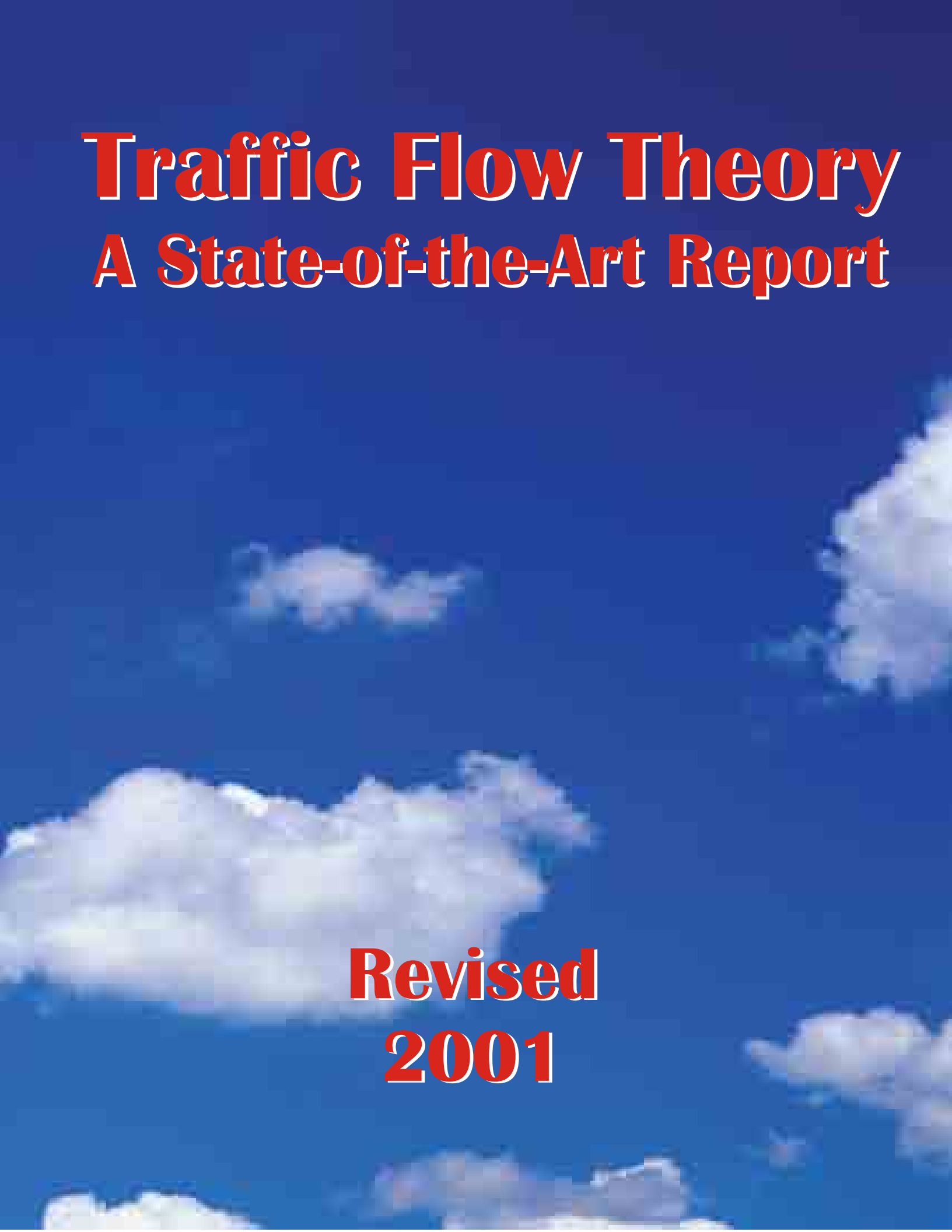

Traffic Flow Theory
A State-of-the-Art Report
Revised
2001

Traffic Flow Theory
A State-of-the-Art Report

Revised
2001

Organized by the Committee on Traffic Flow Theory and Characteristics (AHB45)

TABLE OF CONTENTS

LIST OF FIGURES

2. TRAFFIC STREAM CHARACTERISTICS

3. HUMAN FACTORS

4. CAR FOLLOWING MODELS

5. CONTINUUM FLOW MODELS

6. MACROSCOPIC FLOW MODELS

7. TRAFFIC IMPACT MODELS

8. UNSIGNALIZED INTERSECTION THEORY

9. TRAFFIC FLOW AT SIGNALIZED INTERSECTIONS

10. TRAFFIC SIMULATION

11. KINETIC THEORIES

List of Tables

3. HUMAN FACTORS

4. CAR FOLLOWING MODELS

5. CONTINUUM MODELS

7. TRAFFIC IMPACT MODELS

8. UNSIGNALIZED INTERSECTION THEORY

9. TRAFFIC FLOW AT SIGNALIZED INTERSECTIONS

10. TRAFFIC SIMULATION

11. KINETIC THEORIES

FOREWORD

This publication is an update and expansion of Transportation Research Board Special Report 165, "Traffic Flow Theory," published in 1975. This updating was undertaken on recommendation of the Transportation Research Board's Committee A3A11 on Traffic Flow Theory and Characteristics. The Federal Highway Administration (FHWA) funded a project to develop this report via an Interagency Agreement with the Department of Energy's Oak Ridge National Laboratory (ORNL). The project was carried out by ORNL under supervision of an Advisory Committee that, in addition to the three co-editors, included the following prominent individuals:

Richard Cunard, *TRB Liaison Representative*
Dr. Henry Lieu, *Federal Highway Administration*
Dr. Hani Mahmassani, *University of Texas at Austin*

While the general philosophy and organization of the previous two reports have been retained, the text has been completely rewritten and two new chapters have been added. The primary reasons for doing such a major revision were to bring the material up-to-date; to include new developments in traffic flow theory (e.g., network models); to ensure consistency among chapters and topics; and to emphasize the applications or practical aspects of the theory. There are completely new chapters on human factors (Chapter 3) and network traffic models (Chapter 5).

To ensure the highest degree of reliability, accuracy, and quality in the content of this report, the collaboration of a large number of experts was enlisted, and this report presents their cooperative efforts. We believe that a serious effort has been made by the contributing authors in this report to present theory and information that will have lasting value. Our appreciation is extended to the many authors for their commendable efforts in writing this update, willingly sharing their valuable time, knowledge, and their cooperative efforts throughout the project.

We would also like to acknowledge the time spent by the members of the Advisory Committee in providing guidance and direction on the style of the report and for their reviews of the many drafts of the report.

Additional acknowledgment is made to Alberto Santiago, Chief of State Programs at National Highway Institute of the FHWA (formerly with Intelligent Systems and Technology Division), without whose initiative and support, this report simply would not have been possible. Thanks also go to Brenda Clark for the initial formatting of the report, Kathy Breeden for updating the graphics and text and coordinating the effort with the authors, Phil Wolff for the creation and management of the report's web-site, and to Elaine Thompson for her project management assistance.

Finally, we acknowledge the following individuals who read and reviewed part or all of the manuscript and contributed valuable suggestions: Rahmi Akcelik, Rahim Benekohal, David Boyce, Micheal Brackstone, Werner Brilon, Christine Buisson, Ennio Cascetta, Michael Cassidy, Avishai Ceder, Arun Chatterjee, Ken Courage, Ray Derr, Mike Florian, Fred Hall, Benjamin Heydecker, Ben Hurdle, Shinya Kikuchi, Helmut "Bill" Knee, Haris Koutsopoulos, Jean-Baptiste Lesort, John Leonard II, Fred Mannering, William McShane, Kyriacos Mouskos, Panos Prevedourous, Vladimir Protopopescu, Bin Ran, Tom Rockwell, Mitsuru Saito, and Natacha Thomas.

We believe that this present publication meets its objective of synthesizing and reporting, in a single document, the present state of knowledge or lack thereof in traffic flow theory. It is sincerely hoped that this report will be useful to the graduate students, researchers and practitioners, and others in the transportation profession.

Editors: Dr. Nathan Gartner
University of Massachusetts - Lowell

Dr. Carroll J. Messer
Texas A&M University

Dr. Ajay K. Rathi
Oak Ridge National Laboratory
Project Leader.

INTRODUCTION

BY NATHAN H. GARTNER[1]
CARROLL MESSER[2]
AJAY K. RATHI[3]

[1] Professor, Department of Civil Engineering, University of Massachusetts at Lowell, 1 University Avenue, Lowell, MA 01854.

[2] Professor, Department of Civil Engineering, Texas A&M University, TTI Civil Engineering Building, Suite 304C, College Station, TX 77843-3135.

[3] Senior R&D Program Manager and Group Leader, ITS Research, Center for Transportation Analysis, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6207.

1. INTRODUCTION

It is hardly necessary to emphasize the importance of transportation in our lives. In the United States, we spend about 20 percent of Gross National Product (GNP) on transportation, of which about 85 percent is spent on highway transportation (passenger and freight). We own and operate 150 million automobiles and an additional 50 million trucks, bringing car ownership to 56 per hundred population (highest in the world). These vehicles are driven an average of 10,000 miles per year for passenger cars and 50,000 miles per year for trucks on a highway system that comprises more than 4 million miles. The indices in other countries may be somewhat different, but the importance of the transportation system, and especially the highway component of it, is just the same or even greater. While car ownership in some countries may be lower, the available highway network is also smaller leading to similar or more severe congestion problems.

Traffic flow theories seek to describe in a precise mathematical way the interactions between the vehicles and their operators (the *mobile* components) and the infrastructure (the *immobile* component). The latter consists of the highway system and all its operational elements: control devices, signage, markings, etc. As such, these theories are an indispensable construct for all models and tools that are being used in the design and operation of streets and highways. The scientific study of traffic flow had its beginnings in the 1930's with the application of probability theory to the description of road traffic (Adams 1936) and the pioneering studies conducted by Bruce D. Greenshields at the Yale Bureau of Highway Traffic; the study of models relating volume and speed (Greenshields 1935) and the investigation of performance of traffic at intersections (Greenshields 1947). After WWII, with the tremendous increase in use of automobiles and the expansion of the highway system, there was also a surge in the study of traffic characteristics and the development of traffic flow theories. The 1950's saw theoretical developments based on a variety of approaches, such as car-following, traffic wave theory (hydrodynamic analogy) and queuing theory. Some of the seminal works of that period include the works by Reuschel (1950a; 1950b; 1950c), Wardrop (1952), Pipes (1953), Lighthill and Whitham (1955), Newell (1955), Webster (1957), Edie and Foote (1958), Chandler et al. (1958) and other papers by Herman et al. (see Herman 1992).

By 1959 traffic flow theory had developed to the point where it appeared desirable to hold an international symposium. The First International Symposium on The Theory of Traffic Flow was held at the General Motors Research Laboratories in Warren, Michigan in December 1959 (Herman 1961). This was the first of what has become a series of triennial symposia on The Theory of Traffic flow and Transportation. The most recent in this series, the 12th symposium was held in Berkeley, California in 1993 (Daganzo 1993). A glance through the proceedings of these symposia will provide the reader with a good indication of the tremendous developments in the understanding and the treatment of traffic flow processes in the past 40 years. Since that time numerous other symposia and specialty conferences are being held on a regular basis dealing with a variety of traffic related topics. The field of traffic flow theory and transportation has become too diffuse to be covered by any single type of meeting. Yet, the fundamentals of traffic flow theory, while better understood and more easily characterized through advanced computation technology, are just as important today as they were in the early days. They form the foundation for all the theories, techniques and procedures that are being applied in the design, operation, and development of advanced transportation systems.

It is the objective of this monograph to provide an updated survey of the most important models and theories that characterize the flow of highway traffic in its many facets. This monograph follows in the tracks of two previous works that were sponsored by the Committee on Theory of Traffic Flow of the Transportation Research Board (TRB) and its predecessor the Highway Research Board (HRB). The first monograph, which was published as HRB Special Report 79 in 1964, consisted of selected chapters in the then fledgling Traffic Science each of which was written by a different author (Gerlough and Capelle 1964). The contents included:

Chapter 1. Part I: Hydrodynamic Approaches, by L. A. Pipes. Part II: On Kinematic Waves; A Theory of Traffic Flow on Long Crowded Roads, by M. J. Lighthill and G. B. Whitham.

Chapter 2. Car Following and Acceleration Noise, by E. W. Montroll and R. B. Potts.

Chapter 3. Queuing Theory Approaches, by D. E. Cleveland and D. G. Capelle.

Chapter 4. Simulation of Traffic Flow, by D. L. Gerlough.

Chapter 5. Some Experiments and Applications, by R. S. Foote.

A complete rewriting of the monograph was done by Gerlough and Huber (1975) and was published as TRB Special Report 165 in 1975. It consisted of nine chapters, as follows:

Chapter 1. Introduction.

Chapter 2. Measurement of Flow, Speed, and Concentration.

Chapter 3. Statistical Distributions of Traffic Characteristics.

Chapter 4. Traffic Stream Models.

Chapter 5. Driver Information Processing Characteristics.

Chapter 6. Car Following and Acceleration Noise.

Chapter 7. Hydrodynamic and Kinematic Models of Traffic.

Chapter 8. Queuing Models (including Delays at Intersections).

Chapter 9. Simulation of Traffic Flow.

This volume is now out of print and in 1987 the TRB Committee on Traffic Flow Theory and Characteristics recommended that a new monograph be prepared as a joint effort of committee members and other authors. While many of the basic theories may not have changed much, it was felt that there were significant developments to merit writing of a new version of the monograph. The committee prepared a new outline which formed the basis for this monograph. After the outline was agreed upon, the Federal Highway Administration supported this effort through an interagency agreement with the Oak Ridge National Laboratory. An Editorial Committee was appointed, consisting of N. H. Gartner, C. J. Messer, and A. K. Rathi, which was charged with the editorial duties of the preparation of the manuscripts for the different chapters.

The first five chapters follow similarly titled chapters in the previous monograph; however, they all have been rewritten in their entirety and include the latest research and information in their respective areas. Chapter 2 presents the various models that have been developed to characterize the relationships among the traffic stream variables: speed, flow, and concentration. Most of the relationships are concerned with uninterrupted traffic flow, primarily on freeways or expressways. The chapter stresses the link between theory and measurement capability, since to as large extent development of the first depends on the latter.

Chapter 3, on Human Factors, discusses salient performance aspects of the human element in the context of the person-machine system, i.e. the motor vehicle. The chapter describes first discrete components of performance, including: perception-reaction time, control movement time, responses to traffic control devices, to the movement of other vehicles, to hazards in the roadway, and how different segments of the population differ in performance. Next, the kind of control performance that underlies steering, braking, and speed control -- the primary control functions -- is described. Applications of open-loop and closed-loop vehicle control to specific maneuvers such as lane keeping, car following, overtaking, gap acceptance, lane closures, and sight distances are also described. To round out the chapter, a few other performance aspects of the driver-vehicle system are covered, such as speed limit changes, distractions on the highway, and responses to real-time driver information. The most obvious application of human factors is in the development of Car Following Models (Chapter 4). Car following models examine the manner in which individual vehicles (and their drivers) follow one another. In general, they are developed from a stimulus-response relationship, where the response of successive drivers in the traffic stream is to accelerate or decelerate in proportion to the magnitude of the stimulus at time t after a time lag T. Car following models form a bridge between the microscopic behavior of individual vehicles and the macroscopic characteristics of a single-lane traffic stream with its corresponding flow and stability properties.

Chapter 5 deals with Continuous Flow Models. Because traffic involves flows, concentrations, and speeds, there is a natural tendency to attempt to describe traffic in terms of fluid behavior. Car following models recognize that traffic is made up of discrete particles and it is the interactions between these particles that have been developed for fluids (i.e., continuum models) is concerned more with the over all statistical behavior of the traffic stream rather than with the interactions between the particles. In the fluid flow analogy, the traffic stream is treated

as a one dimensional compressible fluid. This leads to two basic assumptions: (i) traffic flow is conserved, which leads to the conservation or continuity equation, and (ii) there is a one-to-one relationship between speed and density or between flow and density. The simple continuum model consists of the conservation equation and the equation of state (speed-density or flow-density relationship). If these equations are solved together with the basic traffic flow equation (flow equals density times speed) we can obtain the speed, flow and density at any time and point of the roadway. By knowing these basic traffic flow variables, we know the state of the traffic system and can derive measures of effectiveness, such as delays, stops, travel time, total travel and other measures that allow the analysts to evaluate how well the traffic system is performing. In this chapter, both simple and high order models are presented along with analytical and numerical methods for their implementation.

Chapter 6, on Macroscopic Flow Models, discards the microscopic view of traffic in terms of individual vehicles or individual system components (such as links or intersections) and adopts instead a macroscopic view of traffic in a network. A variety of models are presented together with empirical evidence of their applicability. Variables that are being considered, for example, include the traffic intensity (the distance travelled per unit area), the road density (the length or area of roads per unit area of city), and the weighted space mean speed. The development of such models extends traffic flow theory into the network level and can provide traffic engineers with the means to evaluate system-wide control strategies in urban areas. Furthermore, the quality of service provided to motorists can be monitored to assess the city's ability to manage growth. Network performance models could also be used to compare traffic conditions among different cities in order to determine the allocation of resources for transportation system improvements. Chapter 7 addresses Traffic Impact Models, specifically, the following models are being discussed: Traffic and Safety, Fuel Consumption Models, and Air Quality Models.

Chapter 8 is on Unsignalized Intersection Theory. Unsignalized intersections give no positive indication or control to the driver. The driver alone must decide when it is safe to enter the intersection, typically, he looks for a safe opportunity or "gap" in the conflicting traffic. This model of driver behavior is called "gap acceptance." At unsignalized intersections the driver must also respect the priority of other drivers. This chapter discusses in detail the gap acceptance theory and the headway distributions used in gap acceptance calculations. It also discusses the intersections among two or more streams and provides calculations of capacities and quality of traffic operations based on queuing modeling.

Traffic Flow at Signalized Intersections is discussed in Chapter 9. The statistical theory of traffic flow is presented, in order to provide estimates of delays and queues at isolated intersections, including the effect of upstream traffic signals. This leads to the discussion of traffic bunching, dispersion and coordination at traffic signals. The fluid (shock-wave) approach to traffic signal analysis is not covered in this chapter; it is treated to some extent in Chapter 5. Both pretimed and actuated signal control theory are presented in some detail. Further, delay models that are founded on steady-state queue theory as well as those using the so-called coordinate transform method are covered. Adaptive signal control is discussed only in a qualitative manner since this topic pertains primarily to the development of optimal signal control strategies, which is outside the scope of this chapter.

The last chapter, Chapter 10, is on Traffic Simulation. Simulation modeling is an increasingly popular and effective tool for analyzing a wide variety of dynamical problems which are not amenable to study by other means. These problems are usually associated with complex processes which can not readily be described in analytical terms. To provide an adequate test bed, the simulation model must reflect with fidelity the actual traffic flow process. This chapter describes the traffic models that are embedded in simulation packages and the procedures that are being used for conducting simulation experiments.

Consideration was also given to the addition of a new chapter on Traffic Assignment Models. Traffic assignment is the process of predicting how a given set of origin-destination (OD) trip demands will manifest themselves onto a transportation network of links and nodes in terms of flows and queues. It has major applications in both transportation planning models and in dynamic traffic management models which are the bedrock of Intelligent Transportation Systems (ITS). Generally, the assignment process consists of a macroscopic simulation of the behavior of travelers in a network of transportation facilities. At the same time it reflects the interconnection between the microscopic models of traffic behavior that are discussed in this monograph and the overall distribution of traffic demands throughout the network. This is expressed by link cost functions that serve as a basis for any assignment or route choice process. After much deliberation by the editorial and advisory committees

it was decided that the subject cannot be presented adequately in a short chapter within this monograph. It would be better served by a dedicated monograph of its own, or by reference to the extensive literature in this area. Early references include the seminal works of Wardrop (1952), and Beckmann, McGuire and Winsten (1956). Later publications include books by Potts and Oliver (1972), Florian (1976), Newell (1980), and Sheffi (1985). Recent publications, which reflect modern approaches to equilibrium assignment and to dynamic traffic assignment, include books by Patriksson (1994), Ran and Boyce (1994), Gartner and Improta (1995), Florian and Hearn (1995), and Bell and Iida (1997). This is a lively research area and new publications abound.

Research and developments in transportation systems and, concomitantly, in the theories that accompany them proceed at a furious pace. Undoubtedly, by the time this monograph is printed, distributed, and read, numerous new developments will have occurred. Nevertheless, the fundamental theories will not have changed and we trust that this work will provide a useful source of information for both newcomers to the field and experienced workers.

References

Adams, W. F. (1936). *Road Traffic Considered as a Random Series*, J. Inst. Civil Engineers, 4, pp. 121-130, U.K.

Beckmann, M., C.B. McGuire and C.B. Winsten (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven.

Bell, M.G.H. and Y. Iida (1997). *Transportation Network Analysis*. John Wiley & Sons.

Chandler, R. E., R. Herman, and E. W. Montroll, (1958). *Traffic Dynamics: Studies in Car Following*, Opns. Res. 6, pp. 165-183.

Daganzo, C. F., Editor (1993). *Transportation and Traffic Theory. Proceedings*, 12th Intl. Symposium. Elsevier Science Publishers.

Edie, L. C. and R. S. Foote, (1958). *Traffic Flow in Tunnels*, Proc. Highway Research Board, 37, pp. 334-344.

Florian, M.A., Editor (1976). *Traffic Equilibrium Methods*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag.

Florian, M. and D. Hearn (1995). *Network Equilibrium Models and Algorithms.* Chapter 6 in Network Routing (M.O. Ball et al., Editors), Handbooks in OR & MS, Vol. 8, Elsevier Science.

Gartner, N.H. and G. Improta, Editors (1995). *Urban Traffic Networks; Dynamic Flow Modeling and Control*. Springer-Verlag.

Gerlough, D. L. and D. G. Capelle, Editors (1964). *An Introduction to Traffic Flow Theory*. Special Report 79. Highway Research Board, Washington, D.C.

Gerlough, D. L. and M. J. Huber, (1975). *Traffic Flow Theory - A Monograph*. Special Report 165, Transportation Research Board.

Greenshields, B. D. (1935). *A Study in Highway Capacity*. Highway Research Board, Proceedings, Vol. 14, p. 458.

Greenshields, B. D., D. Schapiro, and E. L. Erickson, (1947). *Traffic Performance at Urban Intersections*. Bureau of Highway Traffic, Technical Report No. 1. Yale University Press, New Haven, CT.

Herman, R., Editor (1961). *Theory of Traffic Flow*. Elsevier Science Publishers.

Herman, R., (1992). *Technology, Human Interaction, and Complexity: Reflections on Vehicular Traffic Science*. Operations Research, 40(2), pp. 199-212.

Lighthill, M. J. and G. B. Whitham, (1955). *On Kinematic Waves: II. A Theory of Traffic Flow on Long Crowded Roads*. Proceedings of the Royal Society: A229, pp. 317-347, London.

Newell, G. F. (1955). *Mathematical Models for Freely Flowing Highway Traffic*. Operations Research 3, pp. 176-186.

Newell, G. F. (1980). *Traffic Flow on Transportation Networks*. The MIT Press, Cambridge, Massachusetts.

Patriksson, M. (1994). *The Traffic Assignment Problem; Models and Methods*. VSP BV, Utrecht, The Netherlands.

Pipes, L. A. (1953). *An Operational Analysis of Traffic Dynamics.* J. Appl. Phys., 24(3), pp. 274-281.

Potts, R.B. and R.M. Oliver (1972). *Flows in Transportation Networks*. Academic Press.

Ran, B. and D. E. Boyce (1994). *Dynamic Urban Transportation Network Models; Theory and Implications for Intelligent Vehicle-Highway Systems*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag.

Reuschel, A. (1950a). *Fahrzeugbewegungen in der Kolonne.* Oesterreichisches Ingenieur-Aarchiv 4, No. 3/4, pp. 193-215.

Reuschel, A. (1950b and 1950c). *Fahrzeugbewegungen in der Kolonne bei gleichformig beschleunigtem oder verzogertem Leitfahrzeug.* Zeitschrift des Oesterreichischen Ingenieur- und Architekten- Vereines 95, No. 7/8 59-62, No. 9/10 pp. 73-77.

Sheffi, Y. (1985). *Urban Transportation Networks; Equilibrium Analysis with Mathematical Programming Methods.* Prentice-Hall.

Wardrop, J. G. (1952). *Some Theoretical Aspects of Road Traffic Research.* Proceedings of the Institution of Civil Engineers, Part II, 1(2), pp. 325-362, U.K.

Webster, F. V. (1958). *Traffic Signal Settings.* Road Research Technical Paper No. 39. Road Research Laboratory, London, U.K.

TRAFFIC STREAM CHARACTERISTICS

BY FRED L. HALL[4]

[4]Professor, McMaster University, Department of Civil Engineering and Department of Geography, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4L7.

CHAPTER 2 - Frequently used Symbols

k	density of a traffic stream in a specified length of road
L	length of vehicles of uniform length
c_k	constant of proportionality between occupancy and density, under certain simplifying assumptions
k_i	the (average) density of vehicles in substream *I*
q_i	the average rate of flow of vehicles in substream *I*
$\bar{u}$	average speed of a set of vehicles
A	$A(x,t)$ the cumulative vehicle arrival function over space and time
k_j	jam density, i.e. the density when traffic is so heavy that it is at a complete standstill
u_f	free-flow speed, i.e. the speed when there are no constraints placed on a driver by other vehicles on the road

2. TRAFFIC STREAM CHARACTERISTICS

Author's note: This material has benefited greatly from the assistance of Michael Cassidy, of the University of California at Berkeley. He declined the offer to be listed as a co-author of the chapter, although that would certainly have been warranted. With his permission, and that of the publisher, several segments of this material have been reproduced directly from his chapter, "Traffic Flow and Capacity", which appears as Chapter 6 in ***Handbook of transportation science****, edited by Randolph W. Hall, and published by Kluwer Academic Publishers in 1999. That material appears in italics below, in section 2.1. The numbering of Figures and equations has been altered from his numbers to correspond to numbering within this chapter.*

In this chapter, properties that describe highway traffic are introduced, such as flow, density, and average vehicle speed; issues surrounding their measurements are discussed; and various models that have been proposed for describing relationships among these properties are presented. Most of the work dealing with these relationships has been concerned with uninterrupted traffic flow, primarily on freeways or expressways, but the general relationships will apply to all kinds of traffic flow.

This chapter starts with a section on definitions of key variables and terms. Because of the importance of measurement capability to theory development, the second section deals with measurement issues, including historical developments in measurement procedures, and criteria for selecting good measurement characteristics. The third section discusses a number of the bivariate models that have been proposed in the past to relate key variables. That is followed by a short section on attempts to deal simultaneously with the three key variables. The final, summary section provides links to a number of the other chapters in this monograph.

