow must be the same everywhere.

If flow and density at each location are related to each other in a way that is independent of nearby traffic conditions (i.e., one can define curves of q vs. k for each point along the road) then it should be possible to plot curves depicting the values of k (or v) that would arise at each location for every level of *system* flow. To illustrate this idea, Fig. 4.5 shows a homogeneous road with a bottleneck, at location $\hat{x}$, and an accompanying diagram with two dashed (q,k)-curves. The small curve corresponds to the bottleneck, and the large one to any locations in the homogeneous part, such as x_u and x_d. Intermediate curves (not depicted) should arise at intermediate locations. The collection of curves can be represented mathematically by means of a smooth function of both k and x:

$$q = Q(k,x) \tag{4.22}$$

that gives the flow observed when the (time-independent) density at x is k.

We use $\hat{q}_0$ to denote the capacity of the bottleneck (see Fig. 4.5) and note that this value is the maximum possible system flow. Thus, if (stationary) measurements of flow and density are taken at any location then only the portions of the corresponding curve lying below the

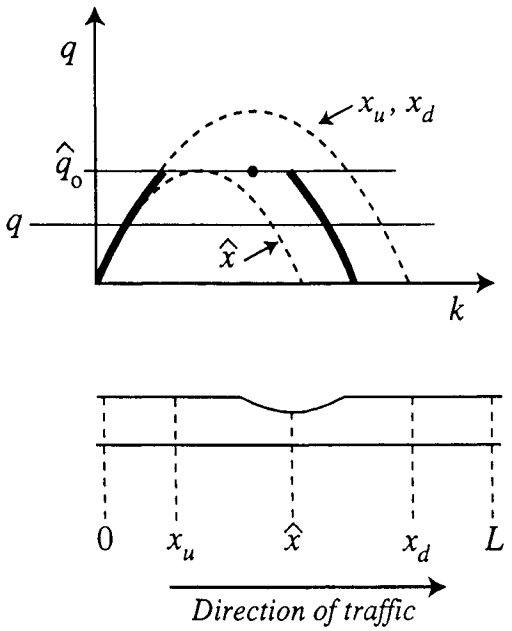


Figure 4.5 Relationship between flow and density on an inhomogeneous road section.

horizontal line at level $\hat{q}_o$ shown on Fig. 4.5 can be observed; i.e., the (q–k)-curve will appear to be truncated. Conversely, a horizontal line across the family of curves of Fig. 4.5 at any feasible height $q < \hat{q}_o$ can be used to identify two possible densities for each location, 'unrestricted' or 'queued', that would be consistent with the given flow. Fortunately, as is explained below, it turns out that only a few combinations of queued and unrestricted states across all locations are 'stable', so that knowledge of q tells much indeed about the overall status of the section; e.g., the number of vehicles in it, their trip times, etc.

It is generally believed and confirmed by observation that an 'unrestricted' state cannot persist for an extended period of time downstream of a 'queued' state if there is nothing to hold the queue back. This should be intuitive to experienced drivers because queues do not develop and grow spontaneously. They are caused by a restriction. In fact, if they exist and the restriction is removed then the front of the queue begins to recede as the queued vehicles advance into the uncongested road ahead. Experienced drivers also know that a bottleneck is a restriction that can separate upstream queued traffic from downstream

unrestricted traffic for a long time. When this happens, experience shows that the bottleneck is serving traffic at its maximum sustainable rate; i.e., that the system flow is $\hat{q}_o$. A bottleneck in this state will be said to be 'active' and to be working at 'capacity'. The same principles and definitions should also apply to the 'walking' transportation mode.

It is also possible, although unlikely, to observe a standing queue that does not grow or dissipate.[9] Then, 'unrestricted' traffic would be observed upstream of the end of the queue and 'queued' traffic downstream. If the possibility of standing queues is disregarded, only three stable states need to be considered in our inhomogeneous road section:

1. unrestricted; with the densities and speeds everywhere obtained from the left portion of the diagram;
2. queued; with the densities and speeds everywhere obtained from the right portion of the diagram; or
3. mixed: with the densities and speeds upstream of the bottleneck obtained from the right portion of the diagram, all the downstream data obtained from the left, and $q = \hat{q}_o$.

We are now prepared to investigate the form of the relationship between q and the aggregate density for the whole highway section ($0 \leq x \leq L$) which we denote k_{agg}.

Our discussion is relevant because manuals often define a single curve for a whole section of highway, even if it is inhomogeneous, based on the average features of the highway. While this is expedient, and may be needed on practical grounds, a direct extrapolation of the ideas presented up to this point without exercising care is rather dangerous. Let us now elaborate on this comment.

In order to 'construct' representation (c) of Fig. 4.4 we shall evaluate k_{agg} for every possible feasible q. Of course only q's satisfying $q \leq \hat{q}_o$ can be considered. If traffic is unrestricted, i.e., on the left branch of the q–k curve everywhere, then the density at x, k(x), is obtained from the inverse function corresponding to said branch at location x, which we denote $Q^{-1}(q, x)$. The definite integral (sum) of these quantities from $x = 0$ to $x = L$ is the total number of vehicles in the section and as a result we can write:

$$k_{agg} = \frac{1}{L}\int_0^L Q^{-1}(q, x)dx. \tag{4.23}$$

Note that this expression is just the average of the values of k that prevail at each specific point for the given (unrestricted) q. In Fig. 4.5,

this is represented by the increasing portion of the darkened curve. Each point on this branch of the curve is obtained by averaging the values of k obtained for the given (unrestricted) q by reading them off the (dashed) local q–k curves at different x.

If traffic is queued, then one would obtain through similar considerations the dark decreasing branch on the right side of the figure; and we note that the two branches so obtained are separated by a gap.

Intermediate values across the gap can sometimes also arise. A state corresponding to the dark dot shown on the figure, for example, would arise whenever the system is in the 'mixed' stable state (iii) mentioned above (with $q = \hat{q}_o$). The location of this point across the diagram is a property of the study section. It is given by the location of the bottleneck $\hat{x}$, since this determines the number of vehicles affected by the queue[10]. Intermediate values could also be observed for $q < \hat{q}_o$ if downstream (standing) queues were to spill back onto our study section, but in these cases the location of the point would be describing the length of the spillback and not an inherent property of our section.

These results are intriguing because they suggest that the diagrams in design manuals should not be smooth at capacity whenever they describe an extended heterogeneous highway section (e.g., over rolling terrain). The 1994 version of the Highway Capacity Manual explicitly states that the procedures and diagrams in Chapter 3 only pertain to homogeneous freeway sections, but the tables and adjustment factors (as those of the 1965 version) are not always consistent with this view.

If we wish to use for design aggregate q vs. k_{agg} curves such as Fig. 4.5 we may ask how the ratio q/k_{agg} should be interpreted. We recall that both q and k_{agg} correspond to the generalized definitions of traffic variables for the rectangle **H** defined by $0 \leq t \leq T$ and $0 \leq x \leq L$. As a result q/k_{agg} can be interpreted as the (generalized) space-mean speed, and Lk_{agg}/q as the average vehicular trip time in the section. In essence, q/k_{agg} is the definition of 'overall travel speed' (1965 Highway Capacity Manual, HCM) or 'average travel speed' (1994 HCM); it is not equal to the 1994 HCM's 'average running speed' which excludes stopped-vehicle delays.

4.2.2 Manuals

This section includes a couple of examples in which the recipes of the 1965 HCM (Chapter 9) are applied. (The 1965 procedures are easy to explain and similar in philosophy to the 1994 procedures.) This will serve to illustrate how engineers go about evaluating the desirability of a particular highway design and will provide a focus for some discussion.

For basic freeway sections without entering/leaving traffic, both versions of the HCM define a relationship between v and the vehicular flow per lane, q/l, where l is the number of lanes. The form of the relationship depends on the freeway's design speed and also (slightly) on l. The relationship is assumed to hold for 'ideal' conditions[11]. Adjustment factors for lane width and percent of heavy vehicles are then applied to the (forecasted) freeway flow per lane to determine an equivalent 'ideal' flow per lane, q^e. This flow is then used with the underlying relationship to obtain a predicted operating speed. Based on the predicted speed, and on the ratio of the predicted equivalent ideal flow to the maximum possible 'ideal' flow per lane, q_0^e (a measure of traffic intensity that is called the 'volume/capacity' ratio), the manual then identifies a 'level-of-service' that ranges from A (excellent) to F (fail). Level of service 'C' or better is achieved if $v \geq 45$ mi/hr and $q^e/q_0^e \leq 0.7$; and level 'B' or better if $v \geq 55$ mi/hr and $q^e/q_0^e \leq 0.5$. These two levels are typical design targets.

Example: Selection of a freeway's number of lanes. Let us find the number of lanes needed to accommodate at level of service 'C' a design flow of 3000 veh/hr. The freeway, passing through rolling terrain with 11 ft lanes and 4 ft shoulders, has a design speed of 65 mi/hr. There is a ditch and a median.

The first step in the solution is converting the 3000 veh/hr into an equivalent ideal flow. For a freeway, design speed and side interference factors are not used. The adjustment factor for lane width w_l (feet) and shoulder width w_s (feet) given in Table 9.2 of the 1965 HCM obeys the approximate expression:

$$F_w \approx 11 + \frac{(12 - w_l)^2}{35} + \left(\frac{i}{l}\right)\frac{(6 - w_s)^2}{170}; \quad 9 \leq w_l \leq 12, 0 \leq w_s \leq 6 \quad (4.24)$$

where i is the number of obstacles (e.g., ditches, medians) at the outside edge of the shoulders (i = 1 or 2). For our design i = 2 and the formula yields $F_w = 1.044$ for $l = 3$, $F_w = 1.04$ for $l = 4$, and $F_w = 1.052$ for $l = 2$. In view of the small differences among these values we choose to work initially with 1.044, which should be close to the actual factor for the number of lanes chosen.

Trucks and buses are treated by converting each one of these vehicles into an 'equivalent' number of passenger cars. The equivalences are determined based on the type of terrain or, for short sections on an incline, based on the length and steepness of the grade. The truck equivalents for level, rolling and mountainous terrain are $E_T = 2, 4, 8$.

For buses they are $E_B = 1.6, 3, 5$. These equivalents are then converted into a multiplicative factor by considering the fractions of trucks and buses in the traffic stream, f_T and f_B, or more precisely, the fractions that would be seen by a *stationary observer*. (The method of observation needs to be specified because if passenger cars travel faster than heavy vehicles then the fraction of heavy vehicles is observer-dependent.) The multiplicative factor is:

$$F_C = (E_T)(f_T) + (E_B)(f_B) + (1)(1 - f_T - f_a)$$
$$= 1 + (E_T - 1)(f_T) + (E_B - 1)(f_B).$$

For our example the factor is $1 + (1)(0.05) + (0.6)(0.05) = 1.08$ passenger car units (or pcu's) per vehicle, and the equivalent flow is: $q^e = 3000\ F_W\ F_C = 3382$ pcu/hr.

To determine the number of lanes one would divide 3382 by a trial 'l' and then, from the curves in the HCM, would determine if the design criteria for level of service 'C' are satisfied. In order to practice working with equations we will instead pretend that the HCM diagram obeys (4.20) with $v_f = 65$ mi/hr and $q_o = 2000$ pcu/hr per lane.[12] For the critical v corresponding to level of service 'C' ($v = 55$, $v/v_f = 55/65$) the right side of (4.20) is 0.13. Since for the model of (4.20) $q_o = 1/4\ k_j\ v_f$, the left side of said equation is $3382/(8000\ l) = 0.42/l$. Equating both sides, $0.42/l = 0.13$, we find the critical number of lanes satisfying the speed criterion; i.e., $l = 3.23$. Because l must be an integer it should be increased to $l = 4$ which will further increase v; i.e., it will continue to satisfy the speed condition for level 'C '. This service level will be achieved if, in addition, the volume-capacity ratio satisfies $q^e/q_o^e \leq 0.7$; i.e., if $3382/8000 = 0.42 \leq 0.7$ in our case[13]. Had this inequality been violated, l would have to be increased one more notch.

Discussion. The neat functional form of (4.24) illustrates that some of the curves and tables in the HCM (and other manuals as well) are the result of simple curve-fitting exercises (such as linear and non-linear regression) in which families of functions have been fit to available data, collected from many different freeways. Unfortunately, the manuals do not present the recipes in a form that makes their accuracy to be self-evident, and this makes it especially important to be very conscious of the types of errors that may arise. Errors in curve-fitting predictions can arise from (i) having chosen the wrong family of functions (specification errors), (ii) poor data fit (low correlation) and (iii) not enough data (low statistical reliability). There is little that can be done to guard

against type (i) errors, except perhaps ensuring that the best possible theories have been used to develop functional forms. This type of error is particularly severe when the desired inputs to the formulas or tables are not representative of the data that were used to generate the table. Type (ii) errors arise when the original data cannot be reliably predicted with the recipe. This is symptomatic of a recipe that does not include sufficient input information even if it is correct. Scatter-plots of the original data against the predictions one would have obtained with the recipe can give a rough idea of the magnitude of likely type (ii) errors (for a facility that it is similar to those in the original data). Examples of possible effects that are not included in these recipes but could be important include: condition of the pavement due to weather, whether the commute direction is into or away from the (raising, setting) sun, tunnel lighting, driver motivation due to proximity to merges, etc. Type (iii) errors can be corrected by gathering more data over a wider variety of facilities.

Because the original data is rarely displayed in manuals (to avoid clutter presumably) the user is given no clue as to the magnitude of type (ii) and (iii) errors. This problem will be accentuated with the introduction of calculation software in newer versions of the HCM. This is unfortunate because, as a result of the specific steps of the recipe, the inexperienced user may come away with a false sense of precision. It should be remembered that 'factor of 2' differences in lane capacities of similar freeways have been measured, and that such large discrepancies cannot be explained with the HCM's (small) adjustment factors.

This author believes that direct observation of a facility, or a facility as similar as possible to the one being studied, is a more reliable predictor of performance than a number found in a book. Nonetheless, a manual, coupled with practical experience and common sense, can be of use when the above course of action cannot be taken.

One should also reiterate, and this will become clearer when we talk about traffic dynamics, that the performance of a freeway section when traffic is heavy is only as efficient as allowed by its 'bottleneck'; this is a fact that remains despite the practical necessity to use input data that averages out some of the freeway's features (e.g., 'rolling terrain' as a proxi for a detailed vertical profile) in capacity analysis. Bottlenecks are normally located at places where the freeway geometry changes such as sag vertical curves, tunnel entrances and lane drops; they deserve to be carefully considered because their impact is felt widely and not just at the bottleneck itself.

An attempt to develop a causal theory of light traffic flow on freeways with an emphasis on stationary conditions was started more than four

decades ago by a number of authors, including Newell (1955 and 1995), Carlesson (1957) and Andrews (1970) among others. These theories predict a slow decline of freeway speeds with flow under unrestricted conditions with a particular form that depends on the number of lanes. (This will be explained in the next subsection, which introduces just the main idea at the core of these theories.) Although the 1994 HCM makes no use of these models, its new improved v–q diagrams now comply more closely with the theoretical predictions than those in the 1965 version. Similar theories have been developed to predict the form of the (more sudden) decline in speed as a function of flow for two-lane bi-directional roads; see for example, Morse and Yaffe (1971) and Daganzo (1975). Clearly, progress over the last 40 years might have been faster if empiricists and theoreticians had communicated better; attention to theory can reduce the dangerous type (i) errors, that may be present in our handbooks' recipes.

4.2.3 Light traffic theory

We are concerned here with a description of unrestricted flow on multi-lane freeways when vehicle delays due to an inability to pass are rare. It is assumed that vehicles enter he freeway section of interest and travel on it at a desired speed that varies from driver to driver. When the headway between successive cars becomes smaller than Δt (a minimum safe headway) we assume that they must occupy separate lanes.

As long as the number of vehicle trajectories crossing a location x_0 within an interval $(t, t + \Delta t)$ is less than the number of lanes, l, we can assume that vehicles can proceed without interference; there is no delay.

However, if on plotting the trajectories that vehicles would like to have followed (traveling at their desired speeds) we find that $(l + 1)$ trajectories intersect our interval $(t, t + \Delta t)$, we must conclude that a delay should have arisen. Then, if we assume that vehicles do not accelerate to make room for those behind, vehicle trajectories can only be spaced properly by forcing one (or more) of the vehicles to cross our location later than they would have liked. Although the specific vehicle that is delayed and the amount of the delay will depend on the relative vehicle velocities and their desired time of arrival at x_0, which would determine when and how the conflict first arose, the details of this process are unimportant for our purposes. We simply recognize that while a conflict is in progress one of the vehicles must reduce speed by

an amount comparable but smaller than the range of desired speeds (i.e., on the order of 5–20 km/hr). For typical desired speeds, on the order of 100–120 km/hr, the resulting increase in vehicle pace while the conflict is in progress will then range from 0.02 to 0.15 min/km. The average of these increments across all possible conflicts will be denoted Δp, which should be a quantity comparable with 0.1 min/km.

If conflicts are so rare that the resolution of a conflict does not trigger additional conflicts (i.e., multi-vehicle queues are rare) then an approximation to the total vehicle delay can be estimated from the frequency and duration of conflicts in a (t, x) region, using the desired vehicle trajectories.

To do this calculation in the easiest possible way we consider the infinitely thin rectangle (A) that includes (t, x) points with $x_0 \leq x \leq x_0 + dx$ and $0 \leq t \leq T$, as shown in Fig. 4.6, and look at the Δt interval behind each vehicle observation. The figure shows that 2 vehicles are found in such interval behind vehicle 1. If $l = 2$, this would mean a delay that would add on average $\Delta p\, dx$ vehicle-min (if we measure time in minutes) to the total time spent by all vehicles in A. Without getting into too much probabilistic reasoning we can advance at this point that the desired vehicle arrival times at location x_0 can be expected to be described by a 'Poisson process' with rate q (see Chapter 6). This means that the fraction of vehicles that would be followed by a conflict involving l other cars (in less than Δt) is approximately

$$\frac{e^{-q\Delta t}}{l!}(q\Delta t)^l.$$

If traffic is light, i.e., if the average number of arrivals per lane in time Δt is much smaller than 1, $(q/l)\Delta t \ll 1$, then the probability of conflicts with more than l cars can be neglected[14].

Let us now see how to write an approximate expression for the total time spent in A by all the vehicles. Letting $t_0(A)$ and $d_0(A)$ represent the total time and distance that would prevail in A if there were no delays, we can write (for $T \to \infty$):

$$t(A) = t_0(A) + \left(e^{-q\Delta t}\frac{(q\Delta t)^l}{l!}\right)(\Delta p dx)(qT) \leq t_0(A)$$

$$+ (q\Delta t)^l\left[\frac{\Delta p}{l!}\right](qT)dx$$

and

$$d(A) = d_0(A) = (qT)dx,$$

The first equality in these expressions expresses the delay due to the conflicts as the product of the probability of a conflict per car observed,

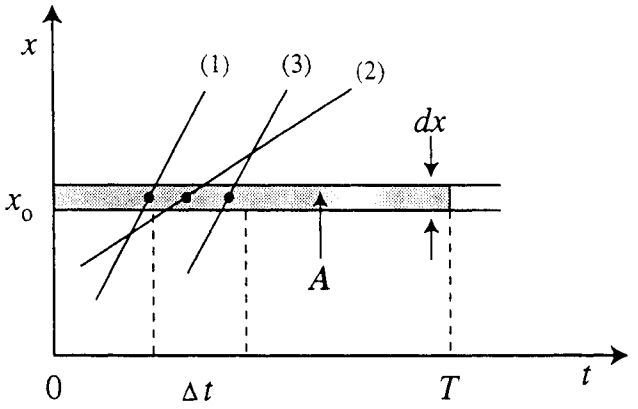


Figure 4.6 Example of a 3-car interaction of the type considered in a first order theory of light traffic flow on freeways.

the average delay per conflict (Δpdx) and the number of cars observed (qT). The ratio t(**A**)/d(**A**) is the generalized pace, which satisfies

$$p \leq p_f + \left|\frac{\Delta p}{l!}\right|(q\Delta t)^l \quad ((q/l)\Delta t \ll 1) \tag{4.25}$$

where p_f denotes the free-flow pace, t_0 (**A**)/d_0 (**A**).

Inasmuch as Δt represents a minimum headway, the term $(q/l)\Delta t$ equals the 'volume/capacity ratio' introduced earlier, q/q_0, and (4.25) can be written more intuitively as:

$$p \leq p_f + \left[\frac{\Delta p l^l}{l!}\right](q/q_0)^l \quad (q/q_0 \ll 1); \tag{4.26}$$

i.e., the pace cannot grow faster than the l^{th} power of the volume/capacity ratio, with a coefficient of proportionality that is comparable with Δp.

In terms of speeds (4.26) is equivalent to:

$$v \geq v_f\left\{1 + \alpha(q/q_0)^l\right\}^{-1} \quad (q/q_0 \ll 1), \tag{4.27}$$

where α is a dimensionless constant, $\alpha = (\Delta p/p_f)\ (l^l/l!)$, that is comparable with 1. Qualitatively, this means that v vs. q curves for multilane freeways should be rather insensitive to flow and that this

insensitivity should be more pronounced for larger l. In all cases, their slope for $q = 0$ should be zero.

Yet, the old (1965) HCM depicted gradually declining relationships. (This seems to be partially rectified in the 1994 HCM.) A possible explanation for this outcome is that the curves in the manual were generated by pooling data from many similar facilities in different parts of the United States and choosing the curves that would best fit the pooled data. Even if one controls for differences in freeway design and motorist population characteristics when pooling the data, chances are that facilities included in the same data set may not exhibit the same v_f. It is then possible for the data on which the curves are based to exhibit a pattern as in Fig. 4.7, where some freeways (whose data are denoted by crosses) might carry less flow at higher speeds than others (dots) for reasons not controlled for. This could happen for example if it turns out that people drive faster when traveling farther, as seems to be the case, and if western US freeways serve longer trips on average; this is the motivation for the labels used in the figure. Although the pattern for each individual freeway may obey (4.27), the combined data would suggest a relationship that declines too fast as is shown in the figure.

It is beyond the scope of this book to extend the theory just presented to non-stationary traffic. It is in fact quite complicated, and to this author's knowledge it has not been done correctly. An exception is perhaps the limiting case with no vehicular interactions (i.e., $q\Delta t/l \to 0$) treated in Section 3.1 of Prigogine and Herman (1971) and in Newell (1995), Chapter II. We now turn our attention to theories of non-stationary traffic in the special case where passing is restricted (e.g.,

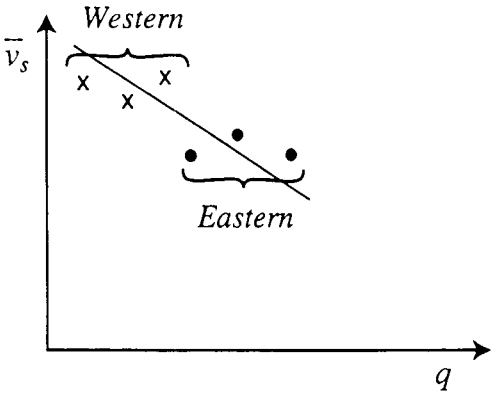


Figure 4.7 Example of an incorrect speed-flow curve obtained by pooling data.

because the road is crowded). The explanation starts with the basic building block of all traffic theories: the conservation law. This law, which applies to both passing and no-passing traffic, simply states that in a road without entrances or exits vehicles can be neither lost nor created; i.e., that if the number of vehicles in a given road section has declined by a given amount in a given time, then the number of vehicles elsewhere must have increased by the same amount. Extended forms of the conservation law can also be defined for facilities with entrances and exits. These extended forms are implicitly or explicitly an integral part of every theory of network traffic dynamics.

4.3 The conservation law

Section 4.3.1 below will show that the conservation law can be expressed in terms of an equation relating the 'point' flows and densities in a road section q(t, x) and k(t, x), and also in terms of the N(t, x) function. Although additional assumptions are needed to specify a complete theory of traffic flow, the conservation law by itself can already be used to make certain predictions. As an illustration of this, Sec. 4.3.1 includes two examples: one yielding the flow past a moving observer in stationary traffic, and another one yielding the (t, x)-path of an interface separating two (t, x)-regions with different stationary traffic states. Section 4.3.2 introduces conservation concepts with exiting and entering traffic.

4.3.1 No entering or exiting traffic

Recall that for a highway without entering or exiting traffic there is a function N(t, x) that gives the latest vehicle label seen at time t and location x, and that the finite differences $N(t, x_1) - N(t, x_2)$ and $N(t_2, x) - N(t_1, x)$ give the number of vehicles observed in the road section (x_1, x_2) at time t, and the number observed during the time interval (t_1, t_2) at location x. Figure 4.8 depicts a few vehicle trajectories and the values of such a function, including those at the corners of a square. Recall as well that the value of the function remains constant on the bands between given vehicle positions such as the darkly shaded region in the figure (corresponding to N = 7), and that vehicles trade position labels if there is passing.

It should be clear from observation of the figure that the aforementioned differences in N represent the number of vehicles crossing the sides of the square (e.g., $m_2 = 6 - 4$, $n_1 = -(4 - 7)$, etc...). However, if a vehicle were to enter the highway near (t_0, x_0) so that a trajectory was

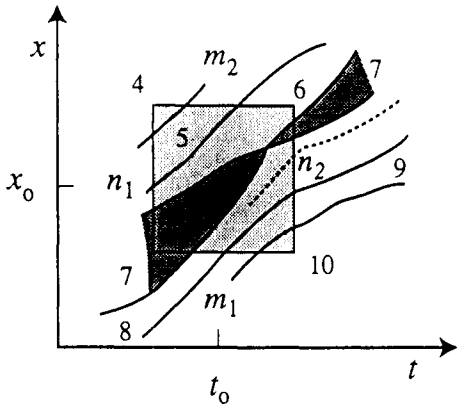


Figure 4.8 Values adopted by the N(t, x) function when there is passing.

created somewhere inside the square (e.g., dotted line) it would be impossible to assign a label to the bottom right corner of the square that would be true to both m_1 and n_2.

This means that our ability to define a function N(t, x) for all t, x in a region of interest implies no entering (or exiting traffic); its mere existence is a conservation condition that ensures vehicles are not created or lost along the road. This is perhaps the simplest and most general way in which the *conservation law* can be stated.

If we are dealing with the usual type of problems where N can be replaced by a smooth $\tilde{N}$ such that instantaneous flows and densities can be defined as $\partial\tilde{N}/\partial t$ and $-\partial\tilde{N}/\partial x$ at all points in time and space, then the existence of $\tilde{N}$ implies a relation between q and k.

If $\tilde{N}$ has second derivatives (flow and density are smooth) then we know that $\partial^2\tilde{N}/\partial x\partial t$ must be equal to $\partial^2\tilde{N}/\partial t\partial x$ which, in view of (4.7) and (4.11), means that $\partial q(t, x)/\partial x = -\partial k(t, x)/\partial t$; i.e., that:

$$\frac{\partial q(t, x)}{\partial x} + \frac{\partial k(t, x)}{\partial t} = 0. \tag{4.28}$$

For smooth q and k, (4.28) also implies the existence of $\tilde{N}(t, x)$. In the literature, Eq. (4.28) is called the *conservation equation*; it ensures that the rates of variation of flow and density in space and time are consistent with the no entering/leaving traffic hypothesis. We show below that Eq.(4.28) can also be derived with a direct argument that sheds additional light on its interpretation.

The reader will notice from Fig. 4.8 that the number of trajectories

entering the rectangle, $m_1 + n_1$, must equal the number leaving, $m_2 + n_2$; in other words: $(m_2 - m_1) + (n_2 - n_1) = 0$. Of course, this must be true of any rectangle you may draw, including very small ones (of dimension dt and dx). Since for small rectangles we are using the approximations m = qdt and n = kdx, the identity $(m_2 - m_1) + (n_2 - n_1) = 0$ becomes:

$$\frac{q_2 - q_1}{dx} + \frac{k_2 - k_1}{dt} = 0$$

after division by *dxdt*. This, of course, is the same as (4.28).

The two forms of the conservation equation just presented also apply if vehicle trajectories are allowed to slant both up and down; e.g., when traffic is bi-directional. Figure 4.9 shows the values of N(t, x) in a small region of the (t, x) plane where the trajectory of a vehicle traveling in the direction of decreasing x crosses two other trajectories. We continue to assume that crossing vehicles trade labels, so that N continues to represent position in the traffic stream. The region with N = 3 is shaded darkly on the figure. Thus, $N(t, x_1) - N(t, x_2)$ still represents the number of vehicles in (x_1, x_2) at time t (for any $x_1 < x_2$). The N-value may increase or decrease with t for a given location, however, because the sign of the change depends on the direction of travel; see for example the result for $x = x_2$ on the figure. As such, $N(t_2, x) - N(t_1, x)$ represents now the *excess* number of vehicles seen traveling *in the direction of positive x* . Therefore, if N can be approximated by $\tilde{N}$ then the partial derivatives of the latter with respect to x and t will now

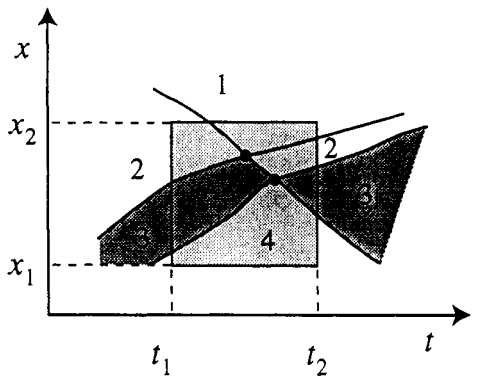


Figure 4.9 Values adopted by the N(t, x) function when there is two-way traffic.

represent the negative of the density (as before) and the *net flow* in the positive direction (not as before). This means that (4.28) still holds, if q is interpreted as the *net positive flow* passing a stationary observer.

A third form of the conservation law can be written in terms of line integrals of the gradient of $\tilde{N}$; i.e., the vector of partial derivatives $(\partial\tilde{N}/\partial t,\ \partial\tilde{N}/\partial x) = (q, -k)$. It is well known from calculus (Green's theorem) that a line integral of the projection of the gradient along any curve joining two points is independent of the curve, and that the result equals the difference in $\tilde{N}$ at the extremes of the curve. In our particular instance, if C is a curve in the (t, x) plane between two points (t_1, x_1) and (t_2, x_2) we have:

$$\tilde{N}(t_2, x_2) - \tilde{N}(t_1, x_1) = \int_C q\,dt - k\,dx, \quad \text{for all } C. \tag{4.29}$$

On taking C to be a closed loop, such as the loop $\mathscr{L}$ around A in Fig. 4.1, or the perimeter of the square in Fig. 4.8, we find:

$$\int_{\mathscr{L}} q\,dt - k\,dx = 0, \quad \text{for any closed loop } \mathscr{L}. \tag{4.30}$$

Evaluation of this integral clockwise and around the square of Fig. 4.9 yields the familiar result: $-n_1 + m_2 + n_2 - m_1 = 0$.

Forms (4.29) and (4.30) of the conservation equation are useful because they apply even if q and k are discontinuous, when (4.28) could not be applied. As we shall see in the second example outlined below, the integral form of the conservation equation can be used to determine the speed of an interface between two stationary states in the (t, x) plane, even if the detailed behavior of traffic inside the interface is not known. The expression is also useful in continuum theories of traffic flow (Sec. 4.4) to trace the paths of these interfaces when conditions are not stationary. It should also be said that both of these feats can also be achieved from the requirement that the N(t, x) function be continuous and that, as we shall see in Sec. 4.4, this most basic form of the conservation law simplifies the theoretical calculations considerably.

Example: Relative flow measured by a moving observer. Suppose that an observer travels a distance L during a time T with a variable speed that averages $v^\circ = L/T$. Then we ask: If traffic is stationary with flow q and density k, how many more vehicles will have passed her than she will have passed? By definition of $\tilde{N}$, the answer to this question is the difference in the vehicle labels at $(t_1 + T, x_1 + L)$ and (t_1, x_1), which is independent of the details of the observer trajectory and can be evalu-

ated with (4.29). To do this easily we choose a curve C that goes first from (t_1, x_1) to $(t_1 + T, x_1)$ (keeping x constant) and then from $(t_1 + T, x_1)$ to $(t_1 + T_1, x_1 + L)$ (keeping t constant). Then, the line integral reduces to:

$$\int_{t_1}^{t_1+T} q dt - \int_{x_1}^{x_1+L} k dx = qT - kL. \tag{4.31}$$

which yields the answer to our question. ■

On dividing this result by T, we obtain a useful formula for the net *relative flow* q° measured by the observer during time T:

$$q^o = q - kv^o. \tag{4.32}$$

Since this expression is independent of T, q° is also the *instantaneous* flow passing the observer. Note that (4.32) holds whether or not the observer maintains a constant speed, and that the relative flow is zero if $v^\circ = \bar{v}_s$. If the observer travels faster, then she will see negative flows (meaning that she will pass more vehicles than will pass her). This result should be intuitive in the case where all vehicles travel at the same speed; then an observer traveling with the speed of traffic will obviously see no flow.

The moving observer results can be displayed neatly for all the possible traffic states of a particular facility by means of a flow versus density diagram. If on a diagram containing the facility's $q = Q(k)$ curve one plots a ray emanating from the origin with slope v° (see Fig. 4.10) then the vertical separation between the curve and the ray (positive or negative) is the relative flow predicted by (4.32). The reader is encouraged to extend these results to the case of an inhomogeneous highway segment, such as that depicted on Fig. 4.5.

Example: velocity of an interface. Suppose that, as occurred in the three parts of Fig. 4.3, traffic is stationary in two adjoining sections of the (t,x) plane. Is there something that can be said about the interface between these two regions when no traffic enters or leaves the highway if the upstream and downstream states are known and different; i.e., if (q^u,k^u) (q^d,k^d)? Yes, and the answer again can be obtained from (4.29) by placing points '1' and '2' on the interface between sections and noting that the result obtained from a curve that is entirely upstream, C^u, must match the result from a downstream curve, C^d.

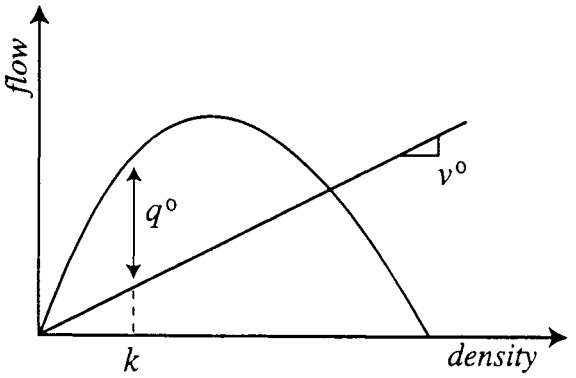


Figure 4.10 Geometrical depiction of the relative flow seen by a moving observer.

If points '1' and '2' are separated by T time units and L distance units (as in the previous example) and the interface is very narrow, then Eq. (4.31) remains valid for both curves and we have:

$$q^u T - k^u L = q^u T - k^u L, \quad \text{or} \quad q^u - k^u U = q^d - k^d U \qquad (4.33)$$

where U = L/T is the velocity (positive or negative) of the interface. Solving for U we find:

$$U = \frac{q^u - q^d}{k^u - k^d}. \qquad (4.34)$$

This result indicates that the interface's velocity is given by the change in flow across the interface over the change in density. ■

Equation (4.34) is a direct consequence of the conservation equation. It holds true independent of the detailed behavior of drivers in each of the two stationary regimes, e.g., despite the presence of passing, 'tail gating' drivers, high frequency oscillations or any other phenomenon that would not rule out stationarity on the relevant scale of observation. The equation applies even if the interface between stationary traffic states has a 'characteristic width' where traffic is not stationary because vehicles are adjusting from one region to the other. This is true because one can always choose points '1' and '2' to be very far apart, in comparison with the characteristic width of the interface, and then choose two curves C_1 and C_2 that are almost entirely outside of the interface so that (4.33) will hold in the limit.

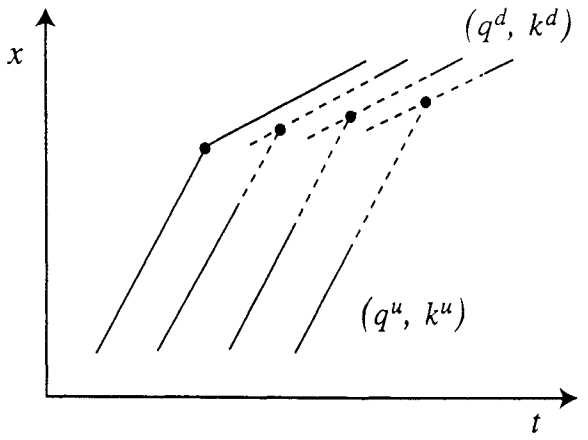


Figure 4.11 Velocity of an interface: geometrical interpretation.