2.1 Definitions and terms

This section provides definitions of some properties commonly used to characterize traffic streams, together with some generalized definitions that preserve useful relations among these properties. Before turning to the definitions, however, a useful graphical tool is introduced, namely trajectories plotted on a time-space diagram. This is followed by definitions of traffic stream properties, time-mean and space-mean properties, and the generalized definitions. Following that is a discussion of the relationship between density, which is often used in the generalized definitions, and occupancy, which is measured by many freeway systems. The final topic covered is three-dimensional representations of traffic streams. The discussion in this section follows very closely or repeats some of that in the chapter by Cassidy (1999), which in turn credits notes composed by Newell (unpublished) and a textbook written by Daganzo (1997).

2.1.1 The Time-Space Diagram.

The Time-Space Diagram. *Objects are commonly constrained to move along a one-dimensional guideway, be it, for example, a highway lane, walkway, conveyor belt, charted course or flight path. Thus, the relevant aspects of their motion can often be described in cartesian coordinates of time, t, and space, x. Figure 2-1 illustrates the trajectories of some objects traversing a facility of length L during time interval T; these objects may be vehicles, pedestrians or cargo. Each trajectory is assigned an integer label in the ascending order that the object would be seen by a stationary observer. If one object overtakes another, their trajectories may exchange labels, as shown for the fourth and fifth trajectories in the figure. Thus, the ℓth trajectory describes the location of a reference point (e.g. the front end) of object ℓ as a function of time t, $x_\ell(t)$.*

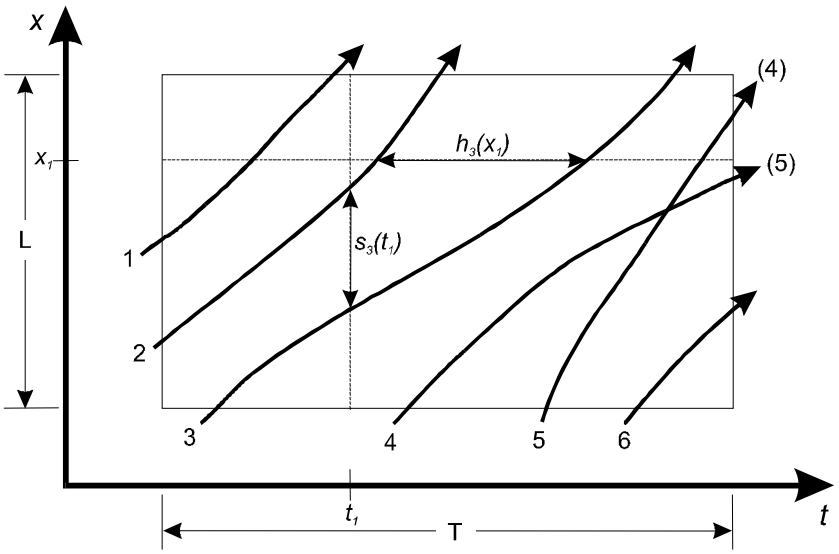


Figure 2-1. Time-space diagram.

The characteristic geometries of trajectories on a time-space diagram describe the motion of objects in detail. These diagrams thus offer the most complete way of displaying the observations that may have actually been measured along a facility. As a practical matter, however, one is not likely to collect all the data needed to construct trajectories. Rather, time-space diagrams derive their (considerable) value by providing a means to highlight the key features of a traffic stream using only coarsely approximated data or hypothetical data from "thought experiments."

2.1.2 Definitions of Some Traffic Stream Properties.

It is evident from Figure 2-1 that the slope of the ℓth trajectory is object ℓ's instantaneous velocity, $v_\ell(t)$, i.e.,

$$v_\ell(t) \equiv dx_\ell(t)/dt\,, \qquad (2.1)$$

and that the curvature is its acceleration. Further, there exist observable properties of a traffic stream that relate to the times that objects pass a fixed location, such as location x_1, for example.

These properties are described with trajectories that cross a horizontal line drawn through the time-space diagram at x_1.

Referring to Figure 2-1, the headway of some ith object at x_1, $h_i(x_1)$, is the difference between the arrival times of i and i-1 at x_1, i.e.,

$$h_i(x_1) \equiv t_i(x_1) - t_{i-1}(x_1) \tag{2.2}$$

Flow at x_1 is m, the number of objects passing x_1, divided by the observation interval T,

$$q(T, x_1) \equiv m/T. \tag{2.3}$$

For observation intervals containing large m,

$$\sum_{i=1}^{m} h_i(x_1) \approx T \tag{2.4}$$

and thus,

$$q(T, x_1) \approx \frac{1}{\frac{1}{m}\sum_{i=1}^{m} h_i(x_1)} = \frac{1}{\bar{h}(x_1)}, \tag{2.5}$$

i.e., flow is the reciprocal of the average headway.

Analogously, some properties relate to the locations of objects at a fixed time, as observed, for example, from an aerial photograph. These properties may be described with trajectories that cross a vertical line in the t-x plane. For example, the spacing of object j at some time t_1, $s_j(t_1)$, is the distance separating j from the next downstream object; i.e.,

$$s_j(t_1) \equiv x_{j-1}(t_1) - x_j(t_1). \tag{2.6}$$

Density at instant t_1 is n, the number of objects on a facility at that time, divided by L, the facility's physical length; i.e.,

$$k(L, t_1) \equiv n/L. \tag{2.7}$$

If the L contains large n,

$$\sum_{j=1}^{n} s_j(t_1) \approx L \tag{2.8}$$

and

$$k(L, t_1) \approx \frac{1}{\frac{1}{n}\sum_{j=1}^{n} s_j(t_1)} = \frac{1}{\bar{s}(t_1)}, \tag{2.9}$$

giving a relation between density and the average spacing parallel to that of flow and the average headway.

2.1.3 Time-Mean and Space-Mean Properties. *For an object's attribute α, where α might be its velocity, physical length, number of occupants, etc., one can define an average of the m objects passing some fixed location x_1 over observation interval T,*

$$\alpha(T, x_1) = \frac{1}{m}\sum_{i=1}^{m}\alpha_i(x_1), \qquad (2.10)$$

i.e., a time-mean of attribute α. If α is headway, for example, $\alpha(T, x_1)$ is the average headway or the reciprocal of the flow.

Conversely, the space-mean of attribute α at some time t_1, $\alpha(L, t_1)$, is obtained from the observations taken at that time over a segment of length L, i.e.,

$$\alpha(L, t_1) = \frac{1}{n}\sum_{j=1}^{n}\alpha_j(t_1). \qquad (2.11)$$

If, for example, α is spacing, $\alpha(L, t_1)$ is the average spacing or the reciprocal of the density.

For any attribute α, there is no obvious relation between its time and space means. The reader may confirm this (using the example of α as velocity) by envisioning a rectangular time-space region $L \times T$ traversed by vehicles of two classes, fast and slow, which do not interact. For each class, the trajectories are parallel, equidistant and of constant slope; such conditions are said to be ***stationary****.*

The fraction of fast vehicles distributed over L as seen on an aerial photograph taken at some instant within T will be smaller than the fraction of fast vehicles crossing some fixed point along L during the interval T. This is because the fast vehicles spend less time in the region than do the slow ones. Analogously, one might envision a closed loop track and note that a fast vehicle passes a stationary observer more often than does a slow one.

2.1.4 Generalized Definitions of Traffic Stream Properties. *To describe a traffic stream, one usually seeks to measure properties that are not sensitive to the variations in the individual objects (e.g. the vehicles or their operators) without averaging-out features of interest. This is the trade-off inherent in choosing between short and long measurement intervals, as previously noted. It was partly to address this trade-off that Edie (1965, 1974) proposed some generalized definitions of flow and density that averaged these properties in the manner described below.*

To begin this discussion, the thin, horizontal rectangle in Figure 2-2 corresponds to a fixed observation point. As per its conventional definition provided earlier, the flow at this point is m/T, where $m = 4$ in the figure. Since this point in space is a region of temporal duration T and elemental spatial dimension dx, the flow can be expressed equivalently as $\frac{m \cdot dx}{T \cdot dx}$. The denominator is the euclidean area of the thin horizontal rectangle, expressed in units of distance $\times$ time. The numerator

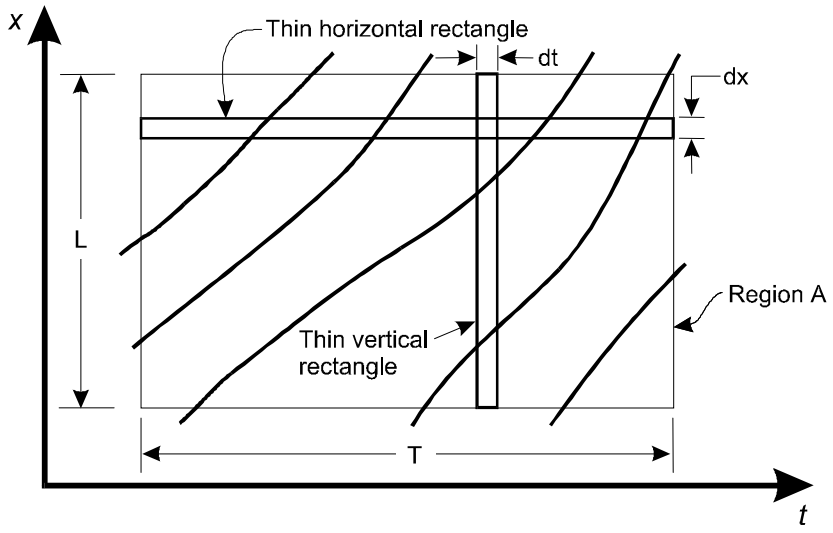


Figure 2-2. Trajectories in time-space region.

is the total distance traveled by all objects in this thin region, since objects cannot enter or exit the region via its elementally small left and right sides.

That flow, then, is the ratio of the distance traveled in a region to the region's area is valid for any time-space region, since all regions are composed of elementary rectangles. Taking, for example, region A in Figure 2-2, Edie's generalized definition of the flow in A, q(A), is d(A) / |A|, where d(A) is the total distance traveled in A and |A| is used to denote the region's area.

As the analogue to this, the thin, vertical rectangle in Figure 2-2 corresponds to an instant in time. As per its conventional definition, density is n / L (where n = 2 in this figure) and this can be expressed equivalently as $\frac{n \cdot dt}{L \cdot dt}$*. It follows that Edie's generalized definition of density in a region A, k(A), is t(A) / |A|, where t(A) is the total time spent in A.*

It should be clear that these generalized definitions merely average the flows collected over all points, and the densities collected at each instant, within the region of interest. Dividing this flow by this density gives d(A) / t(A), which can be taken as the average velocity of objects in A, v(A). The reader will note that, with Edie's definitions, the average velocity is the ratio of flow to density. Traffic measurement devices, such as loop detectors installed beneath the road surface, can be used to measure flows, densities and average vehicle velocities in ways that are consistent with these generalized definitions. Discussion of this is offered in Cassidy and Coifman (1997).

As a final note regarding v(A), when A is taken as a thin horizontal rectangle of spatial dimension dx, the time spent in the region by object i is dx / v_i, where v_i is i's velocity. Thus, for this thin,

horizontal region A, $t(A) = dx\sum_{i=1}^{m}\frac{1}{v_i}$. Given that for the same region, d(A) = m·dx, the generalized mean velocity becomes

$$v(A) = \frac{d(A)}{t(A)} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{v_i}}, \tag{2.12}$$

i.e., the reciprocal of the mean of the reciprocal velocities, or the harmonic mean velocity. The $1/v_i$ is often referred to as the pace of i, p_i, and thus

$$v(A) = \left[\frac{1}{m}\sum_{i=1}^{m} p_i\right]^{-1}. \tag{2.13}$$

Eq. 2.15 applies for regions with L > dx provided that all i span the L and that each p_i (or v_i) is i's average over the L.

It follows that when conditions in a region A are stationary, the harmonic mean of the velocities measured at a fixed point in A is the v(A). By the same token, the v(A) is the space-mean velocity measured at any instant in A (provided, again, that conditions are stationary).

2.1.5 The Relation Between Density and Occupancy. *Occupancy is conventionally defined as the percentage of time that vehicles spend atop a loop detector. It is a commonly-used property for describing highway traffic streams; it is used later in this chapter, for example, for diagnosing freeway traffic conditions. In particular, occupancy is a proxy for density. The following discussion demonstrates that the former is merely a dimensionless version of the latter.*

One can readily demonstrate this relation by adopting a generalized definition of occupancy analogous to the definitions proposed by Edie. Such a definition is made evident by illustrating each trajectory with two parallel lines tracing the vehicle's front and rear (as seen by a detector) and this is exemplified in Figure 2-3. The (generalized) occupancy in the region A, ρ(A), can be taken as the fraction of the region's area covered by the shaded strips in the figure. From this, it follows that the ρ(A) and the k(A) are related by an average of the vehicle lengths. This average vehicle length is, by definition, the area of the shaded strips within A divided by the t(A); i.e., it is the ratio of the ρ(A) to the k(A),

$$average\ vehicle\ length = \frac{\rho(A)}{k(A)} = \frac{area\ of\ the\ shaded\ strips}{|A|}\cdot\frac{|A|}{t(A)}. \tag{2.14}$$

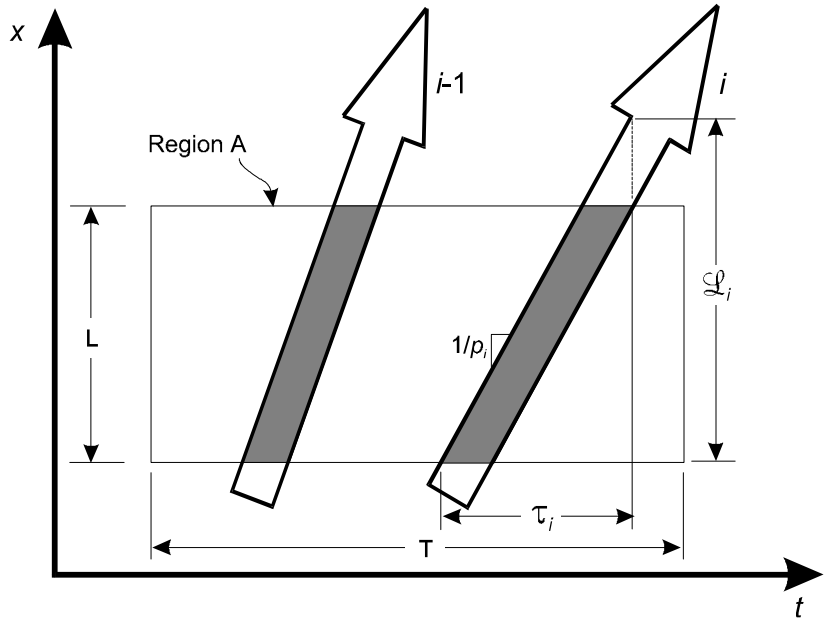


Figure 2-3. Trajectories of vehicle fronts and rears.

Notably, an average of the vehicle lengths also relates the k(A) to ρ, where the latter is the occupancy as conventionally defined (i.e., the percentage of time vehicles spend atop the detector). Toward illustrating this relation, the L in Figure 6-4 is assumed to be the length of road "visible" to the loop detector, the so-called detection zone. The T is some interval of time; e.g. the interval over which the detector collects measurements. The time each ith vehicle spends atop the detector is denoted as τ_i. Thus, if m vehicles pass the detector during time T, the

$$\rho = \frac{\sum_{i=1}^{m} \tau_i}{T}.$$

As shown in Figure 6-4, $\mathscr{L}_i$ is *the summed length of the detection zone and the length of vehicle i. Therefore,*

$$\sum_{i=1}^{m} \tau_i = \sum_{i=1}^{m} \mathscr{L}_i \cdot \frac{1}{v_i} = \sum_{i=1}^{m} \mathscr{L}_i p_i \tag{2.15}$$

if the front end of each i has a constant v_i over the distance $\mathscr{L}_i$. Since

$$\frac{\sum_{i=1}^{m} \tau_i}{T} = \frac{\frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i}{\frac{1}{m} \cdot T} = q(A) \cdot \frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i, \tag{2.16}$$

it follows that

$$\rho = q(A) \cdot \frac{1}{m} \sum_{i=1}^{m} \ell_i \cdot p_i ,$$

$$\rho = q(A) \cdot \frac{1}{v(A)} \left[v(A) \cdot \frac{1}{m} \sum_{i=1}^{m} \ell_i \cdot p_i \right],$$

$$\rho = k(A) \cdot \left[\frac{\sum_{i=1}^{m} \ell_i \cdot p_i}{\sum_{i=1}^{m} p_i} \right], \qquad (2.17)$$

where the term in brackets is the average vehicle length relating ρ to the k(A); it is the so-called average effective vehicle length weighted by the paces. If pace and vehicle length are uncorrelated, the term in brackets in (2.17) can be approximated by the unweighted average of the vehicle lengths in the interval T.

When measurements are taken by two closely spaced detectors, a so-called speed trap, the p_i are computed from each vehicle's arrival times at the two detectors. The ℓ_i are thus computed by assuming that the p_i are constant over the length of the speed trap. When only a single loop detector is available, vehicle velocities are often estimated by using an assumed average value of the (effective) vehicle lengths.

2.1.6 Three-Dimensional Representation of Vehicle Streams. *It is useful to display flows and densities using a three-dimensional representation described by Makagami et al. (1971). For this representation, an axis for the cumulative number of objects, N, is added to the t-x coordinate system so that the resulting surface N(t, x) is like a staircase with each trajectory being the edge of a step. As shown in Figure 2-4, curves of cumulative count versus time are obtained by taking cross-sections of this surface at some fixed locations and viewing the exposed regions in the t-N plane. Analogously, cross-sections at fixed times viewed in the N-x plane reveal curves of cumulative count versus space.*

Figure 2-4 shows cumulative curves at two locations and for two instants in time. The former display the trip times of objects and the time-varying accumulations between the two locations, as labeled on the figure. These cumulative curves can be transformed into a queueing diagram (as described in Chapter 5) by translating the curve at upstream x_1 forward by the free-flow (i.e., the undelayed) trip time from x_1 to x_2. Also displayed in Figure 2-4, the curves of cumulative count versus space show the number of objects crossing a fixed location during the interval $t_2 - t_1$ and the distances traveled by individual objects during this same interval.

If one is dealing with many objects so that measuring the exact integer numbers is not important, it is advantageous to construct the cumulative curves with piece-wise linear approximations; e.g. the curves may be smoothed using linear interpolations that pass through the crests of the steps. The time-dependent flows past some location are the slopes of the smoothed curve of t versus N

constructed at that location (Moskowitz, 1954; Edie and Foote, 1960; Newell, 1971, 1982). Analogously, the location-dependent densities at some instant are the negative slopes of a smoothed curve of N versus x; densities are the negative slopes because objects are numbered in the reverse direction to their motion. This three-dimensional model has been applied by Newell (1993) in work on kinematic waves in traffic (see Chapter 5). In addition, Part I of his paper contains some historical notes on the use of this approach for modelling.

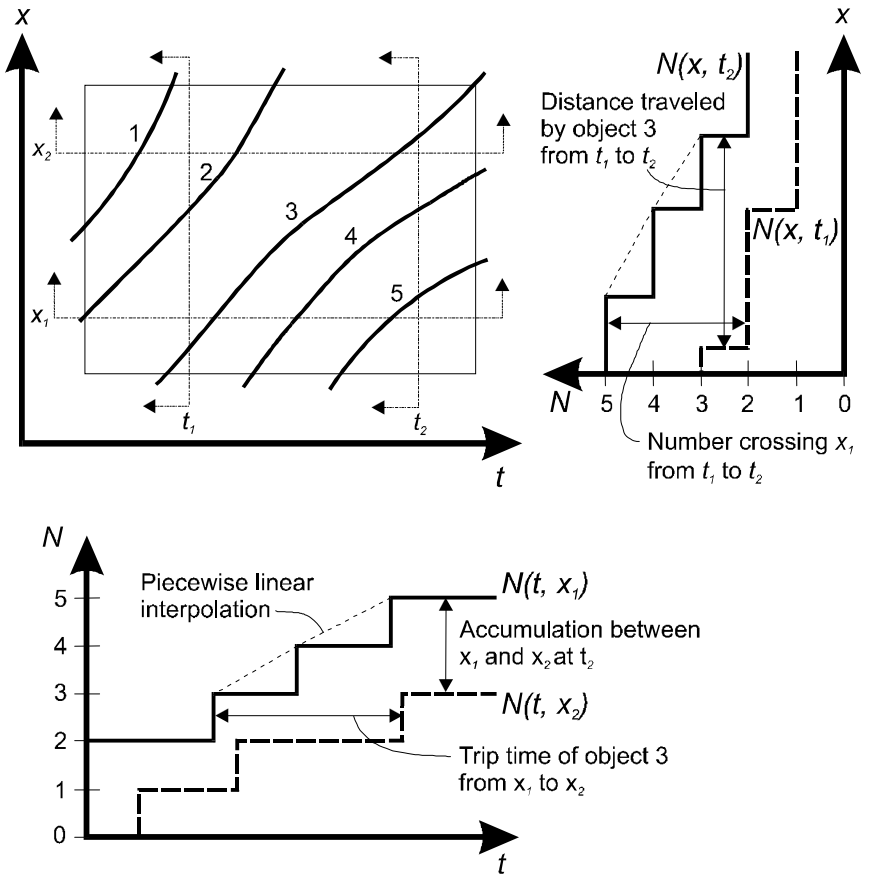


Figure 2-4. Three-dimensional representation.

2.2 Measurement issues

This section addresses issues related to the measurement of the key variables defined in the previous section. Four topics are covered. First, there is a discussion of measurement procedures that have traditionally been used. Second, potential measurement errors arising from the mismatch between the definitions and the measurement methods are discussed. Third, measurement difficulties that can potentially arise from the particular location used for collecting measurements are considered. The final topic is the nature of the time intervals from which to collect data, drawing on the cumulative curves just presented.

2.2.1 Measurement procedures Measurement technology for obtaining traffic data has

changed over the 60-year span of interest in traffic flow, but most of the basic procedures remain largely the same. Five measurement procedures are discussed in this section:

- measurement at a point;
- measurement over a short section (by which is meant less than about 10 meters (m);
- measurement over a length of road [usually at least 0.5 kilometers (km)];
- the use of an observer moving in the traffic stream; and
- wide-area samples obtained simultaneously from a number of vehicles, as part of Intelligent Transportation Systems (ITS).

The first two were discussed with reference to the time-space diagram in Fig 2.2, and the remaining three are also readily interpreted on that diagram. Details on each of these methods can be found in the ITE's Manual of Transportation Engineering Studies (Robertson, 1994). The wide-area samples from ITS are similar to having a number of moving observers at various points and times within the system. New developments such as this will undoubtedly change the way some traffic measurements are obtained in the future, but they have not been in operation long enough to have a major effect on the material to be covered in this chapter.

Measurement at a point by hand tallies or pneumatic tubes, was the first procedure used for traffic data collection. This method is easily capable of providing volume counts and therefore flow rates directly, and can provide headways if arrival times are recorded. The technology for making measurements at a point on freeways changed over 30 years ago from using pneumatic tubes placed across the roadway to using point detectors (May et al. 1963; Athol 1965). The most commonly used point detectors are based on inductive loop technology, but other methods in use include microwave, radar, photocells, ultrasonics, and closed circuit television cameras.

Except for the case of a stopped vehicle, speeds at a 'point' can be obtained only by radar or microwave detectors: dx/dt obviously requires some dx, however small. (Radar and microwave frequencies of operation mean that a vehicle needs to move only about one centimeter during the speed measurement.) In the absence of such instruments for a moving vehicle, a second observation location is necessary to obtain speeds, which moves the discussion to that of measurements over a short section.

Measurements over a short section Early studies used a second pneumatic tube, placed very close to the first, to obtain speeds. More recent systems have used paired presence detectors, such as inductive loops spaced perhaps five to six meters apart. With video camera technology, two detector 'lines' placed close together provide the same capability for measuring speeds. All of these presence detectors continue to provide direct measurement of volume and of time headways, as well as of speed when pairs of them are used.

Most point detectors currently used, such as inductive loops or microwave beams, take up space on the road, and are therefore a short section measurement. These detectors also measure occupancy, which was not available from earlier technology. This variable is available because the loop gives a continuous reading (at 50 or 60 Hz usually), which pneumatic tubes or manual

counts could not do. Because occupancy depends on the size of the detection zone of the instrument, the measured occupancy may differ from site to site for identical traffic, depending on the nature and construction of the detector.

Measurements along a length of road come either from aerial photography, or from cameras mounted on tall buildings or poles. On the basis of a single frame from such sources, only density can be measured. The single frame gives no sense of time, so neither volumes nor speed can be measured. Once several frames are available, as from a video-camera or from time-lapse photography over short time intervals, speeds and volumes can also be measured, as per the generalized definitions provided above.

Despite considerable improvements in technology, and the presence of closed circuit television on many freeways, there is very little use of measurements taken over a long section at the present time. The one advantage such measurements might provide would be to yield true journey times over a lengthy section of road, but that would require better computer vision algorithms (to match vehicles at both ends of the section) than are currently possible. There have been some efforts toward the objective of collecting journey time data on the basis of the details of the 'signature' of particular vehicles or platoons of vehicles across a series of loops over an extended distance (Kühne and Immes 1993), but few practical implementations as yet.

The moving observer method was used in some early studies, but is not used as the primary data collection technique now because of the prevalence of the other technologies. There are two approaches to the moving observer method. The first is a simple floating car procedure in which speeds and travel times are recorded as a function of time and location along the road. While the intention in this method is that the floating car behaves as an average vehicle within the traffic stream, the method cannot give precise average speed data. It is, however, effective for obtaining qualitative information about freeway operations without the need for elaborate equipment or procedures. One form of this approach uses a second person in the car to record speeds and travel times. A second form uses a modified recording speedometer of the type regularly used in long-distance trucks or buses. One drawback of this approach is that it means there are usually significantly fewer speed observations than volume observations. An example of this kind of problem appears in Morton and Jackson (1992).

The other approach was developed by Wardrop and Charlesworth (1954) for urban traffic measurements and is meant to obtain both speed and volume measurements simultaneously. Although the method is not practical for major urban freeways, it is included here because it may be of some value for rural expressway data collection, where there are no automatic systems. While it is not appropriate as the primary mode of data collection on a busy freeway, there are some useful points that come out of the literature that should be noted by those seeking to obtain average speeds through the moving car method.