An alternative, purely geometrical, derivation of (4.34) can be made for the special case where all the vehicles travel at the same speed in each of the two regimes and the interface has zero width. Since the traffic regimes are known it is possible to draw two families of vehicle trajectories, such as the two families of solid lines shown in Fig. 4.11. Then, if one of the vehicle trajectories is completely identified, e.g., the left-most solid line in the figure, we see that the extrapolation of the remaining trajectories (shown by dotted lines) identifies unique points where each vehicle would change speeds. These points uniquely define the interface trajectory. Its slope can be seen clearly to depend on the horizontal (q^{-1}) and vertical (k^{-1}) separation of the trajectories in the figure. The result is again (4.34); can you show it?

4.3.2 Entering and exiting traffic

This section may be skipped without loss of continuity because the concepts about to be presented are not used for the remainder of this book. The ideas should be of value, however, for readers intending to study network traffic dynamics, which is a currently active research area.

Conservation equations can be written if traffic enters and/or leaves at particular points on the highway. We have already seen that in such a case a function N(t, x) (or $\tilde{N}$) does not exist, but we can do something equivalent. If the dotted trajectory in Fig. 8 materializes, then the difference between the number of trajectories exiting the square ($m_2 + n_2$) and those entering ($m_1 + n_1$) will be 1; i.e., the net number of

trajectories generated in the square. This statement, of course, is true of any region of the (t,x) plane enclosed by a loop (a simply connected region). For problems where instantaneous flows and densities can be defined, this observation can be expressed mathematically as an extension of (4.30).

To see this recall that (4.32) represents the relative net flow passing a moving observer. This expression is true for $T \to 0$, so that it applies at all times for an observer moving along any curve (with t increasing) such as the lower or upper branches of $\mathscr{L}$ in Fig. 4.2a, which we denote C^l and C^u.

Thus, the integral of $q^\circ dt = qdt - kdx$ along these curves represents the net number of vehicles entering (for C^l) and leaving (for C^u) the enclosed region A along these curves. The total number of vehicles leaving the region, which must be the total net number generated in A, G(A), is given by the line integral along C^u minus the line integral along C^l; i.e., by the left side of (4.30), with $\mathscr{L}$ traveled clockwise. This allows us to write the generalized conservation relation:

$$\int_{\mathscr{L}} qdt - kdx = G(A), \tag{4.35}$$

which applies even if q and k are discontinuous.

In applications where it would make sense to define g(t, x) as the traffic generation rate (in vehicles per unit time, per unit distance) at point (t, x) we can write:

$$G(A) = \int_A g(t, x)tdx. \tag{4.36}$$

If q and k are smooth, then it is possible to rewrite the conservation law in a differentiated form. To see this, consider an elementary rectangle of dimensions dt and dx as in Fig. 8. Then, G(A) = g(t, x)*dtdx* and we have already seen that the left side of (4.35) equals $\partial q/\partial x + \partial k/\partial t$, where q is the 'net' flow. Therefore,

$$\frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} = g(t, x), \tag{4.37}$$

which generalizes (4.28) to cases with entering and leaving vehicles.

If there are intersections at specific locations, x_i, where traffic enters or leaves the freeway with net generation rates $g_i(t)$ (in vehicles per unit time), then

$$G(A) = \sum_i \int_{C^i} g_i(t)dt \tag{4.38}$$

where C^i is the portion of the curve $x = x_i$ that is in A.

In this particular case it is also possible to write the conservation law

for the singular points where traffic enters or leaves the road in terms of the N(t,x) functions that describe traffic on the highway segments between points of entry or exit; i.e., by defining an N_{i+1} for the highway segment directly downstream of i and upstream of i + 1 (with the convention that $x_1 < x_2 < x_3 ...$). To do this we define for each intersection an (arbitrary) reference time t_i when we choose $N_i = N_{i+1}$ and note that at any other time the N must satisfy:

$$N_{i+1}(t, x_i) - N_i(t, x_i) = G_i(t) - G_i(t_i), \tag{4.39}$$

where $G_i(t)$ is the cumulative number of (net) entries at x_i between time zero and t (i.e., the integral of $g_i(t)$, $G_i(t)$). One may choose the same t_i for all intersections, if desired. Equation (4.39) is simply another form of saying that the difference between the numbers of vehicles to have passed the downstream and upstream sides of location x_i since the reference time must equal the net number to have entered at x_i.

In practical applications where instantaneous flows can be defined the time derivative of (4.39) may be written as:

$$q_{i+1} = q_i + g_i \tag{4.40}$$

where the arguments t and x have been eliminated from the expressions. This equation simply states that the total flow leaving an intersection at a given time must equal the sum of the flows entering it at the same time. The equation is used in network models, and is sometimes called the 'node' conservation equation. It can also be obtained from (4.35) if **A** is taken to be a thin horizontal rectangle straddling the line $x = x_i$ between t and t_i.

The conservation law also applies to situations where different vehicle types move on the road. If 'types' is an inherent vehicle property (e.g., trucks vs. cars, destination, etc...) the modification is simple. We just apply the conservation law, to each of the classes; e.g., by introducing additional subscripts. If the type of a vehicle can change (e.g., we are keeping track of the vehicles by freeway lane), then we need to define the rates at which vehicles change from one type to another per unit time and unit distance (e.g., from j' to j) $g'_{j'j}(x, t)$ and introduce such terms in the formula for calculating the right side of (4.35); essentially g should be replaced in (4.36) by $g_j + \Sigma_{j'} g_{j'j}$. The differentiated form of the conservation equation (4.38) becomes then:

$$\frac{\partial q_j}{\partial x} + \frac{\partial k_j}{\partial t} = g_j + \sum_{j'} g_{j'j}. \tag{4.41}$$

4.4 Dynamic macroscopic models

This section contains an introduction to continuum models of traffic flow dynamics. The presentation is somewhat long because in an attempt to make it self-contained to typical first-year graduate students in transportation engineering, some concepts in the theory of partial differential equations had to be included as part of the explanation. And although the section only covers one model – the simplest and most useful traffic model in this author's opinion – several application recipes have been covered in detail. The ideas are organized into subsections as follows: Section 4.4.1 reviews the kinds of problems that can be solved with continuum models and outlines the assumptions behind the simplest model; Section 4.4.2 describes the concept of a *wave* and then the nature of the solution in terms of such waves; Secs. 4.4.3, 4.4.4 and 4.4.5 then describe solution methods for gradually more complicated problems; and finally Sec. 4.4.6 briefly discusses some numerical approaches, extensions and accuracy issues.

4.4.1 Background

The goal of a theory of highway traffic dynamics is to predict the evolution of traffic into the future from some set of initial conditions (e.g., the positions, types and speeds of all vehicles on the road at time $t = 0$) and some time-varying data (e.g., the times and speeds at which various vehicles, of known type, enter the road at its upstream end, for $t > 0$.) To make such predictions it would be necessary to know how each vehicle (driver) type reacts to a specified set of circumstances in its environment (such as its position and time, its headway, the relative speed with its leading vehicle and the evolution of these data for the recent past.) It should be easy to recognize that driver behavior is quite complicated and that a complete list of relevant stimuli would have to be much longer. In any case, if we knew what drivers did in every circumstance, then it would be possible to predict their future positions, speeds, etc... over small time increments so that one could construct for every car a time-space trajectory much as we did for the example of Fig. 1.2 in Chapter 1. The problem with such an approach is that we can never hope to have all the behavioral information to predict precisely 'what happens if' and, even if we did, we cannot expect to have the detailed individual vehicle data that would be necessary to apply the theory.

It is then logical to ask whether there are some coarse (macroscopic) measures of traffic behavior that can be predicted based on also coarse

information about current traffic conditions and future upstream inputs. For example we might like to be able to predict approximately the cumulative number of cars that pass various road locations as a function of time because we have already seen (Chapter 2) that such knowledge readily yields information about the accumulation of vehicles between locations and vehicular trip times. One would hope to be able to make such predictions based on coarse data such as the cumulative number of vehicles arriving at the upstream end of the road as a function of time, and the capacity of a downstream bottleneck.

It should be noted that the method presented in connection with Fig. 2.3b (Sec. 2.2.1) for predicting the extent of a physical queue upstream of a constant-capacity bottleneck on a homogeneous road qualifies as a traffic flow theory that solves this particular problem because the method can be used to predict the times at which each individual vehicle passes any particular location. The reader may also recall from the discussion of Sec. 2.2.1 that this particular method was not needed to predict vehicular delays because they were independent of spatial considerations, and that this property of delays was also true for time-dependent bottlenecks. The main added value of a more detailed model of traffic dynamics for the latter kind of problems lies in is its ability to predict where the delayed vehicles are stored.

Models of traffic dynamics could for example be used to predict the evolution of traffic on a long inhomogeneous road when both the cumulative flow wishing to enter upstream and the initial distribution of cars along the road are specified. In the notation of Sec. 4.1, we would like to predict $N(t, x)$ for various x's, given that $N(0, x)$ and $N(t, 0)$ are known approximately for $x > 0$ and $t > 0$. We will call this problem 'the semi-infinite highway problem'.

Other problems of potential interest are 'the infinite highway problem' (also called the initial value — or Cauchy — problem in mathematics) in which $N(0, x)$ is specified for all x and we seek $N(t, x)$ for $t > 0$; and 'the 3-detector problem' in which $N(t, x)$ is specified for all t at the locations of hypothetical 'upstream' and 'downstream' detectors x_u and x_d ($x_u < x_d$), and we seek $N(t, x)$ for $x_u \leq x \leq x_d$. We call the latter problem the 3-detector problem because its solution yields the $N(t, x_m)$ for the location x_m of an intermediate detector. We will use the term *boundary* to denote the curve(s) of the (t, x) plane on which the known data (or *boundary conditions*) are specified; e.g., lines $x = x_u$ and $x = x_d$ for the 3-detector problem.

In general we can say that the goal of our theory is to find $N(t, x)$ given some physically meaningful boundary conditions. An ability to make such predictions should help evaluate the efficiency of time-de-

pendent control schemes such as freeway ramp metering, the practicality of emergency evacuation measures, and the ability to monitor the state of a system continuously from detector data (e.g., detecting and diagnosing incidents). It should be noted that the system under consideration can be a passenger corridor, a bicycle lane or a freeway, and that the theory about to be presented is 'mode-abstract'. Although (for historical reasons) we adopt a highway traffic terminology, since this was the context in which the theory has been developed, we advance here that the only requirements of the theory are (i) that it follows objects that obey the conservation law of Sec. 4.3 (a pretty mild requirement) and (ii) that it satisfies one more 'local' criterion.

By this we mean that the traffic conditions at time t and location x, should only depend on the traffic conditions in the interval $(x \pm \Delta x)$ during the time interval $(t - \Delta t, t)$; and that both Δx and Δt can be considered to be 'small', in some sense. In continuum models of traffic flow (a better word than 'macroscopic' to describe what they actually do) it is assumed that Δx and Δt can be replaced by differentials, which opens the door to the world of calculus. This assumption is very appealing because the laws of traffic might then be reduced to a partial differential equation (or a system thereof) that may be studied as elegantly and simply as those of other physical phenomena that are also governed by partial differential equations; e.g., water flow in rivers, gas dynamics, elasticity, etc....

To be able to do this, however, we must be able to say that the variables appearing in the formulation of the model (i.e., N) must be unambiguously defined for a level of description where Δx and Δt can be treated as differentials. While this may be true for physical systems consisting of elements with very many molecules, even if the elements are made very small on a human observation scale, it may not be always appropriate for traffic. We had stated earlier in this chapter that the discontinuous function $N(t, x)$ may be approximated by a smooth $\tilde{N}(t, x)$ in many different ways and that the forms of the conservation law that involve $\tilde{N}$ and its derivatives (k and q) are only relevant for a level of description where all the smooth $\tilde{N}(t, x)$ give equivalent answers. The same can be said now; only if the answers to our questions involve quantities of vehicles evaluated over periods of time and lengths of highway such that all smooth approximations of N give the same answer, can we be confident that a continuum model will be of some use.

In this chapter we will present one such model (Lighthill and Whitham, 1955 and Richards, 1956), which we shall call the LWR model.[15] This model is in fact the simplest dynamic model ever formu-

lated and is intended to describe only *quite coarsely* the evolution of traffic. Because of the model's coarseness and some inaccuracies it is known to possess, (discussed later) various refinements have been proposed. Unfortunately, these refinements have often borrowed lines of thought from physical disciplines such as gas dynamics and the kinetic theory of gases where a more detailed description of the system can be achieved by complicating the partial differential equations that govern it. Unlike in the physical analogies, however, the additional phenomena that these 'improved' models try to capture is measured on space (and time) scales that are comparable with the vehicular spacing (or headway) where the notions of density (or flow) and its derivatives are ill-defined. At that level of description, for example, we may need to recognize that a driver reacts differently to the spacings and relative speeds in front of it than to those behind, which is difficult to do with partial derivatives. It is this author's opinion that detailed models of traffic flow that are formulated as partial differential equations should be regarded with suspicion, and one should not be surprised if they do not have the desired effect. This point will be discussed again in Sec. 4.4.6.

The LWR model arises from the assumption that the stationary relationships of Fig. 4.4 also apply when traffic is not stationary. This means that we take q to be a function of k (and t and x perhaps), $q = Q(k, t, x)$, independently of the flows densities and speeds prevailing upstream and downstream of x, and also independently of these conditions at prior times. That is, we ignore whether the vehicles currently at x have incurred any delays, sudden decelerations, etc. While this may be inaccurate for a detailed description it seems like a reasonable thing to try for long crowded roads examined on a coarse scale.

The next three subsections show how the LWR problem can be solved for (homogeneous) highways whose features do not change in time or space. In these cases the function $Q(k, t, x)$ can be treated as if it had only one argument so that we shall write $q = Q(k)$ instead. The inhomogeneous highway problem is addressed in Sec. 4.4.5. The next two subsections present in some detail an unconventional but practical way of solving the homogeneous highway problem with general data. Because the material is more difficult to read, most first time readers should probably read only the introductory remarks of Sec. 4.4.2 upto the section entitled 'The LWR trajectories and the entropy condition' (where the concept of a wave is introduced) and then proceed directly to Sec. 4.4.4. This latter section shows in a less mathematical way how to solve in the conventional way problems with piecewise constant data.

4.4.2 Solution methods using waves: nature of the solution.

We know from conservation that if two regions of the time-space plane are neighboring they must be separated by an interface that satisfies (4.34); i.e., for traffic states A and B, the interface velocity U_{AB} is given by:

$$U_{AB} = \frac{q_B - q_A}{k_B - k_A} \tag{4.42}$$

This of course is the slope of the segment joining the representative points A and B on the q – k curve; see Fig. 4.12.

For a case where vehicles do not pass, the right side of that figure depicts a few vehicle trajectories that pass through different states: 'A' near the bottom, and 'B' and 'B'' immediately above. If the scales on the axes have been chosen so that parallel segments on the (k, q) and (t, x) planes represent the same velocity, then for the diagram on the right to be consistent with LWR theory the vehicle trajectories in any of the regions (e.g., B) must be parallel to the segment connecting the corresponding state on the q–k curve and the origin. The reader should verify that this is approximately so in Fig. 4.12.[16] Similarly, as a consequence of (4.42), the interfaces between any two states (e.g., A and B) should be parallel to corresponding segments on the (k, q) plane (e.g., AB). The reader should also verify that this is true for all five interface segments of our example. If all the (t, x) lines are drawn with the proper slope, then the vehicular densities and flows appearing in the (t, x) plane after filling it with continuous equidistant trajectories in every region

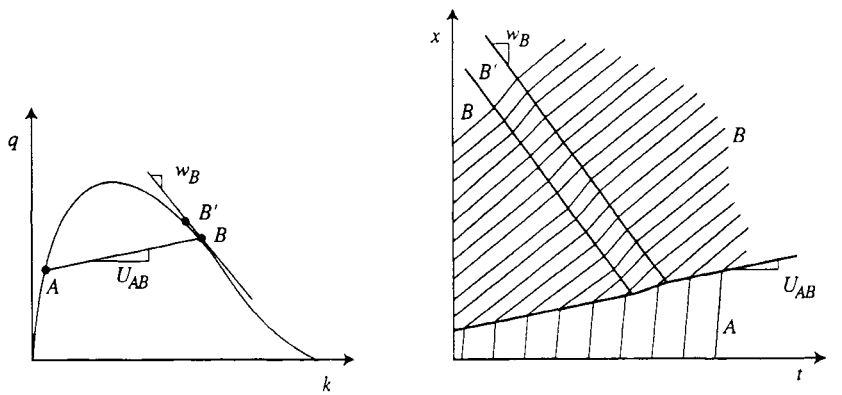


Figure 4.12 A 'wave' propagating through a set of vehicle trajectories.

will automatically be those corresponding to the velocities used for those regions (i.e., those of states A, B and B′ for our example), except for a multiplicative constant. This should be clear after reconsideration of Fig. 4.11 and its discussion.

The solution displayed on Fig. 4.12 could have arisen by specifying that our semi-infinite highway is in state B at time $t = 0$, $k(0, x) = k_B$, except for a little disturbance of slightly higher density somewhere downstream, with $k = k_{B'}$, and a short segment of much lower density near the entrance, with $k = k_A$; and also specifying that the entering flow for $t > 0$ is that of state A. It should be clear from the geometrical construction we have described that: (i) the solution displayed in Fig. 4.12 is the only one that satisfies these specifications and (ii) the solution could be extended to x and $t \to \infty$. These two properties (uniqueness and completeness) are only satisfied if one formulates a problem that makes physical sense. We then say that the problem is *well-posed.*[17]

As an example of an 'ill-posed problem', consider again the same initial data for the semi-infinite highway of Fig. 4.12 but pretend now that state A is located elsewhere on the q–k curve. If point A is moved up slightly so that $q_A > q_B$ (and $k_A < k_B$) the problem becomes ill-posed because the interface between A and B issued at $t = 0$ will slant down and intersect the abscissa at some time $t^* > 0$. Physically this would mean that a queue of slow-moving vehicles had backed-up to the highway entrance so that from that time on we could no longer specify an entering flow; thus the problem is 'ill-posed' for $t > t^*$.

Although rigorous proofs are difficult to provide, it turns out as a rule that the initial value problem is well-posed. Likewise, the semi-infinite highway problem is well-posed if no interfaces reach back to the highway entrance. The conditions that make 'engineering sense' for the 3-detector problem will be introduced later.

In Fig. 4.12 we used an initial state for $t = 0$ that included a small 'disturbance' in density (B′) on a uniform background (B). This disturbance propagates into the time-space solution as a band moving with the slope of segment $\overline{BB'}$ of the (k, q) plane. If we imagine that B′ is very close to B we see that the velocity of propagation of an infinitesimal disturbance must be the slope of the tangent to the q–k curve at B, w_B. We also see from the figure that the (infinitesimal) band of slope w_B contains traffic state B′, with speed ($v_{B'}$), flow ($q_{B'}$) and density ($k_{B'}$). If another small disturbance B″ (very close to B) is introduced somewhere else in the highway, we also see that that disturbance will travel parallel with the first and will therefore not affect the traffic state in the original band. Furthermore, if the value of B′ on the boundary is changed just a

little to B''' (also close to B) the state inside the band would then change to B''' while the state outside would remain unchanged. Thus, we can view the band as a *signal* or *wave* that tells the traffic stream reached by the wave to adopt the state prevailing at the source of the signal; B' or B'''. A simple mathematical proof of this statement is given below. It establishes that the solution to the LWR problem on a homogeneous highway is determined by straight lines issued from the boundary on which q, k and v are constant and equal to the values at the boundary. (For non-homogeneous highways this statement needs to be modified). These lines are called *characteristics* in mathematics books. In the fluids literature they are called 'waves' or 'signals.' The argument is as follows.

Where the solution k(t, x) is smooth (4.28) applies. Then, by virtue of the LWR hypothesis, we can replace q(t, x) by Q(k(t, x)) in (4.28) to obtain the LWR form of the conservation condition:

$$k_t + Q_k k_x = 0 \tag{4.43}$$

where we have used subscripts to denote partial derivatives. This condition must be satisfied everywhere in the (t, x) plane where the solution is smooth[18]. We now note that the left side of (4.43) is the rate at which k(t, x) varies with time along the line with slope $w_b = Q_k(k(t_b, x_b))$ passing through boundary point (t_b, x_b); i.e., (4.43) is the total derivative, $dk/dt = k_t + k_x[dx/dt]$, for $dx/dt = w_b$. Therefore, the LWR conservation condition implies that $dk/dt = 0$ along such a line; i.e., that the density is constant. This, of course, is just a more formal way of restating the graphical argument given earlier.

This result is very useful because if a problem is well-posed and Q(k) is concave, then it turns out that the characteristics emanating from the boundary cover the space of the solution, and that every point in the region of interest is reached by at least one characteristic that determines q, k and v at the point. It should be thus possible to construct a solution by drawing a characteristic from every point on the boundary and assigning the appropriate (k, q, v) to all the points on each line. Because there are some technical problems that need to be solved with this approach, it is instructive to see how the characteristics can be used to obtain vehicle trajectories, and from these the desired solution $\tilde{N}(t, x)$. Some readers may prefer to proceed directly to Sec. 4.4.4 now and return to the skipped material in a second reading.

The LWR trajectories and the 'entropy' condition. The thin lines of Fig. 4.13a depict a family of characteristics (emanating from the t = 0 line) corresponding to a hypothetical infinite highway problem. To be some-

CHAPTER FIVE

Control

In contrast to Chapter 4, which answered 'why' and 'what-if' questions, the current chapter is more 'engineering'. By means of some simple examples it illustrates how engineers and other practitioners exercise control over unscheduled transportation systems. Scheduled systems are discussed in Chapter 7.

Although it is tempting to formulate control problems in terms of mathematical optimization, we will see in this chapter that this is only seriously possible for small 'toy' problems that rarely arise in practice. In our discussion, which starts with simple problems and gradually complicates them, we will see that once a certain level of realism is reached the formulation of the problem must include elements that usually: (i) are very difficult to quantify, (ii) may not be perfectly understood and (iii) may complicate the numerical solution.

In view of this, this chapter will only give a peek into the types of control problems that engineers face in practice. The hope is that armed with this basic knowledge, you will be able better to pass an educated judgment on the practical control tools that you may one day learn, and will thus have some sense on what they can and cannot do.

For unscheduled transportation systems one does not have the option of choosing the vehicle routes, and control over the system can only be exercised by changing the rates at which vehicles are allowed to flow over specific points of the networks (perhaps in a time-varying fashion) as with a traffic signal. In doing this we assume that we can discriminate across certain vehicle types, although not all types. For example, we may treat buses and car-pools differently, but cannot expect to apply different controls to identical cars that go to different destinations.

Flow control can also be applied to scheduled transportation modes and an example involving the flow of containers through a port will be used just to show that a way of thinking that is natural for unscheduled transportation modes is also useful in other contexts. This should illustrate that even if you have an exclusive interest in a particular transportation mode, you will be well served by expanding your horizons since analysis methods developed for one purpose are often transferrable with little modification. And furthermore, the best solution ideas to resolve particular problems are often found, not in the litera-

ture of the transportation mode in question, but in the literature of modes for which the particular problem is most pressing and difficult; e.g. crowdedness reduction under highway traffic, sorting problems under railroads, multi-stop vehicle routing under trucking, efficient vehicle loading under water transportation, and personnel scheduling under air transportation.

The chapter starts with a discussion of control schemes for systems involving only two traffic streams, and then uses this information to introduce some important issues that arise with more complicated systems. The presentation deemphasizes computer heuristic approaches to control, and stresses instead a few basic qualitative ideas that should be more 'fundamental' and lasting.

Section 5.1, below, introduces the concept of intersection capacity and explains how the service between two competing streams should be alternated when conditions vary with time in a predictable way. The focus is on adjusting the settings so as to maximize the combined service rate of the intersection. Sections 5.2 and 5.3 examine in more detail the case of an undersaturated isolated intersection under stationary conditions, including unpredictable (random) fluctuations. The focus there is on obtaining settings that minimize 'delay', and on contrasting the performance of pretimed control (Sec. 5.2) vs. actuated control (Sec. 5.3). Sections 5.4 and 5.5 then deal with systems of intersections: section 5.4 with systems, such as an arterial street, where only one route exists between each origin and destination, and section 5.5 with more general networks.

5.1 Two interacting traffic streams

Section 5.1.1, below, introduces the notions of intersection capacity, 'saturation' flow, oversaturation and undersaturation. It also describes why it is often advantageous to alternate the right of way between two traffic streams instead of giving it permanently to one of them. Sections 5.1.2 and 5.1.3 then show how the allocation should be varied over time in the simple case where the cumulative virtual arrival curves at both approaches are known (deterministic) and slow-varying: Sec. 5.1.2 discusses the case of light traffic and 5.1.3 heavy traffic.

5.1.1 Intersection 'capacity': saturation limit

Suppose that we are analyzing the behavior of traffic at an unsignalized intersection of two one-way streets in which one of the streams has to

yield the right of way to the other. We ask: given the amount and character of traffic on the main (high priority) road, what is the maximum possible flow (the capacity) of the minor road over an extended period of time? The answer to this question, of course, depends on the behavior of drivers at the stop (or yield) sign. In the simplest possible theory one would assume that all drivers are identical and that they enter the intersection if, at decision time, the next main vehicle is not due for at least H_o time units. In the jargon of the field we would say that a driver accepts a gap (i.e. a headway or a portion thereof) if the gap is greater than the driver's *critical gap*, H_o. To evaluate the capacity of the secondary approach we must imagine that a driver is ready to move into position as soon as the prior one enters the intersection, as if there was an infinite queue. In this simplest theory the time interval in between successive vehicles, or *move-up time*, M, would also be assumed to be equal for all. Values of $H_o \approx 7$ secs and $M \approx 2$ secs for a yield sign or $M \approx 4$ secs for a stop sign are comparable with those one would measure in the field.

Suppose now that we have a collection of headways, h_i, e.g.:

$$\{h_i\} = \{6,7,10,2,2,4,9,3,4,8,2,\ldots\}, (secs)$$

and that we wish to evaluate the number of minor street vehicles that could have used them. The solution is simple if one arranges the data as shown in the first two columns of Table 5.1 and fills additional columns for T_i (the elapsed time from the beginning of the experiment to the end of the *ith* main street headway), n_i (the number of minor street vehicles that fit in each headway) and N_i (the accumulated number of fitted vehicles immediately after the *ith* headway). The columns have been filled for our problem with $H_o = 7$ secs and $M = 2$ secs as could be the case for a yield sign. It was assumed that a headway satisfying $h_i - H_o \geq 0$ will include at least 1 vehicle,[1] a headway satisfying $h_i - H_o - M \geq 0$ will include at least 2, and a headway satisfying $h_i - H_o + M - n_i M \geq 0$ will include at least n_i. This means that n_i is the largest integer smaller than $(h_i - H_o + M)/M$, provided that it is positive. Thus, n_i may be expressed as follows: $n_i = \max\{0, [(h_i - H_o + M)/M]^-\}$, where the symbol $[\ ]^-$ denotes the largest integer smaller than the quantity in brackets. Armed with information, one can use a spreadsheet to automate the calculations of a table such as 5.1, and in this way treat large data sets–the reader is encouraged to do so. It is then easy to produce a plot of T_i (on the abscissa) vs. i and N_i, which gives the cumulative number of vehicles observed to enter the intersection from both approaches.

People have developed formulae for the ratio N_i/T_i for $i \to \infty$ as a function of the average flow on the main street if the sequence $\{h_i\}$

Table 5.1: Sequence of headways on a major street and the throughput it allows on a secondary street controlled by a yield sign.

i	h_i	T_i	n_i	N_i
1	6	6	—	—
2	7	13	1	1
3	10	23	2	3
4	2	25	—	3
5	2	27	—	3
6	4	31	—	3
7	9	45	2	5
8	3	48	—	5
9	4	52	—	5
10	8	60	1	6
11	2	62	—	6
12	2	64	—	6

varies in certain random ways. Such formulae allow one to make predictions even before actual data become available. Of course, if actual data exist, and the problem is important enough to deserve the extended observation effort, then the procedure of Table 5.1 may be preferable because it does not assume any particular form of variation in $\{h_i\}$.

In our case, even though only 10 cars arrive on the main road during one minute, they are enough to restrict the outflow of the secondary approach to 6 vehicles during the first minute.[2] This is not atypical of unsignalized intersections since we know from experience that it does not take much main street flow to choke almost completely a minor street. It should also be intuitive that the degree to which this happens depends on the level of bunching in the traffic, and that the lesser the bunching the lesser the secondary flow. For example, if the 10 cars arriving in the first minute of our example had instead arrived regularly with 6 sec headways, then no secondary vehicles would have been served at all.

What happens is that in the sequence of cars observed over time, every switch from a low-priority vehicle to a high-priority one requires a traffic-less interval of H_0 secs, whereas a low-priority to low-priority combination only requires an interval of M secs. A distribution of high-priority traffic that allows large numbers of low-priority vehicles to cross in succession is clearly advantageous. This is the reason why an

on/off control on the main road can be effective. It artificially creates gaps in the main stream to achieve the desired effect.

If in the above example we had bunched the first 10 mainstream vehicles at 2 second intervals at the end of our minute (from t = 42 to t = 60) they would have created a 42 second gap, which 18 vehicles could have used if M = 2 seconds.[3]

It is fair to ask then: if we bunch traffic in this manner, what is the maximum bunching possible? If the main street has only one lane so that the minimum headway on it approximately equals M (M $\approx$ 2 secs), then the total number of vehicles entering the intersection from both approaches in time T will be equal to the effective portion of T (not wasted by switches from low to high priority) divided by M. In essence, if the average cycle between low/high switches lasts C time units there will be T/C switches of this type, and if each switch of this type wastes approximately (H_0 – M) time units, then the maximum total number of vehicles observed in time T will be:

$$(\text{max. } \#) = \frac{1}{M}\left[T - (H_0 - M)\frac{T}{C}\right].$$

Thus, the maximum flow into the intersection from both approaches combined is:

$$(\text{max } \textit{flow}) = \frac{1}{M}\left[1 - \frac{(H_0 - M)}{C}\right]. \tag{5.1}$$

The second term inside the bracket represents the reduction in maximum possible flow that results from the switches. Clearly, the larger the cycles the lesser the waste.

If gaps are introduced in a one-lane traffic stream by means of a control device, the above expression still applies, although the time lost per cycle is no longer related to the gap acceptance parameter H_0. Instead it depends on the duration of the 'all-red' phases during a full signal cycle (and on local driver's habits) and we call it the '*lost time*', L. In traffic lingo, the (maximum) flow corresponding to a headway of size M is called the *saturation flow*, $\mu = 1/M$, and with this new notation (5.1) becomes

$$\textit{capacity} = \mu\left[1 - \frac{L}{C}\right]. \tag{5.2}$$

Although this expression only applies to the case where the saturation flow is the same on both approaches, it is interesting because it shows that by making C large one can approach the maximum theoretical flow μ independently of the ratio in the flows of the two approaches. In contrast, if the intersection is unsignalized the combined maximum

flows will be significantly less than μ unless the flow ratio is very unbalanced. This should be intuitive from the behavior of Table 5.1 where only 18 vehicles (instead of 32) used the intersection in 64 secs.

It is possible to design an intersection, called a *roundabout* or *traffic circle* in various parts of the world,[4] whose capacity in the case of two intersecting one-way one-lane streets is still close to μ independently of the distribution of traffic on the approaches. These intersections are appealing because they increase capacity without stopping traffic to create gaps. Their main disadvantages are: (i) that they take up more physical space than signal controlled intersections (thus, they are only suited for low to medium population density areas where lack of space is not a problem) and (ii) that the ratio of entering flows cannot be exogenously controlled.

When the saturation flows on different approaches are not the same it is no longer meaningful to define an 'intersection capacity' since the maximum flow possible will depend on the flow ratios allowed in from the two (or more) approaches. But before we look at this case it is worthwhile to examine the lost time L in more detail.

The effective green. It has been known for a long time (e.g. Webster, 1958) that when a traffic signal turns green the headways at the stop line do not adopt the value M suddenly, but instead increase toward this value gradually with successive vehicles. This effectively creates an additional lost time due to starting.

Webster used a (t, x) diagram such as the one on the top part of Fig. 5.1 to argue that this additional lost time was given by $t_1 - t_0$; i.e., by the difference between the time that it takes for the first vehicle to accelerate to cruising speed, v_f, (from point P to point Q) and the time that it would need to cover the distance traveled (d) at speed v_f. This is logical because if vehicles accelerated instantaneously and the first one were to do it at point E instead of point P, then the resulting idealized vehicle trajectories would coincide with the original ones, after the latter had reached speed v_f. The vehicle departure times would also match the original times after the first few headways.

We saw in Chapter 1 that under uniform acceleration the time $(t_1 - t_0)$ is one half of the time that it takes to reach cruising speed. For cruising speeds of 30 MPH and typical accelerations this time should be on the order of 10 secs so that the lost time due to acceleration should be about $10/2 - 2/2 = 4$ secs – of course, more on uphill sections and less on downhills.

For the above conclusions to be exact we have to assume, as was done in the construction of Fig. 5.1, that every vehicle trajectory is an exact

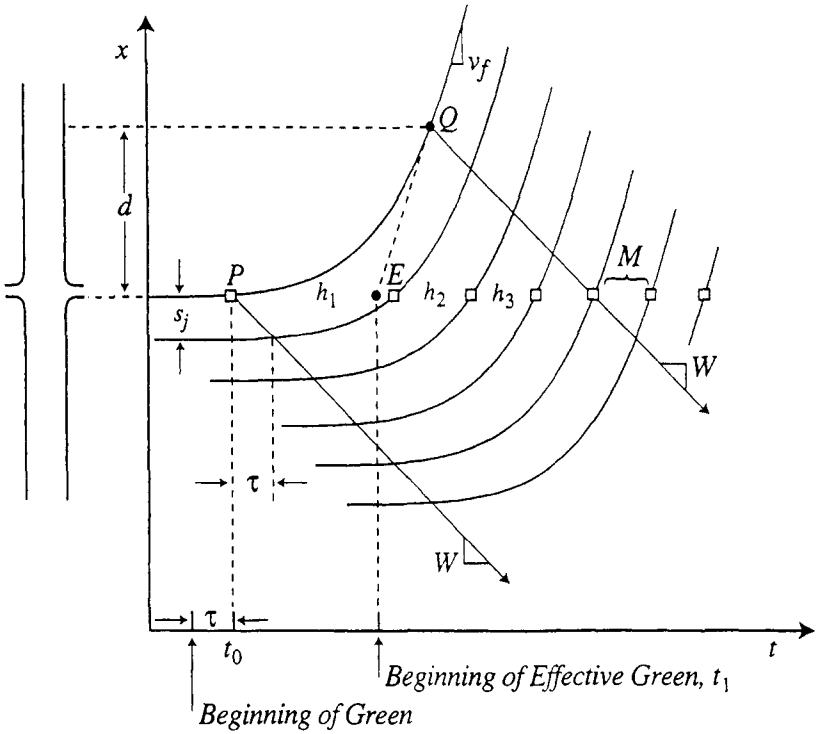


Figure 5.1 Hypothetical set of vehicle trajectories discharging from a traffic signal.

replica of the prior one, translated down by the jam spacing, s_j, and to the right by the reaction time, τ; i.e. that every vehicle does the same as its predecessor precisely τ time units later.[5] If this is accurate, then additional information can be obtained from the figure; e.g. expressions for the headways a given distance downstream of the stop line. Of particular interest is the sequence of headways at the stop line $\{h_i\}$, which declines toward $M = s_j/v_f + \tau$ and reaches this value after the wave passing through Q arrives at the stop line.[6]

The usefulness in what we have described, as opposed to a purely empirical approach based exclusively on observation of headways, is that the causal explanations given allow one to make predictions for situations that have not yet been observed, involving for example changes in grades, the location of the stop line,[7] etc....

Finally note that the number of vehicles that can discharge during the green phase can be related to the saturation flow ($\mu = 1/M$) by the

approximate expression, $G\mu$, where G denotes the length of the green time *starting from* $t = t_1$. This *effective green* phase is $(\tau + t_1 - t_0)$ time units shorter than the actual green, and will be the phase duration to which the symbols G_1, G_2 and G_i will refer in this chapter[8].

Saturation and undersaturation. Let us now consider a pretimed traffic signal with effective green phases G_1 and G_2. We saw earlier in Chapter 3, Eqs. (3.3), that any feasible stationary flows passing through the intersection must satisfy the inequalities:[9]

$$q_1 C \leq \mu_1 G_1;\ q_2 C \leq \mu_2 G_2;\ \text{with } G_1 + G_2 + L = C. \tag{5.3}$$

If this happens, we say that both signal phases, and hence the intersection too, are *undersaturated.*

If for a given set of flows $\{q_1, q_2\}$ we can find green phases that satisfy the above inequalities, we say that the combination $\{q_1, q_2\}$ is *below capacity.* This can be checked easily by eliminating G_1 and G_2 from the above 3 relations. Recognizing from the first two that $G_i/C \geq q_i/\mu_i$ $(i = 1, 2)$ and introducing these inequalities into the third expression we find:

$$y_1 + y_2 + L/C \leq 1 \tag{5.4}$$

where $y_i = q_i/\mu_i$ $(i = 1, 2)$. The dimensionless variables, y, represent the degree of saturation of each approach when the flow is not interrupted; as such, they represent the minimum fraction of time that their approach must have the green light. They will be called here the *minimum cycle shares* or *demand pressures*, although the term 'flow ratio' is often used in the field. Our derivation implies that if (5.4) is satisfied for some C then green phases of the required length can be found. Furthermore, we see from (5.4) that if the *overall demand pressure* satisfies: $y_1 + y_2 < 1$, then one will be able to find a long enough cycle to satisfy (5.4), regardless of L. Thus, this latter inequality (free of all timing variables) defines the theoretical capacity of an intersection controlled by a traffic signal. Notice that 'capacity' is no longer a number but *a condition* that is applied to a set of flows.