The method developed by Wardrop and Charlesworth is based on a survey vehicle that travels in both directions on the road. The formulae allow one to estimate both speeds and flows for one direction of travel. The two formulae are

$$q = \frac{(x + y)}{(t_a + t_w)} \tag{2.18}$$

$$\bar{t} = t_w - \frac{y}{q} \tag{2.19}$$

where

q is the estimated flow on the road in the direction of interest,

x is the number of vehicles traveling in the direction of interest, which are met by the survey vehicle while traveling in the opposite direction,

y is the net number of vehicles that overtake the survey vehicle while traveling in the direction of interest (i.e. those passing minus those overtaken),

t_a is the travel time taken for the trip against the stream,

t_w is the travel time for the trip with the stream, and

$\bar{t}$ is the estimate of mean travel time in the direction of interest.

Wright (1973) revisited the theory behind this method. His paper also serves as a review of the papers dealing with the method in the two decades between the original work and his own. He finds that, in general, the method gives biased results, although the degree of bias is not significant in practice, and can be overcome. Wright's proposal is that the driver should fix the journey time in advance, and keep to it. Stops along the way would not matter, so long as the total time taken is as determined prior to travel. Wright's other point is that turning traffic (exiting or entering) can upset the calculations done using this method. This fact means that the route to be used for this method needs to avoid major exits or entrances. It should be noted also that a large number of observations are required for reliable estimation of speeds and flows; without that, the method has very limited precision.

2.2.2 Error caused by the mismatch between definitions and usual measurements

The overview in the previous section described methods that have historically been used to collect observations on the key traffic variables. As was mentioned within that section, the methods do not always accord with the definitions of these variables, as they were presented in Section 2.1. One of the most common methods for collecting these data currently is based on inductive loops embedded in the roadway. When speed data are also to be collected, a pair of closely spaced loops (often called a speed trap) is needed, a known distance apart. Equivalent systems are based on microwave beams that cover a part of a roadway surface. Cassidy and Coifman (1997) point out the criteria that need to be met for the data from these systems to meet the definitional requirements.

In practical terms, these criteria can be reduced to ensuring that any vehicle entering the speed trap also clears it within the time interval for which the data are obtained – or that the number of vehicles in the time interval is quite large. Neither of these criteria is met by the ordinary implementation of inductive loop speed trap detectors. Whether or not the last vehicle

entering a speed trap will clear it within the interval is a simple random variable. There is no guarantee that this criterion will regularly be met. If the number of vehicles were very large, this would not be an issue, but because of the need for timely information, many systems poll the loop detector controllers at least every minute, with some systems going to 30 second or even 20 second data collection. For intervals of that size, the volume counts on a single freeway lane will be at most 40 (or 25 or 20) vehicles, which is not large enough to overcome the error introduced by missing one of the vehicle's speeds. While the data from these detector systems are certainly good enough for operational decision-making, the data may give misleading results if used directly with these equations because of the mismatch between the collection and the definitions.

2.2.3 Importance of location to the nature of the data

Almost all of the bivariate models to be discussed represent efforts to explain the behaviour of traffic variables over the full range of operation. In turning the models from abstract representations into numerical models with specific parameter values, an important practical question arises: can one expect that the data collected will cover the full range that the model is intended to cover? If the answer is no, then the difficult question follows of how to do curve fitting (or parameter estimation) when there may be essential data missing.

This issue can be explained more easily with an example. At the risk of oversimplifying a relationship prior to a more detailed discussion of it, consider the simple representation of the speed-flow curve as shown in Figure 2.5, for three distinct sections of roadway. The underlying curve is assumed to be the same at all three locations. Between locations A and B, a major entrance ramp adds considerable traffic to the road. If location B reaches capacity due to this entrance ramp volume, there will be a queue of traffic on the mainstream at location A. These vehicles can be considered to be waiting their turn to be served by the bottleneck section immediately downstream of the entrance ramp. The data superimposed on graph A reflect the situation whereby traffic at A had not reached capacity before the added ramp flow caused the backup. There is a good range of uncongested data (on the top part of the curve), and congested data concentrated in one area of the lower part of the curve. The flows for that portion reflect the capacity flow at B less the entering ramp flows.

At location B, the full range of uncongested flows is observed, right out to capacity, but the location never becomes congested, in the sense of experiencing stop-and-go traffic. It does, however, experience congestion in the sense that speeds are below those observed in the absence of the upstream congestion. Drivers arrive at the front end of the queue moving very slowly, and accelerate away from that point, increasing speed as they move through the bottleneck section. This segment of the speed-flow curve has been referred to as queue discharge flow, QDF (Hall et al. 1992). The particular speed observed at B will depend on how far it is from the front end of the queue (Persaud and Hurdle 1988). Consequently, the only data that will be observed at B are on the top portion of the curve, and at some particular speed in the QDF segment.

If the exit ramp between B and C removes a significant portion of the traffic that was observed at B, flows at C will not reach the levels they did at B. If there is no downstream situation similar to that between A and B, then C will not experience congested operations, and

the data observable there will be as shown in Figure 2.5.

None of these locations taken alone can provide the data to identify the full speed-flow curve. Location C can help to identify the uncongested portion, but cannot deal with capacity, or with congestion. Location B can provide information on the uncongested portion and on capacity operation, but cannot contribute to the discussion of congested operations. Location A can provide some information on both uncongested and congested operations, but cannot tell anything about capacity operations. This would all seem obvious enough. A similar discussion appears in Drake et al (1967). It is also explained by May (1990). Other aspects of the effect of location on data patterns are discussed by Hsu and Banks (1993). Yet a number of important efforts to fit data to theory have ignored this key point (for example Ceder and May 1976; Easa and May 1980). They have recognized that location-A data are needed to fit the congested portion of the curve, but have not recognized that at such a location data are missing that are needed to identify capacity. Consequently, discussion of bivariate models will look at the nature of the data used in each study, and at where the data were collected (with respect to bottlenecks) in order to evaluate the theoretical ideas. It is possible that the apparent need for several different models, or for different parameters for the same model at different locations, or even for discontinuous models instead of continuous ones, arose because of the nature (location) of the data each was using.

2.2.4 Selecting intervals from which to extract data

In addition to the location for the data, there are also several issues relating to the time intervals for collecting data. The first is the issue of obtaining representative data. By examining trends on the cumulative curves, one can observe how flows and densities change with time and space. By selecting flows and densities as they appear on the cumulative curves, their values may be taken over intervals that exhibit near-constant slopes. In this way, the values assigned to these properties are not affected by some arbitrarily selected measurement interval(s). Choosing intervals arbitrarily is undesirable because data extracted over short measurement intervals are highly susceptible to the effects of statistical fluctuations while the use of longer intervals may average-out the features of interest. Further discussion and demonstration of this in the context of freeway traffic is offered in (Cassidy, 1998).

There is also an issue of how many observations are needed to obtain good estimates of key variables such as the capacity. *A bottleneck's capacity, q_{max}, is the maximum flow it can sustain for a very long time (in the absence of any influences from restrictions further downstream). It can be expressed mathematically as*

$$q_{\max} \equiv \lim_{T\to\infty}\left(\frac{N_{\max}}{T}\right)^{\wedge}, \qquad (2.20)$$

where N_{max} denotes that the vehicles counted during very long time T discharged through the bottleneck at a maximum rate. The engineer assigns a capacity to a bottleneck by obtaining a value for the estimator q_{max} (since one cannot actually observe a maximum flow for a time period approaching infinity). It is desirable that the expected value of this estimator equal the capacity, $E(q_{max})\tilde{}\ q_{max}$. For this reason, one would collect samples (i.e., counts) immediately downstream of an active bottleneck so as to measure vehicles discharging at a maximum rate. The amount q_{max} can

deviate from q_{max} is controlled by the sample size, N. A formula for determining N to estimate a bottleneck's capacity to a specified precision is derived below.

To begin this derivation, the estimator may be taken as

$$\hat{q}_{\max} = \sum_{m=1}^{M} n_m /(M \cdot \tau), \tag{2.21}$$

where n_m is the count collected in the mth interval and each of these M intervals has a duration of τ. If the $\{n_m\}$ can be taken as independent, identically distributed random variables (e.g. the counts were collected from consecutive intervals with τ sufficiently large), then the variance of q_{max} can be expressed as

$$variance(\hat{q}_{\max}) = \frac{1}{\tau^2}\left[\frac{variance(n)}{M}\right] = \frac{1}{T}\left[\frac{variance(n)}{\tau}\right] \tag{2.22}$$

since q_{max} is a linear function of the independent n_m and the (finite) observation period T is the denominator in (2.21).

The bracketed term variance(n) / τ is a constant. Thus, by multiplying the top and bottom of this quotient by E(n), the expected value of the counts, and by noting that E(n) / τ = q_{max}, one obtains

$$\text{variance}(\hat{q}_{\max}) = \frac{\gamma}{T}, \tag{2.23}$$

where γ is the index of dispersion; i.e., the ratio variance(n)/E(n).

The variance(q_{max}) is the square of the standard error. Thus, by isolating T in (2.23) and then multiplying both sides of the resulting expression by q_{max}, one arrives at

$$q_{\max} \cdot T = \frac{\gamma}{\varepsilon^2}, \tag{2.24}$$

where q_{max} $T \tilde{}$ N, the number of observations (i.e., the count) needed to estimate capacity to a specified percent error ε. Note, for example, that ε = 0.05 to obtain an estimate within 5 percent of q_{max}. The value of γ may be estimated by collecting a presample and, notably, N increases rapidly as ε diminishes.

The expression $N = \gamma/\varepsilon^2$ may be used to determine an adequate sample size when vehicles, or any objects, discharging through an active bottleneck exhibit a nearly stationary flow; i.e., when the cumulative count curve exhibits a nearly constant slope. If necessary, the N samples may be obtained by concatenating observations from multiple days. Naturally, one would take samples during time periods thought to be representative of the conditions of interest. For example, one should probably not use vehicle counts taken in inclement weather to estimate the capacity for fair weather conditions.

2.3 Bivariate models

This section provides an overview of work to establish relationships among pairs of the variables described in the opening section. Some of these efforts begin with mathematical models; others are primarily empirical, with little or no attempt to generalize. Both are important components for understanding the relationships. For reasons discussed above, two aspects of these efforts are emphasized here: the measurement methods used to obtain the data; and the location at which the measurements were obtained.

2.3.1 Speed-flow models

The speed-flow relationship is the bivariate relationship on which there has been the greatest amount of work recently, so is the first one discussed. This section starts from current understanding, which provides some useful insights for interpreting earlier work, and then moves to a chronological review of some of the major models.

The bulk of then recent empirical work on the relationship between speed and flow (as well as the other relationships) was summarized in a paper by Hall, Hurdle, and Banks (1992). In it, they proposed the model for traffic flow shown in Figure 2.6. This figure is the basis for the background speed-flow curve in Figure 2.5.

It is useful to summarize some of the antecedents of the understanding depicted in Figure 2.6. The initial impetus came from a paper by Persaud and Hurdle (1988), in which they demonstrated (Figure 2.7) that the vertical line for queue discharge flow in Figure 2.6 was a reasonable result of measurements taken at various distances downstream from a queue. This study was an outgrowth of an earlier one by Hurdle and Datta (1983) in which they raised a number of questions about the shape of the speed-flow curve near capacity.

Additional empirical work dealing with the speed-flow relationship was conducted by Banks (1989, 1990), Hall and Hall (1990), Chin and May (1991), Wemple, Morris and May (1991), Agyemang-Duah and Hall (1991) and Ringert and Urbanik (1993). All of these studies supported the idea that speeds remain nearly constant even at quite high flow rates. Another of the important issues they dealt with is one that had been around for over thirty years (Wattleworth 1963): is there a reduction in flow rates within the bottleneck at the time that the queue forms upstream? Figure 2.6 shows such a drop on the basis of two studies. Banks (1991a, 1991b) reports roughly a 3% drop from pre-queue flows, on the basis of nine days of data at one site in California. Agyemang-Duah and Hall (1991) found about a 5% decrease, on the basis of 52 days of data at one site in Ontario. This decrease in flow is often not observable, however, as in many locations high flow rates do not last long enough prior to the onset of congestion to yield the stable flow values that would show the drop.

Two empirical studies in Germany support the upper part of the curve in Figure 2.6 quite well. Heidemann and Hotop (1990) found a piecewise-linear 'polygon' for the upper part of the curve (Figure 2.8). Unfortunately, they did not have data beyond 1700 veh/hr/lane, so could not address what happens at capacity. Stappert and Theis (1990) conducted a major empirical study of speed-flow relationships on various kinds of roads. However, they were interested only in estimating parameters for a specific functional form,

$$V = (A - e^{BQ})\, e^{-c} - K\, e^{dQ} \qquad (2.25)$$

where V is speed, Q is traffic volume, c and d are constant "Krummungs factors" taking values between 0.2 and 0.003, and A, B, and K are parameters of the model. This function tended to

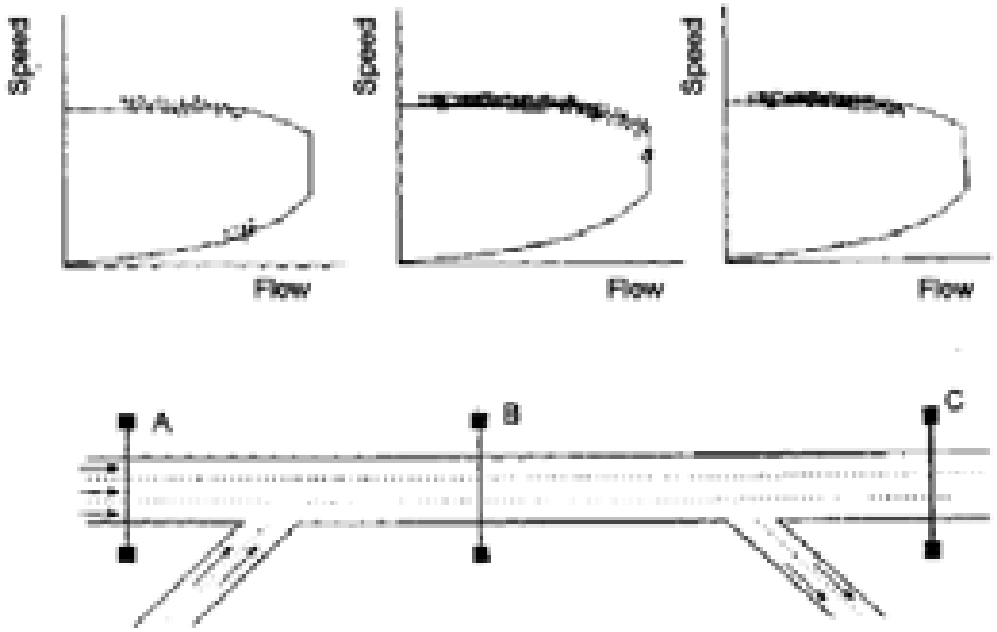


Figure 2. 5
Effect of Measurement Location on Nature of Data
(Similar to figures in May 1990 and Hall et al. 1992).

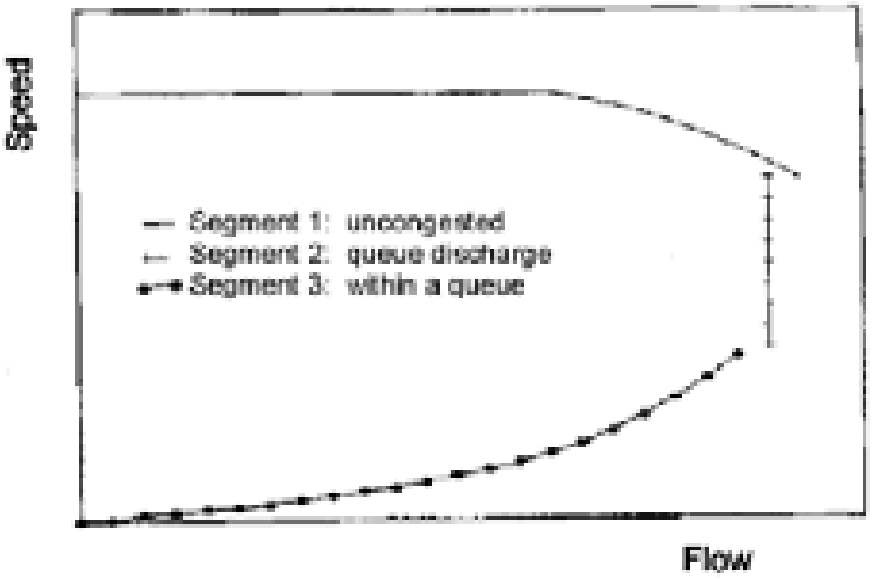


Figure 2. 6
Generalized Shape of Speed-Flow Curve
Proposed by Hall, Hurdle, & Banks
(Hall et al. 1992).

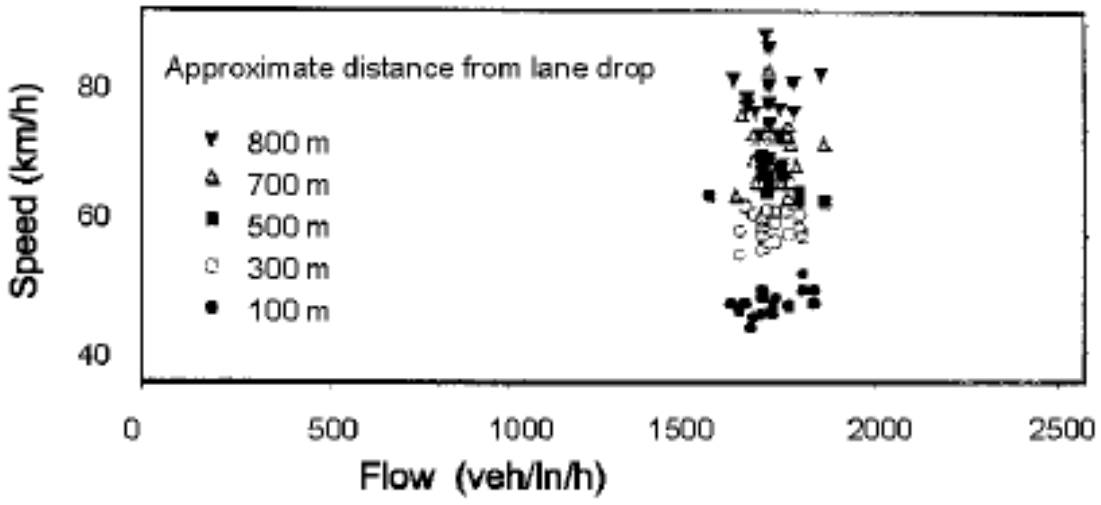


Figure 2. 7
Speed-Flow Data for Queue Discharge Flow at Varied Distances Downstream from the Head of the Queue (Modified from Persaud and Hurdle 1988).

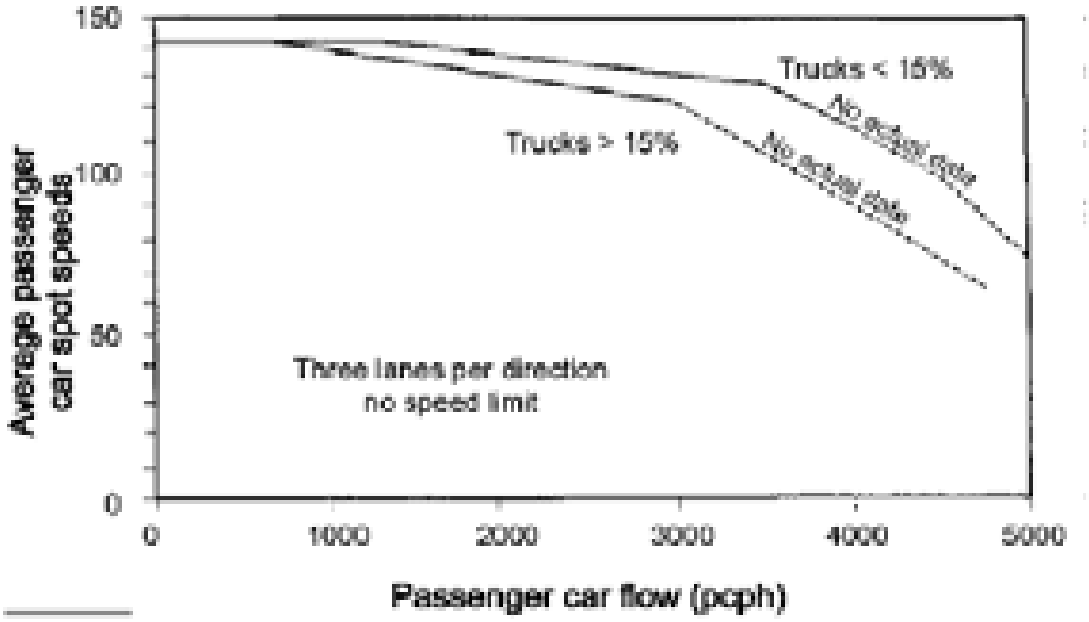


Figure 2. 8
Results from Fitting Polygon Speed-Flow Curves to German Data (Modified and translated from Heidemann and Hotop 1990).

give the kind of result shown in Figure 2.9, despite the fact that the curve does not accord well with the data near capacity. In Figure 2.9, each point represents a full hour of data, and the graph represents five months of hourly data. Note that flows in excess of 2200 veh/h/lane were sustained on several occasions, *over the full hour*.

The problem for traffic flow theory is that these curves are empirically derived. There is not really any theory that would explain these particular shapes, except perhaps for Edie et al.(1980), who propose qualitative flow regimes that relate well to these curves. The task that lies ahead for traffic flow theorists is to develop a consistent set of equations that can replicate this reality. The models that have been proposed to date, and will be discussed in subsequent sections, do not necessarily lead to the kinds of speed-flow curves that data suggest are needed.

It is instructive to review the history of depictions of speed-flow curves in light of this current understanding. Probably the seminal work on this topic was the paper by Greenshields in 1935, in which he derived the following parabolic speed-flow curve on the basis of a linear speed-density relationship together with the equation, flow = speed * density:

$$q = k_j (u - \frac{u^2}{u_f}) \tag{2.26}$$

Figure 2.10 contains that curve, and the data it is based on, redrawn. The numbers against the data points represent the "number of 100-vehicle groups observed", on Labor Day 1934, in one direction on a two-lane two-way road (p. 464). In counting the vehicles on the road, every 10th vehicle started a new group (of 100), so there is 90% overlap between two adjacent groups (p. 451): the groups are not independent observations. Equally important, the data have been grouped in flow ranges of 200 veh/h and the averages of these groups taken prior to plotting. The one congested point, representing 51 (overlapping) groups of 100 observations, came from a different roadway entirely, with different cross-section and pavement, which were collected on a different day.

These details are mentioned here because of the importance to traffic flow theory of Greenshields' work. The parabolic shape he derived was accepted as the proper shape of the curve for decades. In the 1965 Highway Capacity Manual, for example, the shape shown in Figure 2.10 appears exactly, despite the fact that data displayed elsewhere in the 1965 HCM showed that contemporary empirical results did not match the figure. In the 1985 HCM, the same parabolic shape was retained, although broadened considerably. In short, Greenshields' model dominated the field for over 50 years, despite the fact that by current standards of research the method of analysis of the data, with overlapping groups and averaging prior to curve-fitting, would not be acceptable.

There is another criticism that can be addressed to Greenshields' work as well, although it is one of which a number of current researchers seem unaware. Duncan (1976; 1979) has shown that calculating density from speed and flow, fitting a line to the speed-density data, and then converting that line into a speed-flow function, gives a biased result relative to direct estimation

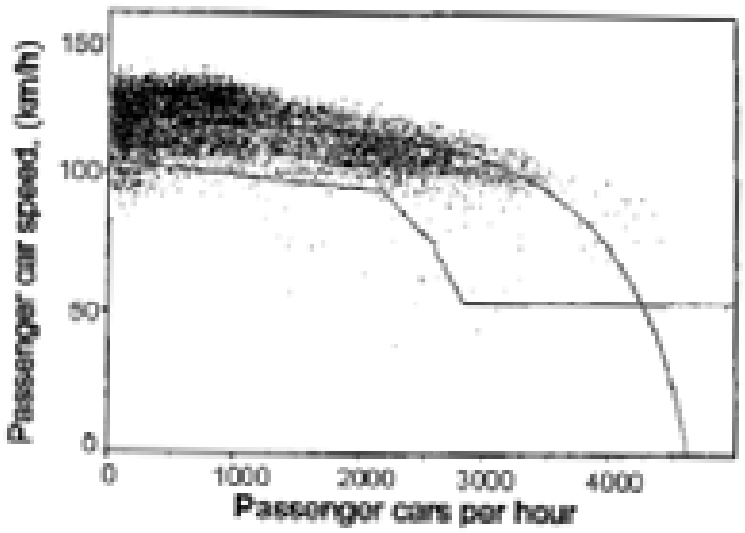


Figure 2. 9
Data for Four-Lane German Autobahns (Two-Lanes per Direction), as reported by Stappert and Theis (1980).

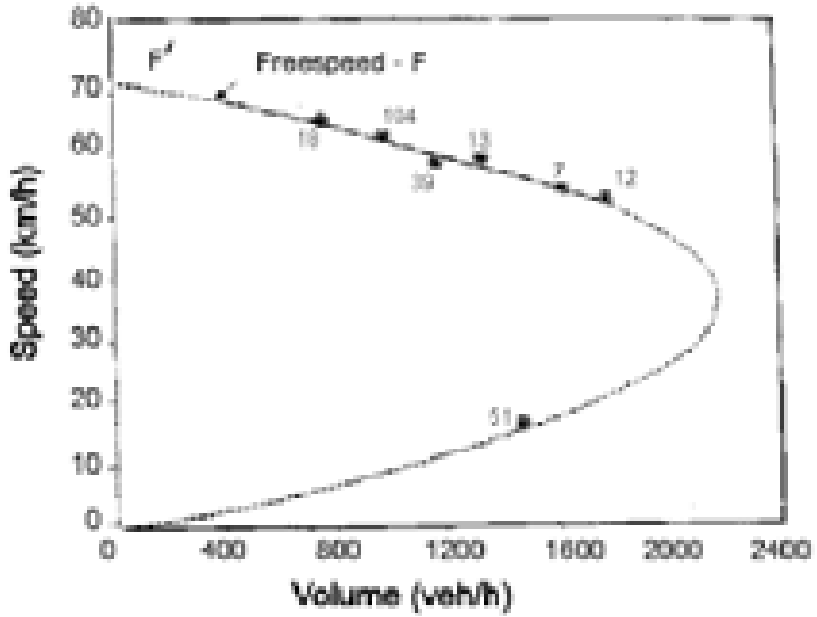


Figure 2.10
Greenshields' Speed-Flow Curve and Data (Greenshields 1935).

of the speed-flow function. This is a consequence of three things discussed earlier: the non-linear transformations involved in both directions, the stochastic nature of the observations, and the inability to match the time and space measurement frames exactly.