5.1.2 Timing plan for variable and deterministic traffic: light traffic case

In this subsection we extend these ideas to the time-dependent case. We shall see how the timing plan of an isolated traffic signal should be changed over time when the arrival rates q_1 and q_2 vary in a known way, in order to avoid oversaturation and reduce delay. As in the foregoing discussion and the rest of this Chapter we only consider a

simple intersection of two one-way streets with no turns. This suffices to illustrate many of the issues and solution approaches without complicating the subject beyond the scope of our book.

In the simplest possible case one might choose a *fixed timing plan* in which C, G_1 and G_2 are constant throughout the study period (e.g. a day). This may make sense if $q_1(t)$ and $q_2(t)$ are never so great that they violate (5.4); then, it *may* be possible to find a single timing pattern (G_1, G_2, C) that will remain undersaturated for all t; i.e. will satisfy (5.3) for all t. These conditions can be written as:

$$y_i(t) \le G_i/C, \quad (i = 1,2)\ \forall t$$

and

$$(G_1/C) + (G_2/C) + (L/C) = 1.$$

The above inequalities are equivalent to:

$$\hat{y}_i \le G_1/C, \quad (i = 1,2)$$

where $\hat{y}_i$ denotes the maximum value of $y_i(t)$ during the day. These inequalities can be substituted into the equality to yield:

$$\hat{y}_1 + \hat{y}_2 + (L/C) \le 1.$$

As before, the condition:

$$\hat{y}_1 + \hat{y}_2 < 1 \tag{5.5a}$$

ensures that a fixed timing plan exists, with cycle time:

$$C \ge L/(1 - \hat{y}_1 - \hat{y}_2).$$

Since the $\hat{y}_i$'s are the maximum cycle shares required by each approach during the day, the denominator of this expression will be called the *slack*. Note that the smaller the slack the longer the cycles necessary to overcome the lost time.

If $q_i(t)$ varies slowly with time so that it remains (nearly) constant within each cycle, then the average delay during any cycle can be approximated by the New Jersey/Clayton formula (see Chapter 3). For any undersaturated flow, the result of this formula for approach *i* is always greater than $\frac{1}{2}R_i^2/C$. Since $R_i > G_j$ for $i \neq j$, the average delay is in turn greater than $\frac{1}{2}G_j^2/C = \frac{1}{2}C(G_j/C)^2 \ge \frac{1}{2}C\hat{y}_j^2$. This last inequality follows from the condition $G_j/C \ge \hat{y}_j$ introduced earlier. Thus, the quantity $\frac{1}{2}C\hat{y}_j^2$ is a time-independent lower bound for the average vehicle delay on approach *i* throughout the study period. If the bound is high, then delays would be high even during periods of very low flow. This is clearly inefficient, since one could have chosen a shorter cycle and shorter signal phases during periods of little flow with an ensuing reduction in delay.

If the daily pattern of variation is predictable, it may be possible to divide the day into a few periods (long compared with the cycle time) with different maximum flows but little flow variation within each period. For example, we may choose to have one timing pattern each for nighttime, off-peak daytime, morning peak and afternoon peak. This would allow the average delay per car experienced in each period to depend on the available slack *for the period*, and thus to be significantly reduced.

In deciding the number of timing plans and the times at which they should come into effect one can look at a diagram of the two saturation levels vs. time, as shown in Fig. 5.2. The diagram shows at a glance how a selection of times (t_1, t_2, t_3...) influences the maximum demand pressures in each time interval, given by the two step curves in the figure. We have chosen to plot the two saturation levels in opposite directions along the ordinate axis because then the separation between steps is the overall pressure, which subtracted from 1 yields the slack.[10] The slack, in turn, can be used to choose C and the green phases for each period. The idea then, is to introduce t_i's where the steps could be made closer. In our case, four discrete steps allowed us to improve from the dashed to the solid lines.

Although it is possible to formulate the problem more rigorously as a delay minimization problem, we chose to present it in this graphical way better to highlight the issues. After all, total delay is only a coarse approximation of what society wants and in many cases the solution is relatively insensitive to the objective function used. This may occur for example if, as often is the case, the curves $q_1(t)$ and $q_2(t)$ increase or decrease significantly only during short, well defined periods of the day, and remain nearly constant at other times. Then, good locations for t_i may become readily apparent.

In closing this discussion about undersaturated intersections we note that variable timing plans may keep an intersection undersaturated even if condition (5.5a) is violated. This could occur for example if the vertical separation between the dashed lines of Fig. 5.1 had exceeded 1 but the maximum vertical separation between the step curves, which is what matters for a variable timing plan, had been less. In other words, one will be able to find a feasible variable timing plan if the maximum vertical separation between the solid data curves satisfies:

$$\max_t\{y_1(t) + y_2(t)\} \leq 1. \tag{5.5b}$$

This, of course, is a less restrictive condition than (5.5a). The improvement is most noticeable if flows are directional so that when $q_1 = q_1^{max}$, $q_2 \ll q_2^{max}$ and when $q_2 = q_2^{max}$, $q_1 \ll q_1^{max}$.

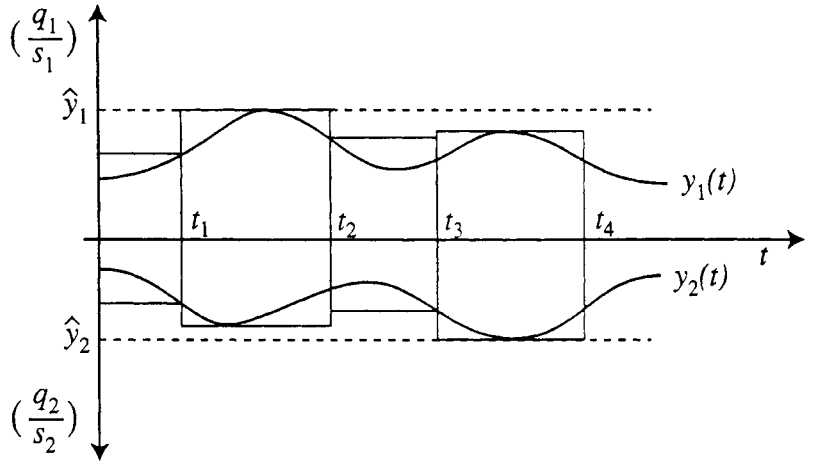


Figure 5.2 Time-dependent traffic patterns, and the selection of a timing plan.

5.1.3 Timing plans for variable and deterministic traffic: heavy traffic case

What if condition (5.5b) is not satisfied? Then there will be times where queues will overflow from one cycle to the next and the selection of a timing plan is more complicated[11]. There is a large body of literature dealing with this problem, with most of the effort aimed at developing delay formulae that incorporate the effects of stochastic fluctuations in the flows. Fortunately these complications can be ignored for a qualitative description of the issues since, as shown in Newell (1971, 1982) stochastic fluctuations only play a minor role during the brief times that the signal is transitioning between under and over-saturation; therefore, a rather good prediction can be obtained by ignoring the stochastic fluctuations altogether. When this is done it is easier to find and evaluate reasonable control strategies.

Although it is beyond the scope of this introductory book to present a detailed timing recipe for a general problem, as is done in Gazis and Potts (1965), we can say that during oversaturated periods one would want to operate with the maximum practical cycle (so as to reduce as much as possible the portion of idle time), favoring the approach with greatest saturation flow.[12] If that approach was undersaturated, one could set the signals to maintain it on the verge of oversaturation and could then allocate the rest of the green to the other approach. This would have the effect of maximizing the sum of the flows that pass

through the intersection during the period of oversaturation thereby reducing the overall delay.

The reader should recognize that the total number of vehicles present at the intersection on both approaches can be studied with the input-output method of Chap. 2, treating both approaches as a single arrival stream with a combined flow, $q_1 + q_2$, and a service rate equal to the combined flow through the intersection. This combined rate is: $\mu = q_1 + \mu_2 [1 - L/C_{max} - q_1/\mu_1]$, if approach (1) is the one with the largest saturation flow, i.e., if $\mu_1 > \mu_2$. The suggested strategy should achieve the desired effect since the combined delay during oversaturation decreases with μ, which has been made as large as possible.

Furthermore, by favoring the efficient approach (1) we are also encouraging users of the less efficient approach (2) to change routes, which may be a sensible thing to do if the alternative routes are reasonable options. Although one may be tempted to temper this strategy in order to be 'fair' to both approaches, this should only be done after carefully weighing the system-wide consequences. We will see later in this chapter that 'fair' signal setting strategies can be unstable and make things worse.

Let us now turn our attention to the effects of unpredictable fluctuations in traffic flow that cannot be anticipated in a timing plan. Stochastic fluctuations have little effect on the average delay over many days if a signal phase is well under the saturation level, or if it is well oversaturated;[13] Random effects only really matter when (undersaturated) traffic is so close to saturation that a fluctuation has a reasonable chance of creating an overflow from one cycle to the next. Such a situation is depicted in Fig. 5.3. Part (a) of the figure shows the average behavior of the signal in terms of an average (virtual) arrival and departure curve and part (b) depicts a fluctuation with the same number of total arrivals as in part (a). We can see clearly that the first two cycles of part (b) are quite undersaturated, and that the rush of arrivals during the third cycle causes an overflow on that cycle and the next. Note that some vehicles remain unserved, even though the same number arrived, and that the total number of vehicle-hours spent in part (b) is greater than in part (a). [The reader can verify graphically that this is also true if the surge of arrivals occurs earlier in the time interval shown.] Clearly, oversaturated cycles have an effect that propagates from one cycle to the next and their occurrence increases delay.

The following section introduces a simple formula that quantifies this delay.We have already stated that this (isolated intersection) problem is

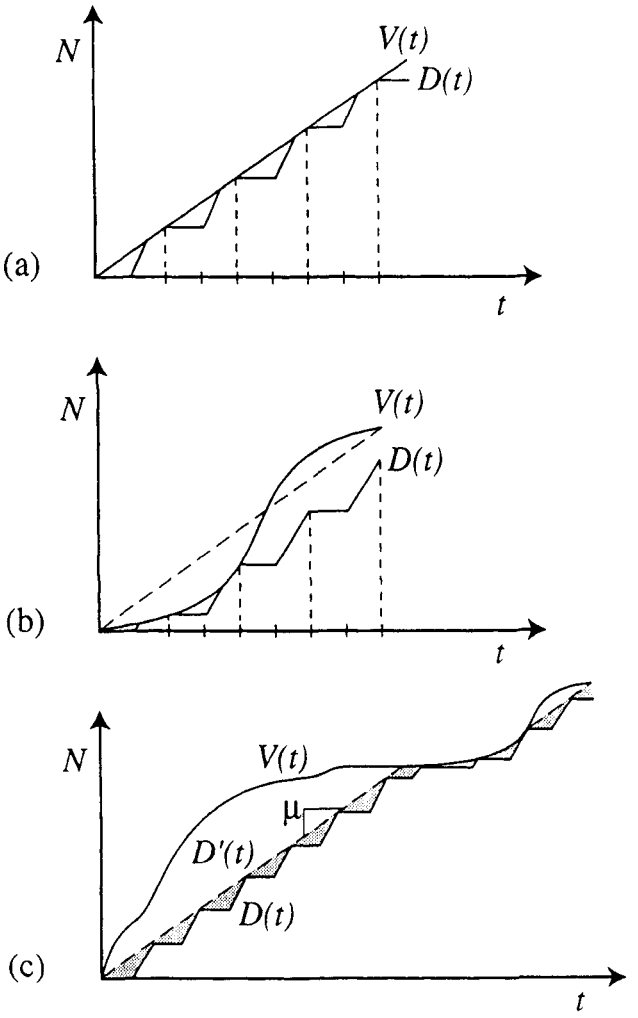


Figure 5.3 Input-output diagrams for a pre-timed traffic signal: (a) undersaturated, stationary arrivals; (b) a fluctuation causing a spillover; (c) a larger fluctuation.

not particularly common or important, but a description of its solution has pedagogical value in that it illustrates how stochastic delays come about physically. Section 5.3 will then show how the delay due to the fluctuations can be reduced with traffic-responsive control schemes.

5.2 The isolated traffic signal with stationary traffic and fluctuations

We consider initially one leg of a pretimed intersection and explain the logic behind a formula that predicts the average delay per car when the arrivals are random but stationary; i.e., when a realization of the arrival process yields no clue as to the time of day. Starting with the elegant early work of Winsten, who contributed Chapter 1 in Beckmann *et al.* (1956), this is a problem that has received a great deal of attention in the queueing literature. Although at this stage the reader may not yet have the probability background needed to understand all the developments that eventually led to its exact solution, she should be able to appreciate the qualitative arguments that led to a well-known approximate formula (Webster, 1958) that has come to be known as 'Webster's formula'. The arguments presented here will be more general than those in the original work since we shall not specify the properties of the arrival process in so much detail.

We have already said that if a signal phase is substantially undersaturated, then stochastic fluctuations matter little and one can use the New Jersey formula, Eq. (2.9), for delay. To do this properly one should use the average arrival rate across many days (without fluctuations) as an input, which we denote q. (The subscript 'i' will be omitted from the notation as long as our discussion refers to a generic phase.) The average delay across all cars can then be written as follows:

$$w_{det} = \mu R^2/2(\mu - q)C = [R^2/2C](1-y)^{-1}, \qquad (5.6)$$

where y is the average demand pressure q/μ and the subscript 'det' stresses that (5.6) is a deterministic approximation. Recall that this formula was based on a linear virtual arrival curve as in Fig. 5.3a.

To measure the saturation level of a given phase it is convenient to introduce the dimensionless parameter $\rho = \bar{q}C/sG = yC/G$, which we call the *degree of saturation*. Note from (5.3) that if $\rho = 1$ then the phase is perfectly saturated. Equation (5.6) holds if $\rho \ll 1$. As $\rho \to 1$, however, the signal will receive on average almost as many vehicles as it can handle. This means that small fluctuations will have more of a chance to cause spillovers, and that the effects of any such fluctuation will be longer lived since the ability of the signal to dissipate a queue is quite limited for ρ close to 1. (Dissipation happens at a rate of about $(1-\rho)\mu G$ vehicles per cycle on average.) In this case, Eq. (5.6) will underpredict delay significantly.

Fig. 5.3c depicts on a larger scale a possible realization of a process

with ρ close to 1 for many cycles, where it can be seen that the majority of the delay can be attributed to fluctuations and the effects of spillovers. This component of delay is the unshaded area between curves V(t) and D'(t). It can be visualized as the delay that would have arisen if the traffic signal, instead of operating in cyclical pulses, had operated at an even rate $\overline{\mu} = sG/C$.

We saw in Chapter 2, Eq.(2.13), that the average per car of such a delay, for random arrival processes observed for a long time, was $\frac{1}{2}(\Delta/q)/(1 - q/\overline{\mu})$, where Δ was a measure of the variability in the arrival and service processes. Since $q/\overline{\mu}$ is the degree of saturation, this expression can be rewritten as follows:

$$w_s \approx \tfrac{1}{2}\Delta/[(1-\rho)q] \quad \text{for } \rho \to 1. \tag{5.7}$$

The subscript 's' is used with w to emphasize that this is the delay due to the *stochastic* fluctuations when the service rate is *smooth*. We will see in Sec. 6.1.3 that if the service rate has no variability the parameter Δ can be written in the form $\gamma\rho$, where γ is the 'index of dispersion' of the arrival process; i.e., a parameter with units of 'vehicles' that measures the size of the wiggles in the arrival curve. It is well known that $\gamma = 1$ vehicles for the Poisson process.

The shaded area is the delay due to the pulses. For $\rho \to 1$ the process of Fig. 5.3c spends most of the time in an overflow regime and (5.6) is a good approximation for the shaded area; thus, we can express the average delay with fluctuations and pulses as the sum of Eqs.(5.6) and (5.7); i.e.,

$$\overline{w} \approx w_{det} + w_s = [R^2/2C](1-y)^{-1} + \tfrac{1}{2}\Delta/(1-\rho)q, \quad \text{for } \rho \to 1.$$

This formula is not accurate for $\rho = 0$, however, because the second term does not vanish for $\rho \to 0$, which it should.[14] An accurate formula for $\rho \to 0$ and $\rho \to 1$ can be obtained by multiplying the second term by any smooth function of ρ that, both, approaches 0 faster than q as $\rho \to 0$, and approaches 1 as $\rho \to 1$. For example, using the factor ρ we obtain:

$$\overline{w} \approx [R^2/2C](1-y)^{-1} + \tfrac{1}{2}\rho\Delta/(1-\rho)q. \tag{5.8}$$

Inasmuch as (5.8) is a monotonic function of ρ (the form of the true dependence between $\overline{w}$ and ρ) that gives reasonable results for $\rho \to 0$ and $\rho \to 1$ one would expect it to give reasonable results for intermediate ρ's. The interpolation factor used (ρ), however, could have depended on the remaining variables of the problem and one should not be surprised to find moderate errors for intermediate ρ's.

In fact, Webster found through simulation experiments (using Pois-

son arrivals and regular services) that an overprediction on the order of 10% occurred when one puts $\Delta = \rho$ in (5.8). By statistical analyses of the simulation output he proposed the following additive correction term $\langle w \rangle$:

$$\langle w \rangle = -0.65\left(\frac{C}{q^2}\right)^{1/3} \rho^{(2+5G/C)}. \tag{5.9}$$

Although an approximate correction to (5.8) for more general arrival processes (with arbitrary γ) has already been obtained,[15] it is educational to see how (5.8) can be extended to this case with very little effort by means of dimensional analysis. The following paragraph, which contains this explanation, can be skipped without loss of continuity.

Because the evolution of the system can be described approximately by a set of equations that are invariant to changes in the units used for measuring vehicular quantity,[16] γ can be used as the unit of choice without changing the numerical prediction for the average waiting time. This modification is equivalent to keeping the original units and replacing the arrival and service processes with ones described by appropriately rescaled parameters; i.e.:

$$q' = q/\gamma, \quad \gamma' = \gamma/\gamma = 1 \quad \text{and} \quad \mu' = \mu/\gamma. \tag{5.10}$$

Since the scaling transformation preserves the character of the arrival process as a process with 'independent increments' (see Chapter 6) and also changes its index of dispersion to 1, it makes it behave like a Poisson process.[17] Thus, Webster's formula applies to the new problem. Its prediction, of course, is the average waiting time for the new as well as the old problem. It is given by (5.8) and (5.9), after substituting q', γ' and μ' for q, γ and μ. The result in the original units is then obtained after replacing q', γ' and μ' by the right sides of (5.10). The reader can verify that these manipulations leave (5.8) unchanged but transform (5.9) into:

$$\langle w \rangle = -0.65C\left(\frac{\gamma}{qC}\right)^{2/3} \rho^{(2+5G/C)} \tag{5.11}$$

which is the sought generalization. Equation (5.11) should be as good an approximation to the general case as (5.9) is to the continuum model with $\gamma = 1$. Note for example that it tends to 0 as $\gamma \to 0$, which is reasonable since all stochastic effects should disappear with smaller fluctuations. Dimensional analysis, however, has told us more; e.g. the *rate* at which (5.11) goes to zero when $\gamma \to 0$.

5.2.1 Warnings and comparisons: the relaxation time

First of all we should emphasize that the delay calculated with any of the prior expressions is not the time in queue but the *extra time caused by the existence of the intersection*, and that this delay is always smaller than the time in queue. This delay interpretation is correct because the V(t) curves in our analysis (Fig. 5.3) represented 'virtual' or desired arrivals. To obtain the time in queue, one would have to use a traffic flow model, such as the simple two-wave speed model of Chapter 4. Of course, the delay calculated in that manner must coincide to the level of accuracy in Webster's approximation with the previous formulae. The remainder of this subsection discusses the applicability of the delay expressions. Because some elementary probabilistic reasoning could not be avoided, some readers may prefer to skip this material and return to it after reading Sec. 6.1.

Not given here, the derivation of Eqs. (5.8) and (5.11) assumes that our queue with stationary arrivals has been observed for a time that is long compared with the system's *relaxation time*, T^*. Briefly introduced in Chapter 2, the relaxation time is the time that a stochastic system needs to 'forget' its present state, in the sense that if one samples the system regularly every T^* time units then past observations cannot be used to predict future ones.

If traffic is so light that there are no oversaturated cycles, then the events in successive cycles (e.g. delays) will be unrelated and we can assume that $T^* = C$. At the other extreme, when the system behaves as in Fig. 5.3c with $\rho \rightarrow 1$, we see that most of the delay is due to the fluctuations and that the pulses do not matter significantly. Therefore, Eq. (2.14) of Chapter 2 applies, and since $\overline{\mu} \approx q$, we can write: $T^* \approx \Delta/[(1-\rho)^2 q]$. This expression can be justified informally without writing many equations from our knowledge, Eq. (2.12), that the average queue length is approximately $\frac{1}{2}\Delta(1-\rho)^{-1}$. Simply note that at the (rare) times when the queue is several times longer than the average, the deviation will be proportional to $\Delta(1-\rho)^{-1}$. Since the average time needed to reduce the queue length by one unit is $[\mu(1-\rho)]^{-1}$, the time to return to the average from an extraordinary position must be proportional to $T^* \approx \Delta/[(1-\rho)^2\mu] \approx \Delta/[(1-\rho)^2 q]$. Thus, the effect of an extraordinary event can only be felt for a time comparable with T^*. This, of course, is the significance of the relaxation time.

A simple argument cannot be given for intermediate values of ρ, but it should be intuitive that a period of observation that is long compared with the largest of C and $\Delta/[(1-\rho)^2 q]$ should do the job in this case.

It is interesting to evaluate the relaxation time when $R \sim \frac{1}{2}$ C and ρ

is so close to 1 that the stochastic delay (5.7) is comparable to the deterministic component (5.6). Under these conditions T^* can be expressed as $T^* \sim C(\mu C/16)$,[18] i.e., as the product of a time constant equal to one cycle and a number comparable with the number of vehicles that can be served in one sixteenth of a cycle. For typical values of μ and C, T^* should last for a few cycles. Therefore Webster's formula applies if the average demand is constant for a time long compared with T^*, e.g., 20 or 30 cycles.

In order to solidify these concepts, it is useful to compare the analytical predictions with those of a spreadsheet simulation that can be applied to observation periods of various lengths. An efficient way of doing this consists in tracking the (continuum) number of vehicles left at the end of the green, n, from cycle to cycle with a recursive formula that involves the number of arrivals, a, in the cycle. The delay can then be retrieved from the simulated sequence of a's and n's.

It is postulated that the following formula is reasonable:

$$n_{next} = \max\{0;\ n_{old} + a - \mu G\}. \tag{5.12}$$

The expression simply says that if more than μG vehicles want service in a cycle (the sum of 'a' and n_{old}) then the excess over μG must overflow; otherwise the overflow is zero.[19]

We also claim that the average delay experienced over all possible arrival patterns in a cycle with a given n_{old} and a given 'a' is approximately equal to the delay that would be obtained if the arrivals were evenly distributed throughout the cycle. This turns out to be:

$$\text{total delay in cycle} \approx (n_{old} + a/2)C - G^2\mu/2 + \tfrac{1}{2}(n_-)^2/(\mu - a/C) \tag{5.13}$$

where n_- is the extra number of cars that could have been served in the cycle, but were not; i.e. $n_- = \max\{0;\ \mu G - (n_{old} + a)\}$. Note that $n_- = 0$ if $n_{next} > 0$ and $n_{next} = 0$ if $n_- > 0$. These two cases are shown in Fig. 5.4 (a and b).

Equation (5.13) is based on a linear V(t) curve, and can be easily derived from the geometry of the cumulative count diagrams displayed in Fig. 5.4. When $n_- = 0$, as in Fig. 5.4a, (5.13) is the exact average. [The reason is that the area between the curves depends linearly on the times at which each of the successive arrivals occurs in the cycle, and the average of a linear function is a linear function of the averages.] The same cannot be said for part (b) but the non-linear effects are minor.

The function @RAND in most spreadsheets returns a different number between zero and 1 each time that the spreadsheet is recalcu-

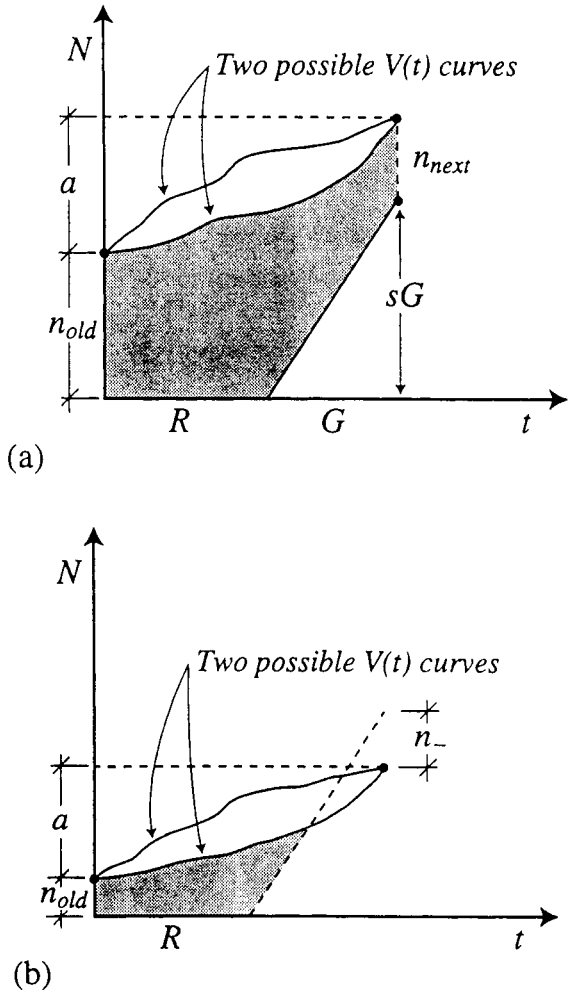


Figure 5.4 Effect of variable arrivals during a traffic cycle: (a) queue does not clear ($n_- = 0$, $n_{next} > 0$); (b) queue clears ($n_- = 0$, $n_{next} = 0$).

lated. And each time all the possible numbers (with a given number of significant digits) have an equal chance of being drawn. This allows us to simulate a random draw for the number of arrivals in each cycle. We shall use:

$$a = qC + \sqrt{3qC}\,(u_1 + u_2 + u_3 + u_4 - 2) \tag{5.14}$$

where u_i represents a random number generated with @RAND. The parenthetical expression in (5.14) would then read in the spreadsheet: @RAND + @RAND + @RAND + @RAND - 2. On average, each u_i is 0.5 and the parenthetical expression is zero; thus, the average 'a' generated in this manner is qC, as desired. The quantity in parenthesis will vary with each calculation of the spreadsheet and its fluctuations must obviously be of a magnitude comparable with 1. This implies that the fluctuations in 'a' are comparable with $(qC)^{1/2}$. An objective measure for their size is their 'root mean square'; i.e. the square root of the average of the squared deviations from the mean, also called the '*standard deviation*' and denoted σ_a. Experimentation with the spreadsheet (or equivalent statistical calculations) would show that the mean squared fluctuation of (5.14), also called the *variance* is: $\sigma_a^2 = qC$. This result is satisfactory because it means that the simulated arrivals behave like a Poisson process; see Chapter 6.

It is now easy to create a three-column spreadsheet that recursively calculates Eqs. (5.14), (5.12) and (5.13). (Our version, called WEBSTER.WK1, has been made available to the public.) With one column allocated to each equation, the spreadsheet should also include, a block of six cells containing the five parameters of our problem and an initial queue n_{orig}. In the columns corresponding to (5.12) and (5.13) the value of n_{old} is taken from the previous row, except for the first row which must use n_{orig}. In this way, each row simulates one traffic cycle. The sum of a range of cells for the column corresponding to (5.13) divided by the sum of the equivalent range corresponding to (5.14) yields $\bar{w}$. This can be compared with the theoretical predictions, which can be included in the spreadsheet as well.

It is instructive to see how the results depend on n_{orig}, and the various parameters of the problem for a fixed number of cycles, N_C. We can verify numerically that if one chooses n_{orig} to be smaller than the maximum queues that are likely to develop ($n_{orig} < 10q\bar{w}$), the results are insensitive to n_{orig} as long as $CN_C \gg T^*$. For $\rho \approx 0.5$ and smaller we also find that the deterministic formula (5.6) is a good approximation and that the simulated results do not vary greatly from day to day (e.g. from one simulation of $N_C \sim 200$ cycles to the next). As ρ tends to 1, however, the stochastic fluctuations increase and for sufficiently large values, when T^* is no longer small relative to CN_C, the results eventually become sensitive to n_{orig}. These effects, of course, are what one would expect in view of the theoretical discussion.

5.2.2 Pretimed control

We are now ready to discuss the control strategy for a signal that

regulates (stationary) traffic at an isolated intersection of two one-way streets. If minimization of total delay can be justified as the objective one could simply replace the deterministic objective function of Chapter 3, Eq. (3.2), by a weighted average of (5.8), perhaps including the correction term (5.11), and then could repeat the optimization process. A similar objective function can be defined for more complicated intersections with two-way traffic and turning movements, provided that each movement is only handled during one of the green phases in the cycle. In such an event the left side of (5.8) would still apply to each movement, and the overall average delay would continue to be a weighed average of (5.8). Here we restrict our attention to the simplest case with two conflicting movements, since this scenario suffices to illustrate the issues[20].

An approximate result of the optimization is:

$$G_i = (C - L)y_i/(y_1 + y_2) \quad (i = 1,2) \tag{5.15a}$$

and

$$C^* = (1.5L + l_0)/[1 - (y_1 + y_2)] \tag{5.15b}$$

where $l_0 = 5$ secs. The reader can verify algebraically that the choice (5.15a) satisfies the constraints $y_i \leq G_i/C$ in (5.3) if C is greater than the minimum value allowed by (5.4):

$$C_{min} = \frac{L}{1 - (y_1 + y_2)}. \tag{5.16}$$

Note that the recommended cycle exceeds C_{min} by about 100% for typical values of L. Delay is not minimized for $C \sim C_{min}$ because then all approaches would be close to saturation and there would be frequent overflows; the stochastic component of (5.8) would be large. The overflows can be eliminated with $C \gg C_{min}$, but this is not recommended either because long phases mean long delays for the vehicles affected; the deterministic component of (5.8) would then be large. If there are two phases but more movements, e.g., because one or both streets are two-way, the same formulas can be used although the results should be less accurate; Webster recommended using the movement with the largest demand pressure for approach i to define y_i. As an exercise, the reader may want to find a timing plan for an intersection without turns, L = 10 secs. and the following N, S, E, W average flows and saturation flows: $\{q_i\} = (600, 500, 1200, 200)$ and $\{\mu_i\} = (1800, 1800, 3600, 1800)$. (Answer: C = 60 secs; $G_1 = G_2 = 25$ secs.)

Although more elaborate approaches can and have been developed to time an isolated intersection, we stop our discussion here. We have already said that delay is a somewhat arbitrary measure of performance

for signal timing since it only measures the impact of traffic on those who are part of it; and it does so imperfectly. [Recall the extreme policies to which delay minimization leads when the objective is applied to the oversaturated case.] Clearly, other criteria could be used. The U.S. Highway Capacity Manual, for example, uses a percentage of oversaturated cycles as one of the criteria for intersection design.

5.3 Actuated control

We have seen in Sec. 5.1.2 that changing a timing plan in response to predictable hourly changes in flow can reduce delay. We will show here that changing it in response to unpredictable fluctuations is also beneficial. As in the pretimed case, we base our discussion on an intersection of two one-way streets with no turns because this is the simplest case that can illustrate the issues. More complicated intersections are appropriate for a course on traffic engineering, in which all the possible configurations can receive the deserved attention.[21]

We recall that the phases and cycles of Webster's recipe are about twice as large as the minimum possible in order to reduce overflow delay, and that this choice roughly doubles the deterministic component of delay. Clearly then, if it were possible to operate with cycles and phases close to the minimum while avoiding overflows, traffic delay would be reduced by more than a factor of 2. We shall see that this goal can be achieved with actuated systems that end the green phase on each approach as soon as its queue dissipates.

Consider Fig. 5.5, which depicts two cumulative (virtual) arrival curves, $V_1(t)$ and $V_2(t)$, starting at an instant ($t = 0$) when the queue at approach 2 has vanished and the queue of approach 1 is $Q_1(0) > 0$. Although it is not necessary we have set the intercept of both arrival curves at $t = 0$ equal to the initial queues; thus, both departure curves start from the origin, as shown. If the saturation flows and the lost times for both approaches (μ_i, L_i) are known the departure curves can be easily constructed for any pair of $V_i(t)$'s. (Note that $L_1 + L_2$ is the combined lost time, L, used in Sec. 5.2.) The complete solution is displayed on Fig. 5.5, where arrowheads indicate the order in which points along the departure curves are obtained.

Starting at the origin '0' the signal turns and a lost time of duration L_1 begins. Both departure curves then remain horizontal until $t = L_1$, which locates point 'A' of the figure. At such time the signal turns green for approach 1 and remains red for 2. This phase is characterized by departure curves slopes μ_1 and 0; it terminates when queue 1 vanishes

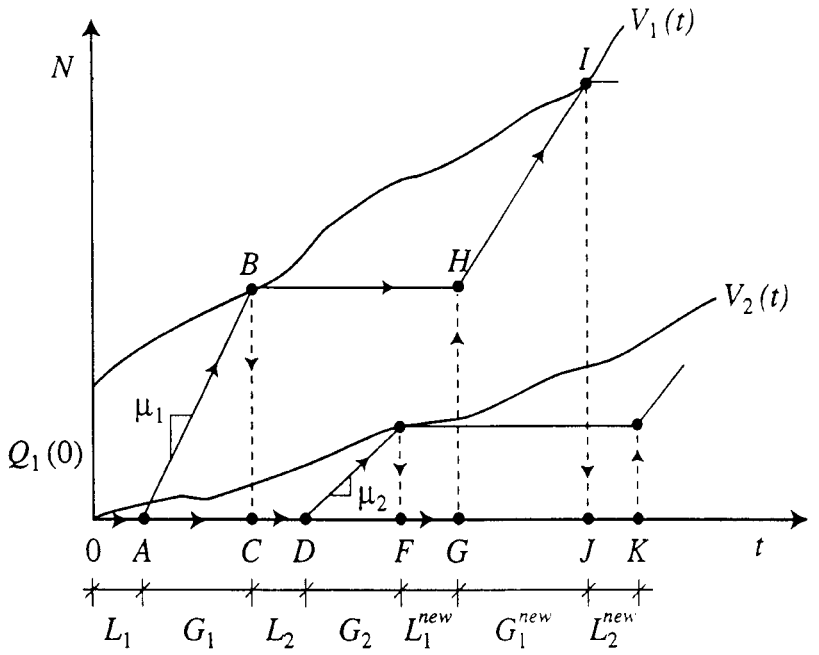


Figure 5.5 The actuated traffic signal. Strategy in which the green phases are terminated as soon as the queue vanishes.

at point 'B' of the figure. The duration of the phase, G_1, is the length of segment AC. The system is now in a position identical to that in which it was at $t = 0$, with the approaches reversed. Therefore, the construction method from now on is a continuation of what we have described with 1 and 2 interchanged. It goes on with a lost time L_2 which causes both departure slopes to equal zero; this identifies point D. Green phase 2 then begins, etc... The reader should try and trace the remaining steps through completion, following the arrows shown on the figure.

As long as the arrival curves rise moderately, every phase in the construction of Fig. 5.5 comes to a definite end and we say that the signal is undersaturated. We will now derive an expression for the average length of each phase and from it will obtain a condition for undersaturation. Note that our analysis is rather general for we are not placing any formal restrictions on the $V_i(t)$ curves.