It is interesting to contrast the emphasis on speed-flow models in recent years with that 20 years ago, as represented in TRB SR 165 (Gerlough and Huber 1975). In that volume, the discussion of speed-flow models took up less than a page of text, and none of the five accompanying diagrams dealt with freeways. (Three dealt with the artificial situation of a test track.) In contrast, five pages and eleven figures were devoted to the speed-density relationship. Speed-flow models are important for freeway management strategies, and will be of fundamental importance for intelligent vehicle/highway systems (IVHS) implementation of alternate routing; hence there is considerably more work on this topic than on the remaining two bivariate topics. Twenty and more years ago, the other topics were of more interest. As Gerlough and Huber stated the matter (p. 61), "once a speed-concentration model has been determined, a speed-flow model can be determined from it." That is in fact the way most earlier speed-flow work was treated (including that of Greenshields). Hence it is sensible to turn to discussion of speed-concentration models, and to deal with any other speed-flow models as a consequence of speed-concentration work, which is the way they were developed.

2.3.2 Speed-concentration models

Greenshields' (1935) linear model of speed and density was mentioned in the previous section. It is:

$$u = u_f(1 - k/k_j) \tag{2.27}$$

where u_f is the free-flow speed, and k_j is the jam density. The measured data were speeds and flows; density was calculated using equation 2.27. The most interesting aspect of this particular model is that its empirical basis consisted of half a dozen points in one cluster near free-flow speed, and a single observation under congested conditions (Figure 2.11). The linear relationship comes from connecting the cluster with the single point: "since the curve is a straight line it is only necessary to determine accurately two points to fix its direction" (p. 468). What is surprising is not that such simple analytical methods were used in 1935, but that their results (the linear speed-density model) have continued to be so widely accepted for so long. While there have been studies that claimed to have confirmed this model, such as that in Figure 2.12a (Huber 1957), they tended to have similarly sparse portions of the full range of data, usually omitting both the lowest flows, and flow in the range near capacity. There have also been a number of studies that found contradictory evidence, most importantly that by Drake, et al. (1965), which will be discussed in more detail subsequently.

A second early model was that put forward by Greenberg (1959), showing a logarithmic relationship:

$$u = u_m \ln(k/kj) \tag{2.28}$$

The paper showed the fit of the model to two data sets, both of which visually looked very reasonable. However, the first data set was derived from speed and headway data on individual vehicles, which "was then separated into speed classes and the average headway was calculated for each speed class" (p. 83). In other words, the vehicles that appear in one data point (speed class) may not even have been travelling together! While a density can always be calculated as the reciprocal of average headway, when that average is taken over vehicles that may well not have been travelling together, it is not clear what that density is meant to represent. The second data set used by Greenberg was Huber's. This is the same data that appears in Figure 2.12a; Greenberg's graph is shown in Figure 2.12b. Visually, the fit is quite good, but Huber reported an R of 0.97, which does not leave much room for improvement.

A third model from the same period is that by Edie (1961). His model was an attempt to deal with a discontinuity that had consistently been found in data near the critical density, which "suggested the existence of some kind of change of state" (p. 72). He proposed two linear models for the two states of flow. The first related density to the logarithm of velocity above the "optimum velocity", i.e. "non-congested flow". The second related velocity to the logarithm of spacing (i.e. the inverse of density) for congested flow. The model was fitted to the same Lincoln Tunnel data as used by Greenberg.

These three forms of the speed-density curve, plus four others, were investigated in an empirical test by Drake et al. in 1967. The test used data from the middle lane of the Eisenhower Expressway in Chicago, one-half mile (800 m) upstream from a bottleneck whose capacity was only slightly less than the capacity of the study site. This location was chosen specifically in order to obtain data over as much of the range of operations as possible. A series of 1224 1-minute observations were initially collected. The measured data consisted of volume, time-mean speed, and occupancy. Density was calculated from volume and time-mean speed. A sample was then taken from among the 1224 data points in order to create a data set that was uniformly distributed along the density axis, as is assumed by regression analysis of speed on density.

The intention in conducting the study was to compare the seven speed-density hypotheses statistically, and thereby to select the best one. In addition to Greenshields' linear form, Greenberg's exponential curve, and Edie's two-part logarithmic model, the other four investigated were a two-part and three-part piecewise linear models, Underwood's (1961) transposed exponential curve, and a bell-shaped curve. Despite the intention to use "a rigorous structure of falsifiable tests" (p. 75) in this comparison, and the careful way the work was done, the statistical analyses proved inconclusive: "almost all conclusions were based on intuition alone since the statistical tests provided little decision power after all" (p. 76). To assert that intuition alone was the basis is no doubt a bit of an exaggeration. Twenty-one graphs help considerably in differentiating among the seven hypotheses and their consequences for both speed-volume and volume-density graphs.

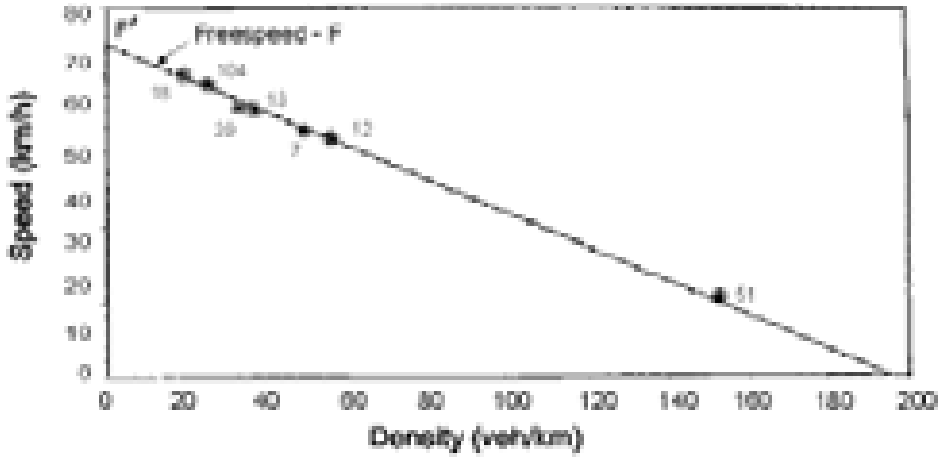


Figure 2.11
Greenshields' Speed-Density Graph and Data (Greenshields 1935).

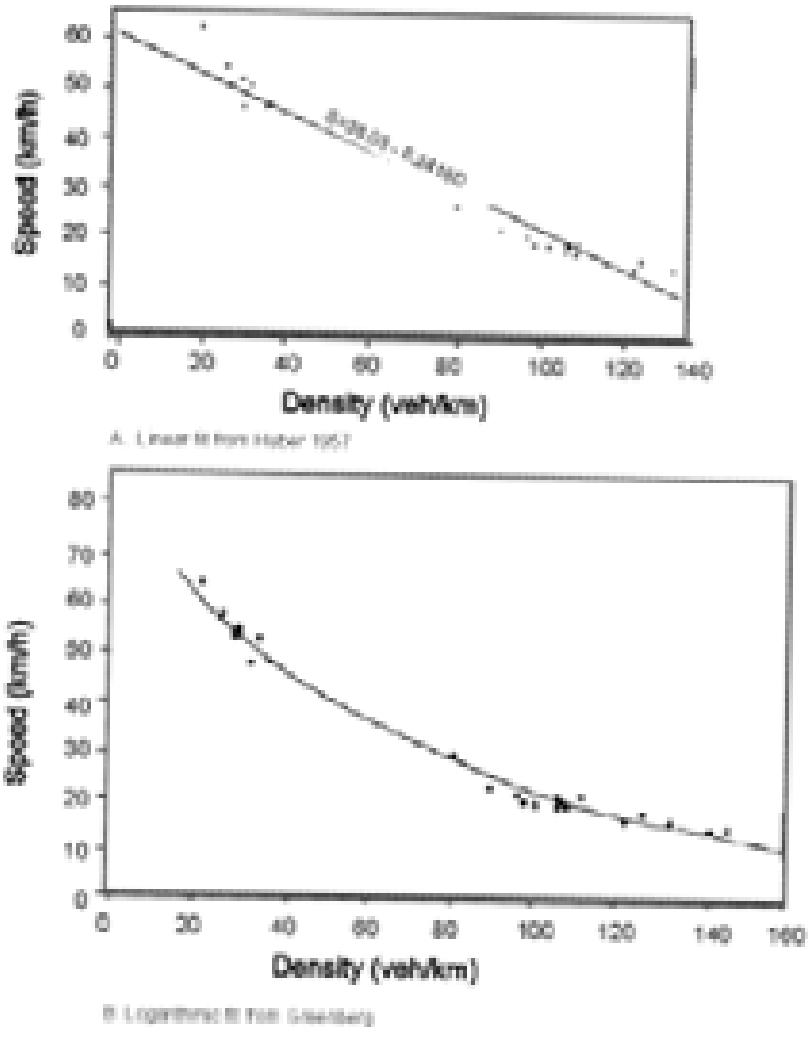


Figure 2.12
Speed-Density Data from Merritt Parkway and Fitted Curves.

Figure 2.13 provides an example of the three types of graphs used, in this case the ones based on the Edie model. Their comments about this model (p. 75) were: "The Edie formulation gave the best estimates of the fundamental parameters. While its R^2 was the second lowest, its standard error was the lowest of all hypotheses." One interesting point with respect to Figure 2.13 is that the Edie model was the only one of the seven to replicate capacity operations closely on the volume-density and speed-volume plots. The other models tended to underestimate the maximum flows, often by a considerable margin, as is illustrated in Figure 2.14, which shows the speed-volume curve resulting from Greenshields' hypothesis of a linear speed-density relationship. (It is interesting to note that the data in these two figures are quite consistent with the currently accepted speed-flow shape identified earlier in Figures 2.5 and 2.6.) The overall conclusion one might draw from the Drake et al. study is that none of the seven models they tested provide a particularly good fit to or explanation of the data, although it should be noted that they did not state their conclusion this way, but rather dealt with each model separately.

There are four additional issues that arise from the Drake et al. study that are worth noting here. The first is the methodological one identified by Duncan (1976; 1979), and discussed earlier with regard to Greenshields' work. Duncan showed that calculating density from speed and flow data, fitting a speed-density function to that data, and then transforming the speed-density function into a speed-flow function results in a curve that does not fit the original speed-flow data particularly well. This is the method used by Drake et al, and certainly most of their resulting speed-flow functions did not fit the original speed-flow data very well. Duncan's 1979 paper expanded on the difficulties to show that minor changes in the speed-density function led to major changes in the speed-flow function, suggesting the need for further caution in using this method of double transformations to get a speed-flow curve.

The second is that of the data collection location, as discussed above. The data were collected upstream of a bottleneck, which produced the kind of discontinuity that Edie had earlier identified. The Drake et al. approach was to try to fit the data as they had been obtained, without considering whether there was a portion of the data that was missing. They had intentionally tried to obtain data from as wide a range as possible, but as discussed above it may not be possible to get data from all three parts of the curve at one location.

The third issue is that identified by Cassidy, relating to the time period for collection of the data. The Drake et al. data came from standard loop detectors, working on fixed time intervals. As a consequence, there will be some measurement error, which may well affect the estimation of the bivariate relationships.

The fourth issue is the relationship between car-following models (see Chapter 5) and the models tested by Drake et al. They noted that four of the models they tested "have been shown to be directly related to specific car-following rules", and cited articles by Gazis and co-authors (1959; 1961). The interesting question to raise in the context of the overall appraisal of the Drake et al. results is whether those results raise doubts about the validity of the car-following models for freeways. The car-following models gave rise to four of the speed-density models tested by Drake et al. The results of the testing suggest that the speed-density models are not

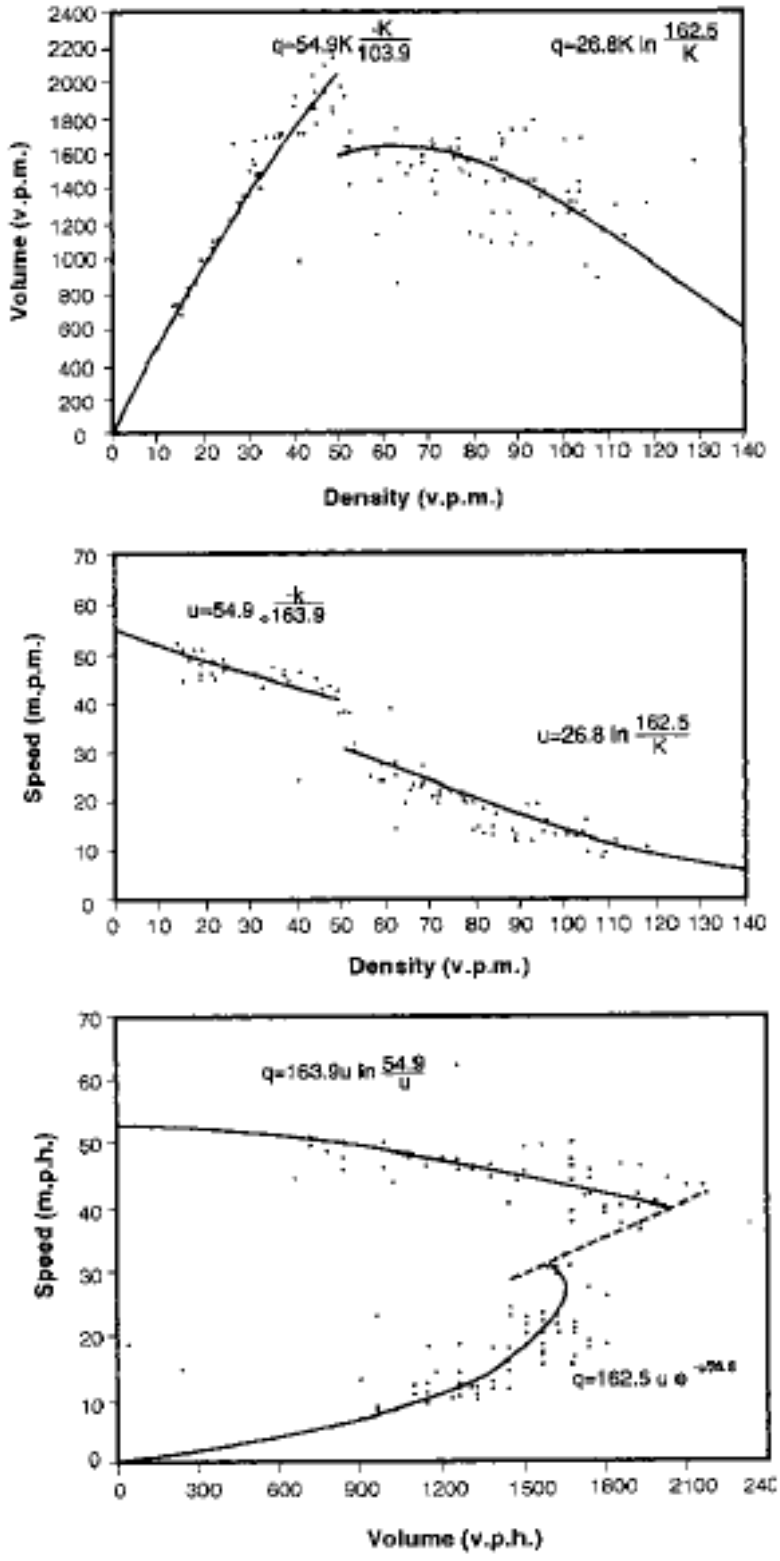


Figure 2.13
Three Parts of Edie's Hypothesis for the Speed-Density Function, Fitted to Chicago Data (Drake et al. 1967).

particularly good. *Modus tollens* in logic says that if the consequences of a set of premises are shown to be false, then one (at least) of the premises is not valid. It is possible, then, that the car-following models are not valid for freeways. (This is not surprising, as they were not developed for this context. Nor, it seems, are they used as the basis for contemporary microscopic freeway simulation models.) On the other hand, any of the three issues just identified may be the source of the failure of the models, rather than their development from car-following models.

2.3.3 Flow-concentration models

Although Gerlough and Huber did not give the topic of flow-concentration models such extensive treatment as they gave the speed-concentration models, they nonetheless thought this topic to be very important, as evidenced by their introductory paragraph for the section dealing with these models (p. 55):

> Early studies of highway capacity followed two principal approaches. Some investigators examined speed-flow relationships at low concentrations; others discussed headway phenomena at high concentrations. Lighthill and Whitham [1955] have proposed use of the flow-concentration curve as a means of unifying these two approaches. Because of this unifying feature, and because of the great usefulness of the flow-concentration curve in traffic control situations (such as metering a freeway), Haight [1960; 1963] has termed the flow-concentration curve "the basic diagram of traffic".

Nevertheless, most flow-concentration models have been derived from assumptions about the shape of the speed-concentration curve. This section deals primarily with work that has focused on the flow-concentration relationship directly. Under that heading is included work that uses either density or occupancy as the measure of concentration.

Edie was perhaps the first to point out that empirical flow-concentration data frequently have discontinuities in the vicinity of what would be maximum flow, and to suggest that therefore discontinuous curves might be needed for this relationship. (An example of his type of curve appears in Figure 2.13 above.) This suggestion led to a series of investigations by May and his students (Ceder 1975; 1976; Ceder and May 1976; Easa and May 1980) to specify more tightly the nature and parameters of these "two-regime" models (and to link those parameters to the parameters of car-following models). The difficulty with their resulting models is that the models often do not fit the data well at capacity (with results similar to those shown in Figure 2.14 for Greenshields' single-regime model). In addition, there seems little consistency in parameters from one location to another. Even more troubling, when multiple days from the same site were calibrated, the different days required quite different parameters.

Koshi, Iwasaki and Ohkura (1983) gave an empirically-based discussion of the flow-density relationship, in which they suggested that a reverse lambda shape was the best description of the data (p. 406): "the two regions of flow form not a single downward concave

curve... but a shape like a mirror image of the Greek letter lamda [sic] (λ)". These authors also investigated the implications of this phenomenon for car-following models, as well as for wave propagation. Data with a similar shape to theirs appears in Figure 2.13; Edie's equations fit those data with a shape similar to the lambda shape Koshi et al. suggested.

Although most of the flow-concentration work that relies on occupancy rather than density dates from the past decade, Athol suggested its use nearly 30 years earlier (in 1965). His work presages a number of the points that have come out subsequently and are discussed in more detail below: the use of volume and occupancy together to identify the onset of congestion; the transitions between uncongested and congested operations at volumes lower than capacity; and the use of time-traced plots (i.e. those in which lines connected the data points that occurred consecutively over time) to better understand the operations.

After Athol's early efforts, there seems to have been a dearth of efforts to utilize the occupancy data that was available, until the mid-1980s. One paper from that time (Hall et al. 1986) that utilized occupancy drew on the same approach Athol had used, namely the presentation of time-traced plots. Figure 2.15 shows results for four different days from the same location, 4 km upstream of a primary bottleneck. The data are for the left-most lane only (the high-speed, or passing lane), and are for 5-minute intervals. The first point in the time-connected traces is the one that occurred in the 5-minute period after the data-recording system was turned on in the morning. In part D of the Figure, it is clear that operations had already broken down prior to data being recorded. Part C is perhaps the most intriguing: operations move into higher occupancies (congestion) at flows clearly below maximum flows. Although Parts A and B may be taken to confirm the implicit assumption many traffic engineers have that operations pass through capacity prior to breakdown, Part C gives a clear indication that this does not always happen. Even more important, all four parts of Figure 2.15 show that operations do not go through capacity in returning from congested to uncongested conditions. Operations can 'jump' from one branch of the curve to the other, without staying on the curve. (This same result, not surprisingly, occurs for speed-flow data.)

Each of the four parts of Figure 2.15 show at least one data point between the two 'branches' of the usual curve during the return to uncongested conditions. Because these were 5-minute data, the authors recognized that these points might be the result of averaging of data from the two separate branches. Subsequently, however, additional work utilizing 30-second intervals confirmed the presence of these same types of data (Persaud and Hall 1989). Hence there appears to be strong evidence that traffic operations on a freeway can move from one branch of the curve to the other without going all the way around the capacity point. This is an aspect of traffic behaviour that none of the mathematical models discussed above either explain or lead one to expect. Nonetheless, the phenomenon has been at least implicitly recognized since Lighthill and Whitham's (1955) discussion of shock waves in traffic, which assumes instantaneous jumps from one branch to the other on a speed-flow or flow-occupancy curve. As well, queueing models (e.g. Newell 1982) imply that immediately upsteam from the back end of a queue there must be points where the speed is changing rapidly from the uncongested branch of the speed-flow curve to that of the congested branch. It would be beneficial if flow-

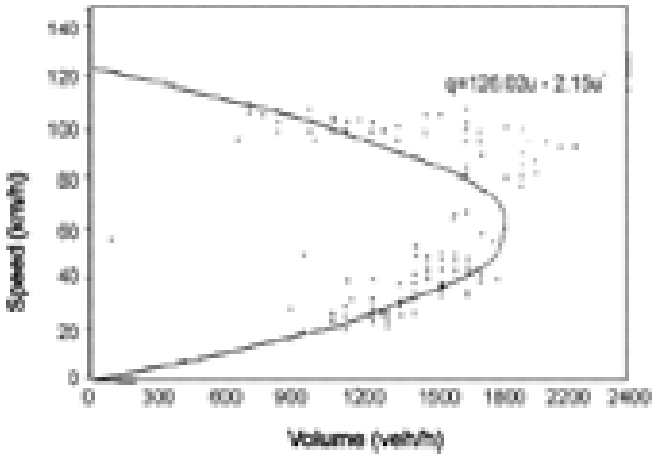


Figure 2.14
Greenshields' Speed-Flow Function Fitted to Chicago Data (Drake et al. 1967).

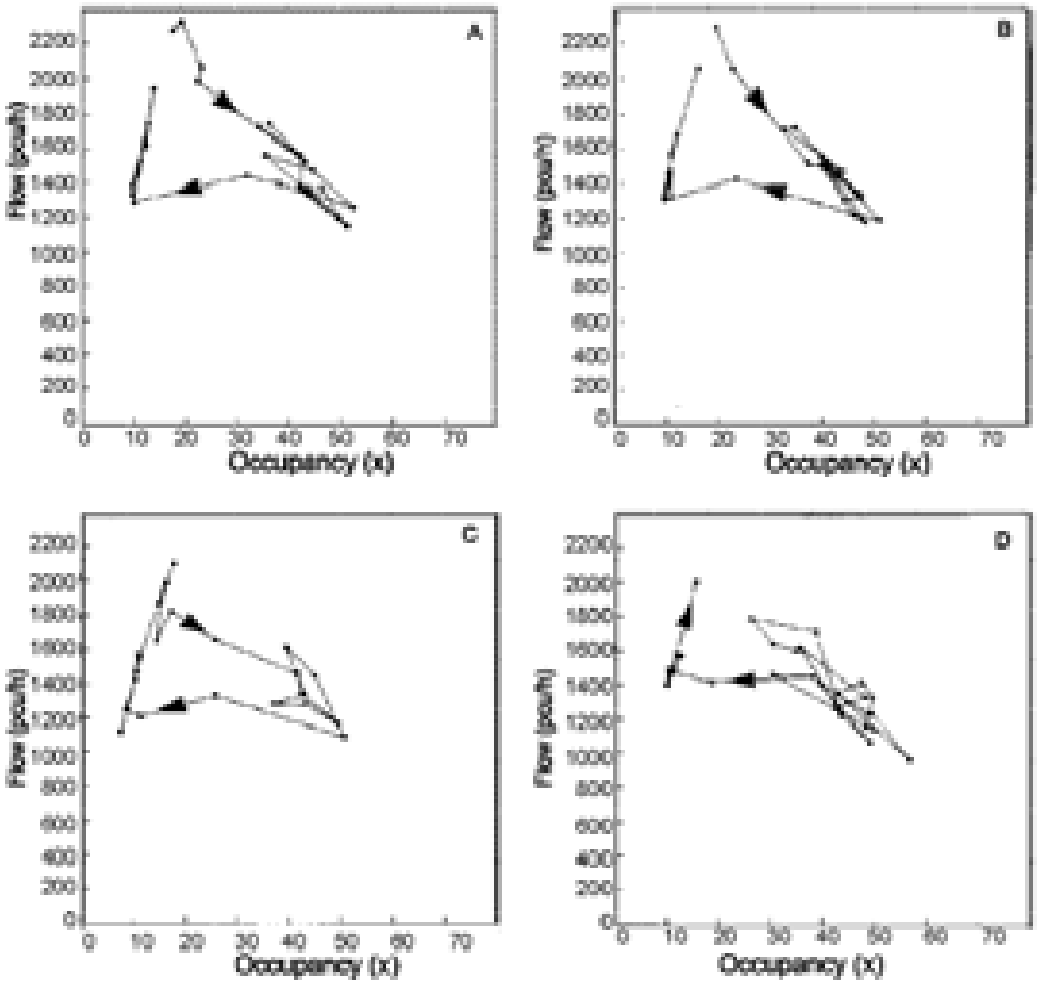


Figure 2.15
Four Days of Flow-Occupancy Data from Near Toronto (Hall et al. 1986).

concentration (and speed-flow) models explicitly took this possibility into account.

One of the conclusions of the paper by Hall et al. (1986) from which Figure 2.15 is drawn is that an inverted 'V' shape is a plausible representation of the flow-occupancy relationship. Although that conclusion was based on limited data from near Toronto, Hall and Gunter (1986) supported it with data from a larger number of stations. Banks (1989) tested their proposition using data from the San Diego area, and confirmed the suggestion of the inverted 'V'. He also offered a mathematical statement of this proposition and a behavioural interpretation of it (p. 58):

> The inverted-V model implies that drivers maintain a roughly constant average time gap between their front bumper and the back bumper of the vehicle in front of them, provided their speed is less than some critical value. Once their speed reaches this critical value (which is as fast as they want to go), they cease to be sensitive to vehicle spacing....

2.4 Three-dimensional models

There has not been a lot of work that attempts to treat all three traffic flow variables simultaneously. Gerlough and Huber presented one figure (reproduced as Figure 2.16) that represented all three variables, but said little about this, other than (1) "The model must be on the three-dimensional surface u = q/k," and (2) "It is usually more convenient to show the model of [Figure 2.16] as one or more of the three separate relationships in two dimensions..." (p. 49). As was noted earlier, however, empirical observations rarely accord exactly with the relationship $q=u\,k$, especially when the observations are taken during congested conditions. Hence focusing on the two-dimensional relationships will not often provide even implicitly a valid three-dimensional relationship.