Let N_C (a large number) be the number of cycles through which the system is observed and T the duration of the experiment. The average cycle therefore lasts T/N_C time units. On dividing the total amount of

time that the signal has been green for approach i, by N_C we obtain the average duration of a green phase for approach i, $\overline{G}_i$. Because each cycle has a total lost time $L = L_1 + L_2$, the total time lost in the N_C cycles is $N_C L$. This means that the duration of the experiment is $N_C(\overline{G}_1 + \overline{G}_2 + L)$. If we now let $\bar{q}_i$ denote the ratio of the number of arrivals on approach i during the experiment to its duration T, which is the conventional definition of average flow in an interval, the total number of arrivals may be expressed as:

$$\text{number of arrivals on } i = N_C\left(\overline{G}_1 + \overline{G}_2 + L\right)\bar{q}_i \quad (i = 1,2). \quad (5.17a)$$

Because all the green phases are saturated, the number of services on i can be expressed as:

$$\text{number of services on } i = N_C \overline{G}_i \mu_i \quad (i = 1,2). \quad (5.17b)$$

Let us now examine what happens to (5.17a) and (5.17b) for large N_C. We choose for the beginning and end of our experiment two instants where one of the queues has just dissipated, and this means that for that approach (5.17a) and (5.17b) are equal. For the other approach the two expressions will differ by an amount that equals the difference between its initial queue on the approach and the number of arrivals during its last red period (see Fig. 5.5). If this quantity is small compared with the left sides of (5.17) when N_C is large, as should happen if the signal is able to serve the demand during the study period,[22] then we can equate the right sides of (5.17a) and (5.17b) for this approach too. The result is a system of two equations that yields $\overline{G}_1$, and $\overline{G}_2$ in terms of $\bar{q}_i$, μ_i, and L. Let us see.

Since the ratio of (5.17a) and (5.17b) is 1 (and we defined $y_i = \bar{q}_i/\mu_i$) we can write:

$$y_i = \overline{G}_i/\overline{C} \quad (i = 1,2) \quad (5.18)$$

where $\overline{C}$ is the average cycle, $(\overline{G}_1 + \overline{G}_2 + L)$. This says that the proposed strategy allocates green time to both approaches in the same ratio as recipe (5.15a). If we now use (5.18) in the identity:

$$\frac{L}{\overline{C}} + \sum_i \frac{\overline{G}_i}{\overline{C}} = 1$$

we obtain the familiar condition for a pretimed signal on the verge of oversaturation (see Eq. (5.4)):

$$\frac{L}{\overline{C}} + \sum_i y_i = 1.$$

Finally, on solving for $\bar{C}$ we find:

$$\bar{C} = L\left(1 - \sum_i y_i\right)^{-1} \tag{5.19}$$

which matches the minimum cycle of a pretimed signal (5.16) as we had conjectured. Thus, the resulting average phases are as in Webster's method with C^* replaced by C_{min}. We note that Equations (5.18) and (5.19) also apply to an intersection with more than 2 phases, but the verification of this statement is left as an exercise for the reader.

Since phase durations have been cut by a factor of 2 and overflows have been eliminated one would expect a proportional reduction in delay or even greater. This is true, but we should warn that the result predicted by the New Jersey/Clayton approximation:

$$w_{det} = \tfrac{1}{2}L[1 + (\bar{q}_1 y_2 + \bar{q}_2 y_1)/[(\bar{q}_1 + \bar{q}_2)(1 - y_1 - y_2)]], \tag{5.20}$$

which assumes that the $V_i(t)$ are linear, underestimates the actual delay in this case. This happens because our phases vary over time, and the longer cycles (with the longer phases and delays) tend to entrap more vehicles. This phenomenon also occurs with waiting times at bus stops, which will be studied later in this book; we will find that the average waiting time of randomly arriving passengers exceeds one half of the average headway (the average waiting time one might expect) because more people experience long headways than short. This phenomenon is called the '*length bias*' effect in statistics and it arises often in transportation applications. For example, it is the reason why the time-mean speed is always greater than the space-mean speed.

An exact calculation of delay for random but stationary traffic can be given, but is beyond the scope of these notes; results for intersections with two-directional traffic (Newell and Osuna, 1969) and for general arrival processes with more than two phases have been obtained (Daganzo, 1990). The problem is of general interest because the mathematical equations also describe the behavior of other systems such as computer token rings, container cranes and shuttle systems (buses, elevators). As an exercise, the reader may want to explore the equivalence between what we have described and an elevator that shuttles between two floors. She may also want to answer the following questions: How is a bus line with n stops equivalent to a hypothetical signal with n successive phases? What are the correspondences between the variables and parameters of the two problems? Being able to see the analogy in situations like these allows one to make the connections between what it is known for different transportation modes and be a more competent analyst.

A spreadsheet simulation similar to the one built for the pretimed case can be constructed, but the logic is more complicated since G_1 and G_2 are not fixed. The one that has been made available to the public (ACTUATED. WK1) uses some 'macros' to construct a sequence of simulated phases and waits for random stationary traffic. Its results illustrate how these vary over time and that $\bar{w}$ is much smaller than the average wait for the pretimed signal, although greater than (5.20), as stated.

The actuated strategy presented here is suitable for isolated traffic signals controlling intersections of major streets with substantial (under-saturated) traffic levels but should be modified in other cases. For example, if traffic becomes oversaturated for an extended period of time, the strategy does not proactively allocate more green to the approach with the highest saturation flow. Clearly, a mechanism that promotes this behavior needs to be introduced, e.g., by placing different limits on the maximum green phases allowed for both approaches.

If traffic is very light on one of the approaches the strategy can also be improved slightly by extending the green on the major approach until a small queue has built on the minor approach. (This is also true for the pre-timed control with deterministic constant flows as the reader can verify by numerical experimentation with the spreadsheet of Chapter 3.) One may in fact want to use a semi-actuated form of control, where the signal is green on the major approach until the minor approach has accumulated a certain number of veh-secs of delay.

If both approaches have very little traffic, e.g., during nighttime when capacity is not an issue, it may be better to show an all-red phase by default and allow approaching vehicles to turn the signal green prior to their arrival so as to avoid a stop. The signal at Colusa and Tacoma in Berkeley, California operates in this form.

It should also be noted that the various strategies mentioned here require different placement of detectors so as to ensure that the pertinent information can be gathered in a timely way. The basic scheme requires the detectors to be somewhat upstream of the stop line so that when a drop in flow has been detected, indicating that the green should be terminated, the last vehicles of the saturated platoon should still be on their way toward the stop line[23]. The semi-actuated scheme requires in addition a detector that is well upstream on the minor street, beyond the reach of any queues, in order to construct a virtual arrival curve and keep track of delay. The 'nighttime' strategy requires upstream detectors as well. The exact locations can be determined through judicious use of the time-space diagram. It should be said too that the strategies we have presented are not always used and that on

CHAPTER SIX

Observation and measurement

Earlier in this book we saw how to predict the behavior of simple, unscheduled transportation systems from their basic laws and relevant data (Chapter 4), and also how to affect the performance of these systems in a positive way with various control schemes (Chapter 5). The present chapter addresses one last topic in our review of unscheduled transportation systems. It presents an introduction to the observation and measurement methods that can be used to obtain relevant input data, estimate the parameters of the basic laws, and monitor a system's performance.

Because the world around us is not perfectly reproducible, an understanding of observation and measurement issues requires working knowledge of the theories of probability, stochastic processes and statistical estimation. It is assumed here that the reader is already familiar with the theory of probability at the level of a typical undergraduate engineering course on the subject but not necessarily with the latter two subjects. As a result, the two first sections of this chapter cover the elementary aspects of stochastic process and estimation theory that are necessary for understanding the results of simple observation and measurement experiments. Section 6.1 covers probability, stochastic processes and simulation, and Section 6.2 the procedures for estimating a single quantity (e.g. the capacity, average flow, etc...) from simple experiments that generate data in the form of a stochastic process. Because some of the material in Secs. 6.1 and 6.2 is not always included in conventional courses, the advanced reader may want to skim these sections before proceeding to Sec. 6.3 and then treat them as if they were an appendix. Section 6.3 discusses observation procedures for traffic streams, and the statistical treatment of the data so generated. Subjects covered include estimation of capacity, interpretation of detector data, use of moving observers for monitoring serial systems with entering and exiting flows and the estimation of O-D tables in various ways.

6.1 Probability and stochastic processes

The goal of this section is to present essential information on stochastic processes and probability theory. Although the reader is assumed to know the latter, it should be useful to highlight some of the points that come up most often in our field; especially, since some of them are often not emphasized enough in formal courses, and sometimes are not even mentioned at all. The explanation here (and in Sec. 6.2) will also be different from others you may have seen in that it will emphasize the experiments that probability laws try to describe rather than the mathematical laws themselves. We shall see that the mathematical laws in a complicated situation are sometimes obvious if one thinks about the experiments in a particular way.

Probability is a measure that describes the results of repeatable experiments that are not perfectly reproducible. It is applied to circumstances (called 'events') which either happen or do not happen each time an experiment is performed; the measure is intended to capture the likelihood of an event's occurrence *before* the experiment is performed. If one *imagines* that the experiment can be repeated any number of times, N, probability is then defined as the limit of the fraction of times that the event would occur in a infinite sequence of experiments (as $N \to \infty$).[1] Note that the word 'imagines' has been emphasized. A probability can be assigned to the events of an experiment that is only performed once if we can imagine a repetition of some sort; it is only in the context of such a repetition (imagined or not) that probability can be interpreted physically and intuitively.[2]

In order to describe a non-reproducible (probabilistic, or random, or stochastic) experiment completely one would have to give the collection of probabilities for all the possible events associated with the experiment. When the result of an experiment can be described by a single number (i.e. the number contains all the information that interests us) the experiment is called a *random variable*. The particular number that is obtained when a random variable is 'realized' is called an outcome; i.e., an outcome of a random variable is a number that describes the result of the random experiment. To describe a random variable, X, completely it suffices to give the probability of each of its possible outcomes, x. (Whenever possible capital letters will be used to describe random variables and lower case for outcomes.) Then, from the rules by which probability is manipulated–the same obvious ways in which one would manipulate frequencies–one can construct the probability of any event of interest.

Suppose that we want to assign a probability to the event 'the number of transit passengers that will use BART tomorrow will not exceed x'. If tomorrow is Tuesday we can *imagine* a long sequence of Tuesdays, or otherwise identical days to tomorrow with slightly different riderships, from which the value x would be drawn. Then, the act of observing on any given day is our random variable X. The notation $\{X \leq x\}$ is used to denote the event of interest, and $\Pr\{X \leq x\}$ is used for its probability. Since we have defined our experiment clearly, a probability of 0.7 would have the unambiguous meaning that our critical value would not be exceeded on 70% of the Tuesdays. *Always think about the experiment.*

What does the weatherman mean when he says that there is 50% chance of rain tomorrow? What set of identical experiments is he basing his percentage on? Although the weatherman does not say so, it seems reasonable to assume that he bases his statement on the weather outcomes experienced over all past days in which a prediction had been made based on 'similar' data. In this example, thus, the experiment is living through another day with similar weather data.

Although events of interest can take many complicated forms (e.g. the number of BART riders is an even number that does not exceed x) it turns out that the probability of any practical event can be calculated if we know the value of either $\Pr\{X \leq x\}$ or $\Pr\{X = x\}$ for all x. Consider first the case where the possible x's can be counted, as happens in our BART ridership example, and where each possible value of x, x_i, has a positive probability: $\Pr\{X = x_i\} > 0$. Random variables of this type are said to be *discrete*. The functions returning the probabilities $\Pr\{X \leq x\}$ and $\Pr\{X = x\}$ for all values of x are respectively called the *cumulative distribution function* (or c.d.f.), $F_X(x)$, and the *probability mass function* (or p.m.f.), $f_X(x)$. It should be obvious from the manipulation of proportions or percentages that the probability of any event is the sum of the probabilities of the outcomes from which it is composed and, thus, that the p.m.f. suffices to describe completely a random variable. It should also be clear on the same grounds that one can calculate the p.m.f. from the c.d.f. (i.e., that $f_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ if we take $x_i > x_{i-1}$ for all i) and thus that the c.d.f. also gives a complete description. This quick derivation illustrates that remembering the connection of probability to frequencies of success with repeatable experiments makes many things rather obvious.

Something similar can be said for *continuous* random variables, where F_X varies continuously without jumps. Then the mass function is zero but one can define a probability density function (p.d.f.) based on the ratio of $F_X(x) - F_X(x - \epsilon)$ and ϵ for small ϵ: $f_X(x) = dF_X(x)/dx$. The

probability of any event of interest can be calculated by 'adding' the probabilities of all the elementary events $(x - \epsilon, x)$ in the event of interest; i.e., by integrating $f_X(x)$ over the values of x in the event of interest. This is the well-known 'area under the curve' rule for calculation of probabilities.

Before proceeding, something must be said about units. The outcomes of a random variable have physical units (e.g. 'riders' in our BART example) but probabilities do not, since they are the ratio of two counts. A formula for a p.m.f. or a c.d.f. must thus be dimensionless. The p.d.f on the other hand is a ratio of $dF_X(x)$ and dx, which has the reciprocal units of X ('riders^{-1}' in our example).

When F_X or f_X are not known in detail it is sometimes useful to know two numbers that can be used to identify a range of x where most observations are likely to occur. Although there are many ways in which this could be done, in practical applications one always uses the *mean* to identify the center of the range and the *standard deviation* to identify its size. These choices are very convenient because these two values, unlike others that could have been used, change in a simple way when random variables are linearly combined. We will return to this idea soon.

The reader will recall that the mean or expectation of X, written m_X or E(X), is defined as the sum (or the integral if X is 'continuous') of $xf_X(x)$ over the possible values of x. This is the 'center of gravity' of the probability distribution along x. As such, the mean has the same units as x. Since $f_X(x)$ is the fraction of observations with $X = x$ in an infinite sequence, we see that the formula for E(X) yields the arithmetic average of the values in such a sequence. Thus, the expectation notation E() is also shorthand for the arithmetic average of an infinite number of observations of the random variable enclosed in parenthesis.[3]

The standard deviation is defined as the square root of the 'moment of inertia', or variance, of the probability distribution. The latter is given by the sum (or the integral) of $(x - m_X)^2 f_X(x)$ over the relevant set of x's, which can also be written as $E((X - m_x)^2)$. It should be clear from this formula that the variance has the same units as x^2. Therefore, the standard deviation has the same units as x and m_X. A range of 3 or 4 standard deviations on each side of the mean contains most of the probability for the distributions that one typically encounters in practice, although the range can be wider for unusual distributions.

The importance of means and variances lies in their behavior for linear combinations of random variables. If Z is the revenue collected on a transit route during one morning commute, a_i is the fare paid by

those patrons boarding at i, and X_i is the ridership from i, we can write:

$$Z = \sum_i a_i X_i \tag{6.1}$$

because this linear expression is deterministically true for the values of X_i and Z obtained on any given day. As such, Z is itself a random variable: the imagined experiment in a general case would consist in observing one set of X_i's and then calculating Z with (6.1). In our particular example we can simply imagine running the bus on another similar day and counting the money. Although it is generally very tedious to find the p.m.f. of Z even if one knows enough about the X_i to be able to calculate it, which is unlikely, the mean of Z can be obtained rather easily with the following relation:

$$E(Z) = \sum_i a_i E(X_i); \tag{6.2}$$

i.e.: *the mean of a linear function of random variables is the same linear function of the means.*[4] Note that the units in this expression are consistent if the units of (6.1) are consistent. In our BART example the left sides of (6.1) and (6.2) have units of (dollars/day), and the terms on the right hand side units of (dollars/trip) (trip/day).

Before introducing a similar expression for the variance we need to recall the concept of *statistical independence*. Two random variables are said to be independent if the probability distribution of one of the variables does not change when one only considers in its definition the subset of (imagined) joint experiments where the other variable takes a specific value (or is in a specific narrow range). The 'filtered' set of experiments is called a *conditional random variable*. In physical terms, independence means that knowing the value of one of the variables does not change the probability distribution of the other. In our BART example we would *not* expect pairs (X_i, X_j) to be independent because external factors (such as football games, weather, etc...) can act to increase or decrease them jointly on different days. (If X_i was higher than average on a given day we would then expect other X_j's also to be higher than usual on that day.) In applications where the X_i's are pairwise independent the variance of (6.1) obeys:

$$\text{var}(Z) = \sum_i a_i^2 \text{ var}(X_i); \tag{6.3}$$

i.e. *the variance of a linear function of independent random variables is (almost) the same linear function of the means.* The only difference is the exponent '2' of the coefficients. The modification is easy to remember because the squared coefficients are needed to make the units come out right: $(\text{dollars/day})^2 = (\text{dollars/trip})^2\ (\text{trip/day})^2$.

6.1.1 *The normal random variable*

The *normal* random variable is the most important in all of statistics because a linear combination of n independent random variables such as (6.1) behaves approximately like a normal random variable for large n. This theorem, called the *central limit theorem*, is true if none of the terms in the sum contribute significantly toward the total variance, e.g., if the X_i are independent and the terms a_i^2 var(X_i) are bounded by the same constant. In practical cases one does not need n to be large for the approximation to be rather good. The result is remarkable because, combined with (6.2) and (6.3), it allows detailed probabilistic statements to be made about a variable comprised of many parts (such as Z) despite an incomplete knowledge of its components.

Before giving an example let us briefly review the most important properties of the normal family of random variables. It is a two parameter family where members are characterized by their mean, m, and variance σ^2; we denote them by $\eta(m, \sigma^2)$ or η for short. The normal p.d.f. is:

$$f_\eta(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right\}. \quad -\infty < x < \infty \qquad (6.4)$$

which defines the well-known Gaussian bell-shaped curve. Note that f_η has the proper units since it is the ratio of an exponential, which can have no units, and $(2\pi)^{1/2}\sigma$ which has the same units as x.

An important property of the normal family is that any linear combination of independent members is itself normal[5]. Therefore, the dimensionless random variable $\eta^* = (\eta(m, \sigma^2) - m)/\sigma$ is normal; we call it the *standard normal* variable and see from (6.2) and (6.3) that its mean and variance are $E(\eta^*) = 0$ and $\text{var}(\eta^*) = 1$. Note that the standard normal density $\phi(x)$ is given by (6.4) with $m = 0$ and $\sigma^2 = 1$.

The normal c.d.f. is not available in closed form, but the standard normal c.d.f., $\Phi(x)$, has been tabulated and can be approximated in various ways. Since the event $\{\eta(m, \sigma^2) \le x\}$ happens if and only if $\{\eta^* \le (x - m)/\sigma\}$, as you can see from the definition of η^*, both probabilities are the same, and this allows us to express the (general) normal c.d.f. in terms of Φ:

$$F_\eta(x) = \Phi((x - m)/\sigma). \qquad (6.5)$$

This is the standarization property of the normal family. It says that the cumulative probabilities of any normal variable can be obtained by reading from the Φ-table the value corresponding to the number of

standard deviations by which x is away from the mean (the normal deviate).

Example: The first 4 columns of Table 6.1 give hypothetical data for the transit morning commute problem described earlier. We seek the probability of collecting more than \$100 during the study period when the fare from stop i is a_i, given that we know (e.g. from a demand study) the mean and the variance of the number of trips made, X_i, from each stop.

Since the total revenue, Z, is given by (6.1) we start by calculating $a_i E(X_i)$ and $a_i^2 \text{var}(X_i)$ for all the stops and then arrange the information on columns 5 and 6 of Table 6.1. *Assuming* that (6.3) holds, we see that the variance of the total revenue is the sum of the entries on the last column. The mean of Z, of course, is the sum of the 5th column. The answers are:

$$E(Z) = 95.4\$ \quad \text{and} \quad \text{var}(Z) = 31.86\ \$^2.$$

Since it was argued earlier that the X_i are unlikely to be independent, we recognize that the result for var (Z) is only approximate. (Can you tell intuitively by imagining the result of several days of data, i.e., several observations of Z, whether var(Z) would be larger or smaller if the X_i were influenced in the same direction by an exogenous cause? An answer to this question should help you understand the sign of the error in our final answer.)

Under our assumptions the largest contribution to the variance is about 1/4 of the total, and the central limit theorem should be roughly

Table 6.1

Train stop	Fare a_i	$E(X_i)$	$\text{Var}(X_i)$	$a_i E(X_i)$	$a_i^2 \text{Var}(X_i)$
1	0.5	10	4	5.00	1.00
2	0.6	12	4	7.20	1.44
3	0.7	10	4	7.00	1.96
4	0.8	14	2	11.20	1.28
5	0.9	10	4	9.00	3.24
6	1	12	4	12.00	4.00
7	1.1	8	4	8.80	4.84
8	1.2	8	2	9.60	2.88
9	1.3	10	2	13.00	3.38
10	1.4	9	4	12.60	7.84

true. Therefore the c.d.f. of Z is approximated by that of the normal variable $\eta(E(Z), var(Z))$ and (6.5) can be used to calculate the sought probability:

$$Pr\{Z > 100\} = 1 - F_{\eta}(100) = 1 - \Phi\left(\frac{100 - 95.4}{\sqrt{31.86}}\right) = 0.21 \quad \blacksquare$$

6.1.2 Stochastic processes

The theory of stochastic (point) processes concerns itself with random fluctuations in the evolution of one or more numerical values over time or space. Earlier chapters showed that certain cumulative curves (the N-curves) described quite comprehensively the behavior of some systems over time and space. Yet, if any such system were to be observed on a different day (or a different study period) one would expect to obtain a similar but not identical set of curves. The variation of the curves across observation efforts (e.g. days) is analogous to the variation of a random variable across experiments.

By analogy to the outcome of a random variable we call the picture of a (set of) curve(s) the *realization* of a *stochastic process*. Fig. 6.1 depicts two realizations $n^{(1)}(t)$ and $n^{(2)}(t)$ of a process N(t). Note that each curve is made up of the contributions (or counts) over non overlapping intervals of time.[6] For curve j these counts are denoted $n^{(j)}(t_i, t_{i+1})$, as shown in the figure.

If one is only interested in asking questions that involve a scale of measurement that is large compared with $\Delta t = t_{i+1} - t_i$, a sufficient description of the process is achieved by giving the (joint) probability density/mass function of the random variables $N_i \equiv N(t_i, t_{i+1})$; i.e., by specifying the fraction of days in which the 'elementary joint event' $\{n_1 < N_1 \leq n_1 + dn_1\ ;\ n_2 < N_2 \leq n_2 + dn_2;...\}$ happens for any $\{n_1, n_2...\}$. One can then answer (at least in principle) any question about the process, i.e., how often it is found in any given condition, by adding the probabilities of all the elementary events exhibiting the condition. This is all there is to the theory of stochastic processes; much of it deals with the machinery for carrying out the computations in special cases when the task is easy. This section describes the special case where the N_i are identically distributed and mutually independent random variables;[7] i.e., what mathematicians would call a stationary or homogeneous[8] *process with independent increments*. One would expect to find stationary processes (i.e., identically distributed N_i) during periods of observation

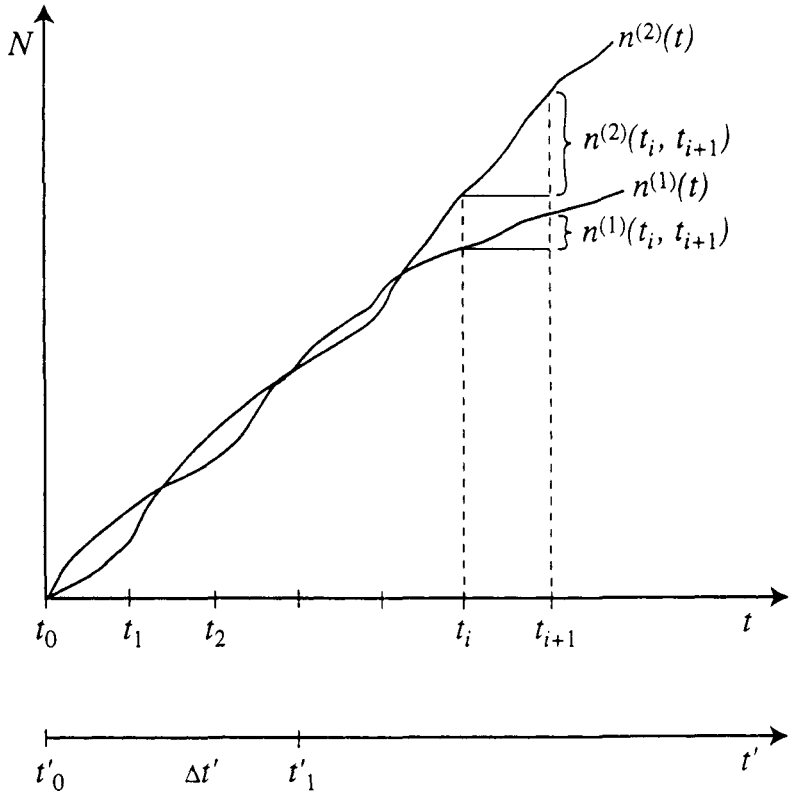


Figure 6.1 Two hypothetical realizations of a counting process.

without rush hours, but we must be careful with the independence condition because it can be broken in subtle ways; e.g., if there are minimum time separations between discrete jumps in N(t) and Δt is not large compared with this minimum time, or if there is an underlying schedule which tends to keep N(t) close to the value on the schedule.

6.1.3 The Brownian process

If the N_i of a process with mutually independent increments are specified to have a normal distribution with a given mean and variance, $m(\Delta t)$ and $\sigma^2(\Delta t)$, the process is completely defined because one can then calculate the probability of any event of interest. A process like that is called a *Brownian* process, or a *Brownian motion* process.[9]

The Brownian process is to stochastic processes like the normal random variable is to general random variables in that the Brownian process is the limiting form (for sufficiently large Δt) of nearly all processes with independent increments.[10] To see this, note from Fig. 6.1 that for any t_i and t_j ($t_i < t_j$) on our lattice we can write:

$$N(t_i, t_j) = N(t_i, t_{i+1}) + N(t_{i+1}, t_{i+2}) + \ldots + N(t_{j-1}, t_j). \quad (6.6)$$

According to the central limit theorem, $N(t_i, t_j)$ will be approximately normal if the number of terms in (6.6) is large enough. Then, if $(t_j - t_i) = \Delta t'$ is used as the new minimum interval, the process becomes approximately Brownian on the new scale of observation.[11]

Equation (6.6) says that the count during any interval (on the lattice) is made up of $(t_j - t_i)/\Delta t$ independent components. By virtue of (6.2) and (6.3) we can thus write:

$$E\big(N(t_i, t_j)\big) = (t_j - t_i)[m(\Delta t)/\Delta t] = (t_j - t_i)\mu \quad (6.7a)$$

$$\mathrm{var}\big(N(t_i, t_j)\big) = (t_j - t_i)[\sigma^2(\Delta t)/\Delta t] = (t_j - t_i)\mu\gamma. \quad (6.7b)$$

The terms containing Δt have been grouped in brackets because these ratios, which are denoted μ and $\mu\gamma$, should be independent of Δt. (To see this note that if Δt is increased by an integer factor, using a subset of the original lattice, then $m(\Delta t)$ and $\sigma^2(\Delta t)$ also increase by the same factor.) This means that both the mean and the variance of a Brownian count are proportional to the length of its interval, as expressed by (6.7).

The proportionality constants, μ and $\mu\gamma$, have units of 'count/parameter' and 'count^2/parameter'. For example, if a process describes the liters of oil entering a reservoir in a port terminal over time, then the units of these factors might be: liters/day and liters^2/day. The first parameter, μ, is the average counting rate of the process, or more simply, its '*rate*.' The second one is the rate at which the variance grows with time; it is usually related to the first rate by a constant γ which is called the *index of dispersion*. The index of dispersion has units of 'count', e.g., liters, and it will change if we change our counting units. Note that γ is also the ratio of (6.7b) and (6.7a). Thus, it is also the variance to mean ratio of the counts in an interval of *any* length (for a Brownian process).

It is quite fortunate that just two constants (μ and γ) should suffice to describe completely the behavior of so many processes on a sufficiently coarse scale of observation because this means that (where a coarse scale can be used) the world can be described in a simple way.

To illustrate this more tangibly, let us return to Eq. (2.12) of Chapter

2, which applies to queueing systems in which the arrival and service processes are independent Brownian processes.[12] The parameter Δ appearing in that expression is the weighed sum of the indices of dispersion of the arrival and service processes, γ_a and γ_s, where the index of dispersion for the arrival process is weighed by the saturation level $\rho = \lambda/\mu$; i.e, $\Delta = \rho\gamma_a + \gamma_s$.

Since crowded systems can be studied on a coarse scale of measurement and queueing systems become crowded as they approach saturation ($\lambda \rightarrow \mu$) we conclude that steady state queues can be studied on a coarse scale if $\lambda \rightarrow \mu$. And since all practical processes with independent increments look Brownian on a coarse scale, it follows that any queue with independent arrival and service processes made up of independent increments can be studied as a Brownian queue when $\lambda \rightarrow \mu$. Thus, we see that Eq.(2.12) can be applied in this case too. It is indeed fortunate that such a simple result should hold for so many systems independently of their detailed probabilistic structure.

One useful property of Brownian processes is that they obey a superposition principle; i.e., any linear combination of independent Brownian processes is itself a Brownian process. The result, which should be intuitive to you, is based on the fact that any linear combination of independent normal variables–e.g., the counts of the various processes in each elementary interval–define a normal variable. In view of this, can you figure out the formulas for the rate and index of dispersion of a linear combination of Brownian processes?

The inverse process: This subsection is provided for completeness and may be skipped without loss of continuity. The reader already familiar or planning to study stochastic processes may appreciate seeing the simple dimensional arguments given here as an alternative to the lengthier and less general mathematical proofs that are given in standard textbooks. Given a process N(t) with *positive* independent increments, one is concerned here with the asymptotic properties of the inverse process T(n), i.e., the process that indicates the 'time' when N(t) reaches n.

Without loss of generality, it is assumed that the units of measurement for time have been chosen so that the rate of N(t) is one, $\mu = 1$. It is also assumed that the scale of observation uses increments sufficiently large for N(t) to be approximately Brownian.

An equivalent mechanism for generating Brownian realizations with $\mu = 1$ consists in sampling two independent sequences of random outcomes from a small positive random variable, i.e., $(x_1, x_2, x_3,...)$ and $(x_1', x_2', x_3'...)$, and then plotting on the (t, n) plane the piecewise linear

curve

$$\left(\sum_{j=1}^{m} x_j, \sum_{j=1}^{m} x'_j \right)$$

with m as the parameter. To see this note that if our random variable is chosen to be very small, in the sense that its maximum possible value is negligible compared with Δt, then the increments in n for each Δt will be nearly independent. This means that the process is approximately Brownian on any scale $\Delta t' \gg \Delta t$. Since $\Delta t'$, Δt and the random variable can be chosen to be as small as one wishes, the approximation can be improved as much as desired on any scale of observation.

What is interesting about this method of generating N(t) is that it is symmetric to interchanges of t and n. This immediately establishes that the inverse process must have the same statistical properties as N(t); i.e., that if it is observed on a scale where each Δn contains many little segments, T(n) will be Brownian with unit rate and *the same index of dispersion* as N(t).

If the time scale is now changed to restore the rate of N(t) to its original value, μ, several things happen; (i) the 'Brownianesses' of N(t) and T(n) are preserved; (ii) the rate of the inverse process changes from 1 to μ^{-1} by purely dimensional reasons; (iii) likewise, γ remains unchanged (dimensions of 'quantity'), and (iv) the index of dispersion of the inverse process (dimensions of 'time') changes from γ to γ/μ. This establishes that the following relations hold for a process with positive independent increments and its inverse, on coarse observation scales:

$$\mu_{\text{inv}} = 1/\mu \quad \text{and} \quad \gamma_{\text{inv}} = \gamma/\mu.$$

6.1.4 The Poisson and Binomial processes

These are two processes that come up very often in applications. Both have independent increments and dimensionless integer counts; they look Brownian on a coarse scale. The Poisson process is taken up first.

We say that N(t) is a stationary Poisson process with rate μ (in occurrences/time or occurrences/distance) if the process has independent increments for any Δt, no matter how small, and the count in any short interval satisfies to first order in Δt:

$$\Pr\{N(t, t+\Delta t) = 1\} = \mu \Delta t \tag{6.8a}$$

$$\Pr\{N(t, t+\Delta t) = 0\} = 1 - \mu \Delta t \tag{6.8b}$$

A random variable like N(t, t + Δt) that is 1 with a certain probability ($\mu\Delta t$ in our case) and 0 otherwise is called a *Bernoulli trial* with probability of 'success' $\mu\Delta t$. The reader may remember (or can deduce from (6.8)) that

$$E(N(t, t+\Delta t)) = \mu\Delta t; \tag{6.9a}$$

this confirms that μ can be interpreted as the rate of the process. She can also check that (to first order in Δt)

$$\text{var}(N(t, t+\Delta t)) = \mu\Delta t(1 - \mu\Delta t) = \mu\Delta t \tag{6.9b}$$

and thus, the index of dispersion of the Poisson process is 1.

Although we already know that on infrequent observation it will look approximately Brownian (normal counts), in this case the exact p.m.f. of the count is available in closed form for intervals of any length. The result[13] is:

$$Pr\{N(t_i, t_j) = i\} = e^{-\mu(t_j - t_i)} \frac{[\mu(t_j - t_i)]^i}{i!} \quad (i = 0, 1, 2 \ldots) \tag{6.10}$$

which is called the Poisson p.m.f. Its mean is $\mu(t_j - t_i)$, which is consistent with the interpretation of μ as the 'rate' of the Poisson process. The reader may have been introduced to this distribution in other courses and may be familiar with its properties. Some of these should now be fairly obvious without any derivation. In particular, if P_1 and P_2 are two independent Poisson random variables with means m_1 and m_2, it should be clear that

1. $P_1 + P_2$ is Poisson with mean $m_1 + m_2$,[14] and
2. P_1 becomes normal as $m_1 \to \infty$ ($m_1 > 20$ is good), and Bernoulli as $m_1 \to 0$ ($m_1 < 0.1$ is good).

The Poisson process also obeys a (limited) superposition principle in that a *sum* of independent Poisson processes is itself Poisson.[15] This means that if the vehicle counting processes on the lanes, i, of a freeway are independent Poisson processes $N_i(t)$ then the aggregate freeway counts $N(t) = \Sigma_i N_i(t)$ form a Poisson process too.

The Poisson process arises often in practice because it can be generated by a superposition of many independent *non-Poisson* processes. This happens for example if each one of these processes has a very small probability of landing a *single* count in the observation interval [0, T) and the probability density of the landing point is evenly distributed in said interval. It should be emphasized that the probabilities of landing a single count do not have to be equal for all the

processes.[16] Clearly then, a Poisson process should be a good approximation for light highway traffic whenever one can imagine every automobile owner in the world generating a simple process whereby (s)he is either counted somewhere in our study period [0, T) with very small probability, or not at all. Inasmuch as all points in [0, T) are equally likely and people have the ability to be seen any time in the interval independently of each other, e.g., there is no congestion and there are no metering devices nearby, we should expect the observed realization to be Poisson. This is the reason why the Poisson process comes up so often in the study of 'rare events' such as accidents and lightning strikes.

The superposition origin of the Poisson process also suggests that a realization with n points in a given interval should look as if the n points had been chosen at random and independently in the interval (e.g., by throwing darts). Obvious on physical grounds, this relationship between the uniform distribution and the Poisson process can also be established mathematically. But the mathematical route is more laborious. Hopefully the above remarks will help you develop a 'feel' for the Poisson process and help you recognize when it might be used to describe a practical situation.

6.1.5 Forgetfulness. Intervals

Any process with independent increments, such as the Brownian or the Poisson process, is 'forgetful' in the following sense: if one looks at the system at $t = t_0$, i.e., the present if one imagines that the parameter of the process is time,[17] then the future of such a process is independent of the past. This means that if one is observing a Poisson process, e.g., the number of airplane accidents in the U.S. as a function of time, the probability of observing an accident in the next time interval is unaffected by how long it has been since the last accident.

The forgetfulness property alone can be used to derive the distribution of the intervals between occurrences (headways) of a Poisson process but the following argument is more direct. Simply, let H be the interval following any randomly chosen observation of a Poisson process (at $t = t_0$) and then write for the c.d.f. of H, $F_H(t)$:

$$1 - F_H(t) = Pr\{H > t\} = Pr\{N(t_0, t_0 + t) = 0\} = e^{-\mu t} \quad (t \geq 0). \quad (6.11)$$

The first equality is simply the definition of $F_H(t)$. The second equality is justified because events prior to t_0 have no relevance, and an interval longer than t is equivalent to having no counts in a period of

duration t following t_0. The last equality follows from (6.10) with i = 0. Note that the mental experiment underlying the definition of Pr{H > t} in (6.11) could consist in looking at *all* the occurrences of the event in a *single* realization of the Poisson process. Viewed in this manner, (6.11) is saying that the distribution of headways must be the same for all (infinite) realizations that come up.[18]

It should be emphasized that H is a continuous random variable with units equal to 'parameter', and that the notation 't' in (6.11) can represent distance or something else depending on the context. Note that a dimensionless combination 'μt' appears in the exponential of (6.11), as must always be the case, and that the result has no dimensions since it is a probability. On taking the derivative of $F_H(t)$ one finds the density:

$$f_H(t) = \mu e^{-\mu t} (t \geq 0). \tag{6.12}$$

which has units of 't^{-1}' as any density should. Thus, if you remember that the p.d.f. and c.d.f. of the intervals are *negative exponential* functions, as this random variable is usually called, you should be able to write the expressions correctly by ensuring that the formulas are dimensionally correct.

It should also be intuitive without any need for derivations that the mean of this random variable must be close to the ratio of the combined length of all the intervals found in a long observation period [0, T), i.e., T, and the number of intervals. To an accuracy of ± 1 the latter number is also the Poisson count in the interval. Because the standard deviation of N(0,T) only increases with $T^{1/2}$, the ratio N(0,T)/T approaches μ as T $\rightarrow \infty$ in every realization. Therefore, it follows that:

$$E(H) = \mu^{-1} \tag{6.13}$$

and the units again come out right[19]. If the above argument sounds familiar you are right! It was used in Chapter 1 to relate the concept of average flow (or density) to that of average headway (or spacing). Equation (6.13) and the expression var(H) = μ^{-2} also follow directly form the formulae for μ_{inv} and γ_{inv}. Applied to the Poisson process, these expressions say that both the rate and the variance to mean ratio of

$$T(n) = \sum_{j=1}^{n} H_j$$

must be μ^{-1}. Therefore, E(H) = μ^{-1} and var(H) = μ^{-2}.

The forgetfulness property of the Poisson process manifests itself through the negative exponential distribution in the following way: the proportion of intervals larger than t that are themselves longer than t + Δt only depends on Δt; i.e., how long one waits for an occurrence is

independent of how long one has already waited. (This is the way it should be because otherwise we would have found a statistical dependence between future and past, which is impossible by assumption.) The validity of this assertion can be verified by writing the relevant proportion with the help of (6.11) as $e^{-\mu(t+\Delta t)}/e^{-\mu t}$, which is indeed just a function of Δt. As one would expect, the result is the same as if the prior wait, t, had been zero: $e^{-\mu \Delta t}$.

A curious consequence of forgetfulness is that if we jump in the realization of a Poisson process at an arbitrary time and measure the forward and backward time intervals to the next and last occurrences, the durations of said intervals, H_f and H_b, must each have the same distribution as H. On the surface this seems like a paradox: why should the interval in which we land (of length $H_b + H_f$) be twice as long on average as a typical headway? The reason lies on the sampling method and can be understood if one goes back to the experiment. In defining the distribution of H we imagine selecting *observations* at random and examining the interval that followed them. In the current case, however, we are selecting random *points* and observing the interval around them. Obviously, this introduces a 'length bias' because long intervals are observed more frequently. For the Poisson process the effect is strong enough to double the mean interval length; can you demonstrate this algebraically?

The effect, which arises with other processes too, illustrates the need for explaining carefully what one means in terms of a sampling experiment when defining a random variable. The already encountered fact that the average vehicular speed measured on a road depends on the method of observation (time-mean speed greater than space-mean speed) can be explained in terms of a 'length-bias'.

Let us now bring up another random variable (called *Erlang* or *gamma*) which describes the time of the Poisson process until the ith occurrence. This material is less applicable and can be skipped. Nonetheless, some readers may find it interesting to see how this family of random variables is physically connected to the Poisson, negative exponential and normal random variables, and how its properties can be derived with relatively little algebra.

If E_i is used to denote the time interval following a randomly chosen observation of a Poisson process realization until the ith observation[20], then using the same logic of (6.11) we can write for the c.d.f. of E_i:

$$1 - F_E(t) = Pr\{E > t\} = Pr\{N(t_0, t_0 + t) \leq i - 1\}$$

$$= \sum_{j=0}^{i-1} e^{-\mu t} \frac{(\mu t)^j}{j!} \quad (t \geq 0) \tag{6.14}$$

where the last equality follows from (6.10). Note that (6.11) is recovered for $i = 1$, as expected, since by definition $E_1 \equiv H$. As in (6.11), μ also appears here in the dimensionless combination (μt).

Note that E_i has been defined as the duration of i consecutive H's and thus it is the sum of i independent negative exponential variables. This is the definition of the Erlang (gamma) variable sometimes found in books. Thus, we see from (6.2) and (6.3), and without using any derivations, that the mean and variance of E_i are i times those of H, i.e.: i/μ and i/μ^2. Furthermore, the central limit theorem guarantees that E_i becomes normal for large i ($i = 20$ is good).

The density of E_i, $f_E(t)$, can be obtained by taking the derivative of (6.14) and changing its sign, or more directly by noting that it must satisfy the following relation: $f_E(t)dt =$ Pr{ $(i-1)$ events in $(0, t)$ and one event in $(t, t + dt)$}. Thus,

$$f_E(t) = \mu \frac{(\mu t)^{i-1}}{i-1!} e^{-\mu t} \quad (t \geq 0). \tag{6.15}$$

6.1.6 Multidimensions

The interpretation of the Poisson process as a superposition of many independent processes with a uniform (small) probability of observation in the interval of interest, suggests how it can be extended to multidimensions. All one has to do is imagine that these independent processes choose a point in a multidimensional region, **R**, rather than in an interval and that the probability of landing in a subregion **A** of **R** is proportional to the 'size' (or measure) of that subregion.

The region **R**, for example, could be the relevant part of the time-space diagram that one would draw to represent a certain road during a given year, and the subregions portions corresponding to specific sections during specific months. (One may do this for a study of accident data.) Just as we associated a measure of 'length' with each interval in the one dimensional case ($t_j - t_i$) we must now associate a measure $|\mathbf{A}|$ with any subregion of interest in **R**. If μ is used to denote the proportionality constant of the landing probability in a subregion we can write for the probability of finding one observation in a small subregion $\Delta\mathbf{A}$, $\Pr\{N(\Delta\mathbf{A}) = 1\} = \mu|\Delta\mathbf{A}|$. And this is our multidimensional extension. The independence of the counts on non-overlapping regions, together with some combinatorics, again produce expression (6.10) for large regions. In the current case we simply have to substitute (t_i, t_j) by **A** and $(t_j - t_i)$ by $|\mathbf{A}|$ in (6.10). The functional form of the count probabilities, means and

variances is also the same if the dimensionless term $\mu(t_j - t_i)$ is replaced by the dimensionless combination $\mu|\mathbf{A}|$.

Note that we have not said anything specific about how one would define the measure $|\mathbf{A}|$. You may think that a measure such as area would do the job, but this is not necessarily so. What we want is a measure that will be proportional to the real world probabilities. In an application, for example, where we are interested in the number of single-vehicle accidents on a long road with entrances and exists (or even a system of roads) it is reasonable to *expect* the probability to be proportional to the vehicle-miles traveled (VMT) in A, rather than to the time-distance size of A, at least if the road is homogeneous and driving conditions such as weather do not vary over time. Thus, $|\mathbf{A}|$ should be chosen as the VMT in A. Alternatively, one can keep the original interpretation of A and allow μ to vary within R. This is called an inhomogeneous Poisson process. In the accident analysis literature, the measure $|\mathbf{A}|$ is called an 'exposure' factor because it multiplies the (constant) accident rate, μ. A great deal of effort is spent by accident specialists studying exposure measures because a reliable one allows them to see whether the accident patterns that develop over time and space depart significantly from what the multi-dimensional Poisson process would predict. These discrepancies help identify dangerous locations (black spots) that need improvement.

Another instance where one would expect a 3-dimensional Poisson process to be a good model of reality is for the occurrence of special purpose telephone calls over a residential geographic area. These calls could be for the request of rarely used special services such as ambulances, taxis or express package collection services. If people act independently and request these services very infrequently, then the superposition principle would guarantee the Poisson law. Of course, one still needs to define the measure for time-space subregions, A, but this is an experimental issue that must be determined by observation. For an A that corresponds to a time slice of small duration (e.g., 15 mins.) and a given geographical sub-area one would expect the measure to be something like the product of the population in the sub-area and the average regional demand in the time-slice. Models of this type have been used in the analysis of emergency delivery, physical distribution, and taxicab operation systems.

6.1.7 The Binomial process

This is a process with independent increments of single counts that arises much in the same way as the Poisson process. But now the (time

or distance) parameter is discrete and is numbered, t = ... −1, 0, 1, 2... Each observation act at a given time is called a *Bernoulli trial.* The process is specified by giving the probability ($0 \leq p \leq 1$) of a single count (or 'success') occurring in any given trial. If by analogy to continuous parameter processes one lets N(t, t + t′) denote the count of successes for trials t to t + t′ − 1 (i.e., not including t + t′), the following is true:

$$\Pr\{N(t, t+1) = 1\} = p \tag{6.16a}$$

$$\Pr\{N(t, t+1) = 0\} = 1 - p \tag{6.16b}$$

which is the direct analog of Eq. (6.8) for discrete time. Thus, the Binomial process (sometimes also called the *random walk*) can be viewed as a sequence of Bernoulli trials with probability of success p. It is a simple exercise to verify that the mean and variance of a Bernoulli trial are p and p(1 − p).

The following two examples involving vehicle sequences use different definitions of success. For vehicles lined up at a red traffic signal (i) one may define success as turning left, while for railcars arriving at a classification yard (ii) one may define it as *not* being sent to a specific classification track. In both cases one might be interested in the average length of a consecutive sequence (or run) of 'failures' because a long run postpones (i) the blockage of a lane by a left turner if the turn is unprotected or (ii) the lost time arising from the switching procedure from one track to another. The length of such a run, G, is the discrete time equivalent of the negative exponential interval, H. It is defined here as the number of trials up to, and including, the first success. Some books define it as the number of failures.

The c.d.f. of G is obtained by noting that G > i if and only if the first i trials were failures. (This is the same logic that was used in the derivation of H.) This happens with probability $(1 - p)^i$ and thus we can write:

$$1 - F_G(i) = Pr\{G > i\} = (1-p)^i \quad (i = 1, 2, 3 \ldots) \tag{6.17a}$$

Note that the range of possible values starts with i = 1.[21] The mass function is obtained by subtracting consecutive values of F_G (or $1 - F_G$), and this leads to

$$f_G(i) = (1-p)^{i-1} - (1-p)^i = p(1-p)^{i-1}. \quad (i = 1, 2, 3.) \tag{6.17b}$$

The easy way to calculate the mean is by adding the complementary c.d.f. (6.17a) for all *i* (this shortcut formula can be used with any non-negative, integer-valued random variable). Because (6.17a) is a geometric series the answer follows from the standard recipe for such series; it is: E(G) = 1/p.

Although the variance can also be calculated by adding a series using similar manipulations, an analogy to the Poisson process is used here instead to guess what the result must be. It should be clear that if p is small and we imagine that the trials are spaced 1 time unit apart, the binomial process will look like a Poisson process with rate $\mu = p$ on a scale of measurement that is large compared with $t = 1$. Thus one would expect (6.17a) and (6.11) to be close approximations to each other with the substitution $\mu \Leftrightarrow p$,[22] and the same thing to happen for the means and variances of H and G. It should be reassuring to see that, indeed, $E(H) = 1/\mu$ and $E(G) = 1/p$ match after the substitution. For the variance, the formula corresponding to $\text{var}(H) = 1/\mu^2$ is $\text{var}(G) = 1/p^2$, but this can only be correct for small p since var(G) must be close to 0 for $p \to 1$. In this case the probability of a run of failures longer than 1 is negligible, and the geometric variable is like a Bernoulli trial. Its variance should be $p(1-p) \approx (1-p)$, for $p \to 1$. The simplest expression that varies like $(1-p)$ for $p \to 1$ and like $1/p^2$ for $p \to 0$ is $(1-p)/p^2$. This guess is in fact the correct expression for var(G). The same expression can also be readily obtained from the Brownian inversion formulae by repeating the steps used earlier to find the variance of the negative exponential distribution. Can you do it?

The random variable $N(t, t+n)$, representing a count in n trials, is called *binomial* and denoted B_n for short. In the above-mentioned two examples, the binomial random variable can be used to study the length of a left turn pocket (if one wishes to segregate left-turning cars out of the through lane) or the length of a classification track to avoid overflow[23]. The probability that n trials result in a specific ordered sequence of i successes and $(n-i)$ failures is $p^i(1-p)^{n-i}$. This is true independent of the order. We know from combinatorics that there are $n!/i!(n-i)!$ different such sequences–this is the number of combinations of i specific positions from n possible slots. Thus, the aggregate probability for all possible sequences is the familiar formula:

$$f_B(i) = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-1} \quad (i = 0, 1, \ldots n). \tag{6.18}$$

Let us now see what can be said about this distribution from its physical relationship to our stochastic processes, without any mathematics. First B_n is a sum of n independent Bernoulli trials B_1, whose mean and variance were p and $p(1-p)$. Therefore, $E(B_n) = np$ and $\text{var}(B_n) = np(1-p)$. Furthermore, for large n the distribution function must be close to the normal distribution with the same mean and variance by virtue of the central limit theorem. This approximation works well if the tails of the normal distribution function do not extend much out of the

range [0, n]; i.e., if $3(np(1-p))^{1/2} \leq \min(np, n(1-p))$. Finally, for the case with low p ($p \sim 1/20$ is very good) the binomial distribution can be approximated by the Poisson distribution with the same mean. The reason is the same as that used earlier to justify the approximation of G by H.

The random variable denoting the number of trials until the i^{th} success is called *negative binomial*, but we do not examine it here because it does not arise much in applications. Its mass function can either be derived from the binomial p.m.f., much as the gamma density was derived from the Poisson, or directly from combinatorial arguments. Since it is a sum of i independent geometric variables its mean and variance are i times larger than those of G, and it will tend to the normal for large i. Can you guess what distribution it will resemble for low values of p?

6.1.8 Simulation

This review of applied probability ideas is ended here with a description of *simulation*; i.e., the act of generating artificial data with a computer to mimic a real life (stochastic) process. Although whole courses can be given on this subject, its basic principles are so simple that they can be summarized in just a few pages. This knowledge should suffice to write simple simulations and interpret their results. The author believes that running simple simulations is important because such an exercise can solidify one's understanding of previously introduced, abstract concepts such as 'realization', 'relaxation time' and 'index of dispersion'; it should also help with concepts not yet introduced, pertaining to data interpretation and statistical theory.

The art of simulation consists in taking a real world system and reducing it to a set of variables that are tracked over time. This set of variables, which shall be called the *state*, must contain at any given time (the present) enough information to make the state's future evolution statistically independent of the state's history.[24] Then, the computer only has to keep the state of the system in memory, updating it as time is advanced. No other information about the system needs to be kept.

In general, the state of a system is not uniquely defined, and one can choose to simulate at various levels of detail. The trick is to choose the simplest, relevant state representation. To achieve relevance, the state of the system must include one (or more) variable(s) from which one can calculate the desired figure(s) of merit. Simplicity is important because it reduces the need for gathering input data, simplifies the

programming task and minimizes the chances that data/programming errors will corrupt the final results. Simplicity also facilitates the interpretation of results.

A simulation produces a stream of numbers that in the end must be summarized into one or just a few figures of merit. Such results do not reveal completely how the solution depends on the parameters of the problem.[25] The significance of this fact was mentioned in Chapter 3, where numerical and analytical optimization approaches were compared. The issues are the same now: one cannot expect computer simulations to yield insight into the way a system should be manipulated to achieve a specific goal. Instead, the role of simulation should be that of an evaluation tool, able to discriminate among a few well defined system configurations. [Although a great deal of effort has been invested to date on simulation models of freeways and arterial surface street networks, it seems fair to say that the impact these tools have had on improving ramp metering or signal optimization policies has been small; see Chapter 5].

Once the state of the system has been defined, developing the simulation is easy. We need to produce a program logic that will change the state of the system and update a '*clock*' according to the rules prevailing in the real world. In the simulation jargon the resulting state changes, occurring at an instant of time, are called *events*. Thus the program logic must specify for every possible state which events are likely to occur next, and when. When the rules are not deterministic, the simulation program must have a way of duplicating the random experiment taking place in the real world. This can be accomplished by generating random numbers with suitable distribution functions with the computer. Although the mechanism that generates the numbers is deterministic, the sequences can be made to have all the statistical properties one would expect of 'true' random numbers.

To see how these ideas can be put to use let us consider a single sever queue with FIFO discipline (see Chapter 2) to which customers arrive as a stationary Poisson process with rate λ. We assume that the service times are mutually independent and identically distributed random variables, and that the arrival and service mechanisms are both, independent of each other and independent of everything else that may be included as part of the state. Our goal is determining the steady state average waiting time for this system.

The advantage of choosing such a simple system for our simulation is twofold: (i) one can construct *and* run the simulation in less than 5 minutes, with access to a computer spreadsheet, and (ii) the output can be compared with the exact analytical solution, which is known. In this

way one can experience in a short time all the steps of building a simulation and interpreting its results. The difference in more complicated cases only lies in the conventions of the particular language chosen for the application.

For our particular problem the number of customers in the system at any given time should be part of the state. Events connected with this definition of the state would be customer arrivals and departures. We know that the occurrence of an arrival in the immediate future after a time t_0 is independent of anything that might have been observed in the past by assumption, because the Poisson process has independent increments. The occurrence of the next departure, however, not only depends on the state (whether the system is empty or not) but also on the amount of time that the current customer has been in service. Therefore, this elapsed service time (or a proxy thereof) should also be part of the state. Consideration of our assumptions shows that the pair of variables, 'number of customers in the system' and 'elapsed service time', is a proper state because any changes to these two variables in $(t_0, t_0 + \Delta t)$ are independent of events prior to t_0.

Sometimes the logic of a simulation can be simplified if some auxiliary variables are introduced in the state, provided the new variables do not introduce dependences with the past. In our case it is convenient for spreadsheet implementation to include as part of the state the service times of all the customers in the queue, or equivalently their projected departure times. One can then step through time one customer at a time, recursively calculating their departure times. The logic of our simulation is based on the observation that a customer's departure time is the sum of its service time and the (clock) time when its service began; the latter being the latest of its arrival time or the departure time of the previous customer.

Letting A_n and D_n be the arrival and departure times of customer n according to the clock, and S_n its service time, we can write:

$$D_n = S_n + \max\{A_n ; D_{n-1}\} \quad (n = 1,2\ldots) \tag{6.19}$$

assuming that successive customers are numbered in order of increasing n and that D_0 is given. In order to iterate (6.19), ways of generating the A_n and S_n are needed. The A_n obey the recursion.

$$A_n = A_{n-1} + H_n \quad (n = 1,2\ldots) \tag{6.20a}$$

$$A_0 = 0 \tag{6.20b}$$

where the headway of customer n, H_n, is a negative exponential random variable which, like S_n, is independent of everything else observed. A

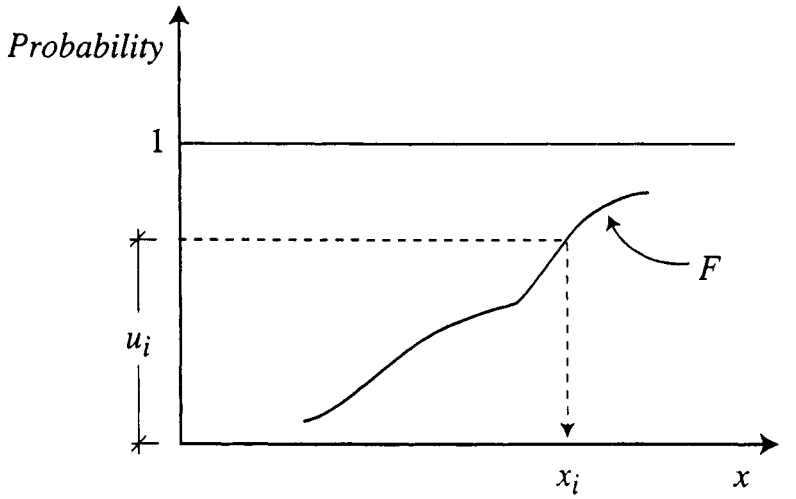


Figure 6.2 The inverse c.d.f. method for generating random numbers.

method for generating observations from a random variable with an arbitrary distribution function would allow us to generate the desired data.

The inverse c.d.f. method described below is quite general although not always easy to implement. It assumes that the user can generate random draws, U, from a continuous random variable that is uniformly distributed between 0 and 1. Most computer languages have functions that return a random U when called; the function @RAND in most spreadsheets is an example. In many cases it is also possible to select from a menu of often used distributions, but we prefer here to stick to fundamentals.

To generate a random draw, x_i, from a random variable with c.d.f. F(x) one simply takes a draw, u_i, from U and then finds the x_i that satisfies $F(x_i) = u_i$ as in Fig. 6.2. The relationship between random variables and U and X is:

$$X = F^{-1}(U). \tag{6.21}$$

That this is a proper recipe is seen by noting that a draw of U can be below the horizontal dashed line of the figure $\{U < u_i\}$ if and only if the corresponding value of X is to the left of the vertical dashed line $\{X < x_i\}$. Since these two events have the same probability we can write, $\Pr\{X \leq x_i\} = \Pr\{U \leq u_i\} = u_i = F(x_i)$, which establishes that F(x) is the c.d.f. of X.

	A	B	C	D	E	F
1	EXAMPLE: SIMULATING A QUEUE					
2						
3	Poisson arrival rate (cust/sec) =			0.3		
4	Service time mean (sec) =			2		
5	Service time range (sec) =			0.5		
6						
7	RESULTS					
8	Avg. delay =	3.76 secs.				
9	Sd. delay =	2.29 secs.				
10						
11	Customer #	Arrival Headway	Arrival Time	Service time	Departure Time	Dwell Time
12						
13						
14	1	0.00	0.00	2.14	2.14	2.14
15	2	0.95	0.95	2.05	4.19	3.24
16	3	0.85	1.80	1.87	6.06	4.26
17	4	8.29	10.09	1.79	11.88	1.79
18	5	0.93	11.03	2.23	14.11	3.08
19	6	1.68	12.71	2.04	16.15	3.43
20	7	1.01	13.72	1.82	17.97	4.25

Figure 6.3 Spreadsheet organization of a simulation.

Example and discussion: Let us see now how the above-mentioned simulation can be implemented in a spreadsheet. The result of this process is displayed in Fig. 6.3.

As a prelude to the explanation, let us first derive a spreadsheet formula that generates a negative exponential draw with mean λ^{-1}, assuming that the arrival rate is stored in cell 'D3'. In this case the curve of Fig. 6.2 is: $u = 1 - \exp\{-\lambda x\}$ for $x \geq 0$; see Eq.(6.11). This equation can be solved for x to yield: $x = -(1/\lambda) \ln(1 - u)$ for $0 \leq u \leq 1$. Hence, the solution implied by Eq.(6.21) becomes: $-(1/\$D\$3) * @LN(1 - @RAND)$. Because $(1 - U)$ is also a uniform variable it is more elegant to write:

$$-@LN(@RAND)/\$D\$3, \tag{6.22}$$

which is the desired formula.

Let us also see how a random variable with a uniform density in the interval $[m \pm r/2]$ can be generated. (This procedure will be used for the service times.) We assume that the value of m is contained in cell 'D4' and the value of r in cell 'D5'. In this case the c.d.f. is $u = 0.5 + (x - m)/r$ for $x \in [m \pm r/2]$. Solving for x we find the intuitive result: $x = m + r(u - 0.5)$ for $0 \leq u \leq 1$. The formula in our spreadsheet would thus read:

$$\$D\$4 + \$D\$5 * (@RAND - 0.5). \tag{6.23}$$

We are now ready to write the simulation. In the implementation of

Fig. 6.3 it is assumed that the service times are uniformly distributed. Rows 1 through 6 contain labels and data, with cells D3, D4 and D5 reserved for the numerical inputs of the procedures that have just been described. Rows 7 through 10 are reserved for result summaries, and rows 11 through 13 for labels of the spreadsheet's columns. Column A contains the customer number n, column B formula (6.22) for the arrival headway H_n, column C formula (6.20a) for the arrival time A_n, column D formula (6.23) for the service time S_n, and column E formula (6.19) for the departure time D_n. To keep track of the customer waits we include a column (F) with a formula for $D_n - A_n$.

Rather than erasing the information for each customer once it has been calculated with a 'macro', it is easiest to keep it in memory. All one has to do is copy the second or third row of the spreadsheet over a range including as many rows as customers. In the implementation that has been made available to the public (as spreadsheet QUEUE.WK1) 1000 customers were simulated and the mean and standard deviation of our last column were written to cells B8 and B9.

It is instructive to plot A_n and D_n vs. n and see how these curves change with each 'recalculation'. One such picture is displayed in Fig. 6.4 for a range of 50 customers.[26] A recalculation, is equivalent to observing another realization of the process with new draws for all the random variables, i.e., viewing another day under the same initial conditions.

For the input data of Fig. 6.3 the relaxation time is on the order of 10^2 secs, or approximately 30 customers, which is much less than the 10^3 customers simulated. Thus, we would expect the simulated average delay (3.76 secs) to be representative of the true mean. The result, however, does not match well the prediction of Eq. (2.13) (2.5 secs.) for the choice of Δ that corresponds to our Poisson input and regular service process ($\Delta = 0.6$).[27] The discrepancy is not due to statistical error in the simulation, which will be shown in Sec. 6.2.3. to be less than ± 0.3 secs with high probability, but to the approximation in the theoretical formula which is usually comparable with the average service time (2 secs in our case). The exact result for this problem is actually 3.5 secs, which is within the error range of the simulation.[28] Repeated recalculations of the spreadsheet reveal that the simulated average delay indeed fluctuates about the theoretical mean. Thus, we have successfully simulated a queue! You may now want to change the input data and repeat the exercise with $\rho \to 1$. You will find that as the relaxation time approaches the length of observation (about 2000 secs) the simulated values underpredict more and more drastically the theoretical equilibrium average. This occurs, because the system has not had the time to 'forget' that it started empty.

Figure 6.4 Graphical result of a simulation run.

The same spreadsheet can be used to simulate a queue in which the arrival process follows an underlying schedule subject to uncertainty. These processes do not have independent increments and therefore the approximate queueing expressions (2.12) and (2.13) no longer apply. For regular schedules the variability in the counts N(t, t') when (t' − t) is large is much less than for processes with independent increments and equally variable headways. Therefore queues and delays are also much smaller. This can be verified with our simulation if we interpret column B as 'lateness' relative to the schedule and then include in the cells of column C a formula with the sum of the scheduled time for the customer and the corresponding cell of column B. For a homogenous schedule with the same arrival rate as before the formula for cell 'C15' should read: + B15 + A15/D3. The reader is encouraged to make this change and note the remarkable difference in the results of the two spreadsheets for $\rho \approx 1$.

6.2 Data interpretation

This section reviews those aspects of statistical theory that are most

relevant to the study of transportation and traffic operations. Whereas the theory of probability/stochastic processes is concerned with the evaluation of event frequencies given an underlying probabilistic model, the goal of estimation is the opposite. One is normally given some observations, or sample data, and from these one must guess (estimate) an underlying feature of the real world.

It goes without saying that the first step in a statistical study must be to define unambiguously the quantity or quantities to be estimated. Yet, perhaps because certain definitions of world quantities are ambiguous in subtle ways this is not always done, even in published works. Therefore, the importance of thoroughly looking at the question that is being asked before embarking in a statistical study cannot be stressed too much. Questions are sometimes improper because they do not specify the population of interest in cases where the answer depends markedly on it; two examples of this type are: 'what is the average car occupancy in Berkeley in a typical day?' and 'what is the average speed on San Pablo Ave?'. The first question is improper because car occupancy may depend on trip length. Therefore, the ratio of all the person-trips to car-trips made in Berkeley will in general be different of the ratio of person-miles to car-miles. The second question is improper for the same reason: it does not specify whether the average is taken across all drivers using the street in a day or across all driver-miles driven in the street. A properly defined statistical goal often has the added benefit of suggesting an experimental approach to sampling, e.g., whether passenger/car counting stations along a freeway corridor should be placed on the entry ramps or on the freeway links.

Because data gathering schemes are problem-dependent they can only be described partially. This will be done later, in Secs. 6.3 and Chapter 7, when addressing particular problems. The goal of the current section is only to present some elementary methods for estimating the average of a properly defined quantity whose experimental measurements fluctuate. The fluctuations could be due to measurement error, real world variations in the quantity, or both. Section 6.2.1 will introduce the main ideas, stressing situations with non-independent observations. Section 6.2.2 shall then discuss an application to the observation of queues and Sec. 6.2.3 some accuracy issues.

6.2.1 Estimation Concepts

For the remainder of this section we show how one can estimate the mean, θ, of a random variable, X, recognizing that the observations of X may be correlated.

Let us consider a sample of N observations (n = 1, 2...N) $\{x_n\}$, and write for each of these:

$$x_n = \theta + \varepsilon_n$$

where $\theta = E(X)$. Note that only the term x_n of this relation is observed. The 'parameter' θ is fixed but unknown, and ϵ_n is both variable and unknown. The latter varies across observations, and in so doing fluctuates around zero. Accordingly, one can take the view that each x_n is a measurement of θ with error ϵ_n.

We can think of a sample as a multidimensional random variable $\{X_n\}$ that takes a numerical value $\{X_n\}$ when one goes to the field and assembles the data. The sample mean $\bar{X} = \Sigma_n X_n/N$ (or any other function of the data) can likewise be considered as a random variable since its value would change across data sets; in other words, the particular value happening for our data set is simply an outcome of this random variable. In statistics, a random variable such as $\bar{X}$ that is used to estimate an unknown parameter is called an *estimator*; and its specific value for the sample at hand an *estimate*. A good estimator is one that produces estimates close to θ for most samples. This will happen if its mean is θ, when the estimator is said to be *unbiased*, and its variance is small. Let us now see how the mean and variance of $\bar{X}$ can be evaluated.

If one can write $\theta = E(X_n)$ for all the observations in the sample, as would be the case if the X_n were identically distributed, then the sample mean turns out to be an unbiased estimator for θ. This is true because:

$$E(\bar{X}) = E(\Sigma_n X_n/N) = \Sigma_n E(X_n)/N = \theta, \tag{6.24}$$

where the linearity of the expectation operation (6.2) justifies the second equality, and the definition of θ as the average of each X_n is the basis for the third one. If the observations X_n can be assumed to be mutually independent with the same variance σ^2, as would be the case if they were counts of a stationary process with independent increments, then the variance of $\bar{X}$ follows from (6.3) in a similar way:

$$\text{var}(\bar{X}) = \Sigma_n \frac{1}{N^2} \text{var}(X_n) = \frac{\sigma^2}{N}. \tag{6.25}$$

Under these conditions we also expect $\bar{X}$ to be normally distributed for large samples because of the central limit theorem. This means that the absolute estimation error in $\bar{X}$, $|\bar{X} - \theta|$, should be less than $2\sigma N^{-1/2}$ for most samples, and that the quantity $\sigma N^{-1/2}$ can be interpreted as a *typical* or *standard error*.[29]

6.2.1.1 Correlated samples

The typical error may be larger or smaller than $\sigma N^{-1/2}$ if the observations in our sample are not independent. This might come about for example if one takes measurements of a steady state queue (e.g., its length) at time intervals that are closely spaced relative to the relaxation time. It should be easy to appreciate that in such an example observations will tend to be serially correlated, with large observations usually followed by large observations and viceversa. Because (6.3) does not apply in such a case, a correction to (6.25) is needed. Equation (6.24) on the other hand still holds true because (6.2) does not rely on independence. Thus, $\overline{X}$ is unbiased but its variance needs to be determined..

To this end note that if the unknown constant θ is subtracted from all the data the means of $\overline{X}$ and X_n become zero but var(X_n) and var($\overline{X}$) $\equiv$ var($\overline{X} - \theta$) remain the same. Therefore, we can assume that $\theta = 0$ in a general derivation of var($\overline{X}$), and the result still is the variance of the error in the general case. For zero mean, a variance is the average of the squares and we can write:

$$\mathrm{var}(\bar{X}) = E(\bar{X}^2) = E\left(\left(\Sigma_n \frac{X_n}{N}\right)\left(\Sigma_m \frac{X_m}{N}\right)\right).$$

After expanding the product of the sums appearing in the last member of the above equalities, and remembering that the average of a linear function is a linear function of the average, we find:

$$\mathrm{var}(\bar{X}) = \frac{1}{N^2} \sum_n \sum_m E(X_n X_m). \tag{6.26}$$

If the expectations in (6.26) only depend on the difference between m and n, as if the sample was obtained by sampling at regular intervals a stationary process, the following series of unknown constants can be defined:

$$E(X_n X_m) = c_l = \sigma^2 \rho_l \quad (l = m - n) \tag{6.27}$$

where the ρ_l are dimensionless.[30] For the present case, where $E(X_n) = E(X_m) = 0$, these constants are called *correlations* and the products $c_l = \sigma^2 \rho_l$ *covariances*. It should be clear that $\rho_0 = 1$ and that ρ_l and ρ_{-l} must be equal. It also turns out that $c_l = \rho_l = 0$ when X_n and X_m are independent.[31]

In applications where ρ_l is negligible if $|l|$ is greater than some integer I, Eq.(6.26) can be rewritten as follows when the sample is so

large that $N/I \gg 1$:[32]

$$\text{var}(\bar{X}) = \frac{1}{N^2} \sum_n \sum_{l=-\infty}^{\infty} \sigma^2 \rho_l = \frac{1}{N^2} \sum_n \sigma^2 R = \sigma^2 \frac{R}{N} \qquad (6.28)$$

where R is used to denote the sum of the correlations: $R = \Sigma_l \rho_l$. Note that it is consistent with (6.25) because for independent variables $R = 1$.

In practical applications one is unlikely to know the quantities σ^2 and/or R appearing in (6.25) or (6.28). Therefore, we now explore the consequences of using certain guesses in their place. Let us take up first the case of an identically distributed sample without correlations, where $\rho_l = 0$ if $l \neq 0$, and again imagine that $\theta = 0$. Then (6.26) becomes:

$$\text{var}(\bar{X}) = \frac{1}{N^2} \Sigma_n E(X_n^2) = \frac{1}{N} \left[\frac{1}{N} \Sigma_n E(X_n^2) \right] = \frac{1}{N} E(X^2) \quad (\textit{for } \theta = 0),$$

where X denotes one generic observation of the sample. The third member of these equalities can be evaluated approximately for large N by inserting the square of each sample value in place of each expectation. The central limit theorem then guarantees that the approximation to the bracketed expression (i.e., the arithmetic average of the sample squares) is close to $E(X^2)$, and thus that the result is close to the fourth member of the above equation. Therefore,:

$$\text{var}(\bar{X}) \approx \frac{1}{N^2} \sum_n x_n^2. \quad (\text{for large } N, \text{ and } \theta = 0)$$

In the general case with $\theta \neq 0$, x_n should be replaced by $(x_n - \theta)$ in the above expression. And since θ is unknown, θ too should be replaced by the sample estimate, $\bar{x} \approx \theta$. Thus, we can use:

$$\text{var}(\bar{X}) \approx \frac{1}{N} \left\{ \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2 \right\}. \quad (\text{for large } N \text{ and } \theta = 0) \quad (6.29)$$

The term in braces, usually denoted by the symbol S^2, is called the *sample variance*. Although S^2 varies across samples, we have just seen that it will likely be close to σ^2 when N is large.

If there are sample correlations, a guess for the correlation sum R appearing in (6.28) is also needed. We study here the case where the correlations have a limited range, $\rho_l \approx 0$ for $|l| > I$, and where the sample is so large that $N/I \gg 1$. As before, one can estimate the covariance c_l from the average of the relevant products in our sample;

i.e., by:

$$s_l = \frac{1}{(N-l)} \sum_{n=1}^{N-l} (x_n - \bar{x})(x_{n+l} - \bar{x}) \approx c_l \quad \text{(for large } N-l\text{)} \quad (6.30)$$

where the substitution $\theta \Leftrightarrow \bar{x}$ has already been made in this expression. (Note that $s_0 = S^2$.) If $\sigma^2 R$ is now approximated by the sum of the relevant s_l we can write:

$$\text{var}(\bar{X}) \approx \left[\sum_{l=-I}^{I} s_l \right] / N \quad \text{(for } N/I \gg I\text{)}. \quad (6.31)$$

In this formula the quantity in brackets may vary across samples significantly if N/I is not large.[33] Thus, one should limit to the extent possible the number of terms included in the numerator of (6.31). Improved estimates of var($\bar{X}$) can be obtained if ρ_l is known to have a particular functional form (e.g., exponential) and the theories of *time-series analysis* and *filtering* can be useful in this respect.