In this context, a paper by Gilchrist and Hall (1989) is interesting because it presents three-dimensional representations of empirical observations, without attempting to fit them to a theoretical representations. The study is limited, in that data from only one location, upstream of a bottleneck, was presented. To enable better visualization of the data, a time-connected trace was used. The projections onto the standard two-dimensional surfaces of the data look much as one might expect. The surprises came in looking at oblique views of the three-dimensional representation, as in Figures 2.17 and 2.18. From one perspective (Figure 2.17), the traditional sideways U-shape that we have been led to expect is quite apparent, and projections of that are easily visualized onto, for example, the speed-flow surface (the face labeled with a 3, on the left side of the ‘box’). From a different perspective (Figure 2.18) that shape is hardly apparent at all, although indications of an inverted 'V' can be seen, which would project onto a flow-occupancy plot on face 1 of the box.. Black and white were alternated for five speed ranges. In both figures, the dark lines at the left represent the data with speeds above 80 km/hr. The light lines closest to the left cover the range 71 to 80 km/hr; the next, very small, area of dark lines contains the range 61 to 70 km/hr; the remaining light lines represent the range 51 to 60 km/hr; and the dark lines to the right of the diagram represent speeds below 50 km/hr.

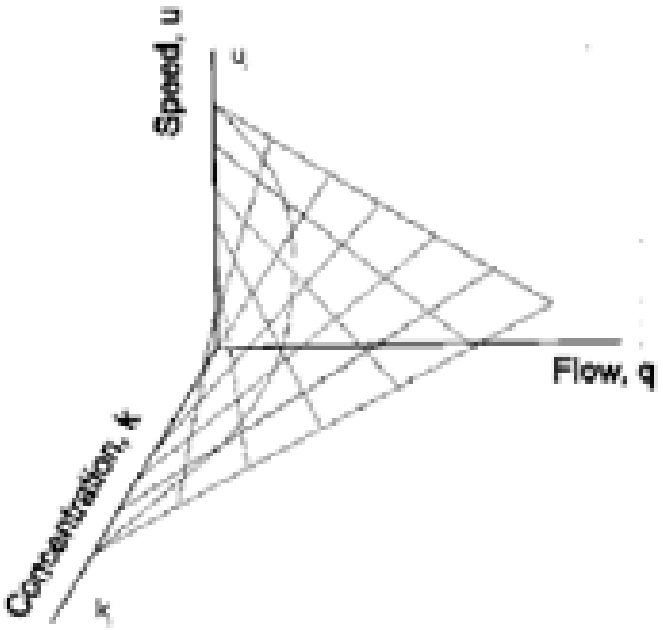


Figure 2.16
The Three-Dimensional Surface for Traffic Operations, as in Transportation Research Board Special Report 165 (Gerlough and Huber 1975).

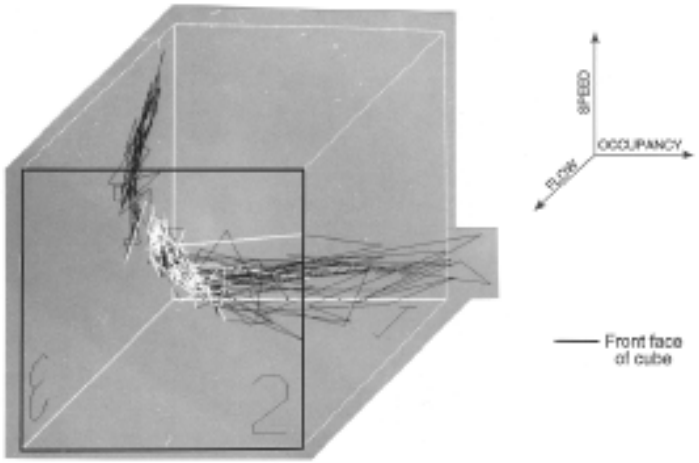


Figure 2.1 7
One Perspective on the Three-Dimensional Speed-Flow-Concentration Relationship (Gilchrist and Hall 1989).

One recent approach to modelling the three traffic operations variables directly has been based on the mathematics of catastrophe theory. (The name comes from the fact that while most of the variables being modelled change in a continuous fashion, at least one of the variables can make sudden discontinuous changes, referred to as catastrophes by Thom (1975), who originally developed the mathematics for seven such models, ranging from two dimensions to eight.) The first effort to apply these models to traffic data was that by Dendrinos (1978), in which he suggested that the two-dimensional catastrophe model could represent the speed-flow curve. A more fruitful model was proposed by Navin (1986), who suggested that the three-dimensional 'cusp' catastrophe model was appropriate for the three traffic variables. The feature of the cusp catastrophe surface that makes it of interest in the traffic flow context is that while two of the variables (the control variables) exhibit smooth continuous change, the third one (the state variable) can undergo a sudden 'catastrophic' jump in its value. Navin suggested that speed was the variable that underwent this catastrophic change, while flow and occupancy were the control variables.

While Navin's presentation was primarily an intuitive one, without recourse to data, Hall and co-authors picked up on the idea and attempted to flesh it out both mathematically and empirically. The initial effort appears in Hall (1987), in which the basic idea was presented, and some data applied. Figure 2.19 is a representation, showing the partially-folded (and torn) surface that is the Maxwell version of the cusp catastrophe surface, approximately located traffic data on that surface, and axes external to the surface representing the general correspondence with traffic variables. Further elaboration of this model is provided by Persaud and Hall (1989). Acha-Daza and Hall (1993) compared the ability of this model with the ability of some of the earlier models discussed above, to estimate speeds from flows obtained using inductive loop detectors and 30-second data, and found the catastrophe theory model to be slightly better than they were. Despite its empirical success, the problem with the model is that it appears to be inconsistent with the basic definitions and relationships with which this chapter opened.

2.5 Summary and links to other chapters

The current status of mathematical models for speed-flow-concentration relationships is in a state of flux. The models that dominated the discourse for nearly 30 years are incompatible with the data currently being obtained, and with currently accepted depictions of speed-flow curves, but no replacement models have yet been developed. The analyses of Cassidy and Newell have shown that the data used to develop most of the earlier models were flawed, as has been described above. An additional difficulty was noted by Duncan (1976; 1979): transforming variables, fitting equations, and then transforming the equations back to the original variables can lead to biased results, and is very sensitive to small changes in the initial curve-fitting. Progress has been made in understanding the relationships among the key traffic variables, but there is considerable scope for better models still.

It is important to note that the variables and analyses that have been discussed in this chapter are closely related to topics in several of the succeeding chapters. The human factors discussed in Chapter 3, for example, help to explain some of the variability in the data that have

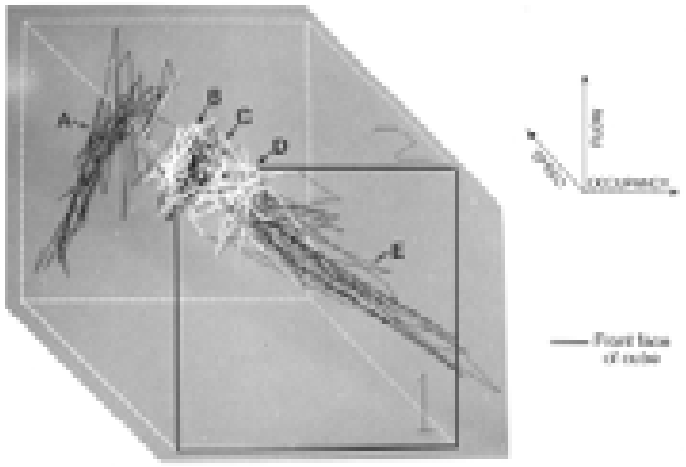


Figure 2.18
Second Perspective on the Three-Dimensional Relationship (Gilchrist and Hall 1989).

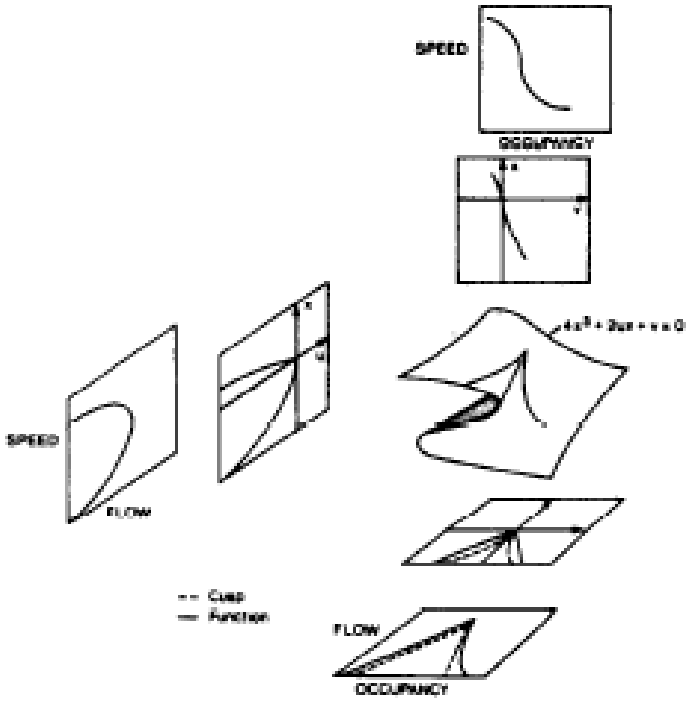


Figure 2.19
Catastrophe Theory Surface Showing Sketch of a Possible Freeway Function, and Projections and Transformations from That (Hall 1987).

been discussed here. As mentioned earlier, some of the bivariate models discussed above have been derived from car-following models, which are covered in Chapter 4. In addition, brief mention was made here of Lighthill and Witham's work, which has spawned a large literature related to the behavior of shocks and waves in traffic flow, covered in Chapter 5. All of these topics are inter-related, but have been addressed separately for ease of understanding.

Acknowledgements

The assistance of Michael Cassidy in developing this latest revision of this chapter was very helpful, as noted in the preface to the chapter. Valuable comments on an early draft of this material were received from Jim Banks, Frank Montgomery and Van Hurdle. The assistance of Richard Cunard and the TRB in bringing this project to completion is also appreciated.

References

Acha-Daza, J.A. and Hall, F.L. 1993. A graphical comparison of the predictions for speed given by catastrophe theory and some classic models, *Transportation Research Record* **1398**, 119-124.

Agyemang-Duah, K. and Hall, F.L. 1991. Some issues regarding the numerical value of freeway capacity, in *Highway Capacity and Level of Service*, Proceedings of the International Symposium on Highway Capacity, Karlsruhe. U. (Brannolte, ed.) Rotterdam, Balkema, 1-15.

Athol, P. 1965. Interdependence of certain operational characteristics within a moving traffic stream. *Highway Research Record* **72**, 58-87.

Banks, J.H. 1989. Freeway speed-flow-concentration relationships: more evidence and interpretations, *Transportation Research Record* **1225**, 53-60.

Banks, J.H. 1990. Flow processes at a freeway bottleneck. *Transportation Research Record* **1287**, 20-28.

Banks, J.H. 1991a. Two-capacity phenomenon at freeway bottlenecks: A basis for ramp metering?, *Transportation Research Record* **1320**, 83-90.

Banks, J.H. 1991b. The two-capacity phenomenon: some theoretical issues, *Transportation Research Record* **1320**, pp. 234-241.

Cassidy, MJ 1998 Bivariate relations in nearly stationary highway traffic. *Transp Res B, 32B,* 49-59.

Cassidy, MJ 1999 Traffic flow and capacity, in *Handbook of transportation science*, RW Hall (ed). Norwell, MA: Kluwer Academic Publishers.

Cassidy , MJ and Coifman, B 1997 Relation among average speed, flow and density and the analogous relation between density and occupancy. *Transportation Research Record 1591*, 1-6.

Ceder, A. 1975. Investigation of two-regime traffic flow models at the micro- and macro-scopic levels. Ph.D. dissertation, University of California, Berkeley, California.

Ceder, A. 1976. A deterministic traffic flow model for the two-regime approach. *Transportation Research Record* **567**, 16-32.

Ceder, A. and May, A.D. 1976. Further evaluation of single- and two-regime traffic flow models. *Transportation Research Record* **567**, 1-15.

Chin, H.C. and May, A.D. 1991. Examination of the speed-flow relationship at the Caldecott Tunnel, *Transportation Research Record* **1320**.

Daganzo, CF 1997 *Fundamentals of transportation and traffic operations*. New York: Elsevier.

Dendrinos, D.S. 1978. Operating speeds and volume to capacity ratios: the observed relationship and the fold catastrophe, *Transportation Research Record* **12**, 191-194.

Drake, J.S., Schofer, J.L., and May, A.D. 1967. A Statistical Analysis of Speed Density Hypotheses, *Highway Research Record* **154**, 53-87.

Duncan, N.C. 1976. A note on speed/flow/concentration relations, *Traffic Engineering and Control*, 34-35.

Duncan, N.C. 1979. A further look at speed/flow/concentration, *Traffic Engineering and Control*, 482-483.

Easa, S.M. and May, A.D. 1980. Generalized procedures for estimating single- and two-regime traffic flow models, *Transportation Research Record* **772**, 24-37.

Edie, L.C. 1961. Car following and steady-state theory for non-congested traffic, *Operations Research* **9**, 66-76.

Edie, LC 1965 Discussion of traffic stream measurements and definitions. *Proc Int Symp. On the Theory of Traffic Flow* J. Almond (ed) Paris: OECD, 139-154.

Edie, LC 1974 Flow theories, in *Traffic Science* DC Gazis (ed) New York: Wiley, 8-20.

Edie, LC and Foote, R 1960 Effect of shock waves on tunnel traffic flow. *Proc Highway Res Board 39*, 492-505.

Edie, L.C., Herman, R., and Lam, T., 1980. Observed multilane speed distributions and the kinetic theory of vehicular traffic, *Transportation Science* **14**, 55-76.

Gazis, D.C., Herman, R., and Potts, R. 1959. Car-following theory of steady-state traffic flow. *Operations Research* **7**, 499-505.

Gazis, D.C., Herman, R., and Rothery, R.W. 1961. Nonlinear follow-the-leader models of traffic flow. *Operations Research* **9**, 545-567.

Gerlough, D.L. and Huber, M.J. 1975. Traffic Flow Theory: a monograph. Special Report 165, Transportation Research Board (Washington D.C.: National Research Council)

Gilchrist, R. 1988. *Three-dimensional relationships in traffic flow theory variables*. Unpublished Master's degree report, Dept. of Civil Engineering, McMaster University, Hamilton, Ontario, Canada.

Gilchrist, R.S. and Hall, F.L. 1989. Three-dimensional relationships among traffic flow theory variables, *Transportation Research Record* **1225**, 99-108.

Greenberg, H. 1959. An analysis of traffic flow, *Operations Research*, **Vol. 7**, 78-85.

Greenshields, B.D. 1935. A study of traffic capacity, *HRB Proceedings* **14**, 448-477.

Haight, F.A. 1960. The volume, density relation in the theory of road traffic, *Operations Research* **8**, 572-573.

Haight, F.A. 1963. *Mathematical Theories of Traffic Flow*. (New York: Academic Press)

Hall, F.L. 1987. An interpretation of speed-flow-concentration relationships using catastrophe theory. *Transportation Research A*, **21A**, 191-201.

Hall, F.L., Allen, B.L. and Gunter, M.A. 1986. Empirical analysis of freeway flow-density relationships, *Transportation Research A*, **20A**, 197-210.

Hall, F.L. and Gunter, M.A. 1986. Further analysis of the flow-concentration relationship, *Transportation Research Record* **1091**, 1-9.

Hall, F.L. and Hall, L.M. 1990. Capacity and speed flow analysis of the QEW in Ontario" *Transportation Research Record* **1287**, 108-118.

Hall, F.L, Hurdle, V.F. and Banks, J.H. 1992. Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. In *Transportation Research Record* **1365**, TRB, National Research Council, Washington, D.C. 12-18.

Heidemann, D. and R. Hotop, R. 1990. Verteilung der Pkw-Geschwindigkeiten im Netz der

Bundesautobahnen- Modellmodifikation und Aktualisierung. *Straße und Autobahn* - Heft 2, 106-113.

HCM 1965. *Special Report 87: Highway Capacity Manual*. Transportation Research Board.

HCM 1985. *Special Report 209: Highway Capacity Manual*. Transportation Research Board.

Hsu, P. and Banks, J.H. 1993. Effects of location on congested-regime flow-concentration relationships for freeways, *Transportation Research Record* **1398**, 17-23.

Huber, M.J. 1957. Effect of temporary bridge on parkway performance. *Highway Research Board Bulletin* **167** 63-74.

Hurdle, V.F. and Datta, P.K. 1983. Speeds and flows on an urban freeway: some measurements and a hypothesis. *Transportation Research Record* **905**, 127-137.

Koshi, M., Iwasaki, M. and Okhura, I. 1983. Some findings and an overview on vehicular flow characteristics. *Proceedings*, 8*th International Symposium on Transportation and Traffic Flow Theory* (Edited by Hurdle, V.F., Hauer, E. and Steuart, G.F.) University of Toronto Press, Toronto, Canada, 403-426.

Kühne, R. and Immes, S. 1993. Freeway control systems for using section-related traffic variable detection. *Proceedings of the Pacific Rim TransTech Conference*, Seattle. New York: Am. Soc. Civil Eng., 56-62.

Lighthill, M.J., and Whitham, G.B. 1955. On kinematic waves: II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society* A229 No. 1178, 317-145.

Makagami, Y., Newell, G.F. and Rothery, R. 1971. Three-dimensional representation of traffic flow, *Transportation Science*, Vol. 5, No. 3, pp. 302-313.

May, A.D. 1990. *Traffic Flow Fundamentals*. Prentice-Hall.

May, A.D., Athol, P., Parker, W. and Rudden, J.B. 1963. Development and evaluation of Congress Street Expressway pilot detection system, *Highway Research Record* **21**, 48-70.

Morton, T.W. and Jackson, C.P. 1992. Speed/flow geometry relationships for rural dual-carriageways and motorways. TRRL Contractor Report 279. 55 pages. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Moskowitz, K 1954 Waiting for a gap in a traffic stream *Proc Highway Res Board 33*, 385-395.

Navin, F. 1986. Traffic congestion catastrophes. *Transportation Planning Technology* **11**, 19-25.

Newell G.F. 1971 *Applications of Queueing Theory*. London: Chapman Hall.

Newell, G.F. 1982. *Applications of Queueing Theory*, second edition. London: Chapman and Hall.

Newell, G.F. 1993. A simplified theory of kinematic waves in highway traffic, *Transportation Research B*, 27*B*: Part I, General theory, 281-287; Part II, Queueing at freeway bottlenecks, 289-303; Part III, Multi-destination flows, 305-313.

Newell, GF unpublished Notes on transportation operations Berkeley, CA: Univesity of California.

Persaud, B.N. and Hall, F.L. 1989. Catastrophe theory and patterns in 30-second freeway data -- implications for incident detection, *Transportation Research A* **23A**, 103-113.

Persaud, B.N. and Hurdle, V.F. 1988. Some new data that challenge some old ideas about speed-flow relationships. *Transportation Research Record* **1194**, 191-198.

Ringert, J. and Urbanik, II, T. 1993. Study of freeway bottlenecks in Texas, *Transportation Research Record* **1398**, 31-41.

Robertson, D. 1994. *Manual of Transportation Engineering Studies*, fifth edition. Arlington, VA: Institute of Transportation Engineers.

Stappert, K.H. and Theis, T.J. 1990. Aktualisierung der Verkehrsstärke-/Verkehrsgeschwindigkeitsbeziehungen des BVWP-Netzmodells. Research Report VU18009V89 Heusch/Boesefeldt, Aachen.

Thom, R. 1975. *Structural Stability and Morphogenesis*. English translation by D.H. Fowler of *Stabilite Structurelle et Morphogenese*. Benjamin: Reading, Massachusetts.

Underwood, R.T. 1961. Speed, volume, and density relationships: Quality and theory of traffic flow. Yale Bureau of Highway Traffic, pp. 141-188, as cited in Drake et al. 1967.

Wardrop, J.G. and Charlesworth, G. 1954. A method of estimating speed and flow of traffic from a moving vehicle. Proceedings of the Institution of Civil Engineers, Part II, Vol. 3, 158-171.

Wattleworth, J.A. 1963. Some aspects of macroscopic freeway traffic flow theory, *Traffic Engineering* **34**: 2, 15-20.

Wemple, E.A., Morris, A.M. and May, A.D. 1991. Freeway capacity and level of service concepts, in *Highway Capacity and Level of Service*, Proc. of the International Symposium on Highway Capacity, Karlsruhe (U. Brannolte, Ed.) Rotterdam: Balkema, 439-455.

Wright, C. 1973. A theoretical analysis of the moving observer method, *Transportation Research* **7**, 293-311.

HUMAN FACTORS

BY RODGER J. KOPPA[5]

[5]Associate Professor, Texas A&M University, College Station, TX 77843.

CHAPTER 3 - Frequently used Symbols

ξ	=	parameter of log normal distribution ~ standard deviation
λ	=	parameter of log normal distribution ~ median
σ	=	standard deviation
Φ	=	value of standard normal variate
a_{GV}	=	maximum acceleration on grade
a_{LV}	=	maximum acceleration on level
A	=	movement amplitude
C_r	=	roadway curvature
$C_{(t)}$	=	vehicle heading
CV	=	coefficient of variation
d	=	braking distance
D	=	distance from eye to target
$E_{(t)}$	=	symptom error function
f	=	coefficient of friction
F_s	=	stability factor
g	=	acceleration of gravity
$g_{(s)}$	=	control displacement
G	=	gradient
H	=	information (bits)
K	=	gain (dB)
l	=	wheel base
L	=	diameter of target (letter or symbol)
LN	=	natural log
M	=	mean
MT	=	movement time
N	=	equiprobable alternatives
PRT	=	perception-response time
$R_{(t)}$	=	desired input forcing function
RT	=	reaction time (sec)
s	=	Laplace operator
SR	=	steering ratio (gain)
SSD	=	stopping sight distance
t	=	time
T_L	=	lead term constant
T_l	=	lag term constant
T_N	=	neuro-muscular time constant
u	=	speed
V	=	initial speed
W	=	width of control device
Z	=	standard normal score

3.
HUMAN FACTORS

3.1 Introduction

In this chapter, salient performance aspects of the human in the context of a person-machine control system, the motor vehicle, will be summarized. The driver-vehicle system configuration is ubiquitous. Practically all readers of this chapter are also participants in such a system; yet many questions, as will be seen, remain to be answered in modeling the behavior of the human component alone. Recent publications (IVHS 1992; TRB 1993) in support of Intelligent Transportation Systems (ITS) have identified study of "Plain Old Driving" (POD) as a fundamental research topic in ITS. For the purposes of a transportation engineer interested in developing a molecular model of traffic flow in which the human in the vehicle or an individual human-vehicle comprises a unit of analysis, some important performance characteristics can be identified to aid in the formulation, even if a comprehensive transfer function for the driver has not yet been formulated.

This chapter will proceed to describe first the discrete components of performance, largely centered around neuromuscular and cognitive time lags that are fundamental parameters in human performance. These topics include perception-reaction time, control movement time, responses to the presentation of traffic control devices, responses to the movements of other vehicles, handling of hazards in the roadway, and finally how different segments of the driving population may differ in performance.

Next, the kind of control performance that underlies steering, braking, and speed control (the primary control functions) will be described. Much research has focused on the development of adequate models of the tracking behavior fundamental to steering, much less so for braking or for speed control.

After fundamentals of open-loop and closed-loop vehicle control are covered, applications of these principles to specific maneuvers of interest to traffic flow modelers will be discussed. Lane keeping, car following, overtaking, gap acceptance, lane closures, stopping and intersection sight distances will also be discussed. To round out the chapter, a few other performance aspects of the driver-vehicle system will be covered, such as speed limit changes and distractions on the highway.

3.1.1 The Driving Task

Lunenfeld and Alexander (1990) consider the driving task to be a hierarchical process, with three levels: (1) Control, (2) Guidance, and (3) Navigation. The *control* level of performance comprises all those activities that involve second-to-second exchange of information and control inputs between the driver and the vehicle. This level of performance is at the control interface. Most control activities, it is pointed out, are performed "automatically," with little conscious effort. In short, the control level of performance is *skill based*, in the approach to human performance and errors set forth by Jens Rasmussen as presented in *Human Error* (Reason 1990).

Once a person has learned the rudiments of control of the vehicle, the next level of human performance in the driver-vehicle control hierarchy is the rules-based (Reason 1990) *guidance* level as Rasmussen would say. The driver's main activities "involve the maintenance of a safe speed and proper path relative to roadway and traffic elements ." (Lunenfeld and Alexander 1990) Guidance level inputs to the system are dynamic speed and path responses to roadway geometrics, hazards, traffic, and the physical environment. Information presented to the driver-vehicle system is from traffic control devices, delineation, traffic and other features of the environment, continually changing as the vehicle moves along the highway.

These two levels of vehicle control, control and guidance, are of paramount concern to modeling a corridor or facility. The third (and highest) level in which the driver acts as a supervisor apart, is *navigation*. Route planning and guidance while enroute, for example, correlating directions from a map with guide signage in a corridor, characterize the navigation level of performance. Rasmussen would call this level *knowledge-based behavior*. Knowledge based behavior will become increasingly more important to traffic flow theorists as Intelligent Transportation Systems (ITS) mature. Little is currently known about how enroute diversion and route changes brought about by ITS technology affect traffic flow, but much research is underway. This chapter will discuss driver performance in the conventional highway system context, recognizing that emerging ITS technology in the next ten years may radically change many driver's roles as players in advanced transportation systems.

At the control and guidance levels of operation, the driver of a motor vehicle has gradually moved from a significant prime mover, a supplier of forces to change the path of the vehicle, to an information processor in which strength is of little or no consequence. The advent of power assists and automatic transmissions in the 1940's, and cruise controls in the 1950's moved the driver more to the status of a manager in the system. There are commercially available adaptive controls for severely disabled drivers (Koppa 1990) which reduce the actual movements and strength required of drivers to nearly the vanishing point. The fundamental control tasks, however, remain the same.

These tasks are well captured in a block diagram first developed many years ago by Weir (1976). This diagram, reproduced in Figure 3.1, forms the basis for the discussion of driver performance, both discrete and continuous. Inputs enter the driver-vehicle system from other vehicles, the roadway, and the driver him/herself (acting at the *navigation* level of performance).