If you are going to analyze time-series data, it is also useful to develop a feel for the 'look' of data sets with various levels of (auto) correlations and how these translate into increased errors. To this end, Figs. 6.5, 6.6 and 6.7 depict three data sets arising from a model with $E(X_n) = 0$, $\text{var}(X_n) = 1$ and $\rho_l = \rho^l$ (for $\rho = 0$, 0.4 and 0.8). The figures also apply to other values of $E(X_n)$ and $\text{var}(X_n)$ if we just imagine the ordinate axis to give the data in terms of number of standard deviations away from the mean. In comparing the time series data (part (a) of each figure) we see that as ρ is increased the data tend to carve out a narrower swath on the picture. Similarly, when we look at the lag-1 scatterplots (part (b) of each figure) we also see that the observations tend to bunch more closely around the diagonal for the higher ρ's. For our particular form of the auto-correlations the correlation sum is $R = \Sigma_l \rho^l = (1+\rho)/(1-\rho)$, which yields R = 1, 2.33 and 5 for each one of the figures. This means that correlations such as those seen in Fig. 6.6 increase the standard error in $\bar{x}$ by a factor of $(2.33)^{1/2} \approx 1.5$ and those of Fig. 6.7 increase it by a factor of $5^{1/2} \approx 2.2$. Alternatively, for the same level of accuracy, one would have to increase the sample size by factors of 2.33 and 5 respectively. These results indicate that effect of correlations should be insignificant if the 'eye' cannot see them.[34]

6.2.2 Illustration: observation of stationary processes and queues

This subsection may be skipped on a first reading. It shows that if a

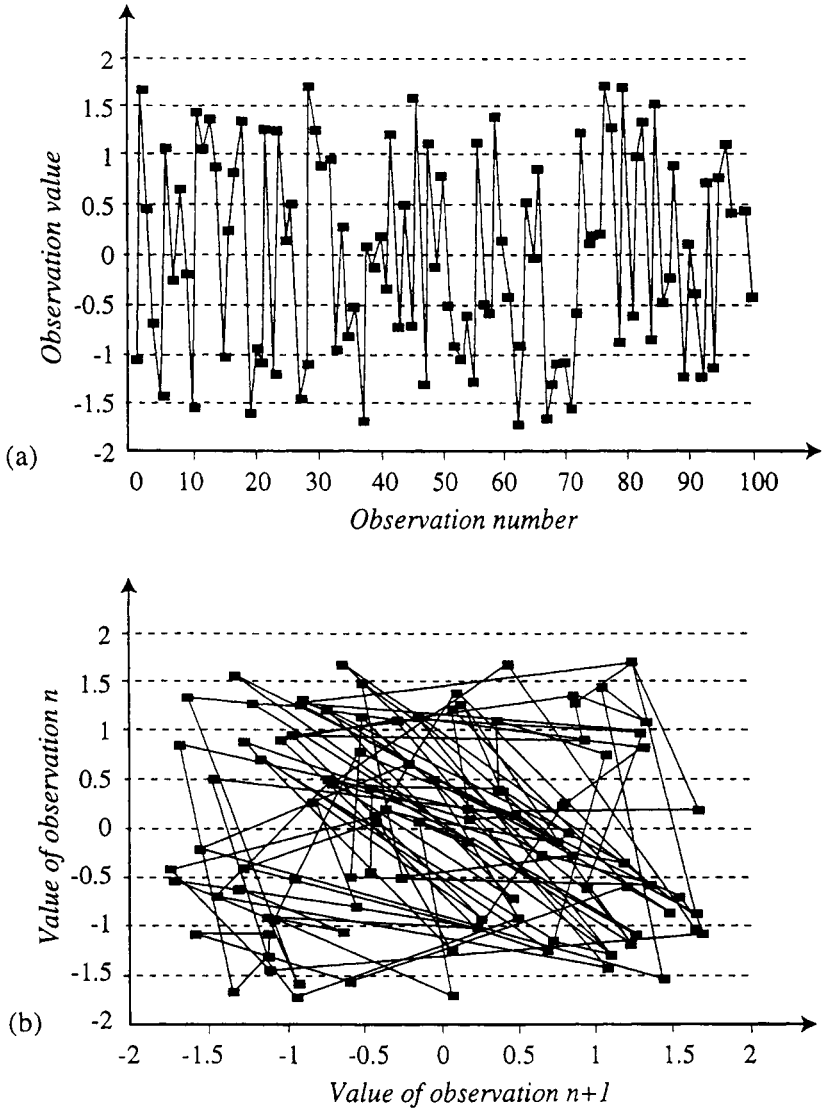


Figure 6.5 Identically distributed serial observations with zero correlation: (top) 'time' series; (bottom) lag-1 scatterplot.

stationary stochastic process in continuous time {X(t)} is observed for a fixed amount of time T, then the sample variance of N evenly spaced observations, $\{X_n = X(n\Delta t)\}$ where $n = 1,2...\ N$ and $\Delta t = T/N$, ap-

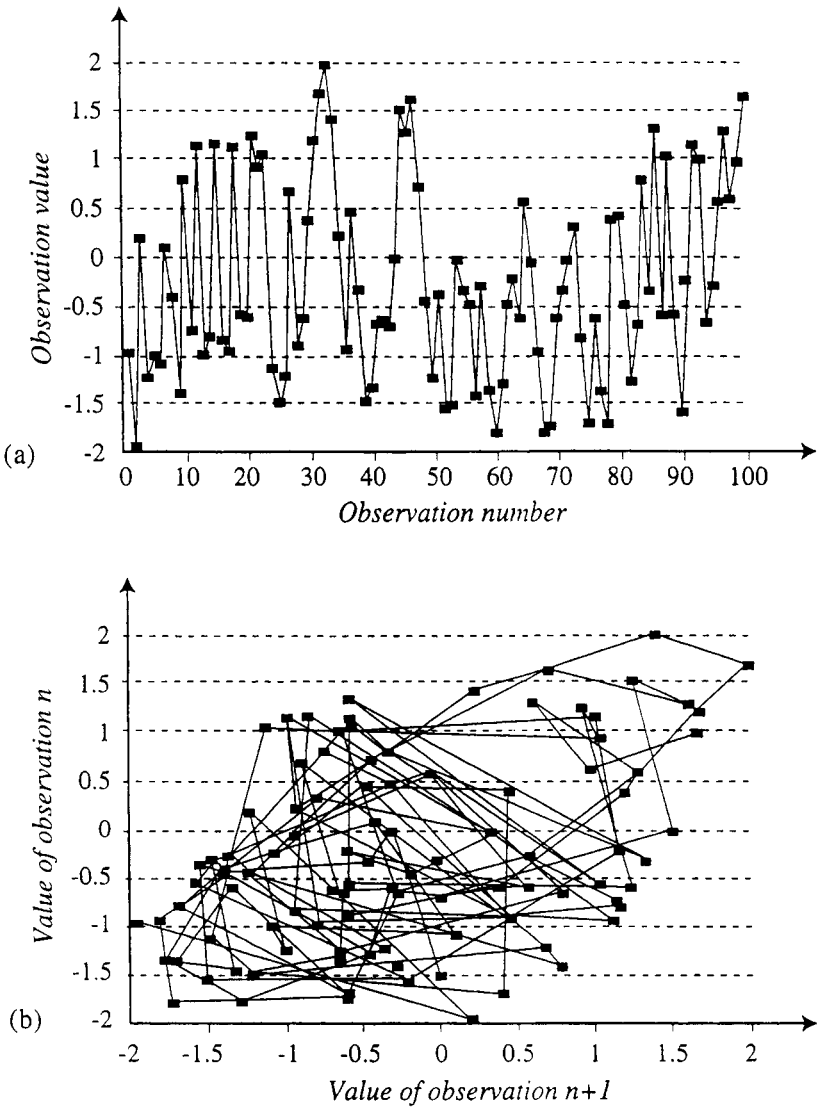


Figure 6.6 Identically distributed observations with an exponentially decaying correlation and lag-1-correlation = 0.4: (top) 'time' series; (bottom) lag-1 scatterplot.

proaches a non-zero limit as $N \to \infty$ and $\Delta t \to 0$. This means that the mean E(X) of a stationary process cannot be estimated arbitrarily

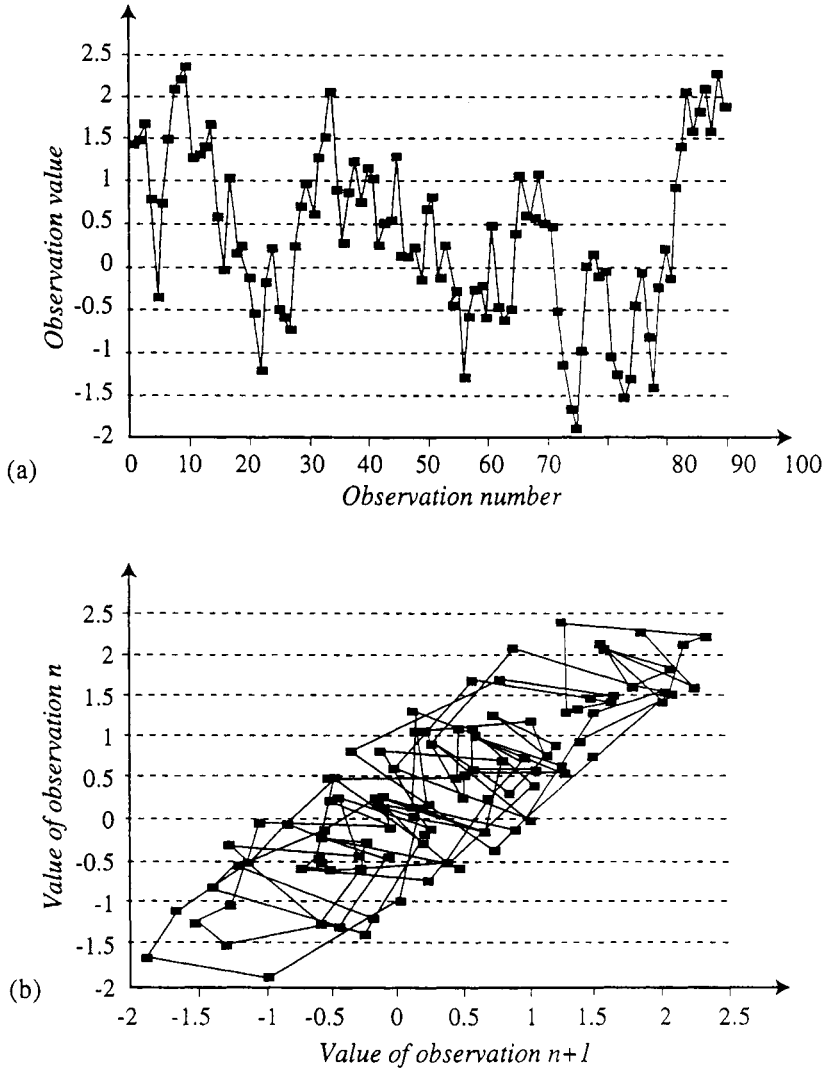


Figure 6.7 Identically distributed observations with an exponentially decaying correlation and lag-1 correlation = 0.8: (top) 'time' series; (bottom) lag-1 scatterplot.

accurately by just increasing the sampling frequency if the period of observation is fixed. The estimation error can be reduced as much as desired by increasing N for a given Δt, however. A formula for the

CHAPTER SEVEN

Scheduled transportation systems

Let us now turn our attention to 'scheduled transportation systems.' By this we mean systems with fixed routes and schedules whose moving components are operated by an external authority. Examples are subways, airlines, the postal system, railroads, and shipping lines. Agency-operated systems with flexible routes and schedules such as ambulances, taxicabs, paratransit and couriers are not discussed here.[1]

Scheduled systems should be further classified as being either passenger or freight conveying because the items using the system (e.g., people, parcels, containers, etc.) choose their own routes in the first case but not in the latter. For freight transportation systems the external authority typically chooses the routes and schedules of both the vehicles and the items, and its overall goal is well defined: e.g., minimizing the overall logistics cost for a given transportation task. Optimization techniques can then be used for system design and control. For passenger transportation the situation is more aptly described as a game in which the authority chooses the system structure, and then passengers find their best individual routes and times of departure. Involving numerous players, the authority's overall goal is more difficult to quantify in this case, and the system design/control game is just as complicated as for unscheduled transportation modes (see Sec. 5.5). Therefore, one needs to watch out for the same type of pitfalls.

As in the case of unscheduled transportation modes, it is beyond the scope of this book to examine the overall structure of these systems. Now too, only an elementary description of the system's basic building blocks will be given. This will be accomplished rather briefly because all the relevant methodologies (time-space diagrams, cumulative plots, simulation, etc.) have already been introduced.

The presentation is further abbreviated by not covering issues that are simple extensions of things the reader already knows. Most notably, the chapter does not describe recipes for estimating the vehicular capacity of a right-of-way or the feasibility of a schedule because the details vary greatly from mode to mode and an elementary application of the time-space diagram readily yields the desired results in almost all

cases.[2] This is true, for example, if one wishes to determine the maximum vehicular flow that can be accommodated by a link or series of links when vehicles must undergo certain motions (e.g., stop at certain location for specified time duration) while satisfying the particular minimum separation rules of the system in question. This would have to be done in order to find the maximum number of BART trains that can operate per hour on a given route, the maximum number of airplanes that can use an airport runway for both landings and departures, and the maximum number of ships that can move through a system of locks. The (t, x) diagram is also useful to determine whether a particular timetable satisfies the safety requirements. This was illustrated in Sec. 1.3.3. with an example that showed how to develop a two-way schedule on two 'single track' links connected by a two-way 'siding.'

In view of this, this chapter only describes how passengers/freight interact with a scheduled system, and what an agency must do to operate on a chosen schedule. Only simple systems in which route choice is not an issue are treated. The chapter also introduces some basic ideas on control, design and evaluation that the reader may find intriguing. Section 7.1 examines customer waits caused by the discreteness of a schedule. Section 7.2 shows how the number and size of the vehicles on a multi stop route should be chosen in order to achieve a given service frequency, and then explains why, when and how it is necessary to build a fair deal of 'slack' into the schedule. Section 7.3 shows how the necessary performance parameters may be estimated from data, and Sec. 7.4 discusses some design and evaluation issues. Rather basic, the latter ideas also apply to unscheduled transportation systems.

7.1 Passenger waiting time

This section considers the time spent waiting by the items using a scheduled transportation system. It shows for example that the average passenger waiting time at a stop with stationary arrivals and uneven headways is always *greater* than one half of the average headway. The discussion is usually phrased in terms of 'passengers,' 'buses' and 'bus stops,' as in the above statement, but it should be remembered that everything said applies to other transportation modes if these words are replaced by pertinent ones, for example 'containers,' 'ships' and 'ports.'

In addition to waiting, a customer's trip usually includes some in-vehicle travel and some access (walking). The in-vehicle travel time is not studied now because it is just a weighed sum of the vehicular trip

times between relevant stops, which will be discussed in Sec. 7.2. The access time is not examined separately either because it is related to the location of the stops much as the waiting time is related to the schedule,[3] and because it can be calculated easily if the distribution of the population along the line is given. The waiting time on the other hand involves enough complicating issues, e.g., deviations from a schedule, transfer passenger waits and schedule coordination, to deserve a more detailed treatment.

Before starting the technical discussion it should be noted that the observed wait at a passenger's inbound stop and any transfer points does not include any delay that the passenger may accrue at the destination because the bus does not arrive there exactly when the passenger wants. Although important this 'exit' delay is often ignored in passenger transportation studies because (i) travelers cannot be observed easily after leaving the system, and (ii) they usually prefer not to wait for their scheduled activity at the exit bus stop. Transit patrons may avoid exit delays if the activity at the trip end does not require a precise arrival time; e.g., for the trip to work of an employee that has 'flex-time' or the reverse trip home if the worker does not have scheduled activities such as picking up children from the nursery at a specific time. Otherwise we would expect this 'exit' delay to be comparable with a headway and similar to the inbound 'waiting' experienced by passengers that do not know the schedule.[4]

Waiting time for uninformed passengers: The formula for the mean (inbound) waiting time can be derived with the help of the cumulative count diagrams of Fig. 7.1. The top part of the figure displays the special case where the passengers arrive at a constant rate λ, without fluctuations. Arrows on the time axis display the bus departure times and the headways $\{h_k\}$. This construction ignores the bus loading time, and assumes that nobody is left behind after a bus departure. This is why the departure curve increases in steps at the times indicated by the arrows all the way up to the arrival curve.

The total waiting time, e.g., in passenger-hours, accumulated in the time period corresponding to the first 'K' headways can now be written as the sum of the triangular areas in the figure; i.e.:

$$W(T) = \sum_{k=1}^{K} \frac{1}{2} \lambda h_k^2, \tag{7.1}$$

where T denotes the duration of the period, $T = \Sigma_k h_k$. Similarly, the total number of passengers served, N(T), can be expressed as:

$$N(T) = \sum_{k=1}^{K} \lambda h_k. \tag{7.2}$$

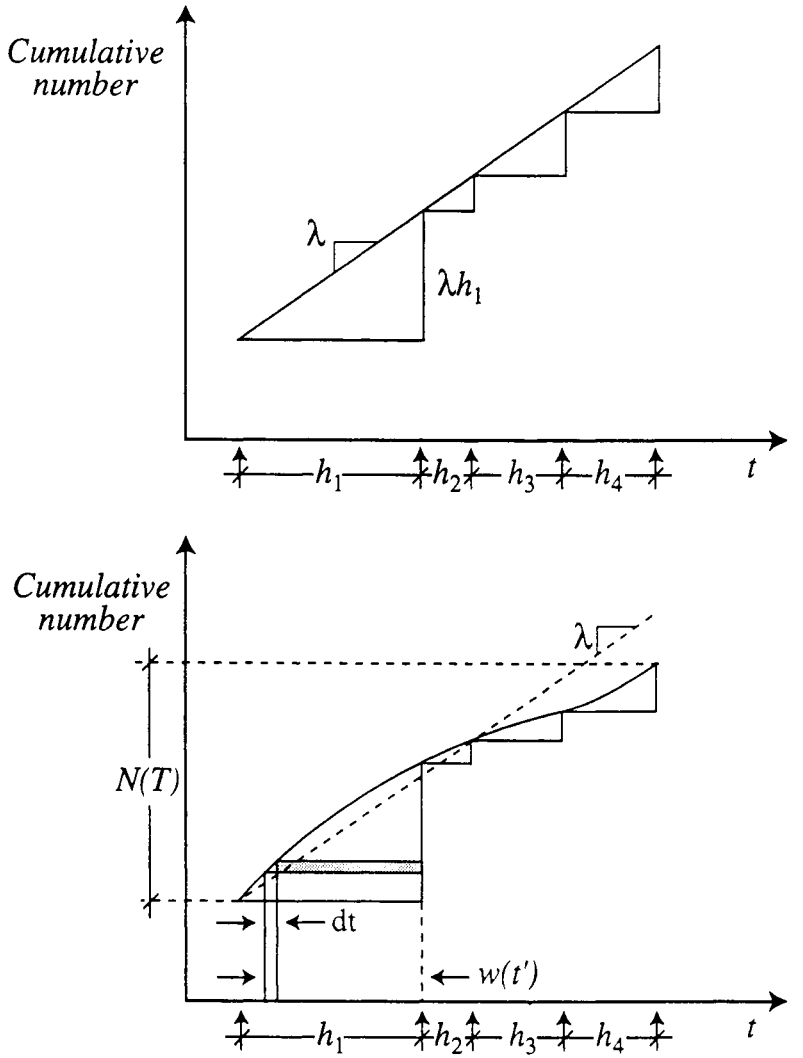


Figure 7.1 Cumulative curves of passenger count at a bus stop. Top: constant arrival rate; bottom: variable arrival rate.

The average waiting time of the observed patrons, $\bar{w}$ is given by the ratio W(T)/N(T); i.e.:

$$\bar{w} = \tfrac{1}{2}\Sigma_k h_k^2 / \Sigma_k h_k. \tag{7.3a}$$

Note that it is independent of λ.

If an over-bar is used to denote the arithmetic mean, then (7.3a) may also be written as:

$$\bar{w} = \tfrac{1}{2}\left[\overline{h^2}/\bar{h}\right]. \tag{7.3b}$$

And since the average of the squares is the sum of the sample variance and the square of the average we can also write:[5]

$$\bar{w} = \tfrac{1}{2}\left[\bar{h} + S_h^2/\bar{h}\right] \geq \tfrac{1}{2}\bar{h}, \tag{7.4}$$

where S_h^2 is the sample variance of the $\{h_k\}$. This shows that $\bar{w}$ can equal $\frac{1}{2}\bar{h}$ only if $S_h^2 = 0$; i.e., only if the service is perfectly regular with $h_k \equiv \bar{h}$.

An explanation for inequality (7.4) may be found in the fact that longer headways entrap more passengers and therefore must receive a larger weight in the computation of average wait. This is yet another example of the so-called 'length-biased sampling' mentioned in earlier chapters. The length-bias effect is one of the reasons why transportation agencies and firms try to operate on regular schedules.

It will now be shown that the same formulas apply to the average wait across many observation periods (e.g., days) despite fluctuations in the arrival process, provided that this process is stationary in the observation interval of interest, $(t_0, t_0 + T)$. To this end, consider Fig. 7.1 (bottom), which shows by means of a curved line a realization of the process on one day, and by means of a slanted dashed line its average across many days. Let us also use $N(t_1, t_2)$ to denote a random variable that gives the number of arrivals in a given time interval (t_1, t_2) for every realization (day). As before, W(T) and N(T) are used for the (random) total wait and total number of passengers in the complete observation interval, $(t_0, t_0 + T)$. We stress that both W(T) and N(T) may vary from day to day but that they are deterministically related to the arrival curve. In particular, W(T) can be expressed as the sum of the areas of horizontal slices of the cumulative count diagram such as the shaded region on the bottom part of the figure. This area, which corresponds to the arrivals in $(t', t' + dt)$, is:

$$\text{area of slice for } (t', t' + dt) = w(t')N(t', t' + dt) \tag{7.5}$$

where $w(t')$ is a function that gives the waiting time that would be experienced by an arrival at $t = t'$. This function is a property of the schedule and is not random; the only part of (7.5) that varies from day to day is $N(t', t' + dt)$. Therefore, W(T) is the following linear combination of the random counts:

$$W(T) = \int_{t_0}^{t_0+T} w(t')N(t', t' + dt); \tag{7.6}$$

and N(T) is also linear in the counts:

$$N(T) = \int_{t_0}^{t_0+T} N(t', t' + dt) \tag{7.7}$$

The linearity of the expectation operation (6.2) allows us to replace the N(t', t' + dt) in these formulae by their averages, which equal λdt, in a calculation of the averages of W(T) and N(T) across many days. This, of course, is the same as replacing the realized arrival curve by its straight-line average across days in the computations, which would yield (7.1) and (7.2) again. If we now denote by E(w) the expected wait of a randomly selected passenger, i.e., the ratio of the total wait after an infinite number of days to the total number of passengers served, we see on dividing the numerator and denominator of this ratio by the number of days that E(w) is also the ratio of the expectations of W(T) and N(T). Therefore, (7.3) and (7.4) still hold for the expected wait of a randomly chosen passenger.

In the above derivation the assumption of stationarity was only used to replace N(t', t' + dt) by λdt in (7.6) and (7.7). In a general case one would simply have to replace it by the time-dependent expectations, $\lambda(t)dt$, which again would be equivalent to replacing the solid curve in Fig. 7.1 (bottom) by its average across many days and repeating the construction of the figure. This means that the average delay in the general case can be calculated rather simply and without much input data. For example, in the special case where $\lambda(t)$ varies slowly from headway to headway and where the headways are similarly distributed during the day, it turns out that (7.4) still holds approximately, even if the total variation in λ during the observation period is very large.[6]

The c.d.f. of the waiting time across all customers, $F_w(w)$, can be obtained from the given schedule and $\lambda(t)$ by evaluating the fraction of all the customers that arrive at a time for which the wait does not exceed w. If the expected demand rate is independent of time, this fraction equals the proportion of the time during a long observation interval in which an arrival would wait w or less. This happens for a time w inside headways longer than w, and for the whole duration of the headway in headways shorter than w. Therefore, the total time is $\Sigma_k \min\{w, h_k\}$, and $F_w(w)$ is:

$$F_w(w) = \Sigma_k \min\{w, h_k\} / \Sigma_k h_k \tag{7.8}$$

if the expected demand is stationary.

It has been assumed so far that vehicles are large enough to accommodate everyone who arrives in any headway. If overflows due to fluctuations arise from headway to headway, the relationship between

W(T) and N(T) and the arrival counts is no longer linear and the delay for a realization with typical fluctuations will tend to be larger than the delay without fluctuations. This statement should not come as a surprise after our discussion of the (analogous) traffic signal problem with fluctuations given in Sec. 5.2.[7]

Advertised schedules: For scheduled transportation systems with infrequent service, such as airlines, the observed wait can be reduced greatly by advertising the schedule and sticking to it. Passengers and/or freight may then choose to arrive just in time to meet their vehicle.[8] It is shown below that if passengers choose their arrival times so as to minimize their wait, then their average waiting time is proportional to the average vehicular deviation from the schedule. This general result explains the importance of transportation service punctuality.

It is assumed in the explanation that vehicles are not allowed to arrive before their scheduled time and that the deviations from the schedule can be modeled as draws from a random variable S, called the *schedule lateness*. As we shall see in Sec. 7.2, the lateness of successive departures at a given stop tend to be negatively correlated, and this will be taken into consideration. It is also assumed that S never exceeds the length of a scheduled headway and that its density function, $f_S()$, is non-increasing; i.e., that longer latenesses are less probable than short ones, as shown in Fig. 7.2. It is finally assumed that each passenger chooses an arrival time τ in a scheduled headway $[0, h_k]$ so as to minimize his/her expected wait, $E(w(\tau))$, taking into account the possibility of missing the 'bus.'[9] In order to stress that this expectation is a function of τ, it will be written $w_{avg}(\tau)$.

Let us now see how to determine the chosen τ by looking at the derivative of $w_{avg}(\tau)$ in the usual way, and then evaluate the corresponding w_{avg}. A perturbation argument is used to write this derivative directly, without first writing the formula for $w_{avg}(\tau)$ in terms of integrals. The result, Eq. (7.9) below, can also be derived in the standard more laborious way.

If a passenger delays her arrival time from τ to $\tau + d\tau$, as shown in Fig. 7.2, and bus $(k-1)$ arrives outside the time interval $(\tau, \tau + d\tau)$ then the passenger has won her bet and saves $d\tau$ units of wait. This is true whether she was early or late to begin with. Otherwise, with a probability given by the shaded area $f_S(\tau)d\tau$, the passenger wait increases by $\delta(\tau)$. This is the expected time until the next bus arrival, k, when it is known that bus $k-1$ arrived in $(\tau, \tau + d\tau)$. It follows that the expected change in wait is (to a first order in $d\tau$):

$$dw_{avg}(\tau) = -d\tau + \delta(\tau) f_s(\tau) d\tau,$$

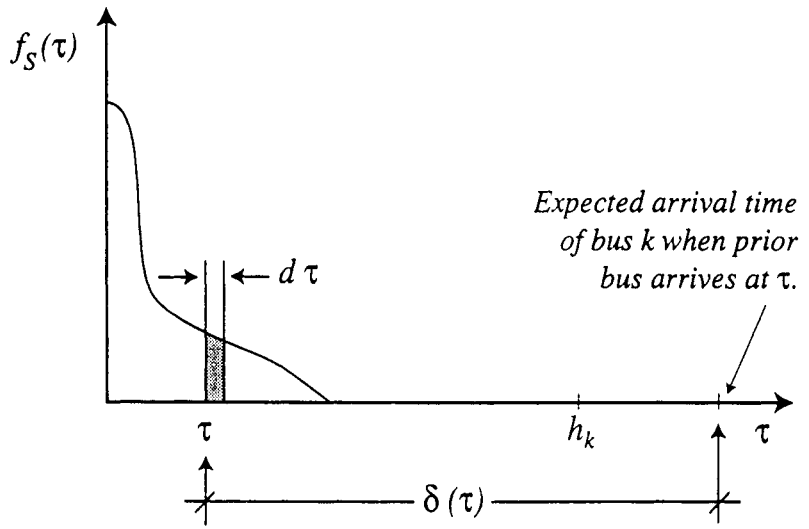


Figure 7.2 Relationship between the lateness density function, and the expected delay of a passenger arriving at time τ.

or

$$dw_{avg}(\tau)/d\tau = -1 + \delta(\tau)f_s(\tau). \tag{7.9}$$

Note that the function $\delta(\tau)$ decreases if the lateness of bus k is independent of S_{k-1}, and also if an increased lateness for bus $(k-1)$ induces a *reduced* lateness for k as may happen if buses have a tendency to pair (see Sec. 7.2). As a result, the second derivative of $w_{avg}(\tau)$ satisfies:

$$d^2w_{avg}(\tau)/d\tau^2 \equiv \delta df_s/d\tau + f_s d\delta/d\tau \leq 0.$$

The inequality is justified because δ and f_S are nonnegative and decreasing. This means (see Chap. 3) that $w_{avg}(\tau)$ is concave, and therefore that it must have a global minimum at one of the extreme points of the interval. Since the wait at these two points is equal, $w_{avg}(0) = w_{avg}(h_k) = E(S)$, either point is acceptable. According to this theory, thus, passengers should arrive shortly before the scheduled time and experience $E(w) \approx E(S)$.[10] Bowman and Turnquist (1981) compared the waiting time predictions of a model similar to the one just presented with those observed in the Chicago transit system and reported good agreement with the data.

Transfers: Another important case where one cannot assume that the arrivals are independent of the schedule arises when they come from another system that is synchronized with ours. In that case it may be reasonable to describe the passenger cumulative arrival curve at our (transfer) stop by a step function that increases by amounts $\{n_k\}$ at times $\{a_k\}$. These times should match the departure times of the connecting system $\{d'_k\}$, except for a delay that should be comparable to the walking time, t_w, from one vehicle to the other; i.e., $a_k \approx t_w + d'_k$. If one is also given the departure schedule of our bus line $\{d_k\}$, then it is possible to calculate the total and average delay of the transferring passengers with the usual graphical construction.

Fig. 7.3 is the result for a case where vehicles have sufficient capacity to clear the queue after every departure. In this case the delay on any given day may be expressed as:

$$W(T) = \Sigma_k n_k w_k \tag{7.10}$$

where w_k is the time elapsed between a_k and the next departure.

If vehicles adhere to the schedules so that the w_k are fixed, the only random variables in Eq. (7.10) are the n_k. Since (7.10) is a linear function of the n_k, it also applies to the average W(T) if the random n_k are replaced by their averages.

Of particular interest is the case where the two schedules are regular with headways h′ for the arriving batches and h for the departing trains because if h′ and h are both integer multiples of a time quantity, h″, then it is possible to synchronize the schedules to reduce delay. Fig. 7.3 displays the case with $h' \equiv h''$ and $h = 2h''$. We can see at once from the picture that if the positive offset between the two schedules (defined as the smallest w_k, w_{min}) is reduced a little by displacing one of the schedules relative to the other, then all the w_k are reduced by the same amount. Thus, a minimum delay is obtained when the offset is zero. For the example of Fig. 7.3 this selection yields $w_0 = w_2 = w_4 = \ldots = h' = h/2$ and $w_1 = w_3 = w_5 = \ldots = 0$. The minimum average delay is then $h/4$ if the n_k, or their averages across realizations, are independent of k.

It is a simple matter to show from these graphical arguments that one should always choose a zero offset, and that the resulting average delay is:

$$E(w) = \tfrac{1}{2}(h - h''), \tag{7.11}$$

if the $E(n_k)$ are independent of k. By contrast, if connecting passengers had arrived independently of the schedule the average delay would have been $h/2$. Therefore, we see that coordination can save $\frac{1}{2}h''$ time units, with a maximum benefit achieved for $h'' = h$; i.e., when the arrival

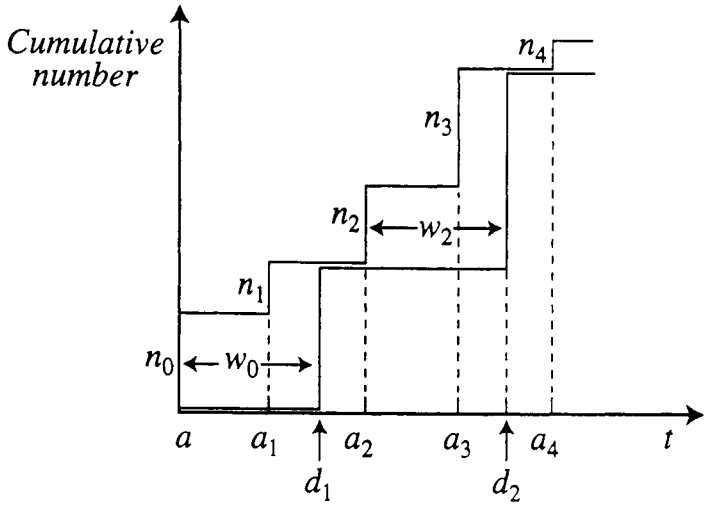


Figure 7.3 Cumulative arrival and departure curves of passengers transferring between two transit lines.

headway is an integer multiple of the departure headway. Waiting can then be eliminated altogether.

In order to ensure that significant savings are achieved at transfer terminals of scheduled systems involving many lines with different headways it makes sense to pick operating headways for each line from a menu of the form $\{\tilde{h}2^p\}$ with p integer. This ensures that every headway is either an integer multiple or submultiple of every other headway. Hence, if the offsets between all line pairs are set to zero, e.g., by forcing all the schedules to coincide at some time, one of the transfer directions between any two lines will always have zero wait. Remarkably, this menu still retains enough flexibility for choosing cost-effective service frequencies.[11]

An interesting optimization problem that cannot be addressed here in detail consists in finding the offset between two schedules $\{d_k\}$ and $\{d'_k\}$ that minimizes the total wait for two-way transfers, while recognizing that people take t_w time units to walk in either direction; i.e., that $a_k = d'_k + t_w$ and $a'_k = d_k + t_w$. A simple practical solution can be obtained for systems with long headways relative to t_w if the scheduler has the option to delay the vehicles of both lines for a short while in order to allow the connections to be made. In this case one should distinguish between the times when a vehicle, k, opens its doors, d_k, and the retarded times, r_k, at which it closes them. If $(\epsilon, \epsilon') \geq t_w$ are the

vehicle delays in this scenario, then $r_k = d_k + \epsilon$ and $r'_k = d'_k + \epsilon'$. Therefore, if the door-opening schedules $\{d_k\}$ and $\{d'_k\}$ are synchronized with zero offset as was just explained, the actual schedule pairs $\{a_k\}$, $\{r_k\}$ and $\{a'_k\}$, $\{r'_k\}$ will be similarly synchronized, albeit with (small) offsets $\epsilon - t_w$ and $\epsilon' - t_w \approx 0$. The net result is that all transferring passengers save a time comparable with $\frac{1}{2}h''$ (experiencing virtually no wait if $h = h' = h''$) but through passengers experience $(\epsilon, \epsilon') \approx t_w$ units of increased trip time. Obviously, the strategy will be appealing if the proportion of transfers is large and $h'' \gg t_w$, i.e., for systems with long headways such as airlines and other intercity transit and freight transportation systems.

Let us now change our focus from passengers to vehicles.

7.2 Multi stop routes

Here we shall determine the type and number of vehicles that are needed to meet a target schedule on a given route (Sec. 7.2.1), and how they should be operated so as to ensure that the service is punctual (Sec. 7.2.2). We will see among other things that the three main determinants of fleet and vehicle size are the route 'cycle time', the operating frequency and the O-D table, and that a certain amount of 'slack' must be built into the typical schedule so as to avoid instabilities.

7.2.1 The vehicle fleet needed for a given task

This subsection examines the performance of a single closed-loop route with n stops that is served with regular headways by V vehicles of capacity C. We are particularly interested in the values of V and C that are needed to serve a given set of origin-destination demands with a given headway, h. As in Sec. 7.1 a bus route will be used as our metaphor, but the results are general. They apply to any transportation mode with a regular schedule on a fixed route; e.g., to rapid transit systems, sea routes covered by liner-type container ships and to some trucking, railroad and airline services. It will be assumed that the vehicles do not skip stops for lack of demand, although this excludes bus routes with low demand from the domain of application. A more general derivation could be given but this is beyond the scope of an introductory book. It will also be assumed that vehicles keep to their schedule, although this may require a control scheme such as the one described in Sec. 7.2.2.

Fig. 7.4 depicts the (t, x) trajectories of three buses that serve a 3-stop loop. The space axis ranges from $x = 0$ to $x = L$ and contains the

locations of the three stops, x_i ($i = 1, 2, 3$). Because the route is a closed loop, each vehicle trajectory reappears at $x = 0$ immediately after crossing $x = L$. Buses cruise at an average speed v_i on the link leading to stop i, and take T_C time units to complete a cycle.[12]

Accurate bus trajectories near each stop would have to include curved sections corresponding to acceleration and deceleration portions, but this is only shown for bus 3 in the neighborhood of $x = x_1$. Note that the trajectory of bus 3 at future times remains the same if we replace the curved sections near x_1 by the piecewise linear dotted lines shown; i.e., if we imagine that bus 3 changes speed instantaneously and remains stopped for a longer time. Because it is simpler to work with piecewise linear trajectories, the rest of the diagram has been constructed in this manner. The times t_i shown in the figure, which will become an important part of the calculations, should therefore be interpreted as the delay experienced by a bus that stops at i.

This delay includes two parts: a variable component that elapses while passengers are entering and/or leaving the bus, and a fixed part τ_{0i} that arises from: (i) the acceleration and deceleration maneuvers, (ii) the time needed to open and close the doors, and (iii) the time spent waiting for the passengers to be seated. Components (ii) and (iii) of the fixed delay are shown by darker sections on the figure.