The fundamental display for the driver is the visual field as seen through the windshield, and the dynamics of changes to that field generated by the motion of the vehicle. The driver attends to selected parts of this input, as the field is interpreted as the visual world. The driver as system manager as well as active system component "hovers" over the control level of performance. Factors such as his or her experience, state of mind, and stressors (e.g., being on a crowded facility when 30 minutes late for a meeting) all impinge on the supervisory or monitoring level of performance, and directly or indirectly affect the control level of performance. Rules and knowledge govern driver decision making and the second by second psychomotor activity of the driver. The actual control

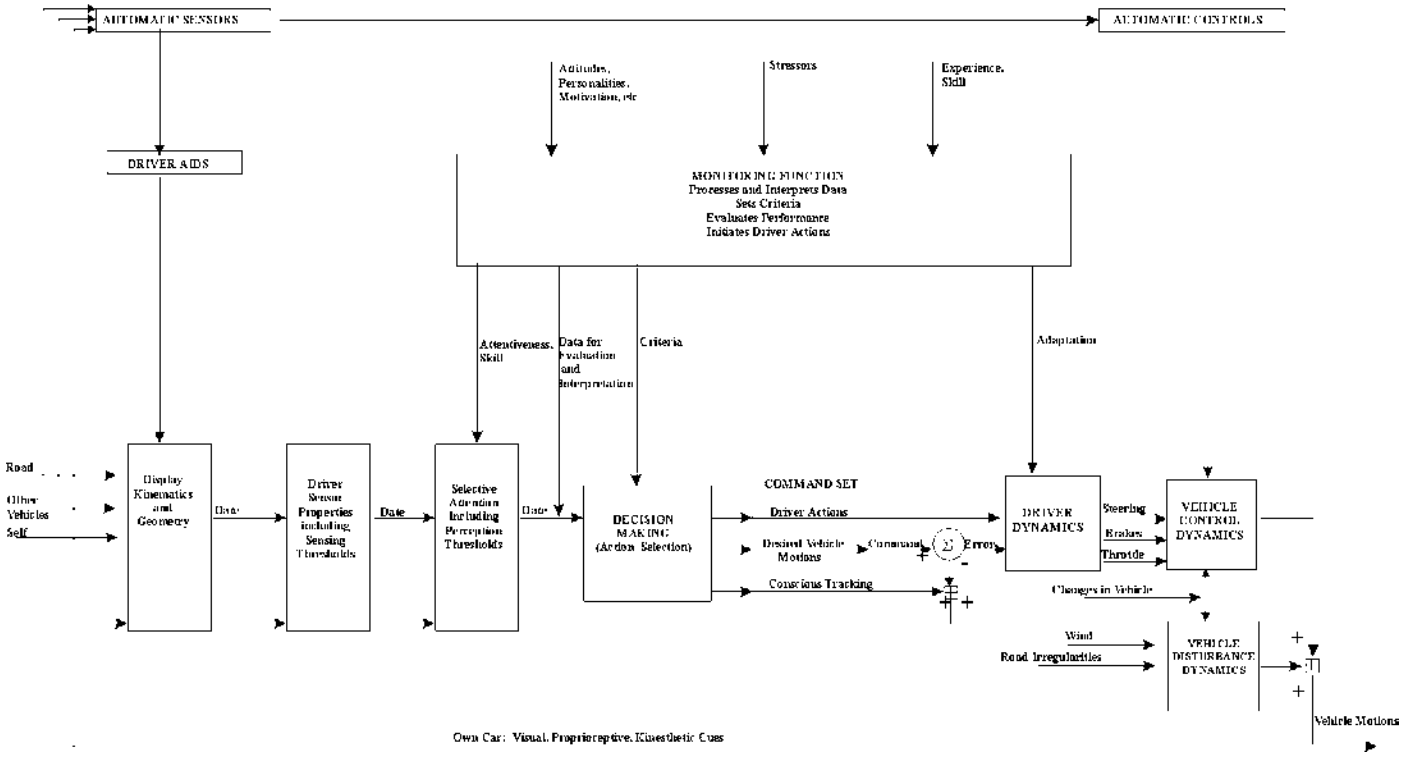


Figure 3.1
Generalized Block Diagram of the Car-Driver-Roadway System.

movements made by the driver couple with the vehicle control at the interface of throttle, brake, and steering. The vehicle, in turn, as a dynamic physical process in its own right, is subject to inputs from the road and the environment. The resolution of control dynamics and vehicle disturbance dynamics is the vehicle path.

As will be discussed, a considerable amount of information is available for some of the lower blocks in this diagram, the ones associated with braking reactions, steering inputs, andvehic le control dynamics. Far less is really known about the higher-order functions that any driver knows are going on while he or she drives.

3.2 Discrete Driver Performance

3.2.1 Perception-Response Time

Nothing in the physical universe happens instantaneously. Compared to some physical or chemical processes, the simplest human reaction to incoming information is very slow indeed. Ever since the Dutch physiologist Donders started to speculate in the mid 19th century about central processes involved in choice and recognition reaction times, there have been numerous models of this process. The early 1950's saw Information Theory take a dominant role in experimental psychology. The linear equation

$$RT = a + bH \qquad (3.1)$$

Where:

RT = Reaction time, seconds
H = Estimate of transmitted information
H = $log_2 N$, if N equiprobable alternatives
a = Minimum reaction time for that modality
b = Empirically derived slope, around 0.13 seconds (sec) for many performance situations

that has come to be known as the Hick-Hyman "Law" expresses a relationship between the number of alternatives that must be sorted out to decide on a response and the total reaction time, that is, that lag in time between detection of an input (stimulus) and the start of initiation of a control or other response. If the time for the response itself is also included, then the total lag is termed "response time." Often, the terms "reaction time" and "response time" are used interchangeably, but one (reaction) is always a part of the other (response).

Underlying the Hick-Hyman Law is the two-component concept: part of the total time depends upon choice variables, and part is common to all reactions (the intercept). Other components can be postulated to intervene in the choice variable component, other than just the information content. Most of these models have then been *chaining* individual components that are presumably orthogonal or uncorrelated with one another. Hooper and McGee (1983) postulate a very typical and plausible model with such components for braking response time, illustrated in Table 3.1.

Each of these elements is derived from empirical data, and is in the 85th percentile estimate for that aspect of time lag. Because it is doubtful that any driver would produce 85th percentile values for each of the individual elements, 1.50 seconds probably represents an extreme upper limit for a driver's perception-reaction time. This is an estimate for the simplest kind of reaction time, with little or no decision making. The driver reacts to the input by lifting his or her foot from the accelerator and placing it on the brake pedal. But a number of writers, for example Neuman (1989), have proposed perception-reaction times (PRT) for different types of roadways, ranging from 1.5 seconds for low-volume roadways to 3.0 seconds for urban freeways. There are more things happening, and more decisions to be made per unit block of time on a busy urban facility than on a rural county road. Each of those added factors increase the PRT. McGee (1989) has similarly proposed different values of PRT as a function of design speed. These estimates, like those in Table 3.1, typically include the time for the driver to move his or her foot from the accelerator to the brake pedal for brake application.

Table 3.1
Hooper-McGee Chaining Model of Perception-Response Time

Component	Time (sec)	Cumulative Time (sec)
1) Perception		
Latency	0.31	0.31
Eye Movement	0.09	0.4
Fixation	0.2	1
Recognition	0.5	1.5
2) Initiating Brake Application	1.24	2.74

Any statistical treatment of empirically obtained PRT's should take into account a fundamental if not always vitally important fact: the times cannot be distributed according to the normal or gaussian probability course. Figure 3.2 illustrates the actual shape of the distribution. The distribution has a marked positive skew, because there cannot be such a thing as a negative reaction time, if the time starts with onset of the signal with no anticipation by the driver. Taoka (1989) has suggested an adjustment to be applied to PRT data to correct for the non-normality, when sample sizes are "large" --50 or greater.

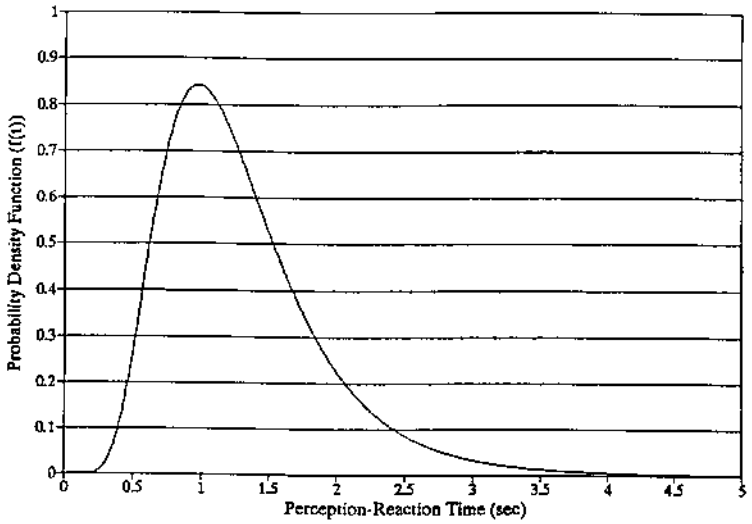


Figure 3.2
Lognormal Distribution of Perception-Reaction Time.

The log-normal probability density function is widely used in quality control engineering and other applications in which values of the observed variable, *t*, are constrained to values equal to or greater than zero, but may take on extreme positive values, exactly the situation that obtains in considering PRT. In such situations, the natural logarithm of such data may be assumed to approach the normal or gaussian distribution. Probabilities associated with the log-normal distribution can thus be determined by the use of standard-score tables. Ang and Tang (1975) express the log-normal probability density function *f(t)* as follows:

$$f(t) = \frac{1}{\sqrt{2\Pi}\ \xi t}\exp\left[\left(\frac{LN(t)-\lambda}{\xi}\right)^2\right] \qquad (3.2)$$

where the two parameters that define the shape of the distribution are λ and ξ. It can be shown that these two parameters are related to the mean and the standard deviation of a sample of data such as PRT as follows:

$$\xi^2 = LN\left(1+\frac{\sigma^2}{\mu^2}\right) \qquad (3.3)$$

The parameter λ is related to the median of the distribution being described by the simple relationship of the natural logarithm of the median. It can also be shown that the value of the standard normal variate (equal to probability) is related to these parameters as shown in the following equation:

$$\lambda = LN\left(\frac{\mu}{\sqrt{1+\sigma^2/\mu^2}}\right) \qquad (3.4)$$

$$\Phi\left(\frac{LN(t)-\lambda}{\xi}\right) = 0.5,\ 0.85,\ etc. \qquad (3.5)$$

and the standard score associated with that value is given by:

$$\frac{LN(t)-\lambda}{\xi} = Z \qquad (3.6)$$

Therefore, the value of *LN(t)* for such percentile levels as 0.50 (the median), the 85th, 95th, and 99th can be obtained by substituting in Equation 3.6 the appropriate *Z* score of 0.00, 1.04, 1.65, and 2.33 for *Z* and then solving for *t*. Converting data to log-normal approximations of percentile values should be considered when the number of observations is reasonably large, over 50 or more, to obtain a better fit. Smaller data sets will benefit more from a tolerance interval approach to approximate percentiles (Odeh 1980).

A very recent literature review by Lerner and his associates (1995) includes a summary of brake PRT (including brake onset) from a wide variety of studies. Two types of response situation were summarized: (1) The driver does not know when or even if the stimulus for braking will occur, i.e., he or she is surprised, something like a real-world occurrence on the highway; and (2) the driver is aware that the signal to brake will occur, and the only question is when. The Lerner et al. (1995) composite data were converted by this writer to a log-normal transformation to produce the accompanying Table 3.2.

Sixteen studies of braking PRT form the basis for Table 3.2. Note that the 95th percentile value for a "surprise" PRT (2.45 seconds) is very close to the AASHTO estimate of 2.5 seconds which is used for all highway situations in estimating both stopping sight distance and other kinds of sight distance (Lerner et al. 1995).

In a very widely quoted study by Johansson and Rumar (1971), drivers were waylaid and asked to brake very briefly if they heard a horn at the side of the highway in the next 10 kilometers. Mean PRT for 322 drivers in this situation was 0.75 seconds with an SD of 0.28 seconds. Applying the Taoka conversion to the log normal distribution yields:

50th percentile PRT = 0.84 sec
85th percentile PRT = 1.02 sec
95th percentile PRT = 1.27 sec
99th percentile PRT = 1.71 sec

Table 3.2
Brake PRT - Log Normal Transformation

	"Surprise"	"Expected"
Mean	1.31 (sec)	0.54
Standard Dev	0.61	0.1
λ	0.17 (no unit)	-0.63 (no unit)
ξ	0.44 (no unit)	0.18 (no unit)
50th percentile	1.18	0.53
85th percentile	1.87	0.64
95th percentile	2.45	0.72
99th percentile	3.31	0.82

In very recent work by Fambro et al. (1994) volunteer drivers in two age groups (Older: 55 and up; and Young: 18 to 25) were suddenly presented with a barrier that sprang up from a slot in the pavement in their path, with no previous instruction. They were driving a test vehicle on a closed course. Not all 26 drivers hit the brakes in response to this breakaway barrier. The PRT's of the 22 who did are summarized in Table 3.3 (Case 1). None of the age differences were statistically significant. Additional runs were made with other drivers in their own cars equipped with the same instrumentation. Nine of the 12 drivers made stopping maneuvers in response to the emergence of the barrier. The results are given in Table 3.3 as Case 2. In an attempt (Case 3) to approximate real-world driving conditions, Fambro et al. (1994) equipped 12 driver's own vehicles with instrumentation. They were asked to drive a two-lane undivided secondary road ostensibly to evaluate the drivability of the road.

Table 3.3
Summary of PRT to Emergence of Barrier or Obstacle

Case 1. Closed Course, Test Vehicle			
12	Older:	Mean = 0.82 sec;	SD = 0.16 sec
10	Young:	Mean = 0.82 sec;	SD = 0.20 sec
Case 2. Closed Course, Own Vehicle			
7	Older:	Mean = 1.14 sec;	SD = 0.35 sec
3	Young:	Mean = 0.93 sec;	SD = 0.19 sec
Case 3. Open Road, Own Vehicle			
5	Older:	Mean = 1.06 sec;	SD = 0.22 sec
6	Young:	Mean = 1.14 sec;	SD = 0.20 sec

A braking incident was staged at some point during this test drive. A barrel suddenly rolled out of the back of a pickup parked at the side of the road as he or she drove by. The barrel was snubbed to prevent it from actually intersecting the driver's path, but the driver did not know this. The PRT's obtained by this ruse are summarized in Table 3.4. One driver failed to notice the barrel, or at least made no attempt to stop or avoid it.

Since the sample sizes in these last two studies were small, it was considered prudent to apply statistical tolerance intervals to these data in order to estimate proportions of the driving population that might exhibit such performance, rather than using the Taoka conversion. One-sided tolerance tables published by Odeh (1980) were used to estimate the percentage of drivers who would respond in a given time or shorter, based on these findings. These estimates are given in Table 3.4 (95 percent confidence level), with PRT for older and younger drivers combined.

The same researchers also conducted studies of driver response to expected obstacles. The ratio of PRT to a totally unexpected event to an expected event ranges from 1.35 to 1.80 sec, consistent with Johansson and Rumar (1971). Note, however, that one out of 12 of the drivers in the open road barrel study (Case 3) *did not appear to notice the hazard at all.* Thirty percent of the drivers confronted by the artificial barrier under closed-course conditions also did not respond appropriately. How generalizable these percentages are to the driver population remains an open question that requires more research. For analysis purposes, the values in Table 3.4 can be used to approximate the driver PRT envelope for an unexpected event. PRT's for expected events, e.g., braking in a queue in heavy traffic, would range from 1.06 to 1.41 second, according to the ratios given above (99th percentile).

These estimates may not adequately characterize PRT under conditions of complete surprise, i.e., when expectancies are greatly violated (Lunenfeld and Alexander 1990). Detection times may be greatly increased if, for example, an unlighted vehicle is suddenly encountered in a traffic lane in the dark, to say nothing of a cow or a refrigerator.

Table 3.4
Percentile Estimates of PRT to an Unexpected Object

Percentile	Case 1 Test Vehicle Closed Course	Case 2 Own Vehicle Closed Course	Case 3 Own Vehicle Open Road
50th	0.82 sec	1.09 sec	1.11 sec
75th	1.02 sec	1.54 sec	1.40 sec
90th	1.15 sec	1.81 sec	1.57 sec
95th	1.23 sec	1.98 sec	1.68 sec
99th	1.39 sec	2.31 sec	1.90 sec

Adapted from Fambro et al. (1994).

3.3 Control Movement Time

Once the lag associated with perception and then reaction has ensued and the driver just begins to move his or her foot (or hand, depending upon the control input to be effected), the amount of time required to make that movement may be of interest. Such control inputs are overt motions of an appendage of the human body, with attendant inertia and muscle fiber latencies that come into play once the efferent nervous impulses arrives from the central nervous system.

3.3.1 Braking Inputs

As discussed in Section 3.3.1 above, a driver's braking response is composed of two parts, prior to the actual braking of the vehicle: the perception-reaction time (PRT) and immediately following, movement time (MT).

Movement time for any sort of response was first modeled by Fitts in 1954. The simple relationship among the range or amplitude of movement, size of the control at which the control movement terminates, and basic information about the minimum "twitch" possible for a control movement has long been known as "Fitts' Law."

$$MT = a + b \ Log_2 \left(\frac{2A}{W} \right) \qquad (3.7)$$

where,

- a = minimum response time lag, no movement
- b = slope, empirically determined, different for each limb
- A = amplitude of movement, i.e., the distance from starting point to end point
- W = width of control device (in direction of movement)

The term

$$Log_2 \left(\frac{2A}{W} \right) \qquad (3.8)$$

is the "Index of Difficulty" of the movement, in binary units, thus linking this simple relationship with the Hick-Hyman equation discussed previously in Section 3.3.1.

Other researchers, as summarized by Berman (1994), soon found that certain control movements could not be easily modeled by Fitts' Law. Accurate tapping responses less than 180 msec were not included. Movements which are short and quick also appear to be preplanned, or "programmed," and are open-loop. Such movements, usually not involving visual feedback, came to be modeled by a variant of Fitts' Law:

$$MT = a + b \sqrt{A} \qquad (3.9)$$

in which the width of the target control (W) plays no part.

Almost all such research was devoted to hand or arm responses. In 1975, Drury was one of the first researchers to test the applicability of Fitts' Law and its variants to foot and leg movements. He found a remarkably high association for fitting foot tapping performance to Fitts' Law. Apparently, all appendages of the human body can be modeled by Fitts' Law or one of its variants, with an appropriate adjustment of a and b, the empirically derived parameters. Parameters a and b are sensitive to age, condition of the driver, and circumstances such as degree of workload, perceived hazard or time stress, and pre-programming by the driver.

In a study of pedal separation and vertical spacing between the planes of the accelerator and brake pedals, Brackett and Koppa (1988) found separations of 10 to 15 centimeters (cm), with little or no difference in vertical spacing, produced control movement in the range of 0.15 to 0.17 sec. Raising the brake pedal more than 5 cm above the accelerator lengthened this time significantly. If pedal separation (= A in Fitts' Law) was varied, holding pedal size constant, the mean MT was 0.22 sec, with a standard deviation of 0.20 sec.

In 1991, Hoffman put together much of the extant literature and conducted studies of his own. He found that the Index of Difficulty was sufficiently low (<1.5) for all pedal placements found on passenger motor vehicles that visual control was

unnecessary for accurate movement, i.e., movements were ballistic in nature. *MT* was found to be greatly influenced by vertical separation of the pedals, but comparatively little by changes in *A*, presumably because the movements were ballistic or open-loop and thus not correctable during the course of the movement. *MT* was lowest at 0.20 sec with no vertical separation, and rose to 0.26 sec if the vertical separation (brake pedal higher than accelerator) was as much as 7 cm. A very recent study by Berman (1994) tends to confirm these general *MT* evaluations, but adds some additional support for a ballistic model in which amplitude A does make a difference.

Her *MT* findings for a displacement of (original) 16.5 cm and (extended) 24.0 cm, or change of 7.5 cm can be summarized as follows:

	Original pedal	Extended pedal
Mean	0.20 sec	0.29 sec
Standard Deviation	0.05 sec	0.07 sec
95 percent tolerance level	0.32 sec	0.45 sec
99 percent tolerance level	0.36 sec	0.51 sec

The relationship between perception-reaction time and *MT* has been shown to be very weak to nonexistent. That is, a long reaction time does not necessarily predict a long *MT*, or any other relationship between these two times. A recent analysis by the writer yielded a Pearson Product-Moment Correlation Coefficient (r) value of 0.17 between these two quantities in a braking maneuver to a completely unexpected object in the path of the vehicle (based on 21 subjects). Total PRT's as presented in Section 3.3.1 should be used for discrete braking control movement time estimates; for other situations, the modeler could use the tolerance levels in Table 3.5 for *MT*, chaining them after an estimate of perception (including decision) and reaction time for the situation under study (95 percent confidence level). See Section 3.14 for a discussion on how to combine these estimates.

3.3.2 Steering Response Times

Summala (1981) covertly studied driver responses to the sudden opening of a car door in their path of travel. By "covert" is meant the drivers had no idea they were being observed or were participating in the study. This researcher found that neither the latency nor the amount of deviation from the pre-event pathway was dependent upon the car's prior position with respect to the opening car door. Drivers responded with a ballistic "jerk" of the steering wheel. The mean response latency for these Finnish drivers was 1.5 sec, and reached the half-way point of maximum displacement from the original path in about 2.5 sec. The

Table 3.5
Movement Time Estimates

Source	N	Mean (Std)	75th Sec	90th Sec	95th Sec	99th Sec
Brackett (Brackett and Koppa 1988)	24	0.22 (0.20)	0.44	0.59	0.68	0.86
Hoffman (1991)	18	0.26 (0.20)	0.50	0.66	0.84	1.06
Berman (1994)	24	0.20 (0.05)	0.26	0.29	0.32	0.36

3.4 Response Distances and Times to Traffic Control Devices

The driving task is overwhelmingly visual in nature; external information coming through the windshield constitutes nearly all the information processed. A major input to the driver which influences his or her path and thus is important to traffic flow theorists is traffic information imparted by traffic control devices (TCD). The major issues concerned with TCD are all related to distances at which they may be (1) detected as objects in the visual field; (2) recognized as traffic control devices: signs, signals, delineators, and barricades; (3) legible or identifiable so that they may be comprehended and acted upon. Figure 3.3 depicts a conceptual model for TCD information processing, and the many variables which affect it. The research literature is very rich with data related to target detection in complex visual environments, and a TCD's target value also depends upon the driver's predilection to look for and use such devices.

3.4.1 Traffic Signal Change

From the standpoint of traffic flow theory and modeling, a major concern is at the stage of legibility or identification and a combination of "read" and "understand" in the diagram in Figure 3.3. One of the most basic concerns is driver response or lag to

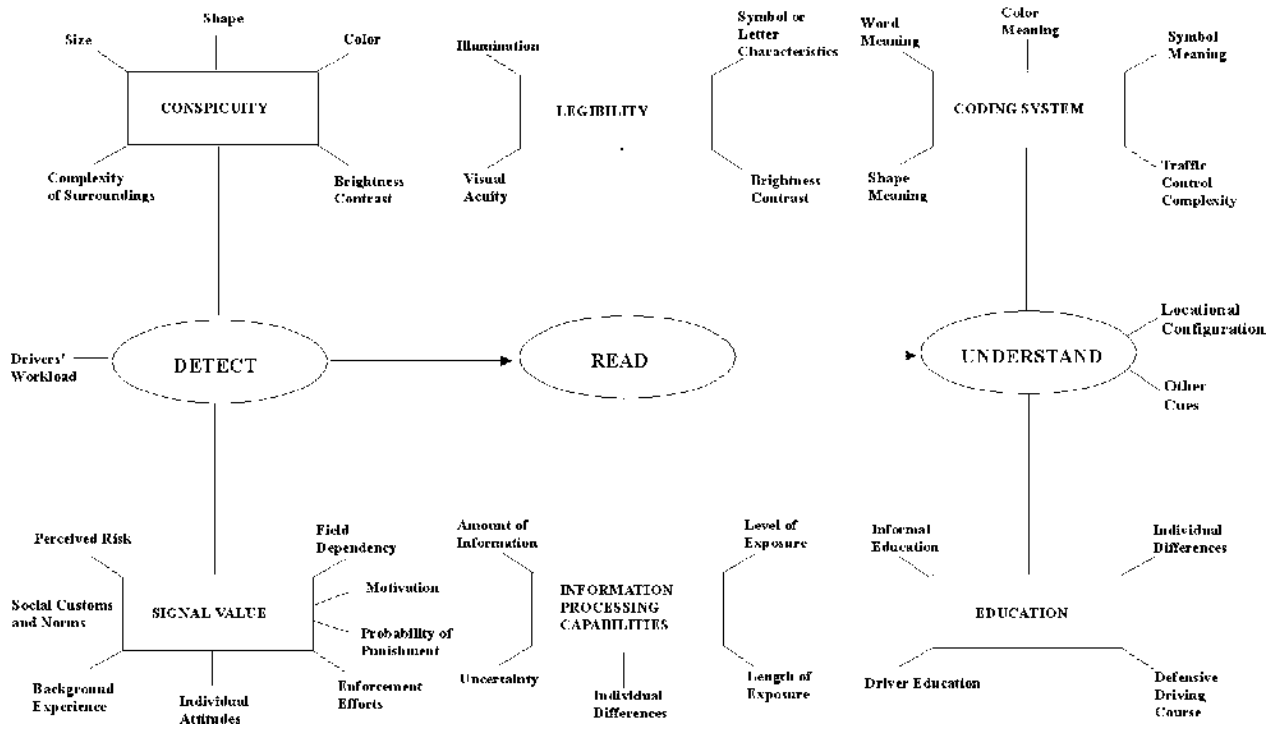


Figure 3.3
A Model of Traffic Control Device Information Processing.

changing traffic signals. Chang et al. (1985) found through covert observations at signalized intersections that drivers response lag to signal change (time of change to onset of brake lamps) averaged 1.3 sec, with the 85th percentile PRT estimated at 1.9 sec and the 95th percentile at 2.5 sec. This PRT to signal change is somewhat inelastic with respect to distance from the traffic signal at which the signal state changed. The mean PRT (at 64 kilometers per hour (km/h)) varied by only 0.20 sec within a distance of 15 meters (m) and by only 0.40 sec within 46 m.

Wortman and Matthias (1983) found similar results to Chang et al. (1985) with a mean PRT of 1.30 sec, and a 85th percentile PRT of 1.5 sec. Using tolerance estimates based on their sample size, (95 percent confidence level) the 95th percentile PRT was 2.34 sec, and the 99th percentile PRT was 2.77 sec. They found very little relationship between the distance from the intersection and either PRT or approach speed ($r^2 = 0.08$). So the two study findings are in generally good agreement, and the following estimates may be used for driver response to signal change:

Mean PRT to signal change =	1.30 sec
85th percentile PRT =	1.50 sec
95th percentile PRT =	2.50 sec
99th percentile PRT =	2.80 sec

If the driver is stopped at a signal, and a straight-ahead maneuver is planned, PRT would be consistent with those values given in Section 3.3.1. If complex maneuvers occur after signal change (e.g., left turn yield to oncoming traffic), the Hick-Hyman Law (Section 3.2.1) could be used with the *y* intercept being the basic PRT to onset of the traffic signal change. Considerations related to intersection sight distances and gap acceptance make such predictions rather difficult to make without empirical validation. These considerations will be discussed in Section 3.15.

3.4.2 Sign Visibility and Legibility

The psychophysical limits to legibility (alpha-numeric) and identification (symbolic) sign legends are the resolving power of

Table 3.6
Visual Acuity and Letter Sizes

Snellen Acuity	Visual angle of letter or symbol		Legibility Index
SI (English)	'of arc	radians	m/cm
6/3 (20/10)	2.5	0.00073	13.7
6/6 (20/20)	5	0.00145	6.9
6/9 (20/30)	7.5	0.00218	4.6
6/12 (20/40)	10	0.00291	3.4
6/15 (20/50)	12.5	0.00364	2.7
6/18 (20/60)	15	0.00436	2.3

the visual perception system, the effects of the optical train leading to presentation of an image on the retina of the eye, neural processing of that image, and further processing by the brain. Table 3.6 summarizes visual acuity in terms of visual angles and legibility indices.

The exact formula for calculating visual angle is

$$\triangle = 2 \arctan\left(\frac{L}{2D}\right) \qquad (3.10)$$

where, L = diameter of the target (letter or symbol)
D = distance from eye to target in the same units

All things being equal, two objects that subtend the same visual angle will elicit the same response from a human observer, regardless of their actual sizes and distances. In Table 3.6 the Snellen eye chart visual acuity ratings are related to the size of objects in terms of visual arc, radians (equivalent for small sizes to the tangent of the visual arc) and legibility indices. Standard transportation engineering resources such as the *Traffic Control Devices Handbook* (FHWA 1983) are based upon these fundamental facts about visual performance, but it should be clearly recognized that it is very misleading to extrapolate directly from letter or symbol legibility/recognition sizes to *sign* perceptual distances, especially for word signs. There are other expectancy cues available to the driver, word length, layout, etc. that can lead to performance better than straight visual angle computations would suggest. Jacobs, Johnston, and Cole (1975) also point out an elementary fact that 27 to 30 percent of the driving population cannot meet a 6/6 (20/20) criterion. Most states in the U.S. have a 6/12 (20/40) static acuity criterion for unrestricted licensure, and accept 6/18 (20/60) for restricted (daytime, usually) licensure. Such tests in driver license offices are subject to error, and examiners tend to be very lenient. Night-time static visual acuity tends to be at least one Snellen line worse than daytime, and much worse for older drivers (to be discussed in Section 3.8).

Jacobs, et al. also point out that the sign size for 95th percentile recognition or legibility is 1.7 times the size for 50th percentile performance. There is also a pervasive notion in the research that *letter* sign legibility distances are half *symbol* sign recognition distances, when drivers are very familiar with the symbol (Greene 1994). Greene (1994), in a very recent study, confirmed these earlier findings, and also notes that extreme variability exists from trial to trial *for the same observer* on a given sign's recognition distance. Presumably, word signs would manifest as much or even more variability. Complex, fine detail signs such as Bicycle Crossing (MUTCD W11-1) were observed to have coefficients of variation between subjects of 43 percent. Coefficient of Variation (*CV*) is simply:

$$CV = 100 \cdot (\text{Std Deviation/Mean}) \qquad (3.11)$$

In contrast, very simple symbol signs such as T-Junction (MUTCD W2-4) had a CV of 28 percent. Within subject variation (from trial to trial) on the same symbol sign is summarized in Table 3.7.

Before any reliable predictions can be made about legibility or recognition distances of a given sign, Greene (1994) found that six or more trials under controlled conditions must be made, either in the laboratory or under field conditions. Greene (1994) found percent differences between high-fidelity laboratory and field recognition distances to range from 3 to 21 percent, depending upon sign complexity. These differences consistent with most researchers, were all in the direction of laboratory distances being greater than actual distances; the laboratory tends to overestimate field legibility distances. Variability in legibility distances, however, is as great in the laboratory as it is under field trials.

With respect to visual angle required for recognition, Greene found, for example, that the Deer Crossing at the mean recognition distance had a mean visual angle of 0.00193 radian, or 6.6 minutes of arc. A more complex, fine detail sign such as Bicycle Crossing required a mean visual angle of 0.00345 radian or 11.8 minutes of arc to become recognizable.

With these considerations in mind, here is the best recommendation that this writer can make. For the purposes of predicting driver comprehension of signs and other devices that require interpretation of words or symbols use the data in Table 3.6 as "best case," with actual performance expected to be somewhat to much worse (i.e., requiring closer distances for a given size of character or symbol). The best visual acuity that can be expected of drivers under optimum contrast conditions so far as static acuity is concerned would be 6/15 (20/50) when the sizable numbers of older drivers is considered [13 percent in 1990 were 65 or older (O'Leary and Atkins 1993)].

Table 3.7
Within Subject Variation for Sign Legibility

Sign		Young Drivers		Older Drivers	
		Min CV	Max CV	Min CV	Max CV
WG-3 2 Way Traffic		3.9	21.9	8.9	26.7
W11-1	Bicycle Cross	6.7	37.0	5.5	39.4
W2-1	Crossroad	5.2	16.3	2.0	28.6
W11-3	Deer Cross	5.4	21.3	5.4	49.2
W8-5	Slippery	7.7	33.4	15.9	44.1
W2-5	T-Junction	5.6	24.6	4.9	28.7

3.4.3 Real-Time Displays and Signs

With the advent of Intelligent Transportation Systems (ITS), traffic flow modelers must consider the effects of changeable message signs on driver performance in traffic streams. Depending on the design of such signs, visual performance to them may not differ significantly from conventional signage. Signs with active (lamp or fiber optic) elements may not yield the legibility distances associated with static signage, because Federal Highway Administration, notably the definitive manual by Dudek (1990).

3.4.4 Reading Time Allowance

For signs that cannot be comprehended in one glance, i.e., word message signs, allowance must be made for reading the information and then deciding what to do, before a driver in traffic will begin to maneuver in response to the information. Reading speed is affected by a host of factors (Boff and Lincoln 1988) such as the type of text, number of words, sentence structure, information order, whatever else the driver is doing, the purpose of reading, and the method of presentation. The USAF resource (Boff and Lincoln 1988) has a great deal of general information on various aspects of reading sign material. For purposes of traffic flow modeling, however, a general rule of thumb may suffice. This can be found in Dudek (1990): "Research...has indicated that a minimum exposure time of one second per short word (four to eight characters) (*exclusive of prepositions and other similar connectors*) or two seconds per unit of information, whichever is largest, should be used for unfamiliar drivers. On a sign having 12 to 16 characters per line, this minimum exposure time will be two seconds per line." "Exposure time" can also be interpreted as "reading time" and so used in estimating how long drivers will take to read and comprehend a sign with a given message.

Suppose a sign reads:

Traffic Conditions
Next 2 Miles
Disabled Vehicle on I-77
Use I-77 Bypass Next Exit

Drivers not familiar with such a sign ("worst case," but able to read the sign) could take at least 8 seconds and according to the Dudek formula above up to 12 seconds to process this information and begin to respond. In Dudek's 1990 study, 85 percent of drivers familiar with similar signs read this 13-word message (excluding prepositions) with 6 message units in 6.7 seconds. The formulas in the literature properly tend to be conservative.

3.5 Response to Other Vehicle Dynamics

Vehicles in a traffic stream are discrete elements with motion characteristics loosely coupled with each other via the driver's processing of information and making control inputs. Effects of changes in speed or acceleration of other elements as perceived and acted on by the driver of any given element are of interest. Two situations appear relevant: (1) the vehicle ahead and (2) the vehicle alongside (in the periphery).

3.5.1 The Vehicle Ahead

Consideration of the vehicle ahead has its basis in thresholds for detection of radial motion (Schiff 1980). Radial motion is change in the apparent size of a target. The minimum condition for perceiving radial motion of an object (such as a vehicle ahead) is the symmetrical magnification of a form or texture in the field of view. Visual angle transitions from a near-linear to a geometric change in magnitude as an object approaches at constant velocity, as Figure 3.4 depicts for a motor vehicle approaching at a delta speed of 88 km/h. As the rate of change of visual angle becomes geometric, the perceptual system triggers a warning that an object is going to collide with the observer, or, conversely, that the object is pulling away from the observer. This phenomenon is called *looming*. If the rate of change of visual angle is irregular, that is information to the perceptual system that the object in motion is moving at a changing velocity (Schiff 1980). Sekuler and Blake (1990) report evidence that actual looming detectors exist in the human visual system. The relative change in visual angle is roughly equal to the reciprocal of "time-to-go" (time to impact), a special

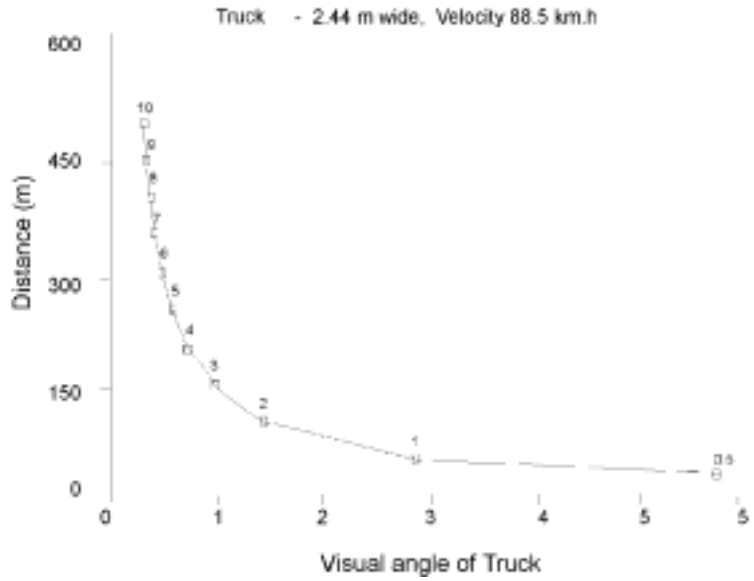


Figure 3.4
Looming as a Function of Distance from Object.

case of the well-known Weber fraction, $S = \Delta I/I$, the magnitude of a stimulus is directly related to a change in physical energy but inversely related to the initial level of energy in the stimulus.

Human visual perception of acceleration (as such) of an object in motion is very gross and inaccurate; it is very difficult for a driver to discriminate acceleration from constant velocity unless the object is observed for a relatively long period of time - 10 or 15 sec (Boff and Lincoln 1988).

The delta speed threshold for detection of oncoming collision or pull-away has been studied in collision-avoidance research. Mortimer (1988) estimates that drivers can detect a change in distance between the vehicle they are driving and the one in front when it has varied by approximately 12 percent. If a driver were following a car ahead at a distance of 30 m, at a change of 3.7 m the driver would become aware that distance is decreasing or increasing, i.e., a change in relative velocity. Mortimer notes that the major cue is rate of change in visual angle. This threshold was estimated in one study as 0.0035 radians/sec.

This would suggest that a change of distance of 12 percent in 5.6 seconds or less would trigger a perception of approach or pulling away. Mortimer concludes that "...unless the relative velocity between two vehicles becomes quite high, the drivers will respond to changes in their headway, or the change in angular size of the vehicle ahead, and use that as a cue to determine the speed that they should adopt when following another vehicle."

3.5.2 The Vehicle Alongside

Motion detection in peripheral vision is generally less acute than in foveal (straight-ahead) vision (Boff and Lincoln 1988), in that a greater relative velocity is necessary for a driver "looking out of the corner of his eye" to detect that speed change than if he or she is looking to the side at the subject vehicle in the next lane. On the other hand, peripheral vision is very blurred and motion is a much more salient cue than a stationary target is. A stationary object in the periphery (such as a neighboring vehicle exactly keeping pace with the driver's vehicle) tends to disappear for all intents and purposes *unless it moves with respect to the viewer* against a patterned background. Then that movement will be detected. Relative motion in the periphery also tends to look slower than the same movement as seen using fovea vision. Radial motion (car alongside swerving toward or away from the driver) detection presumably would follow the same pattern as the vehicle ahead case, but no study concerned with measuring this threshold directly was found.

3.6 Obstacle and Hazard Detection, Recognition, and Identification

Drivers on a highway can be confronted by a number of different situations which dictate either evasive maneuvers or stopping maneuvers. Perception-response time (PRT) to such encounters have already been discussed in Section 3.3.1. But before a maneuver can be initiated, the object or hazard must first be detected and then recognized as a hazard. The basic considerations are not greatly different than those discussed under driver responses to traffic control devices (Section 3.5), but some specific findings on roadway obstacles and hazards will also be discussed.

3.6.1 Obstacle and Hazard Detection

Picha (1992) conducted an object detection study in which representative obstacles or objects that might be found on a roadway were unexpectedly encountered by drivers on a closed course. Six objects, a 1 x 4 board, a black toy dog, a white toy dog, a tire tread, a tree limb with leaves, and a hay bale were placed in the driver's way. Both detection and recognition distances were recorded. Average visual angles of detection for these various objects varied from the black dog at 1.8 minimum of arc to 4.9 min of arc for the tree limb. Table 3.8 summarizes the detection findings of this study.

At the 95 percent level of confidence, it can be said from these findings that an object subtending a little less than 5 minutes of arc will be detected by all but 1 percent of drivers under daylight conditions provided they are looking in the object's direction. Since visual acuity declines by as much as two Snellen lines after nightfall, to be dtected such targets with similar contrast would

Table 3.8
Object Detection Visual Angles (Daytime)
(Minutes of Arc)

Object	Mean	Tolerance, 95th confidence		
		STD	95th	99th
1" x 4" Board, 24" x 1"*	2.47	1.21	5.22	6.26
Black toy dog, 6" x 6"	1.81	0.37	2.61	2.91
White toy dog, 6" x 6"	2.13	0.87	4.10	4.84
Tire tread, 8" x 18"	2.15	0.38	2.95	3.26
Tree Branch, 18" x 12"	4.91	1.27	7.63	8.67
Hay bale, 48" x 18"	4.50	1.28	7.22	8.26
All Targets	3.10	0.57	4.30	4.76
frontal viewing plan dimensions				

have to subtend somewhere around 2.5 times the visual angle that they would at detection under daylight conditions.

3.6.2 Obstacle and Hazard Recognition and Identification

Once the driver has detected an object in his or her path, the next job is to: (1) decide if the object, whatever it is, is a potential hazard, this is the recognition stage, followed by (2) the identification stage, even closer, at which a driver actually can tell what the object is. If an object (assume it is stationary) is small enough to pass under the vehicle and between the wheels, it doesn't matter very much what it is. So the first estimate is primarily of size of the object. If the decision is made that the object is too large to pass under the vehicle, then either evasive action or a braking maneuver must be decided upon. Objects 15 cm or less in height very seldom are causal factors in accidents (Kroemer et al. 1994).

The majority of objects encountered on the highway that constitute hazard and thus trigger avoidance maneuvers are larger than 60 cm in height. Where it may be of interest to establish a visual angle for an object to be discriminated as a hazard or non-hazard, such decisions require visual angles on the order of at least the visual angles identified in Section 3.4.2 for letter or symbol recognition, i.e., about 15 minutes of arc to take in 99 percent of the driver population. It would be useful to reflect that the full moon subtends 30 minutes of arc, to give the reader an intuitive feel for what the minimum visual angle might be for object recognition. At a distance somewhat greater than this, the driver decides if an object is sizable enough to constitute a hazard, largely based upon roadway lane width size comparisons and the size of the object with respect to other familiar roadside objects (such as mailboxes, bridge rails). Such judgements improve if the object is identified.

3.7 Individual Differences in Driver Performance

In psychological circles, variability among people, especially that associated with variables such as gender, age, socio-economic levels, education, state of health, ethnicity, etc., goes by the name "individual differences." Only a few such variables are of interest to traffic flow modeling. These are the variables which directly affect the path and velocity the driven vehicle follows in a given time in the operational environment. Other driver characteristics which may be of interest to the reader may be found in the *NHTSA Driver Performance Data* book (1987, 1994).

3.7.1 Gender

Kroemer, Kroemer, and Kroemer-Ebert (1994) summarize relevant gender differences as minimal to none. Fine finger dexterity and color perception are areas in which women perform better than men, but men have an advantage in speed. Reaction time tends to be slightly longer for women than for men the recent popular book and PBS series, *Brain Sex* (Moir and Jessel 1991) has some fascinating insights into why this might be so. This difference is statistically but not practically significant. For the purpose of traffic flow analysis, performance differences between men and women may be ignored.

3.7.2 Age

Research on the older driver has been increasing at an exponential rate, as was noted in the recent state-of-the-art summary by the Transportation Research Board (TRB 1988). Although a number of aspects of human performance related to driving change with the passage of years, such as response time, channel capacity and processing time needed for decision making, movement ranges and times, most of these are extremely variable, i.e., age is a poor predictor of performance. This was not so for visual perception. Although there are exceptions, for the most part visual performance becomes progressively poorer with age, a process which accelerates somewhere in the fifth decade of life.

Some of these changes are attributable to optical and physiological conditions in the aging eye, while others relate to changes in neural processing of the image formed on the retina. There are other cognitive changes which are also central to understanding performance differences as drivers age. Both visual and cognitive changes affecting driver performance will be discussed in the following paragraphs.

CHANGES IN VISUAL PERCEPTION

Loss of Visual Acuity (static) - Fifteen to 25 percent of the population 65 and older manifest visual acuities (Snellen) of less than 20/50 corrected, owing to senile macular degeneration (Marmor 1982). Peripheral vision is relatively unaffected, although a gradual narrowing of the visual field from 170 degrees to 140 degrees or less is attributable to anatomical changes (eyes become more sunk in the head). Static visual acuity among drivers is not highly associated with accident experience and is probably not a very significant factor in discerning path guidance devices and markings.

Light Losses and Scattering in Optic Train - There is some evidence (Ordy et al. 1982) that the scotopic (night) vision system ages faster than the photopic (daylight) system does. In addition, scatter and absorption by the stiff, yellowed, and possibly cateracted crystalline lens of the eye accounts for much less light hitting the degraded retina. The pupil also becomes stiffer with age, and dilates less for a given amount of light impingement (which considering that the mechanism of pupillary size is in part driven by the amount of light falling on the retina suggests actual physical atrophy of the pupil--senile myosis). There is also more matter in suspension in the vitreous humor of the aged eye than exists in the younger eye. The upshot is that only 30 percent of the light under daytime conditions that gets to the retina in a 20 year old gets to the retina of a 60 year old. This becomes *much worse at night* (as little as 1/16), and is exacerbated by the scattering effect of the optic train. Points of bright light are surrounded by halos that effectively obscure less bright objects in their near proximity. Blackwell and Blackwell (1971) estimated that, because of these changes, a given level of contrast of an object has to be increased by a factor of anywhere from 1.17 to 2.51 for a 70 year old person to see it, as compared to a 30 year old.

Glare Recovery - It is worth noting that a 55 year old person requires more than 8 times the period of time to recover from glare if dark adapted than a 16 year old does (Fox 1989). An older driver who does not use the strategy to look to the right and shield his or her macular vision from oncoming headlamp glare

is literally driving blind for many seconds after exposure. As described above, scatter in the optic train makes discerning *any* marking or traffic control device difficult to impossible. The slow re-adaptation to mesopic levels of lighting is well-documented.

Figure/Ground Discrimination - Perceptual style changes with age, and many older drivers miss important cues, especially under higher workloads (Fox 1989). This means drivers may miss a significant guideline or marker under unfamiliar driving conditions, because they fail to discriminate the object from its background, either during the day or at night.

CHANGES IN COGNITIVE PERFORMANCE

Information Filtering Mechanisms - Older drivers reportedly experience problems in ignoring irrelevant information and correctly identifying meaningful cues (McPherson et al. 1988). Drivers may not be able to discriminate actual delineation or signage from roadside advertising or faraway lights, for example. Work zone traffic control devices and markings that are meant to override the pre-work TCD's may be missed.

Forced Pacing under Highway Conditions - In tasks that require fine control, steadiness, and rapid decisions, forced paced tasks under stressful conditions may disrupt the performance of older drivers. They attempt to compensate for this by slowing down. Older people drive better when they can control their own pace (McPherson et al. 1988). To the traffic flow theorist, a sizable proportion of older drivers in a traffic stream may result in vehicles that lag behind and obstruct the flow.

Central vs. Peripheral Processes - Older driver safety problems relate to tasks that are heavily dependent on central processing. These tasks involve responses to traffic or to *roadway conditions* (emphasis added) (McPherson et al. 1988).

The Elderly Driver of the Past or Even of Today is Not the Older Driver of the Future - The cohort of drivers who will be 65 in the year 2000, which is less than five years from now, were born in the 1930's. Unlike the subjects of gerontology studies done just a few years ago featuring people who came of driving age in the 1920's or even before, when far fewer people had cars and traffic was sparse, the old of tomorrow started driving in the 1940's and after. They are and will be more affluent, better educated, in better health, resident in the same communities they lived in before becoming "older drivers," and they have driven under modern conditions and the urban environment since their teens. Most of them have had classes in driver education and defensive driving. They will likely continue driving on a routine basis until almost the end of their natural lives, which will be happening at an ever advancing age. The cognitive trends briefly discussed above are very variable in incidence and in their actual effect on driving performance. The future older driver may well exhibit much less decline in many of these performance areas in which central processes are dominant.

3.7.3 Driver Impairment

Drugs - Alcohol abuse in isolation and combination with other drugs, legal or otherwise, has a generally deleterious effect on performance (Hulbert 1988; Smiley 1974). Performance differences are in greater variability for any given driver, and in generally lengthened reaction times and cognitive processing times. Paradoxically, some drug combinations can improve such performance on certain individuals at certain times. The only drug incidence which is sufficiently large to merit consideration in traffic flow theory is alcohol.

Although incidence of alcohol involvement in *accidents* has been researched for many years, and has been found to be substantial, very little is known about incidence and levels of impairment in the driving population, other than it must also be substantial. Because these drivers are impaired, they are over-represented in accidents. Price (1988) cites estimates that 92 percent of the adult population of the U.S. use alcohol, and perhaps 11 percent of the total adult population (20-70 years of age) have alcohol abuse problems. Of the 11 percent who are problem drinkers, seven percent are men, four percent women. The incidence of problem drinking drops with age, as might be expected. Effects on performance as a function of blood alcohol concentration (BAC) are well-summarized in Price, but are too voluminous to be reproduced here. Price also summarizes effects of other drugs such as cocaine, marijuana, etc. Excellent sources for more information on alcohol and driving can be found in a Transportation Research Board Special Report (216).

Medical Conditions - Disabled people who drive represent a small but growing portion of the population as technology advances in the field of adaptive equipment. Performance

studies and insurance claim experience over the years (Koppa et al. 1980) suggest tha t such driver's performance is indistinguishable from the general driving population. Although there are doubtless a number of people on the highways with illnesses or conditions for which driving is contraindicated, they are probably not enough of these to account for them in any traffic flow models.

3.8 Continuous Driver Performance

The previous sections of this chapter have sketched the relevant discrete performance characteristics of the driver in a traffic stream. Driving, however, is primarily a continuous dynamic process of managing the present heading and thus the future path of the vehicle through the steering function. The first and second derivatives of location on the roadway in time, velocity and acceleration, are also continuous control processes through modulated input using the accelerator (really the throttle) and the brake controls.

3.8.1 Steering Performance

The driver is tightly coupled into the steering subsystem of the human-machine system we call the motor vehicle. It was only during the years of World War II that researchers and engineers first began to model the human operator in a tracking situation by means of differential equations, i.e., a transfer function. The first paper on record to explore the human transfer function was by Tustin in 1944 (Garner 1967), and the subject was antiaircraft gun control. The human operator in such tracking situations can be described in the same terms as if he or she is a linear feedback control system, even though the human operator is noisy, non-linear, and sometimes not even closed-loop.

3.8.1.1 Human Transfer Function for Steering

Steering can be classified as a special case of the general pursuit tracking model, in which the two inputs to the driver (which are somehow combined to produce the correction signal) are (1) the desired path as perceived by the driver from cues provided by the roadway features, the streaming of the visual field, and higher order information; and (2) the perceived present course of the vehicle as inferred from relationship of the hood to roadway features. The exact form of either of these two inputs are still subjects of investigation and some uncertainty, even though nothing can be more commonplace than steering a motor vehicle. Figure 3.5 illustrates the conceptual model first proposed by Sheridan (1962). The human operator looks at both inputs, *R(t)* the desired input forcing function (the road and where it seems to be taking the driver), and *E(t)* the system error function, the difference between where the road is going and what *C(t)* the vehicle seems to be doing. The human operator can look ahead (*lead*), the prediction function, and also can correct for perceived errors in the path. If the driver were trying to drive by viewing the road through a hole in the floor of the car, then the prediction function would be lost, which is usually the state of affairs for servomechanisms. The two human functions of prediction and compensation are combined to make a control input to the vehicle via the steering wheel which (for power steering) is also a servo in its own right. The control output from this human-steering process combination is fed back (by the perception of the path of the vehicle) to close the loop. Mathematically, the setup in Figure 3.5 is expressed as follows, if the operator is reasonably linear:

$$g_{(s)} = \frac{Ke^{-ts}(1 + T_L s)}{(1 + T_L s)(1 + T_N s)} + R \qquad (3.12)$$

Sheridan (1962) reported some parameters for this Laplace transform transfer function of the first order. *K*, the gain or sensitivity term, varies (at least) between +35 *db* to -12 *db*. Gain, how much response the human will make to a given input, is the parameter perhaps most easily varied, and tends to settle at some point comfortably short of instability (a phase margin of 60 degrees or more). The exponential term e^{-ts} ranges from 0.12 to 0.3 sec and is best interpreted as reaction time. This delay is the dominant limit to the human's ability to adapt to fast-changing conditions. The *T* factors are all time constants, which however may not stay constant at all. They must usually be empirically derived for a given control situation.