The fixed delay τ_{0i} varies greatly from mode to mode, being comparable with one hour for airlines and on the order of seconds for buses. The subscript 'i' is used in connection with this parameter because it can change across stops. Accurate estimates for τ_{0i} can be obtained if one knows the acceleration and jerk characteristics of the vehicle in question as well as the guideway geometry near the stop.

It will be assumed below that the variable portion of the bus delay is proportional to the number of passengers or items using the stop, although this may not be very accurate for passenger systems in which separate doors are used for boarding and alighting, or for freight systems (such as container seaports) whose cranes may 'double-cycle.' In these instances it may be better to express the variable delay as a term proportional to the maximum of the times needed to load inbound and unload outbound passengers/items.

7.2.1.1 The stationary, deterministic problem

In a general case, the origin-destination demand data for our problem is a set of cumulative count curves that give the number of people, $N_{ij}(t, t')$, who arrive at origin i for destination j in the time interval (t, t'). It is assumed for now that the $N_{ij}(t, t')$ are known and that the time-dependent flows $q_{ij}(t)$ corresponding to the $N_{ij}(t, t')$ vary little during time

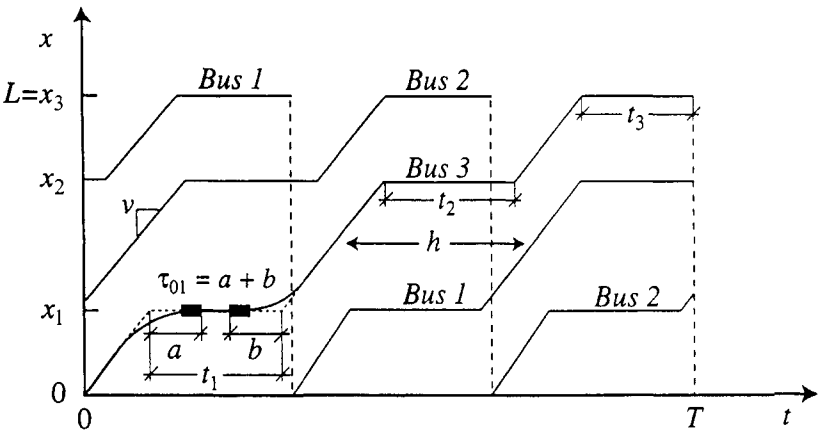


Figure 7.4 Idealized trajectories of three buses operating on a closed loop.

intervals comparable with the bus round trip time. The effects of rapidly varying $q_{ij}(t)$ will be examined in the next subsection and those of uncertainty in Sec. 7.2.2.

If the O-D flows do not vary much during a study period that is long compared with the round trip time then a stationary approximation may be used in which the O-D flows are treated as constants, $q_{ij}(t) \approx q_{ij}$. This simplification is also useful if flows vary significantly during the study period but not much within one round trip time. Then, one can break the study period into stationary subperiods that can be analyzed separately; e.g., the morning rush, midday, etc. (One could also recognize just one subperiod and set each O-D flow equal to the maximum observed, but this approach would be less precise.) The analysis method described below can be applied to any stationary (sub)period.[13]

First note from Fig. 7.4 that the vehicle cycle time can be put in the form:

$$T_C = T_m + \Sigma_{i=1}^{n} t_i = T_m + \Sigma_{i=1}^{n} (\tau_{0i} + \tau_1 M_i) \qquad (7.12)$$

where T_m is the time spent cruising (i.e., $T_m = \Sigma_{i=1}^{n} L_i/v_i$ where L_i is the length of the link leading to stop i), M_i is the number of passenger boarding and alighting movements at i, and τ_1 is the average time per passenger movement. Because under stationary conditions the number of boardings and alightings in every bus trip must be equal, the total number of movements experienced in one round trip $\Sigma_i M_i$ should be twice the number of boardings. Therefore, if we use $q = \Sigma_{ij} q_{ij}$ for the

total system flow, then $\Sigma_i M_i = 2qh$ and (7.12) becomes:

$$T_C = T_f + 2q\tau_1 h. \tag{7.13}$$

where $T_f = T_m + \Sigma_i \tau_{0i}$ denotes the fixed component of the bus round trip time. Equation (7.13) is quite useful because it expresses T_C as a function of readily obtainable data. (We will see in Sec. 7.3.1. that q is much easier to estimate than $\{q_{ij}\}$.)

It is now a simple matter to obtain the required fleet size. It should be clear from inspection of Fig. 7.4, and from the explanation of closed loop systems included in Sec. 1.3.2, that the number of vehicles, V, must satisfy:

$$V \geq [T_C/H]^+, \tag{7.14}$$

where the operation '$[\ ]^+$' denotes rounding up to the next integer.

Next, we should ensure that a vehicle's capacity exceeds its maximum number of occupants. In our stationary, deterministic approximation the maximum number of occupants is the same on every run, and is always reached on the inter-stop link that carries the maximum flow. This location is called the *critical link* or the *maximum load point*. It can be found by assigning the O-D flows to the links l as explained in Chap. 5 and looking for the largest link flow, q*. The assignment recipe for the link flows, q_l, may be expressed compactly with the help of indicator constants Δ_{ijl} that are 1 if a trip from ij uses link l and 0 otherwise:

$$q_l = \Sigma_{ij} q_{ij} \Delta_{ijl} \quad \text{(for all } l\text{)} \tag{7.15}$$

and

$$q^* = \max\{q_l\}. \tag{7.16}$$

An alternative way of obtaining $\{q_l\}$, which does not require knowledge of the O-D table and therefore is much more practical, can be used if one knows the flow past one location of the route, e.g., at a terminus where the flow is zero. The procedure only requires knowledge of the input and output flows at each stop, $F_i = \Sigma_j q_{ij}$ and $G_j = \Sigma_i q_{ij}$. It can be stated very simply if we assign to each link the same label as the stop to which it leads (i.e., $l = i$ if l leads to i) and use l' to denote the link visited after l. Then, by virtue of the flow conservation law, we can write:

$$q_{l'} = q_l + (F_l - G_l) \quad \text{(for all } l\text{)} \tag{7.17}$$

which can be used recursively to calculate all the link flows q_l.

Finally, note that the critical flow and the location of the maximum load point can also be estimated from *observed* link flows if the system is already in operation and is undersaturated. This is the most direct and accurate method.

The $\{q_l\}$ are important because if the system is not oversaturated (our design goal) then each flow must equal the actual number of passengers that pass each link per unit time. Since the flows are steady the bus occupancy on link l must be: $o_l = q_l\, h$. This should not be a problem if $C \geq o_l = q_l h$ for all l. Therefore, our condition for undersaturation is:

$$C \geq q^* h. \tag{7.18}$$

Equations (7.13), (7.14) and (7.18) are the sought recipe, which as of yet ignores stochastic fluctuations.

If the storage capacity of a stop's platform, P_i, is comparable or smaller than C, one should also make sure that the maximum number of accumulated passengers/items does not exceed P_i. Fortunately, said accumulation can also be expressed as a function of easily obtainable data; e.g., if passengers leave the 'bus' before others board, then the maximum accumulation is simply: $(F_i + G_i)h$ – in this case one would then have to check that h satisfied $P_i \geq (F_i + G_i)h$ for the relevant i, in addition to (7.18).

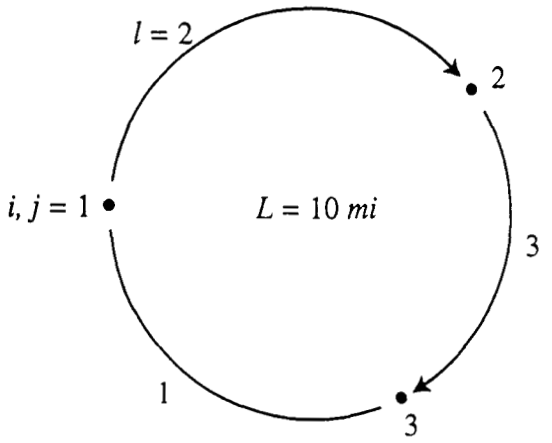
Example: We seek the fleet size and vehicle capacity needed to operate the system of Fig. 7.5 with a 4 minute headway. The route is 10 miles long and has three stops. In a real problem encompassing a 10-mile route one would have more than 3 stops but this would complicate the numerical illustrations unnecessarily. To give the results a semblance of realism we have chosen $\tau_{0i} = 3$ min for $i = 1, 2, 3$ so that the fixed stop delay accrued during a round trip (9 min) will be in line with what would be expected for a line with many stops. We have also chosen an O-D table that would correspond to a busy route serving 700 passengers per hour; the entries of an O-D table with many stops would be much smaller.

The assignment procedure, Eq. (7.15), yields for the link flows:

$$(q_1, q_2, q_3) = (350, 420, 450)\text{pax/hour}$$

and according to (7.16) $q^* = 450$; thus, $l = 3$ is the critical link. The required bus capacity is therefore $C \geq 30$ passengers, as per (7.18). The required fleet size is obtained from (7.13) and (7.14). The former yields $T_C = (10/15) + (3 \times 5 \times 10^{-2}) + (2 \times 700 \times 10^{-3} \times 4/60) = (.666) + (.150) + (0.093) = 0.909$ hrs, i.e., 55 min. Therefore, $V = [55/4]^+ = 14$ buses.

As an additional exercise you may want to apply Eq. (7.17) to our data in order to identify the critical link assuming that only the F_i and G_j are known. Although the link can be identified, you will notice that q^* cannot be determined because our problem has no terminus. ■



q_{ij}		j = 1	2	3	F_i
i	1	–	100	200	300
	2	200	–	50	250
	3	30	120	–	150
	G_j	230	220	250	700

q

Figure 7.5 Route geometry and O-D table of a simple example.

Equations (7.13), (7.14) and (7.18) can also be used to find the fleet size needed and the headway that can be provided with vehicles of a given size. A quick calculation for lines with many vehicles can be made by ignoring the rounding up operation in (7.14). Then, on multiplying

(7.18) and (7.14) we find that $CV \geq q^* T_C$; i.e., that the total seating capacity of our fleet must exceed the system demand during a complete bus cycle. If T_C is insensitive to h, as it is in our case ($T_C \approx 1$ hour), this expression says that the number of vehicles can be halved if we double their size. Although this choice would cut operating costs in half (approximately), it would also double the headway and decrease the level of service. Section 7.2.4. discusses some delicate issues that arise when one attempts to resolve such tradeoff from a 'welfare' perspective.

7.2.1.2. Time-dependent O-D's

The vehicle size recipe, Eq.(7.18), can be easily extended to O-D tables that vary rapidly compared with a 'bus' cycle in the special case where the bus schedule and the bus trip times between all O-D pairs are fixed and known.

To this end, it is convenient to number all the bus cycles (or *runs*) consecutively (k = 1, 2, 3,...) in the order seen from a fixed location. Because it is assumed that buses do not pass one another, this is the same order in which runs would be seen from all other locations. For the example of Fig. 7.4, each continuous portion of a bus trajectory could be defined as a separate run, with fresh run labels issued at x = 0. Note that runs and buses are not the same thing: in the example the run number of a particular bus increases by 3 on crossing location x = L.

From now on, t_{ik} will be used to denote the time at which run k arrives at stop i and T_{ij} to denote the bus trip time from stop i to stop j. The latter are defined to include the stop time at the origin, i, but not at the destination, j. The two sets of variables are related in the sense that for every i, j and k the difference $t_{jk} - T_{ij}$ must be an arrival time for stop i.[14]

The best way to solve the problem is by working directly with the cumulative curves of O-D counts, $N_{ij}(t, t')$. If the bus fleet can carry everyone without overflow, i.e., every bus clears the queues, we can say that the passengers riding toward stop j in run k must include all those that arrived at the other stops in time to catch that bus. More formally, if we assign the same label to a link and to the stop at which it ends (i.e., $l = i$ if l leads to i, as in Fig. 7.5) and then let $N_l(t, t')$ denote the number of passengers that wish to exit link l in the time interval (t, t'], including t' but not t, we can write:

$$o_{lk} \equiv N_l(t_{lk-1}, t_{lk}) = \Sigma_{ij} N_{ij}(t_{lk-1} - T_{il}, t_{lk} - T_{il}) \Delta_{ijl} \quad \text{(for all } lk\text{)}, \tag{7.19}$$

where the identity on the left introduces the shorthand notation o_{lk} for the number of passengers actually leaving l on run k; i.e., for the bus

occupancy on that particular run and location. The second equality is true because the bus occupancy must equal the number of passengers who: (i) show up at all other stops in time to meet our bus, and (ii) wish to travel on this link. Clearly then, the particular schedule can be met without overflows if the largest occupancy calculated with (7.19) does not exceed the bus capacity. This condition is the time-dependent analog of (7.18).

A transit/carrier, of course, has the flexibility to design a schedule where vehicles are brought and withdrawn from service as the demand varies. This has the effect of changing the schedules but does not change the philosophy behind (7.19); i.e., one would still calculate the occupancy data by assigning people to bus runs in a sensible way and would then check that none of the o_{lk} exceed the bus capacity. Section 7.3 describes how the $N_l()$ link count data can be obtained.

The effects of oversaturation can be predicted from the $N_{ij}(t, t')$ if one introduces an assumption about the number of people that are allowed to board the bus at each stop when queues exist. In cases without exogenous controls it is reasonable to assume that people boarding at an oversaturated stop will fill the vehicle if all cannot get on board, and that no destinations are favored in this process. One could then predict the cumulative number of people that would have departed from i for destination j in the time interval (0, t], $D_{ij}(0, t)$, with a simulation. Since the buses have definite schedules, the simulation could simply step through time from (system wide) scheduled stop to scheduled stop, updating a state that only needs to include: (i) the numbers of passengers currently present in each run k stratified by destination j, $b_{kj}(t)$, and (ii) the $D_{ij}(0,t)$.

The simulation procedure can also be used (with a revised logic) to evaluate "metering' strategies that one may try in an attempt to reduce the transient queues at the oversaturated stops as rapidly as possible. Not yet solved completely, this optimization problem is rather similar to that faced by traffic engineers in charge of roundabouts or congested freeways where the goal is to restrict entry so as to favor those origins that send the least flow through the most heavily demanded links.

7.2.2 Management of headway and occupancy fluctuations

This subsection considers the effects of imperfectly known O-D counts $N_{ij}(t, t')$. It explores the effect that fluctuations in the realized counts on any given day, i.e., the deviations from a predicted average, have on both occupancy and on-time performance. The effects and prevention

of unpunctuality are examined after a brief description of the fluctuations in occupancy that can be expected for a 'bus' that adheres to a schedule. Fluctuations in occupancy are important because their magnitude influences the bus capacity required to avoid oversaturation.

If one lets let $l'k'$ denote the (critical) link-run combination with the maximum average flow and $O_{l'k'}$ the corresponding (random) bus occupancy, it makes sense as a first approximation to choose the bus capacity with the rule: $C \geq o' + 2\sigma'$, where o' and σ' are the mean and standard deviation of $O_{l'k'}$. Since $O_{l'k'}$ is a sum of the $N_{ij}(t, t')$ counts indicated by (7.19) its mean equals the sum of the means of said counts. This means that o' can be calculated with Eq.(7.19) if the predicted (average) count curves are used as inputs in this expression. Although one could also relate σ' to the covariances of the O-D counts, it is more pragmatic to estimate it directly from past experience with similar systems. The choice $(\sigma')^2 \approx o'$ may be reasonable in many cases, if historical data are not available.[15]

The proposed rule ensures that our system would fail to carry the desired loads on the critical link-run combination infrequently. A more refined criterion for the selection of C would use the probability of overflow, p_f. The following shows how p_f can be approximated.

If the occupancy on the critical link-run combination is a normal random variable, a reasonable first approximation when the average occupancy is large, then one knows the overflow probability on the critical link-run and this result can be used as a proxy for p_f; i.e.:

$$p_f \approx \Phi((o' - C)/\sigma') \tag{7.20a}$$

where Φ is the standard normal c.d.f. We recognize that this expression will tend to underestimate the true probability of overflow since it ignores that fluctuations may cause non-critical links to overflow. The approximation is very good, however, because it is unusual for a non-critical link to overflow when the critical link-run does not. This statement is true because the occupancies on all the links for a given run are usually highly correlated variables due to the sharing of trips.

If desired, an upper bound to p_f (or a lower bound to $1 - p_f$) can be obtained by ignoring the correlation and calculating the probability that none of the links overflows. This is illustrated now for the stationary case. If one uses o_l and σ_l for the mean and variance of O_l, where l denotes a generic (non-critical) link, the result is:

$$(1 - p_f) \geq \prod_l (1 - \Phi_l) \tag{7.20b}$$

where the symbol Φ_l has been used instead of $\Phi((o_l - C)/\sigma_l)$. This bound will usually be close to (7.20a) because in most cases only a few

Φ_1 will be close enough to $\Phi_{1'}$ to be significantly different form zero and contribute to the product. For example, for the average occupancies of the previous example (23.3, 28, 30) a reasonable choice for C is $C = 30 + 2(30)^{1/2} \approx 41$ seats if one assumes that the occupancies are Poisson random variables. Since $\sigma_1 = o_1^{1/2}$, our data yield $(\Phi_1, \Phi_2, \Phi_3) = (.0002,.0070,.0222)$ and $(.97) \leq (1 - p_f) \leq (.98)$. This type of approximation suffices for most practical purposes because small changes in C usually change the probabilities by more than the difference between the two bounds.[16]

It has been assumed in the above that the vehicles adhere to their schedule and that the variations in occupancy only arise from the fluctuations in demand. Yet, fluctuations in occupancy can also be induced by day-to-day variations in the schedule because buses that follow a long headway will tend to collect more passengers and those that follow a short headway will tend to collect fewer. These effects can be expressed by a simple formula when: (i) the number of passenger trips that want to use the critical link is a stationary process with independent increments and index of dispersion γ, and (ii) the headway of a given run is relatively constant across its stops but varies across days with a mean h and standard deviation σ_H. Then it is possible to show that: $(\sigma')^2 = \gamma(o') + (o'\sigma_H/h)^2$.[17] The first term of this expression captures the effect of demand variability and the second term that of unpunctuality. When $\gamma = 1$ (our first simple guess) we see that the unpunctuality effect can be ignored if the expected occupancy is small compared with the square of the ratio of the mean headway to its standard deviation. For example, if the average occupancy on a link is 25 passengers then the deviations from the schedule can be ignored if they are typically small compared with 1/5 of a headway.

7.2.2.1. Schedule instability and control[18]

While variable vehicle occupancies can be handled by choosing larger vehicles, variable headways are much more difficult to manage. This happens because a scheduled system with multiple vehicles is inherently unstable. If a vehicle has to wait for a longer time than usual at a stop to collect a large number of arrivals, it will fall behind schedule. The delay will increase its headway relative to the prior vehicle, which will in turn increase the expected number of passengers collected at future stops and the related stop times. As a result, the vehicle will tend to fall further and further behind schedule, and at a more rapid rate the further behind it is. At the same time, the bus trailing our vehicle will find shorter headways and will tend to gain on the schedule. For this bus the shorter its headway, the stronger the tendency to gain further.

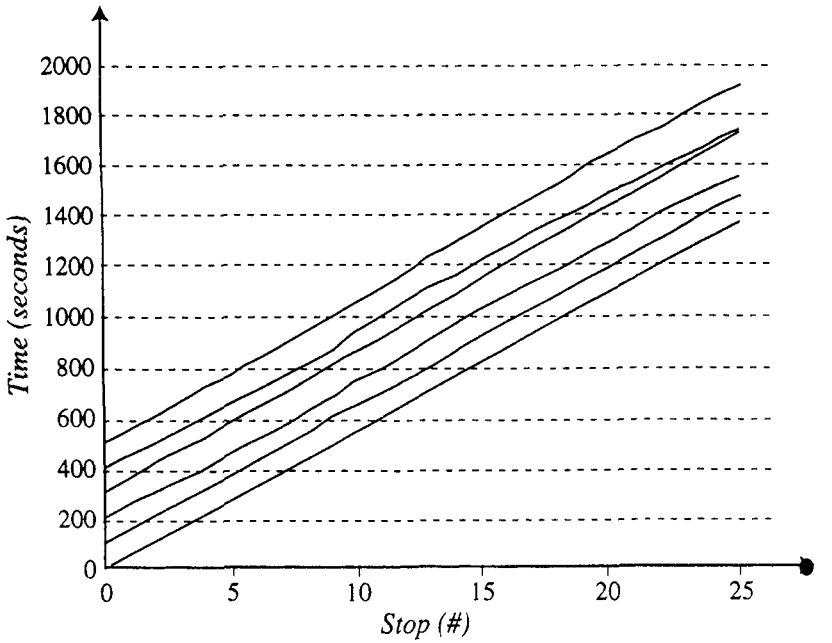


Figure 7.6 Simulated space-time trajectories of six buses susceptible to 'pairing'.

The end result of this process is a pair of buses. If you have ever waited for a long time for a bus, only to find that two of them came together at long last, you have been a victim of this bunching phenomenon.

Fig. 7.6 depicts the simulated space-time trajectories of six buses running on a line with 25 stops where the first bus keeps to the schedule but the following ones are susceptible to bunching. Note that the first bus is the one on the bottom of the picture because time has been plotted as the ordinate. You can see that as the buses progress the headways become irregular and buses 4 and 5 eventually pair. Pairing is less rapid for buses 1 and 2 because bus 1 keeps to its schedule, although these buses could also pair. The reader can check this for himself by running a simulation. This is fairly easy to do as is shown below.

Fig. 7.7 displays the spreadsheet data corresponding to Fig. 7.6. Rows 3 to 6 contain the input data. The passenger arrival processes are assumed to be independent, stationary and Poisson, with an average demand rate $F_i = \Sigma_j q_{ij}$ that is equal for all the stops; this value, F_i

	A	B	C	D	E	F	G	H
1	EXAMPLE: BUS PAIRING SIMULATION							
2								
3	Demand rate (pax/sec/stop) =				0.05	Slack on (10,20)=		0.00
4	Non-stop travel time (secs./stop)=				50.00	Std dev (sec.)=		5.00
5	Lost time per passenger (sec.)=				1.00			
6	Desired headway (seconds)=				100.00			
7								
8	Stop		Time in seconds for bus numbers....					
9	(#)		1.00	2.00	3.00	4.00	5.00	6.00
10								
11	0.00		0.00	100.00	200.00	300.00	400.00	500.00
12	1.00		55.00	159.05	261.38	363.61	459.13	561.85
13	2.00		110.00	219.25	311.01	426.83	511.04	617.19
14	3.00		165.00	268.66	354.06	483.35	561.92	670.24
15	4.00		220.00	324.57	406.91	533.83	610.40	731.23
16	5.00		275.00	381.41	464.83	591.57	659.76	778.56
17	6.00		330.00	429.10	524.45	647.63	710.59	835.28
18	7.00		385.00	489.11	570.46	702.96	767.75	890.53
19	8.00		440.00	543.70	630.13	757.33	814.25	946.81
20	9.00		495.00	604.40	686.03	811.30	866.42	1002.54
21	10.00		550.00	658.10	752.41	871.24	950.00	1056.84
22	11.00		605.00	709.76	796.11	922.37	1007.07	1120.58
23	12.00		660.00	757.22	858.54	981.64	1071.89	1180.77
24	13.00		715.00	803.53	919.28	1039.82	1128.63	1247.50
25	14.00		770.00	864.88	974.83	1095.18	1170.72	1302.96
26	15.00		825.00	922.53	1033.39	1158.07	1230.08	1368.33
27	16.00		880.00	980.94	1085.41	1219.63	1285.33	1424.84
28	17.00		935.00	1031.78	1137.19	1278.73	1341.66	1472.97
29	18.00		990.00	1089.46	1185.31	1336.23	1388.94	1531.68
30	19.00		1045.00	1141.50	1242.88	1387.74	1440.12	1596.23
31	20.00		1100.00	1200.00	1300.00	1450.42	1500.00	1647.89
32	21.00		1155.00	1255.08	1353.61	1504.08	1543.27	1707.02
33	22.00		1210.00	1309.40	1417.61	1559.25	1591.30	1755.88
34	23.00		1265.00	1366.94	1467.02	1615.65	1641.86	1818.05
35	24.00		1320.00	1427.64	1514.46	1675.84	1694.97	1874.03
36	25.00		1375.00	1483.96	1556.95	1733.26	1744.30	1923.80
37								

Figure 7.7 Spreadsheet organization of the bus trajectory simulation.

= 0.05 pax/sec, is included in cell 'E3.'[19] We also imagine that everyone travels past stop 25 because such a simple O-D pattern suffices to illustrate the bunching effect and also simplifies the simulation. The travel time between stops, excluding passenger boarding, is assumed to be equally distributed on all the links with a known mean (m_i = 50 sec, contained in cell 'E4') and standard deviation (σ_i = 5 sec, contained in cell 'H4'). (Again, this could be generalized as explained in the previous

footnote.) The lost time per passenger movement is assumed to be $\tau_1 = 1$ sec, and the desired headway, $h_d = 100$ sec. These values are included in cells 'E5' and 'E6.' We ignore for now the data contained in cell 'H3.'

Still in the form of data, column 'A' includes the stop number, and row 11 the times at which the buses are released into the system. Our goal is filling columns C through H with the arrival times of buses 1 through 6 at the different stops. To this end, range 'C12.. H36' contains formulae from which the arrival times are calculated. On column C the formula corresponds to the (regular) schedule of the first bus, which is assumed to travel between stops in the average time $(m_i + F_i h_d \tau_1)$ without any deviations. If the formula is written in cell 'C12' as: + C11 + E5* E3* E6 + E4, it can be copied to fill the column. A similar formula can then be written in cell 'D12' and copied to the range 'D12.. H36', but the mean travel time, E4, must now be replaced by a draw from a (normal) random variable with mean E4 and standard deviation H4.[20] Likewise, the mean number of arrivals (E3* E6) must be replaced by those that would materialize in the actual headway (D11-C11) from a process with index of dispersion, 1. Again, this can be approximated by generating another random draw from a normal variable with mean E3*(D11-C11) and standard deviation equal to the square root of this value. You may now run the simulation for different values of the input parameters and explore how they influence the pairing tendency.

A schedule can be met without pairing if one designs it with longer vehicle trip times than would be possible on average, and then makes sure that vehicles never depart before the scheduled time. The difference between the scheduled and average trip times is called the *slack*. Airlines and railroads use variations of this approach. Since vehicles adhere to the schedule the expected travel time between stops is $m_i + \tau_1 E(M_i)$, where M_i is the number of passenger movements in a regular headway, and the standard deviation of the trip time is: $(\sigma_i^2 + \tau_1^2 \text{var}(M_i))^{1/2}$. Late departures can be largely avoided by introducing slack between every two consecutive stops in an amount equal to a small multiple of this standard deviation. This means that the total vehicular round trip time would be increased over the average uncontrolled time by a small multiple of $\Delta T = \Sigma_i (\sigma_i^2 + \tau_1^2 \text{var}(M_i))^{1/2}$. Let us now see how the amount of slack can be related to some easily observable data.

If σ_i^2 and $\text{var}(M_i)$ do not vary much across i, we can write $\Delta T \approx (n^{1/2})(\Sigma_i \sigma_i^2 + \tau_1^2 \text{var}(M_i))^{1/2}$, where n is the number of stops.[21] If we now introduce σ_{tot}^2 for the variance of the round trip time excluding

passenger loading/unloading (i.e., $\sigma_{tot}^2 = \Sigma_i \sigma_i^2$ since inter-stop trip times can be expected to be independent), and then approximate the sum of var(M_i) by the variance of the total number of passenger movements in a scheduled round trip time, T_r, i.e., by *four* times the variance of the number of passenger arrivals in T_r, which we denote $(\sigma_{N(Tr)})^2$, we can write:

$$\Delta T \approx (n^{1/2})\left(\sigma_{tot}^2 + 4(\tau_1 \sigma_{N(Tr)})^2\right)^{1/2}. \qquad (7.21)$$

As a first approximation in some cases we may use $(\sigma_{N(Tr)})^2 \approx qT_r$ in this expression. Then, if we expect ΔT to be small in comparison with the (known) average cycle time (T_C) excluding slack, so that $T_r \approx T_C$, we may substitute qT_C for $(\sigma_{N(Tr)})^2$ in (7.21), in order to determine ΔT. This information can then be used to select an appropriate amount of slack, since the slack should be a small multiple of ΔT; e.g., $2\Delta T$. If ΔT is not small, or more precision is desired in the calculation, we could use $(\sigma_{N(Tr)})^2 \approx q(T + 2\Delta T)$ in (7.21) and then solve for ΔT.

It is interesting to see from (7.21) that the amount of slack increases with $n^{1/2}$. Slack translates directly into higher costs for the operating agency because of the larger fleet sizes needed to accommodate long cycle times; see (7.14). It also induces longer travel times for those on board. Thus, when ΔT is a significant part of the total cycle time it may make sense to reduce the number of points at which control is exercised.

If one could exercise it every k^{th} stop and still maintain regular schedules at the intermediate stops, then one would expect (7.21) still to apply, with a factor $(n/k)^{1/2}$ instead of $(n)^{1/2}$. And this is what many well-run transit agencies try to do. First they devise a target schedule for the stops in between each consecutive pair of control points assuming on time departures from the control point and slightly optimistic bus trip times following the control point. Then they stick to it in the sense they never dispatch a bus *before* the target time. Because the schedule is optimistic, buses tend to run late and rarely have to be delayed in order to avoid early departures. The scheme works if enough slack is added at the control points to be reasonably sure that a late-running bus following one on schedule (the worst possible case) can arrive at the control point in time to depart on schedule.[22]

Although the calculations needed to evaluate the probability of arriving too late at the control point can become complicated, simple modifications to a spreadsheet such as that of Fig. 7.7 can also be used to evaluate the robustness of proposed schedules. As an exercise, see if you can reprogram the spreadsheet of Fig. 7.7 to test and refine a

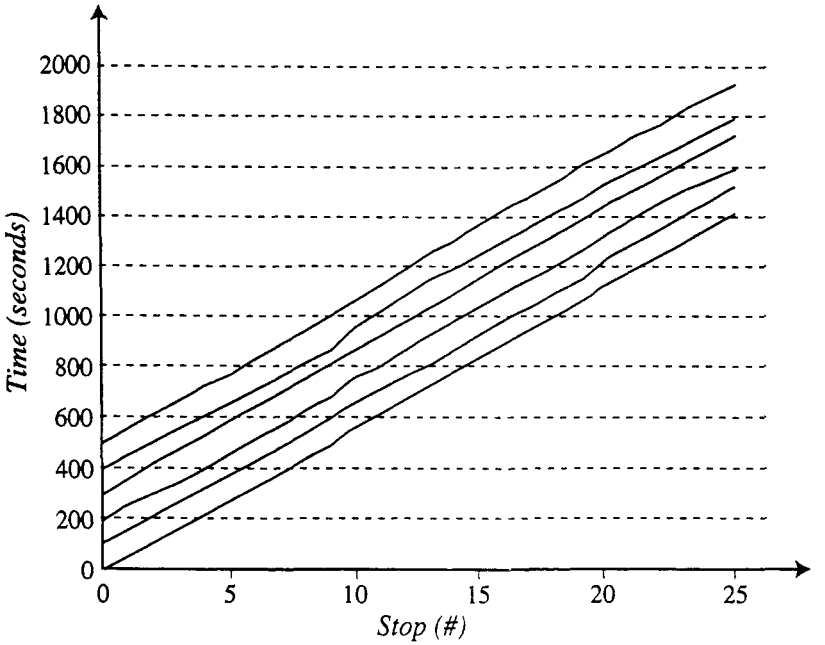


Figure 7.8 Simulated trajectories when control is exercised at stops 10 and 20.

strategy that will avoid pairing while minimizing the scheduled bus trip time from stop 0 to stop 24. Fig. 7.8 shows a realization of the process with the modified logic, after introducing 15 seconds of slack at stations 10 and 20, and using optimistic bus trip times of 50 seconds per stop.[23] In practical situations the control points should be located where the bus is lightly loaded and/or would have to wait for another reason. Major transfer nodes are logical choices.

The specific formulas that have been presented do not apply to systems in which vehicles are allowed to skip stops because then the term $\Sigma_i \tau_{0i}$ in (7.12) is no longer fixed. For suburban bus routes and elevators, which are examples of these kinds of systems, this term is the main contributor to the variance of the vehicle cycle. Fortunately, pairing can still be avoided for most of these systems, with schemes similar to the one just described. Exceptions are systems such as elevators in which the stop time is a significant part of the round trip time. Then, pairing can be so forceful that control strategies to avoid it are a worse remedy than the problem.

7.3 Observation issues

The expressions of Secs. 7.1 and 7.2 include decision variables h, V and C, as well as parameters that need to be estimated. Some of these, such as L_i and n, can be measured rather accurately and others, such as the O-D table, can be estimated with methods already explained, e.g., from passenger surveys, input and output counts, etc. In view of this we only discuss here the estimation of those parameters that can benefit from different experimental setups, and de-emphasize the statistical procedures which would not be new. Section 7.3.1. explains how the time-dependent demand on the system's links $N_l(t, t')$ can be obtained, and Sec. 7.3.2. does the same for the vehicle operating characteristics (v_i, τ_{0i} and τ_1).

7.3.1 Link flow estimation

Knowledge of the time-dependent demand on the system's links is important because fleet size decisions depend on the O-D table only through the N_l. Fortunately, this information is much easier to obtain than the O-D table.

The method about to be described applies to undersaturated systems that adhere to their schedule. It consists in determining the *actual* cumulative number of passengers E_{lk} that have exited each link immediately after the passage of each run, k, i.e., at the times t_{lk} when each run k arrives at stop $i = l$. Since the system is punctual and undersaturated we know that the desired departures in the time interval $(t_{lk}, t_{lk'})$ must equal those that have been observed; i.e., that $N_l(t_{lk}, t_{lk'}) = E_{lk'} - E_{lk}$. This means that the cumulative number of desired link departures can be approximated by a smooth curve passing through the observed points (t_{lk}, E_{lk}).

The E_{lk} can be obtained either by stationary observers working at each stop, e.g., by automatic turnstiles, or by observers on the 'buses.' In the first case we would record at each stop $(i = l)$ the number of passengers boarding and alighting each run k, f_{lk} and g_{lk}. These data can then be used with the flow conservation relation

$$E_{l+1,k} = E_{lk} + f_{lk} - g_{lk}$$

to obtain the complete collection of the exit counts $\{E_{lk}\}$, provided one knows the exit counts from a reference link such as a terminus or a location where the bus occupancies can be readily seen and counted. For example, if $i = 0$ is a terminus where everyone gets off the bus then

$E_{0k} = \Sigma_{k' \le k}\, g_{1k'}$, and the known set of exit counts at the terminus $\{E_{0k}\}$ suffices to obtain the complete set of counts.

Alternatively, an observer in each bus (e.g., the driver) can record the bus occupancy on each link l during each run k. It is then possible to obtain from the complete collection of occupancies $\{o_{lk}\}$ the E_{lk}, by adding the observed occupancies for a given l up to a given k; i.e.,

$$E_{lk} = \sum_{k' \le k} o_{lk'} \quad \text{(for all } l, k\text{)}.$$

The preferred data gathering method should depend on the relative magnitude of the number of stops on the route and the number of vehicles serving it. Repeated observations of the $\{E_{lk}\}$ in successive days may be treated statistically as usual in both cases.

The observed data can be used to calculate the maximum bus occupancy $o^* = \max_{lk}\{o_{lk}\}$ in a day and the observed variations in o^* across days can then help us make decisions about vehicle size.[24] Different schedules may also be evaluated similarly. One would assign the smoothed N_l curves estimated for a given day to the proposed set of runs–using the definitional identity on the left side of (7.19)–and then would check whether the resulting occupancies (and their variation across days) are acceptable.

If the vehicle size can be changed across runs, as happens with train systems such as the BART rapid transit system in the San Francisco Bay Area, then one may choose to work with run-based occupancies $o_k^* = \max_l\{o_{lk}\}$ to decide which runs should have long or short trains. If one is interested in a time of day where the demand may be considered to be stationary, less data are necessary. Then, one does not need to distinguish across runs and the variations in o_k^* within a given study period may suffice to determine an appropriate vehicle size.

If the system is oversaturated the approaches just described are not very useful because we can no longer claim that bus occupancies are a reflection of current demand. Although it is possible to determine the cumulative input flows at each stop from data collected by observers stationed at the stops, one would need additional information regarding the O-D table in order to obtain the *desired* demand curves $N_{ij}(t)$ and $N_i(t)$. The problem is similar to that described in Sec. 6.3.3.

7.3.2 Trip time estimation

Let us now see how to estimate the constants, τ_{0i}, τ_1, and $(1/v_i)$ that are used to describe a vehicle's motion. To this end note that the dotted extension of the vehicle trajectories shown in Fig. 7.4, and the darker

sections also shown in the figure are idealizations used for illustrative purposes only; in reality there are no observable events along a vehicle's trajectory that allow us to measure τ_{0i}, τ_1, and $(1/v_i)$ directly.

If we let M_{i-1}, L_i and $T_{i-1,i}$ be the (random) number of passenger movements at stop $i-1$ for a specific bus run, the length of the route link preceding stop i, and the (random) time spent by the bus on the link from the start of stop $(i-1)$ to the start of stop (i), we expect these variables to be (approximately) related by

$$T_{i-1,i} \approx \tau_i' + \tau_{1,i-1} M_{i-1} + (1/v_i) L_i, \tag{7.22}$$

where τ_i' represents the time wasted accelerating from stop $i-1$ and decelerating to stop i; i.e: $\tau_i' \approx (\tau_{0i-1} + \tau_{0i})/2$. The second and third terms of (7.22) represent respectively the loading/unloading time at stop $i-1$ and the (cruising) trip time from $i-1$ to i.[25] The expression is reasonable and consistent with our earlier assumptions since the sum of (7.22) across stops yields (7.12).

The unknown constants of (7.22) can be estimated quite accurately if observable events can be used to divide $T_{i-1,i}$ into two parts: one that depends only on L_i and another that depends only on M_{i-1}. One can argue that opening and closing doors are such events since the part of $T_{i-1,i}$ in which the vehicle doors are closed T_i^c should only depend on L_i, while the part in which they are open T_i^o should only depend on M_{i-1}. Therefore we can write:

$$T_i^o \approx \alpha_i + \tau_{1,i-1} M_{i-1} \tag{7.23a}$$

and

$$T_i^c \approx \beta_i + (1/v_i) L_i, \tag{7.23b}$$

where α_i and β_i are two constants such that $\alpha_i + \beta_i = \tau_i'$. Each of these equations can be estimated by least squares. Equation (7.22) is then obtained from the sum of (7.23a) and (7.23b).

7.4 Design and evaluation

Although design and evaluation issues are somewhat outside the scope of an introductory book on 'operations', it seems worthwhile to present here some basic ideas that appear not to be widely known.

7.4.1 Design

The most basic decision in the design of scheduled transportation systems, perhaps, is deciding on a desirable schedule for a given route. We have seen already that in cases where the demand does not change

rapidly with time, one should attempt to operate a system with regular headways. Therefore, the selection of h is discussed first.

This problem was already addressed in Chapter 3 as an example that ran through Secs. 3.2.1 and 3.2.2. The choice involved a trade-off between waiting time for customers, captured by the first term of Eq. (3.9a), and the agency's operating cost, captured by the second term. [Note that, except for the notation used in Chap. 3, these terms correspond to formulae (7.4) and (7.14) of the present chapter].

For 'closed' systems in which both types of costs are paid by the same entity (e.g., a firm carrying its own products on its vehicle fleet) it is possible in principle to assign numerical values to the constants α and β weighting the two terms of (3.9a) and to derive a 'best' design. (The constants α and β are usually difficult to quantify, however.) Although the assignment of values to α and β is rather questionable for 'open' systems in which an agency carries other people's items (see Sec. 7.4.2), the terms of a *logistic cost function* such as (3.9a) do reveal how the various players in the game are affected by changes in h, and this can help the agency reach a desired balance between the two types of cost.

Two properties often found in logistic cost functions are: (i) that deviating from the 'optimum' decision variables substantially does not change the value of the objective function appreciably, *although its various terms may change considerably*, and (ii) that that the objective value is not very sensitive to the weights assigned to the terms (e.g., α and β). To see that this is the case with (3.9a) let us consider its dimensionless counterpart (3.11). The latter is 'insensitive' because it increases by only 6% when the decision variable z is changed by a factor of $2^{1/2}$ from the optimum ($z^* = 2^{1/2}$). And, since (3.9a) is just a dimensional version of (3.11), the same statement can be made about (3.9a) for all possible combinations of its parameters α, β, etc.[26] It is this author's experience that insensitivity to the decision variables and the data is the rule rather than the exception in design problems.

If (3.9a) referred to a closed system, this insensitivity would mean that the chosen headway could be varied by a factor of 2 in an interval around the optimum h and this would not affect the bottom line significantly. This kind of flexibility in the choice of the decision variables can be quite useful in practice. It allows one for example to coordinate the schedules of various routes serving a terminal by using a headway menu of the form $\{\bar{h}2^p\}$. Open systems require more thought. One must recognize that changes in the decision variables can induce large changes in the various terms of the objective function (even if the sum total remains nearly constant) and that these changes represent

transfer payments that should not be ignored in an evaluation.[27] Some of the tricky issues that arise as a result are discussed in Sec. 7.4.2.

At the next level of complexity, one may be interested not just in scheduling, but also in route structure and layout. For a single route, the spacing between stops influences the access distance (which increases with the spacing) and also the vehicular trip times (which decrease with the spacing). Vehicular trip times in turn influence the operating cost, see (7.14), and the in-vehicle travel times, T_{ij}, experienced by the goods or passengers carried. As in the previous case, it is now also possible to express the various measures of performance of the system (such as the operating cost and the mean access, waiting and riding times) in terms of our decision variables: the headway and the spacing. In this case, too, a weighted average of these measures is insensitive to the input data and to the precise values of the headways and spacings chosen, provided they are reasonable.

There is an extensive literature on vehicle routing and scheduling that addresses not just the simple problems just described but also those with time- and space-dependent decision variables and data. Although most of the works are numerical in scope, some are designed to use less data by incorporating analytical principles. The interested reader may want to take a look at the paper by Clarens and Hurdle (1975) which introduced a clever design technique for space-dependent problems using continuum approximations. Analysis techniques using the building blocks of this chapter have been developed for complex problems involving transfers and network design.[28]

7.4.2 Evaluation

For both scheduled and unscheduled transportation systems, it has been an objective of this book to present tools that can be used to estimate basic system performance measures such as the vehicle-miles or vehicle-hours of travel in a system, and even the detailed vehicle trajectories in cases where these can be predicted. If one can estimate how these basic measures change when the system is redesigned or controlled in a particular way, then one should be able to predict the changes in the measures normally used for evaluation purposes (e.g., delay, cost, air pollution, noise, etc.) because these are directly related to the former. It should be clear from the discussion in this and earlier chapters that transportation user costs such as delay are linked to the basic measures. In this chapter for example we explained how passenger

delay depended on the vehicular schedule, and in earlier chapters how it could be derived from the N-curves.

Transportation impacts on its non-users, such as air pollution, noise and to a certain extent energy consumption can also be linked to the vehicle trajectories, and in cases with many vehicles where these cannot be obtained individually (such as freeway traffic) to the macroscopic characteristics of traffic derivable from the N-curves. Chapters 30-32 of Homburger et.al. (1992) supply the necessary charts and formulae; this reference also includes more detailed information sources.

7.4.2.1. Some remarks on welfare maximization

The ability to predict, albeit coarsely, how a given system (scheduled or not) distributes its impacts/outputs, i, across all the population segments, j, of users and non-users is important because the change in those impacts when we alter the character of our transportation system, e.g., by expanding its size, changing its control scheme, putting a price on it or doing some other such thing, should be the basis for society's final choice. We warn, however, that serious difficulties arise if we try to anticipate what society wants by constructing an overall *welfare* measure from the impact levels affecting different groups of people. This means that welfare-based approaches for the evaluation and selection of open/public systems must be taken with a grain of salt.

Let y_{ij} be the amount of output i affecting population segment j; e.g., the amount of time (i) spent by middle-income Berkeley residents (j) on the facility. (We assume that the segmentation is so fine that the y_{ij} do not vary significantly across the people in a segment and that all the people in a segment are alike, in the sense that they value the y_{ij} similarly.) Then, if the system in question has been thoroughly analyzed, one could arrange all the outputs as shown in Table 7.1.[29]

Although such a table would describe how everyone in the population is affected by the system, we shall soon see that there is no practical way in which the y_{ij} can be combined to form a measure of society's welfare.

In conventional micro-economics one essentially assumes that welfare, i.e., our society's objective function U, is the sum of the levels of *satisfaction* or *utility* experienced by each individual, and then it is postulated that individuals only derive utility from their own consumption of the various outputs;[30] i.e., $U = \Sigma_j n_j u_j$ if we use u_j to denote the level of utility achieved by population segment j and n_j to denote the size of that population group. It is also assumed here (for simplicity only) that the u_j can be expressed as a linear function of the y_{ij}'s experienced by the segment in question; i.e.:

$$u_j \approx \Sigma_i \alpha_{ij} y_{ij} \tag{7.24}$$

Table 7.1. A hypothetical distribution of impacts.

j =	1	2	3	4	...
i	Berk poor	Berk rich	N. Oak poor	N. Oak rich	
0, cost	y_{01}	y_{02}	y_{03}	...	
1, time	y_{11}	y_{12}	y_{13}	...	
2, fuel	y_{21}	y_{22}	y_{23}	...	
3, noise	y_{31}	y_{32}	y_{33}	...	
:	:	:	:		
:	:	:	:		

where α_{ij} is the 'value' (positive or negative) of one unit of the i^{th} output to the individuals in the j^{th} population segment. We have not specified a unit for this 'value' (economists use the word 'util'), but this should not be a problem as long as we use the same unit for all the j's. Let us assume that this has been done in (7.24).

Although it should be intuitive that one can determine through experiment the ratio between any two weights, α_{ij}, for any specific group of people (j), e.g., α_{ij}/α_{mj}, there is no way in which the α_{ij}'s themselves may be determined. [We can determine the ratio between the coefficients of time, i = 1, and money, i = 0, for example by seeing how people choose between cheap but slow, and fast but expensive transportation options.] Since the α's are unknowable it is often proposed to use the dollar denominated version of (7.24) (which is knowable) as a substitute for (7.24); i.e.:

$$\$_j = u_j/|\alpha_{0j}| \approx \Sigma_i\left[\alpha_{ij}/|\alpha_{0j}|\right] y_{ij} \tag{7.25}$$

and then to evaluate society's benefit by the sum of (7.25) for all people:

$$\$ = \Sigma_j n_j \$_j \tag{7.26}$$

where n_j is the number of people in group j. This measure is called *consumers' surplus*.

It should be intuitive that, for the same token, one could have chosen to refer utilities to time, and therefore instead of (7.25) we could have proposed:

$$B_j = u_j/|\alpha_{1j}| \approx \Sigma_i\left[\alpha_{ij}/|\alpha_{1j}|\right] y_{ij} \tag{7.27}$$

as an equivalent measure of 'time benefit', whose aggregation across classes would have led to a time-denominated consumers' surplus

$$B = \Sigma_j n_j B_j \tag{7.28}$$

that has as much right to be valid as (7.26).

Each one of our three welfare measures (U, $ and B) weighs utilities in a different way, although according to assumption only the first one is right. If we were so fortunate that the α_{0j} and/or the α_{1j} were independent of j, as if all the population groups had the same need/desire for time and money, then there wouldn't be a problem because the u_j's and the $\$_j$'s (or the B_j's) would only differ by a positive multiplicative constant that does not have to be known for the purposes of determining which system design has the highest U. Unfortunately, the lack of 'need/desire variations' cannot be verified because the α's cannot be determined through experiment. The most one can hope to check is for the 'value of time' ratios $r_j = |\alpha_{0j}|/|\alpha_{1j}|$ to be independent of j; i.e., that there are no relative 'taste variations' across the population. If this happens then Equations (7.25) and (7.27) would only differ by a constant factor (r) for all population groups, as would (7.26) and (7.28). This is somewhat gratifying because then either definition of surplus would yield the same result when comparing alternatives.[31] Unfortunately, even in this fortuitous case the choice cannot be expected to maximize the 'true' welfare $U = \Sigma_j u_j$.

In most cases, however, $ and B do not coincide. In fact, we should not be surprised if in a particular application all three measures recommend a different alternative. This possibility is illustrated by the following example, in which alternative 1 (transit) is either equal, better or worse than alternative 2 (highway) depending on the measure used.

Example: Imagine a world with only two population subgroups, j = 1 (rich) and j = 2 (poor), for whom we have to supply one of the following two transportation alternatives: 1 (TRANSIT, slow and cheap) and 2 (HIGHWAY, fast and expensive). Let us assume that our analysis of these alternatives yields the matrices of y_{ij}'s included in Table 7.2 (in some reasonable units of cost and time).

Let us now assume that the utilities for poor and rich people are:

$$u_1 = -10y_{01} - y_{11} \quad \text{and} \quad u_2 = -y_{02} - 10y_{12}$$

Therefore, the dollar- and time-denominated utilities are:

$$\$_1 = -y_{01} - .1y_{11} \quad \text{and} \quad \$_2 = -y_{02} - 10y_{12}$$

$$B_1 = -10y_{01} - y_{11} \quad \text{and} \quad B_2 = -.1y_{02} - y_{12}$$

Table 7.2. Hypothetical impact matrices for two transportation projects

1, TRANSIT j =	1	2
i	Poor	Rich
0, cost	0	0
1, time	1	1

2, HIGHWAY j =	1	2
i	Poor	Rich
0, cost	1	1
1, time	0	0

A simple algebraic substitution yields the overall measures U, $ and B in Table 7.3 (where we assumed that $n_1 = n_2 = 100$).

You can clearly see that the winning alternative depends on the measure used. Unfortunately, there is no way of resolving this discrepancy objectively. ■

Matters are further complicated because the ratios of α's appearing in (7.25) and (7.27) represent how people value their own experience, and because in many cases some people derive 'utility' from the travel of others; e.g., if a person wants good public transportation so other people's children have access to libraries and schools. This means that it is not possible to calculate what society wants in terms of transportation

Table 7.3. Evaluation results for the example

	U	$	B
1, TRANSIT	−1100	−1010	−200
2, HIGHWAY	−1100	−200	−1010

alternatives and that the selection process should be left to the public. In the end the evaluation of an 'open' system boils down to politics, which the transportation professionals should inform; e.g., by providing the matrix of y_{ij}'s for the alternatives under consideration in as clear and accurate a way as possible. An understanding of the transportation and traffic operations field should facilitate this task.

Notes

1. Background material on the design and operation of these types of systems may be found in Larson and Odoni (1981) and Daganzo (1991).
2. This author believes that the basic ideas of Chap. 1 are all that is needed in order to understand the rationale for the formulas appearing in more specialized publications (perhaps even to derive them) given the particular details.
3. In one case we are interested in the average distance of our customers to the nearest point on the distance line, and in the other case in their separation from the nearest (future) point on the time line.
4. This should be clear by symmetry since in both cases we look for the average 'distance' of randomly distributed points on the time line to the nearest time on the bus schedule (in either the forward or backward direction). Of course, there is no a priori reason to expect that the ensemble of possible distributions of demand points and scheduled times should favor either the forward or backward direction.
5. This statement follows from the identity: $\Sigma_k h_k^2 = \Sigma_k (h_k - \bar{h})^2 + K(\bar{h})^2$.
6. The reason for this is that one can partition the observation period into non-overlapping intervals of near-constant λ with many headways for which (7.4) applies approximately. Since (7.4) only depends on the headways, it should be the same for every interval, and therefore for the overall study period.
7. The bus stop delay problem with regular headways is a special case of the pretimed traffic signal problem in which the signal cycle is the headway, the green phase is the time when the bus doors are open and the saturation flow is the rate at which passengers board the bus.
8. Note, however, that this does not eliminate the exit delay for trips in which this is an issue. This is one of the inherent disadvantages of discretely dispatched vehicles vis-a-vis the automobile.
9. This behavioral model is not reasonable for trips in which there is a deadline, but the qualitative results are similar in both cases.
10. If vehicles have a (small) probability of arriving early so that f_L is not decreasing the analysis following (7.9) will lead to a different arrival time. This is also true if passengers choose their arrival times imprecisely. None of these complications should change the fact that E(w) is comparable with E(S) and significantly smaller than h_k.

11. This is discussed in Sec. 7.4.1. A menu of the form $\{\tilde{h}p\}$ is less restrictive, but it is also less beneficial for coordination. It may be appropriate for lines with few transfers.
12. The average cruising speed may include fluctuations exogenous to the system, e.g., due to variable traffic conditions, and may also vary across links due to physical reasons, such as links with different vertical profiles and lengths. None of these variations in speed are shown in the figure.
13. It was proposed in Hauer (1969).
14. If fresh run labels are not issued between i an j then $t_{jk} - T_{ij} = t_{ik}$. This should be clear from inspection of Fig. 7.4, where we can see that the identity holds if $j > i$. If $j < i$, however, then we also see from the figure that the run number of the bus increases by 3 during its trip from i to j so that the relationship becomes $t_{jk} - T_{ij} = t_{ik-3}$.
15. If a significant portion of the trips on the system are repeated every day, then the variability should be smaller. If, on the other hand, a significant fraction of the arrivals are affected in groups by exogenous causes such as sporting events and unpunctual feeder vehicles then the variability should be higher. We do not theorize any further about this issue because a definite answer can only be obtained through observation.
16. In our case, for $C = 42$ seats we obtain $1 - p_f > .985$ and for $C = 45$ seats $(1 - p_f) = 1.00$.
17. If the headway is not constant during a run then the formula is an upper bound because the fluctuations in occupancy are smaller. Note as well that the formula applies approximately to the non-stationary case if, as is usually the case in public transportation applications, the expected demand varies slowly during a headway.
18. Newell and Potts (1964), Potts and Tamlin (1964) and Newell (1974) describe analytically the bus schedule instability phenomenon, and present methods of headway control aimed at preventing it. Gamse and Newell (1982) have explored the elevator problem and showed that it is more serious. These references are recommended as further reading for the reader interested in schedule control problems.
19. Different demand rates could be incorporated into the simulation by including a column of data next to the stop number.
20. In our implementation, which has been made available to the public as spreadsheet 'PAIRING.WK1', the formula $2*(@RAND + @RAND + @RAND - 1.5)*\sigma + m$ with σ = \$H\$4 and m = \$E\$4 was used to approximate a normal draw with mean m and standard deviation σ, since the sum of 3 uniform random variables has a c.d.f. that is quite close to the normal.
21. The approximation is justified upon noting that the exact expression for $(\Delta T/n)^2$ is the square of the arithmetic mean of $\{a_i\}$, where $a_i \equiv (\sigma_i^2 + \tau_i^2 \text{var}(M_i))^{1/2}$, and that the approximation for $(\Delta T/n)^2$ is the mean of the squares. You can easily verify that the approximation is exact if σ_i^2 and $\text{var}(M_i)$ are independent of i.
22. The advanced reader may see that the headway of the late-running bus as a

function of the distance (x) from the last control point can be approximated by a stochastic process with independent increments, and more specifically by a Brownian process with a drift that *increases* with the current headway. Although this process is unstable (the headway tends to ∞ as $x \rightarrow \infty$ with probability 1 because the drift becomes positive for sufficiently large headways), the probability that the headway will violate the schedule for any *finite* x (e.g., the distance to the next control point) is always small if the slack is large.

23. These effectively allow buses to run as early as they can. Spreadsheet 'PAIRING.WK1' allows the user to include slack at stops 10 and 20 by changing the data in cell 'H3'.
24. Note that the occupancies are available even if one has used the stationary observer method because $o_{lk} = E_{lk} - E_{lk-1}$.
25. The subscript $i - 1$ used in connection with τ_l can be deleted if one believes that the average time per passenger movement does not vary significantly across stops, as has been assumed so far in this chapter.
26. A similar dimensional argument can be made to show that moderate errors in the parameters of (3.9a) have even less of an impact on the final objective.
27. Note for example that changing h by a factor r changes the two terms of (3.9a) by the same factor (increasing one and decreasing the other).
28. A survey of this literature can be found in Langevin, Mbaraga and Campbell (1996).
29. To make comparisons easy, one should use the same denominator for all categories of output; e.g., quantity per year, or quantity per life of the project, or quantity per year per family in the region.
30. This assumption ignores basic human traits such as generosity and envy. Its validity can be debated for goods such as transportation that provide a basic societal function. We accept it here just to show that difficulties with the concept of 'welfare' would arise even in a very simple world.
31. There is an economic theory specifying the conditions under which $r_j = r$ but in our particular case the theory does not apply; after all, we do know that different people have different values of time.

References

Andrews, F.C. (1970) A statistical theory of traffic flow on highways, I- Steady state flow in low density limit. II-Three car interactions and the onset of queueing. *Trans. Res.*, **8**, 359–365.

Ashok, K. and M.E. Ben-Akiva (1993). Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. in Transportation and Traffic Theory, *Proc. 12th Int. Symp. on Transportation and Traffic Theory*, (C.F. Daganzo, editor), pp. 465–484, Elsevier, New York, N.Y.

Avriel, M. (1976), *Non-linear programming: Analysis and methods*, Prentice-Hall, Englewood Cliffs, N.J.

Beckmann, M., C.B. McGuire and C.B. Winsten (1956), *Studies in the economics of transportation*, Yale Univ. Press, New Haven, Connecticut.

Bowman, L.A. and M.A. Turnquist, (1981). Service frequency, schedule reliability and passenger wait times at transit stops. *Trans. Res.*, **15A**, 465–471.

Braess, D. (1968). Über ein paradox de verkehrsplanung, *Unternehmenstorchung*, **12**, 258–268.

Bui, D.D., P. Nelson, S.L. Narasimhan, (1992). Computational realizations of the entropy condition in modeling congested traffic flow. FHWA Report, FHWA/TX-92/1232-7, U.S. Dept. of Transportation, Washington, D.C.

Carlesson, L. (1957). Eu matematisk modell for landsvagstrafik. *Nordisk Matematisk Tidskrift*, **5**, 175–180.

Cassidy, M. (1996). Reexamining bivariate relations in highway traffic. Presented at the 1996 annual meeting of the Trans. Res. Board, Washington D.C. See also: Bivariate relations in near stationary highway traffic. *Trans. Res.* **B** (in press).

Cassidy, M.J. and Coifman, B. (1996). The relation between average speed, flow and density and the analogous relation between density and occupancy. Institute of Transportation Studies Research Report UCB-ITS-RR-96-89, University of California, Berkeley, CA.

Cassidy, M. and J. Windover (1995). Methodology for assessing dynamics of freeway traffic flow. *Trans. Res. Rec.*, **1484**, 73–79.

Cassidy, M. and J. Windover (1996). Driver memory: Motorist selection and retention of individualized highways in highway traffic. Research Report UCB-ITS-RR-96-4, Institute of Transportation Studies, Univ. of California, Berkeley, CA. *Trans. Res.* (in press).

Clarens, G. and V.F. Hurdle (1975). An operating strategy for a commuter bus system. *Trans. Sci.*, **9**, 1–20.

Clayton, A.J.H. (1940). Road traffic calculations. *J. Inst. Civ. Engrs.*, 16(7), 247–284; (8) 588–594.

Cox, D.R. and W.L. Smith (1971), *Queues*, Chapman Hall, London, U.K.

Cremer, M. and H. Keller (1988). Dynamic identification of flows from traffic counts at complex intersections. *Proc. 8th Int. Symp. on Transp. and Traffic Theory*, (V.F. Hurdle, E. Hauer and G.N. Steuart, editors), pp. 121–142, Univ. of Toronto Press, Toronto, Canada.

Daganzo, C.F. (1975). Probabilistic structure of two-lane road traffic. *Trans. Res.*, **9**, 339–346.

Daganzo, C.F. (1983). Derivation of delays based on input output analysis. *Trans. Res.*, 17A, (Includes closing remarks by P. Michalopoulos), 341–342.

Daganzo, C.F. (1990). Some properties of polling systems. *Queu. Sys. Theory. App.*, **6**, 137–154.

Daganzo, C.F. (1991), *Logistics Systems Analysis*, Springer-Verlag, New York, N.Y. (and 2nd edition, 1996.)

Daganzo, C.F. (1993). The cell-transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Institute of Transportation Studies, Research Report, UCB-ITS-RR-93-7, Univ. of California, Berkeley, CA; and in abbreviated form, *Trans. Res.*, **28B**(4), 269–287.

Daganzo, C.F. (1993a). A finite difference approximation for the kinematic wave model. Institute of Transportation Studies, Research Report, UCB-ITS-RR-93-11, Univ. of California, Berkeley, CA; and in *Trans. Res.*, **29B**(4), 261–276.

Daganzo, C.F. (1994). The cell-transmission model. Part II: Network traffic, PATH working paper UCB-ITS-PWP-94-12, Univ. of California, Berkeley, CA *Trans. Res.*, **29B**(2), 79–94.

Daganzo, C.F. (1995). Properties of link travel time functions under dynamic loads. *Trans. Res.* **29B**(2), 95–98.

Daganzo, C.F. (1995a). The nature of freeway gridlock and how to prevent it, Institute of Transportation Studies, Research Report UCB-ITS-RR-95-1, Univ. of California, Berkeley, CA; and in *Transportation and Traffic Theory*, pp. 629–646, J.B. Lesort, editor, Pergamon-Elsevier, New York, N.Y.)

Daganzo, C.F. (1996). Requiem for high order fluid approximations of traffic flow. *Trans. Res.* **29B**, 277–287.

Daganzo, C.F. (1996a). Effect of queue spillovers on transportation networks without a route choice. Institute of Transportation Studies, Research Report, UCB-ITS-RR-96-1, Univ. of California, Berkeley, CA.

Daganzo, C.F. and W.H. Lin (1995). The effect of modeling assumptions on the behavior of queues in a single corridor. *Trans. Res. Rec.*, **1453**, 66–74.

Del Castillo, J.M. (1994). Freeway data fitting: appendix C in 'A theory of car following'. PhD thesis, Engineering School, University of Sevilla, Spain.

Del Castillo, J.M. (1996). A car-following model based on the Lighthill and Whitham theory. in *Transportation and Traffic Theory*, pp. 517–538, (J.B. Lesort, editor) Pergamon-Elsevier, New York, N.Y.

Edie, L.C. (1963). Discussion of traffic stream measurements and definitions. *Proc. 2nd Int. Symp. on the Theory of Traffic Flow*, (J. Almond, editor), pp. 139–154, OECD, Paris, France.

Feller, W. (1970), *An introduction to probability theory and its application*, Vol. II (2nd. edition), Wiley, New York, N.Y.

Foster, J. (1962). An investigation of the hydrodynamic model for traffic flow with particular reference to the effect of various speed-density relationships. *Proc. 1st Conf. Aus. Road Res. Board*, pp. 229–257.

Gamse, B. and G.F. Newell (1982). An analysis of elevator operations in moderate height buildings I and II. *Trans. Res.*, **16B**(4), 303–319.

Gazis, D.C. and R. Herman (1993). The moving and 'phantom' bottlenecks. *Trans. Sci.*, **26**(3), pp. 223–229.

Gazis, D.C. and R.B. Potts (1965). The over-saturated intersection. *Proc. 2nd Int. Symp. on the Theory of Road Traffic Flow*, (J. Almond, editor), pp. 221–237, OECD, Paris, France.

Greenshields, B.D., D. Shapiro and E.L. Erickson (1947). Traffic performance at urban street intersections. *Technical Report 1*, Yale Bureau of Highway Traffic, New Haven, Connecticut.

Haefner, L.E. (1986) *Introduction to transportation systems*, Holt, Rinehart and Winston, New York, N.Y.

Hauer, E. (1969). Fleet selection for public transportation routes. Ph.D. thesis, Dept. of Civil Engineering, Univ. of California, Berkeley, CA.

Hay, W.W. (1977) *An introduction of transportation engineering* (2nd edition), Wiley, New York, N.Y.

Herman, R., T. Lam and R.W. Rothery (1971). The starting characteristics of automobile platoons. in *Traffic Flow and Transportation*, Proc. 5th Int. Symp. on the Theory of Traffic Flow and Transportation, (G.F. Newell, editor) pp. 1–17, American Elsevier, New York, N.Y.

Herman, R., E.W. Montroll, R.B. Potts, and R.W. Rothery (1959). Traffic dynamics: Analysis of stability in car following. *Opns. Res.*, **7**, pp. 86–106.

Heydecker, B.G. and Addison, J.D. (1996). An exact expression of dynamic equilibrium. in *Transportation and Traffic Theory* (J.B. Lesort, editor) pp. 359–384, Pergamon-Elsevier, New York, N.Y.

Highway Capacity Manual. (1965), *Highway Research Board*, Special Report 87, National Research Council, Washington, D.C.

Highway Capacity Manual (2nd ed., rev.) (1994), *Transportation Research Board Special Report*, 209, National Research Council, Washington, D.C.

Hillier, F.S. and Lieberman, G.J. (1995) *Introduction to operations research* (6th edition), Mc-Graw Hill, New York, N.Y.

Homburger, W.S., J.H. Kell, and D.D. Perkins (1992), *Fundamentals of Traffic Engineering*, 13th edition, Course Notes, UCB-ITS-CN-92-1, Institute of Transportation Studies, Univ. of California, Berkeley, CA.

Hurdle, V.F., Merlo, M.I. and Robertson, D. (1997). A study of speed vs. flow relationships on individual freeway lanes. Presented at the 1997 annual meeting of the Trans. Res. Board, Washington D.C.

Koshi, M., M. Kuwahara and H. Akahane (1992). Capacity of sags and tunnels in Japanese motorways. *ITE Journal*, May issue, pp. 17–22.

Kuwahara, M. and Akamatsu, T. (1993). Dynamic equilibrium assignment with queues for a one-to-many OD pattern. in *Transportation and Traffic Theory* (C.F. Daganzo, editor) pp. 185–204, Elsevier, New York, N.Y.

Langevin, A., Mbaraga, P. and Campbell, J.F. (1996). Continuous approximation models in freight distribution: an overview. *Trans Res.* B **30**, 163–188.

Larson, R. and A. Odoni, (1981), *Urban operation research*, Prentice-Hall, Englewood Cliffs, N.J.

Lax, P.D. (1954). Weak solutions of non-linear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.* 7, 159–193.

Lax, P.D. (1973), *Hyperbolic systems of conservation laws and the mathematical theory of shockwaves*, SIAM Regional Conference Series in Applied Mathematics. J.W. Arrowsmith Ltd., Bristol, U.K.

Lebacque, J.P. (1993). Les modèles macroscopiques de trafic," *Annales de Ponts*, (3rd trim) **67**, 28–45.

LeVeque, R.J. (1992), *Numerical methods for conservation laws*, (2nd edition), Birkhauser-Verlag, Boston, MA.

Lighthill, M.J. and G.B. Whitham (1955). On kinematic waves. I flow movement in long rives. II A theory of traffic flow on long crowded roads. *Proc. Roy. Soc.*, A. **229**, 281–345.

Lin, W.H. and C.F. Daganzo (1997). A simple detection scheme for delay-inducing freeway incidents. *Trans. Res.* A **31**, 144–155.

Linsley, R.K. and J.B. Franzini (1955), *Elements of hydraulic engineering*, McGraw- Hill, New York, N.Y.

Lovell, D. (1997). Traffic control on metered networks without route choice, PhD thesis, Dept. of Civil and Environmental Eng., U. of California, Berkeley, CA.

Luke, J.C. (1972). Mathematical models for landform evolution. *J.Geophys. Res.*, **77**, 2460–2464.

Makigami, Y., G.F. Newell, and R. Rothery (1971). Three-dimensional representation of traffic flow. *Trans. Sci.*, **5**, 302–313.

Mannering, F.L. and Kilareski, W.P. (1990) *Principles of highway engineering and traffic analysis*, Wiley, New York, N.Y.

Manual of Traffic Engineering Studies (1964), 3rd edition, (D.E. Cleveland, editor), Institute of Traffic Engineers (now Transportation Engineers), Washington, D.C.

Matson, T.M. (1929). The principles of traffic signal timing. *Proc. Nat. Saf. Council*, pp. 109–128. Chicago, IL.

May, A.D. (1964). Experimentation with manual and automatic ramp control. *High. Res. Rec.*, **59**, 9–38.

Morse, P.M. and Yaffe, H.J. (1971). A queueing model for car passing. *Trans. Sci.* **5**, 48–63.

Moskowitz, K (1954). Waiting for a gap in a traffic stream. *Proc. Highway Res. Board*, **33**, 385–395.

Moskowitz, K. (1965). Discussion of 'freeway level of service as influenced by volume and capacity characteristics' by D.R. Drew and C. J. Keese. *Highway Res. Rec.* 99, 43–44.

Newell, G.F. (1955). Mathematical models of freely flowing highway traffic. *J. Opns. Res. Soc. Am.* **3**, 176–188.

Newell, G.F. (1961). Non-linear effects in the dynamics of car-following. *Opns. Res.*, **9**(2), pp. 209–229.

Newell, G.F. (1965). Approximation methods for queues with application to the fixed-cycle traffic light. SIAM Review **7**(4), 223–240.

Newell, G.F. (1971), *Applications of queueing theory*, Chapman Hall, London.

Newell, G.F. (1974). Control of pairing vehicles on a public transportation route, two vehicles, one control point. *Trans. Sci.*, **8**(3), 248–269.

Newell, G.F. (1982), *Applications of queueing theory* (2nd edition), Chapman Hall, London.

Newell, G.F. (1989). Theory of highway traffic signals. Institute of Transportation Studies, Course Notes, UCB-ITS-CN-89-1, Univ. of California, Berkeley, CA.

Newell, G.F. (1993). A simplified theory of kinematic waves in highway traffic, I general theory, II queuing at freeway bottlenecks, III multi-destination flows. *Trans. Res.*, **27B**, 281–313.

Newell, G.F. (1993a). A moving bottleneck. Institute of transportation Studies, Research Report, UCB-ITS-RR-93-3, Univ. of California, Berkeley, CA.

Newell, G.F. (1994). Flow upstream of a bottleneck. Institute of transportation Studies, Research Report, UCB-ITS-RR-94-4, Univ. of California, Berkeley, CA.

Newell, G.F. (1995), *Theory of highway traffic flow: 1945 to 1965*, Institute of Transportation Studies Special Report, University of California, Berkeley.

Newell, G.F. and E.E. Osuna (1969). Properties of vehicle-actuated signals: I. One-way streets. *Trans. Sci.*, 3, 30–52, and. II Two-way streets. *Trans. Sci.*, **3**, 99–125.

Newell, G.F. and Potts, R.B. (1964). Maintaining a bus schedule. *Proc. Aus. Road Res. Board*, **2**(1), 388–393.

Newman, L., A. Dunnet and J. Meirs (1969). Freeway ramp control-What it can and cannot do. *Traffic Engineering*, (June), 14–25.

Oris, W.J. (1987), *1-2-3 for scientists and engineers*, SYBEX, San Francisco, CA.

Payne, H. (1971). Models of freeway traffic control. *Math. Models Publ. Sys. Simul. Council Proc.*, **28**(1), 51–61.

Potts, R.B. and Tamlin, E.A. (1964). Pairing of buses. *Proc. Aus. Road Res. Board* 2, pp. 3–9.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1986), *Numerical recipes*, Cambridge Univ. Press, Cambridge, U.K.

Prigogine, I. (1959). A Boltzmann-like approach to the statistical theory of traffic flow. *Proc. Symp. on Theory of Traffic Flow*, (R. Herman, editor), pp. 158–164, American Elsevier, New York, N.Y.

Prigogine, I. and R. Herman (1971), *Kinetic Theory of Vehicle Traffic*, Elsevier, New York, N.Y.

Richards, P.I. (1956). Shockwaves on the highway. *Opns. Res.*, **4**, 42–51.

Rothery, R. (1968). Car-following - a deterministic model for single lane traffic flow. Ph.D. thesis, Faculte des Sciences Service Chimie-Physique II, Univ. Libre de Bruxelles, Belgium.

Rothery, R., R. Silver, R. Herman, and C. Turner (1964). Analysis of experiments on single-lane bus flow. *Opns. Res.*, **12**(6), 913–933.

Sloan, W.G. (1927). Discussion, report of committee on highway traffic analysis. *Proc. Highway Research Board*, **7**, 259–268.

Smeed, R.J. (1967). Some circumstances in which vehicles will reach their destinations earlier by starting later. *Trans. Sci.*, **1**, 308–317.

Smith, M.J. (1979). Traffic control and route choice; a simple example. *Trans. Res.*, **13B**, 289–295.

Stohr, T.F. (1966). The history of traffic signal theory. Graduate Student Report, ITS Library, University of California, Berkeley.

Wardrop, J.G. (1952). Some theoretical aspects of road traffic research. *Proc. Inst. Civ. Eng., Part II*, 1(2), 325–362; Discussion, 362–378.

Wattleworth, J.A. (1963). Peak period control of a freeway system–some theoretical considerations. Chicago Area Expressway Surveillance Project, Report 9. (see also *Highway Res. Rec.* 89, 1–25, 1965, with D.S. Berry.)

Webster, F.V. (1958). Traffic signal settings. Road Research Technical Paper, No. 39, Road Research Laboratory, Ministry of Transport, HMSO, London, U.K.

Whitham, G.B. (1974), *Linear and non-linear waves*, Wiley, New York, N.Y.

Woodside, C.M., B. Pagurek and G.F. Newell (1980). A diffusion approximation for correlation in queues. *J. Appl. Prob.*, **17**, 1033–1047.

AUTHOR INDEX

Y

SUBJECT INDEX

C

D

F

G

M

N

O

P

Q

S