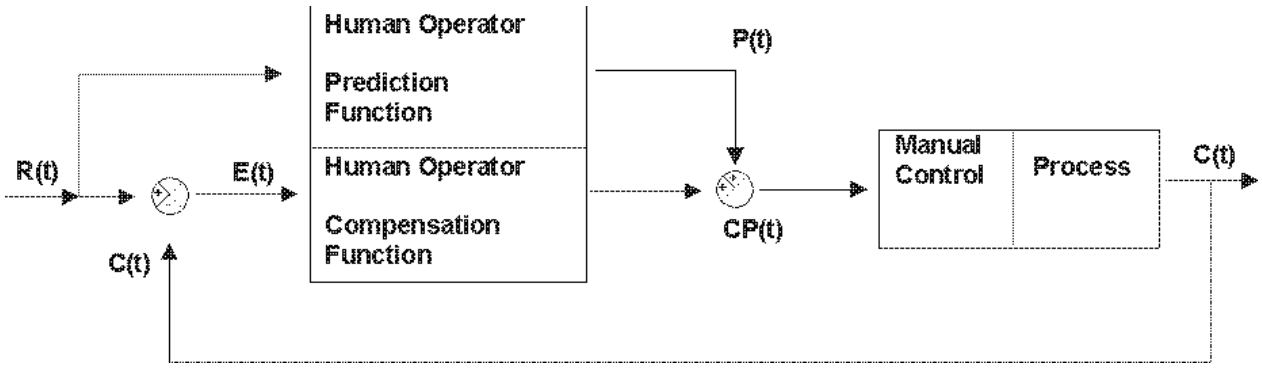


Figure 3.5
Pursuit Tracking Configuration (after Sheridan 1962).

Sheridan reported some experimental results which show T_L (lead) varying between 0 and 2, T_l (lag) from 0.0005 to 25, and T_N (neuromuscular lag) from 0 to 0.67. R, the remnant term, is usually introduced to make up for nonlinearities between input and output. Its value is whatever it takes to make output track input in a predictable manner. In various forms, and sometimes with different and more parameters, Equation 3.10 expresses the basic approach to modeling the driver's steering behavior. Novice drivers tend to behave primarily in the *compensatory* tracking mode, in which they primarily attend to the difference, say, between the center of the hood and the edge line of the pavement, and attempt to keep that difference at some constant visual angle. As they become more expert, they move more to pursuit tracking as described above. There is also evidence that there are "precognitive" open-loop steering commands to particular situations such as swinging into an accustomed parking place in a vehicle the driver is familiar with. McRuer and Klein (1975) classify maneuvers of interest to traffic flow modelers as is shown in Table 3.9.

In Table 3.9, the entries under driver control mode denote the order in which the three kinds of tracking transition from one to the other as the maneuver transpires. For example, for a turning movement, the driver follows the dotted lines in an intersection and aims for the appropriate lane in the crossroad in a pursuit mode, but then makes adjustments for lane position during the latter portion of the maneuver in a compensatory mode. In an emergency, the driver "jerks" the wheel in a precognitive (open-loop) response, and then straightens out the vehicle in the new lane using compensatory tracking.

Table 3.9
Maneuver Classification

Maneuver	Driver Control Mode		
	Compen-satory	Pursuit	Precognitive
Highway Lane Regulation	1		
Precision Course Control	2	1	
Turning; Ramp Entry/Exit	2	1	
Lane Change	2		1
Overtake/Pass	2	1	
Evasive Lane Change	2		1

3.8.1.2 Performance Characteristics Based on Models

The amplitude of the output from this transfer function has been found to rapidly approach zero as the frequency of the forcing function becomes greater than 0.5 *Hz* (Knight 1987). The driver makes smaller and smaller corrections as the highway or wind gusts or other inputs start coming in more frequently than one complete cycle every two seconds.

The time lag between input and output also increases with frequency. Lags approach 100 msec at an input of 0.5 Hz and increase almost twofold to 180 msec at frequencies of 2.4 Hz. The human tracking bandwidth is of the order of 1 to 2 Hz. Drivers can go to a precognitive rhythm for steering input to better this performance, if the input is very predictable, e.g., a "slalom" course. Basic lane maintenance under very restrictive conditions (driver was instructed to keep the left wheels of a vehicle on a painted line rather than just lane keep) was studied very recently by Dulas (1994) as part of his investigation of changes in driving performance associated with in-vehicle displays and input tasks. Speed was 57 km/h. Dulas found average deviations of 15 cm, with a standard deviation of 3.2 cm. Using a tolerance estimation based on the nearly 1000 observations of deviation, the 95th percentile deviation would be 21 cm, the 99th would be 23 cm. Thus drivers can be expected to weave back and forth in a lane in straightaway driving in an envelope of +/- 23 cm or 46 cm across. Steering accuracy with degrade and oscillation will be considerably more in curves. since such driving is mixed mode, with rather large errors at the beginning of the maneuver, with compensatory corrections toward the end of the maneuver. Godthelp (1986) described this process as follows. The driver starts the maneuver with a lead term before the curve actually begins. This precognitive control action finishes shortly after the curve is entered.

Then a stage of steady-state curve driving follows, with the driver now making compensatory steering corrections. The steering wheel is then restored to straight-ahead in a period that covers the endpoint of the curve. Road curvature (perceived) and vehicle speed predetermines what the initial steering input will be, in the following relationship:

$$g_s = \frac{SRl\ (1+F_s u^2)C_r}{1000} \qquad (3.13)$$

where,

C_r = roadway curvature
SR = steering ratio
F_s = stability factor
l = wheelbase
u = speed
g_s = steering wheel angle (radians)

Godthelp found that the standard deviation of anticipatory steering inputs is about 9 percent of steering wheel angle g_s. Since sharper curves require more steering wheel input, inaccuracies will be proportionately greater, and will also induce more oscillation from side to side in the curve during the compensatory phase of the maneuver.

3.9 Braking Performance

The steering performance of the driver is integrated with either braking or accelerator positioning in primary control input. Human performance aspects of braking as a continuous control input will be discussed in this section. After the perception-response time lag has elapsed, the actual process of applying the brakes to slow or stop the motor vehicle begins.

3.9.1 Open-Loop Braking Performance

The simplest type of braking performance is "jamming on the brakes." The driver exerts as much force as he or she can muster, and thus approximates an instantaneous step input to the motor vehicle. Response of the vehicle to such an input is out of scope for this chapter, but it can be remarked that it can result in one or more wheels locking and consequent loss of control at speeds higher than 32 km/h, unless the vehicle is equipped with antiskid brakes (ABS, or Antilock Brake System). Such a model of human braking performance is assumed in the time-hallowed AASHTO braking distance formula (AASHTO 1990):

$$d = \frac{V^2}{257.9f} \qquad (3.14)$$

where,

d = braking distance - meters
V = Initial speed - km/h
f = Coefficient of friction, tires to pavement surface, approximately equal to deceleration in g units

Figure 3.6 shows what such a braking control input really looks like in terms of the deceleration profile. This maneuver was on a dry tangent section at 64 km/h, under "unplanned" conditions (the driver does not know when or if the signal to brake will be given) with a full-size passenger vehicle not equipped with ABS. Note the typical steep rise in deceleration to a peak of over 0.9 g, then steady state at approximately 0.7 g for a brakes locked stop. The distance data is also on this plot: the braking distance on this run was 23 m feet. Note also that the suspension bounce produces a characteristic oscillation after the point at which the vehicle is completely stopped, just a little less than five seconds into the run.

Figure 3.7 shows a braking maneuver on a tangent with the same driver and vehicle, this time on a wet surface. Note the characteristic "lockup" footprint, with a steady-state deceleration after lockup of 0.4 g.

From the standpoint of modeling driver input to the vehicle, the open-loop approximation is a step input to maximum braking effort, with the driver exhibiting a simple to complex PRT delay prior to the step. A similar delay term would be introduced prior to release of the brake pedal, thus braking under stop-and-go conditions would be a sawtooth.

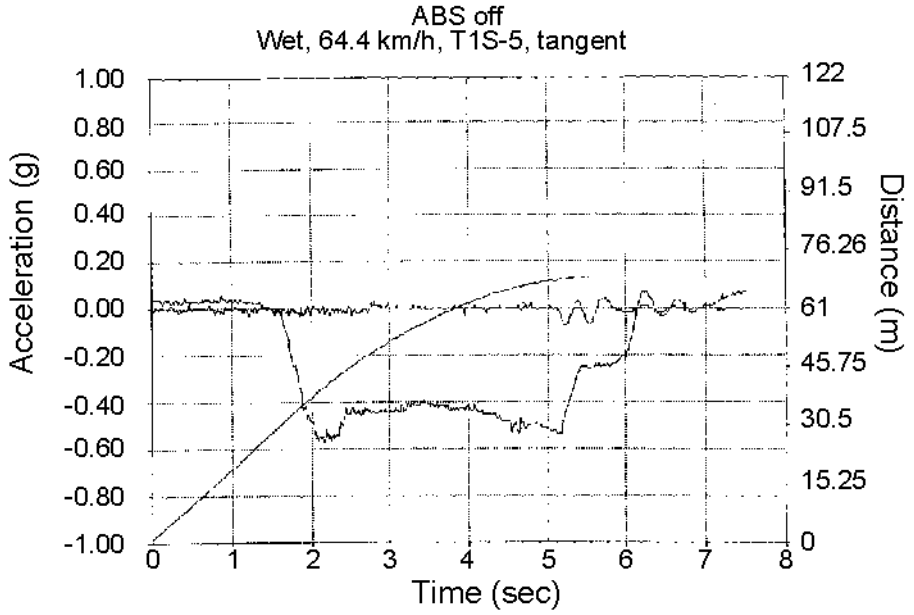


Figure 3.6
Typical Deceleration Profile for a Driver without Antiskid Braking System on a Dry Surface.

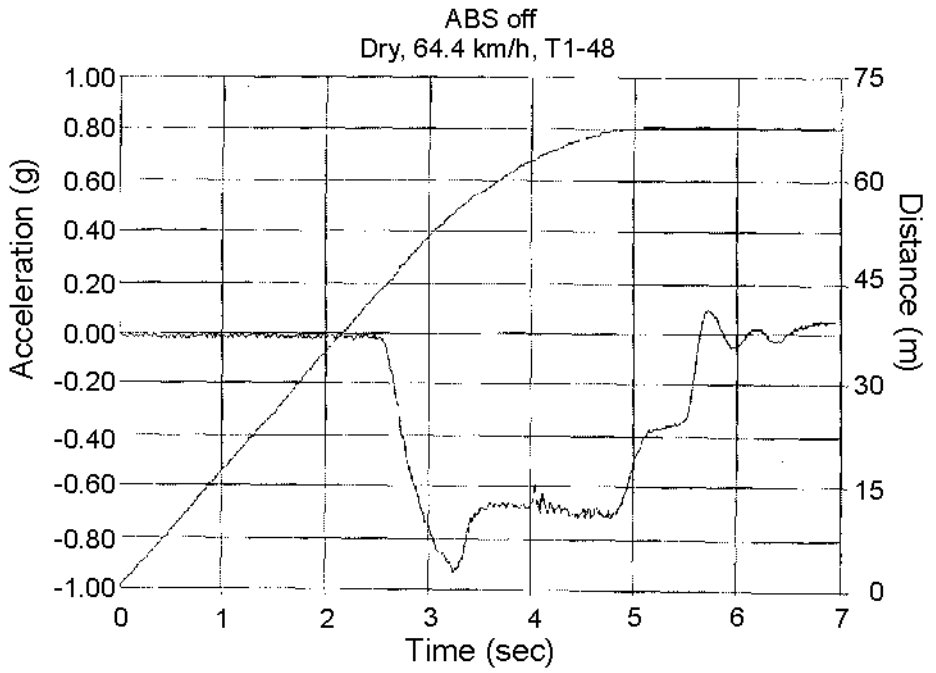


Figure 3.7
Typical Deceleration Profile for a Driver without Antiskid Braking System on a Wet Surface.

3.9.2 Closed-Loop Braking Performance

Recent research in which the writer has been involved provide some controlled braking performance data of direct application to performance modeling (Fambro et al. 1994). "Steady state" approximations or fits to these data show wide variations among drivers, ranging from -0.46 g to -0.70 g.

Table 3.10 provides some steady-state derivations from empirical data collected by Fambro et al. (1994). These were all responses to an unexpected obstacle or object encountered on a closed course, in the driver's own (but instrumented) car.

Table 3.11 provides the same derivations from data collected on drivers in their own vehicle in which the braking maneuver was anticipated; the driver knows that he or she would be braking, but during the run were unsure when the signal (a red light inside the car) would come.

The ratio of unexpected to expected closed-loop braking effort was estimated by Fambro et al. to be about 1.22 under the same pavement conditions. Pavement friction (short of ice) played very little part in driver's setting of these effort levels. About 0.05 to 0.10 g difference between wet pavement and dry pavement steady-state *g* was found.

Table 3.10
Percentile Estimates of Steady State Unexpected Deceleration

Mean	-0.55g
Standard Deviation	0.07
75th Percentile	-0.43
90th	-0.37
95th	-0.32
99th	-0.24

Table 3.11
Percentile Estimates of Steady-State Expected Deceleration

Mean	-0.45g
Standard Deviation	0.09
75th Percentile	-0.36
90th	-0.31
95th	-0.27
99th	-0.21

3.9.3 Less-Than-Maximum Braking Performance

The flow theorist may require an estimate of "comfortable" braking performance, in which the driver makes a stop for intersections or traffic control devices which are discerned considerably in advance of the location at which the vehicle is to come to rest. Driver input to such a planned braking situation approximates a "ramp" (straight line increasing with time from zero) function with the slope determined by the distance to the desired stop location or steady-state speed in the case of a platoon being overtaken. The driver squeezes on pedal pressure to the brakes until a desired deceleration is obtained. The maximum "comfortable" braking deceleration is generally accepted to be in the neighborhood of -0.30 g, or around 3 m/sec^2 (ITE 1992).

The AASHTO Green Book (AASHTO 1990) provides a graphic for speed changes in vehicles, in response to approaching an intersection. When a linear computation of decelerations from this graphic is made, these data suggest decelerations in the neighborhood of -2 to -2.6 m/sec^2 or -0.20 to -0.27 g. More recent research by Chang et al. (1985) found values in response to traffic signals approaching -0.39 g, and Wortman and Matthias (1983) observed a range of -0.22 to -0.43 g, with a mean level of -0.36 g. Hence controlled braking performance that yields a *g* force of about -0.2 g would be a reasonable lower level for a modeler, i.e., almost any driver could be expected to change the velocity of a passenger car by at least that amount, but a more average or "typical" level would be around -0.35 g.

3.10 Speed and Acceleration Performance

The third component to the primary control input of the driver to the vehicle is that of manual (as opposed to cruise control) control of vehicle velocity and changes in velocity by means of the accelerator or other device to control engine RPM.

3.10.1 Steady-State Traffic Speed Control

The driver's primary task under steady-state traffic conditions is to perform a tracking task with the speedometer as the display, and the accelerator position as the control input. Driver response to the error between the present indicated speed and the desired speed (the control signal) is to change the pedal position in the direction opposite to the trend in the error indication. How much of such an error must be present depends upon a host of factors: workload, relationship of desired speed to posted speed, location and design of the speedometer, and personal considerations affecting the performance of the driver at the moment, etc. Drivers in heavy traffic use relative perceived position with respect to other vehicles in the stream as a primary tracking cue (Triggs, 1988). A recent study (Godthelp and Shumann 1994) found errors between speed desired and maintained to vary from -0.3 to -0.8 m/sec in a lane change maneuver; drivers tended to lose velocity when they made such a maneuver. Under steady-stage conditions in a traffic stream, the range of speed error might be estimated to be no more than +/- 1.5 m/sec (Evans and Rothery 1973), basically modeled by a sinusoid. The growing prevalence of cruise controls undoubtedly will reduce the amplitude of this speed error pattern in a traffic stream by half or more.

3.10.2 Acceleration Control

The performance characteristics of the vehicle driver are the limiting constraints on how fast the driver can accelerate the

vehicle. The actual acceleration rates, particularly in a traffic stream as opposed to a standing start, are typically much lower than the performance capabilities of the vehicle, particularly a passenger car. A nominal range for "comfortable" acceleration at speeds of 48 km/h and above is 0.6 m/sec^2 to 0.7 m/sec^2 (AASHTO 1990). Another source places the nominal acceleration rate drivers tend to use under "unhurried" circumstances at approximately 65 percent of maximum acceleration for the vehicle, somewhere around 1 m/sec^2 (ITE 1992). If the driver removes his or her foot from the accelerator pedal (or equivalent control input) drag and rolling resistance produce deceleration at about the same level as "unhurried" acceleration, approximately 1 m/sec^2 at speeds of 100 km/h or higher. In contrast to operation of a passenger car or light truck, heavy truck driving is much more limited by the performance capabilities of the vehicle. The best source for such information is the *Traffic Engineering Handbook* (ITE 1992).

3.11 Specific Maneuvers at the Guidance Level

The discussion above has briefly outlined most of the more fundamental aspects of driver performance relevant to modeling the individual driver-vehicle human-machine system in a traffic stream. A few additional topics will now be offered to further refine this picture of the driver as an active controller at the guidance level of operation in traffic.

3.11.1 Overtaking and Passing in the Traffic Stream

3.11.1.1 Overtaking and Passing Vehicles (4-Lane or 1-Way)

Drivers overtake and pass at accelerations in the sub-maximal range in most situations. Acceleration to pass another vehicle (passenger cars) is about 1 m/sec^2 at highway speeds (ITE 1992). The same source provides an approximate equation for acceleration on a grade:

$$a_{GV} \doteq a_{LV} - \frac{G_g}{100} \qquad (3.15)$$

where,

a_{GV} = max acceleration rate on grade
a_{LV} = max acceleration rate on level
G = gradient (5/8)
g = acceleration of gravity (9.8 m/sec^2)

The maximum acceleration capabilities of passenger vehicles range from almost 3 m/sec^2 from standing to less than 2 m/sec^2 from 0 to highway speed.

Acceleration is still less when the maneuver begins at higher speeds, as low as 1 m/sec^2 on some small subcompacts. In Equation 3.15, overtaking acceleration should be taken as 65 percent of maximum (ITE 1992). Large trucks or tractor-trailer combinations have maximum acceleration capabilities on a level roadway of no more than 0.4 m/sec^2 at a standing start, and decrease to 0.1 m/sec^2 at speeds of 100 km/h. Truck drivers "floorboard" in passing maneuvers under these circumstances, and maximum vehicle performance is also typical driver input.

3.11.1.2 Overtaking and Passing Vehicles (Opposing Traffic)

The current AASHTO *Policy on Geometric Design* (AASHTO 1990) provides for an acceleration rate of 0.63 m/sec^2 for an initial 56 km/h, 0.64 m/sec^2 for 70 km/h, and 0.66 m/sec^2 for speeds of 100 km/h. Based upon the above considerations, these design guidelines appear very conservative, and the theorist may wish to use the higher numbers in Section 3.11.1.1 in a sensitivity analysis.

3.12 Gap Acceptance and Merging

3.12.1 Gap Acceptance

The driver entering or crossing a traffic stream must evaluate the space between a potentially conflicting vehicle and himself or herself and make a decision whether to cross or enter or not. The time between the arrival of successive vehicles at a point is the time gap, and the critical time gap is the least amount of successive vehicle arrival time in which a driver will attempt a merging or crossing maneuver. There are five different gap acceptance situations. These are:

(1) Left turn across opposing traffic, no traffic control
(2) Left turn across opposing traffic, with traffic control (permissive green)
(3) Left turn onto two-way facility from stop or yield controlled intersection
(4) Crossing two-way facility from stop or yield controlled intersection
(5) Turning right onto two-way facility from stop or yield controlled intersection

Table 3.12 provides very recent design data on these situations from the Highway Capacity Manual (TRB 1985). The range of gap times under the various scenarios presented in Table 3.12 is from a minimum of 4 sec to 8.5 sec. In a stream traveling at 50 km/h (14 m/sec) the gap distance thus ranges from 56 to 119 m; at 90 km/h (25 m/sec) the corresponding distances are 100 to 213 m.

3.12.2 Merging

In merging into traffic on an acceleration ramp on a freeway or similar facility, the Situation (5) data for a four lane facility at 90 km/h with a one second allowance for the ramp provides a baseline estimate of gap acceptance: 4.5 seconds. Theoretically as short a gap as three car lengths (14 meters) can be accepted if vehicles are at or about the same speed, as they would be in merging from one lane to another. This is the minimum, however, and at least twice that gap length should be used as a nominal value for such lane merging maneuvers.

3.13 Stopping Sight Distance

The minimum sight distance on a roadway should be sufficient to enable a vehicle traveling at or near the design speed to stop before reaching a "stationary object" in its path, according to the AASHTO *Policy on Geometric Design* (AASHTO 1990). It goes on to say that sight distance should be at least that required for a "below-average" driver or vehicle to stop in this distance.

Previous sections in this chapter on perception-response time (Section 3.3.1) and braking performance (Section 3.2) provide the raw materials for estimating stopping sight distance. The time-honored estimates used in the AASHTO Green Book (AASHTO 1990) and therefore many other engineering resources give a flat 2.5 sec for PRT, and then the linear deceleration equation (Equation 3.13) as an additive model. This approach generates standard tables that are used to estimate stopping sight distance (SSD) as a function of coefficients of friction and initial speed at inception of the maneuver. The empirically derived estimates now available in Fambro et al. (1994) for both these parts of the SSD equation are expressed in percentile levels of drivers who could be expected to (1) respond and (2) brake in the respective distance or shorter. Since PRT and braking distance that a driver may achieve in a given vehicle are not highly correlated, i.e., drivers that may be very fast to initiate braking may be very conservative in the actual braking maneuver, or may be strong brakers. PRT does not predict braking performance, in other words.

Very often, the engineer will use "worst case" considerations in a design analysis situation. What is the "reasonable" worst case for achieving the AASHTO "below average" driver and vehicle? Clearly, a 99th percentile PRT and a 99th percentile braking distance gives an overly conservative 99.99 combined percentile

Table 3.12
Critical Gap Values for Unsignalized Intersections

Maneuver	Control	Average Speed of Traffic			
		50 km/h		90 km/h	
		Number of Traffic Lanes, Major Roadway			
		2	4	2	4
1	None	5.0	5.5	5.5	6.0
2	Permissive Green[1]	5.0	5.5	5.5	6.0
3	Stop	6.5	7.0	8.0	8.5
3	Yield	6.0	6.5	7.0	7.5
4	Stop	6.0	6.5	7.5	8.0
4	Yield	5.5	6.0	6.5	7.0
5[2]	Stop	5.5	5.5	6.5	6.5
5[2]	Yield	5.0	5.0	5.5	5.5

[1]During Green Interval
[2] If curve radius >15 m or turn angle <60° subtract 0.5 seconds.
If acceleration lane provided, subtract 1.0 seconds.

All times are given in seconds.

All Maneuvers: if population >250,000, subtract 0.5 seconds
if restricted sight distance, add 1.0 seconds
Maximum subtraction is 1.0 seconds
Maximum critical gap ≤ 8.5 seconds

level--*everybody* will have an SSD equal to or shorter than this somewhat absurd combination. The combination of 90th percentile level of performance for each segment yields a combined percentile estimate of 99 percent (i.e., 0.10 of each distribution is outside the envelope, and their product is 0.01, therefore 1.00 - 0.01 = 0.99). A realistic worst case (99th percentile) combination to give SSD would, from previous sections of this chapter be:

PRT: 1.57 sec (Table 3.4)

Braking deceleration: -0.37 g (Section 3.10.2)

For example, on a dry level roadway, using Equation 3.12, at a velocity of 88 km/h, the SSD components would be:

PRT: 1.57 x 24.44 = 38.4 m

Braking Distance: 82.6 m

SSD: 38 + 83 = 121 m

For comparison, the standard AASHTO SSD for a dry level roadway, using a nominal 0.65 for *f*, the coefficient of friction, would be:

PRT:	2.50 x 24.44 = 61.1 m
Braking Distance:	47.3 m
SSD:	61 + 47 = 108 m

These two estimates are comparable, but the first estimate has an empirical basis for it. The analyst can assume other combinations of percentiles (for example, 75th percentile performance in combination yields an estimate of the 94th percentile). It is always possible, of course, to assume different levels of percentile representation for a hypothetical driver, e.g., 50th percentile PRT with 95th percentile braking performance.

3.14 Intersection Sight Distance

The AASHTO *Policy on Geometric Design* (AASHTO 1990) identifies four differences cases for intersection sight distance considerations. From the viewpoint of traffic flow theory, the question may be posed, "How long is a driver going to linger at an intersection before he or she begins to move?" Only the first three cases will be discussed here, since signalized intersections (Case IV) have been discussed in Section 3.5.1.

3.14.1 Case I: No Traffic Control

The driver initiates either acceleration or deceleration based upon his or her perceived gap in intersecting traffic flow. The principles given in Section 3.13 apply here. PRT for this situation should be the same as for conditions of no surprise outlined in Section 3.3.1. AASHTO (1990) gives an allowance of three seconds for PRT, which appears to be very conservative under these circumstances.

3.14.2 Case II: Yield Control for Secondary Roadway

This is a complex situation. McGee et al. (1983) could not find reliable data to estimate the PRT. A later, follow-up study by Hostetter et al. (1986) considered the PRT to stretch from the time that the YIELD sign first could be recognized as such to the time that the driver either began a deceleration maneuver or speeded up to clear the intersection in advance of cross traffic. But decelerations often started 300 m or more from the intersection, a clear response to the sign and not to the traffic ahead. PRTs were thus in the range of 20 to over 30 sec. with much variability and reflect driving style rather than psychophysical performance.

3.14.3 Case III: Stop Control on Secondary Roadway

Hostetter et al. (1986) note that "for a large percentage of trials at intersections with reasonable sign distance triangles, drivers completed monitoring of the crossing roadway before coming to a stop." Their solution to this dilemma was to include three measures of PRT. They start at different points but terminate with the initiation of an accelerator input. One of the PRT's starts with the vehicle at rest. The second begins with the first head movement at the stop. The third begins with the last head movement in the opposite direction of the intended turn or toward the shorter sight distance leg (for a through maneuver). None of the three takes into account any processing the driver might be doing prior to the stop at the intersection.

Their findings were as follows (Table 3.13):

Table 3.13
PRTs at Intersections

	4-way		T-Intersection	
	Mean	85th	Mean	85th
PRT 1	2.2 sec	2.7	2.8	3.1
PRT 2	1.8	2.6	1.9	2.8
PRT 3	1.6	2.5	1.8	2.5

Thus a conservative estimate of PRT, i.e., time lag at an intersection before initiating a maneuver, would be somewhat in excess of three seconds for most drivers.

3.15 Other Driver Performance Characteristics

3.15.1 Speed Limit Changes

Gstalter and Hoyos (1988) point out the well-known phenomenon that drivers tend to adapt to sustained speed over a period of time, such that the perceived velocity (not looking at the speedometer) lessens. In one study cited by these authors, drivers drove 32 km at speeds of 112 km/h. Subjects were then instructed to drop to 64 km/h. The average speed error turned out to be more than 20 percent *higher* than the requested speed. A similar effect undoubtedly occurs when posted speed limits change along a corridor. In the studies cited, drivers were aware of the "speed creep" and attempted (they said) to accommodate for it. When drivers go from a lower speed to a higher one, they also adapt, such that the higher speed seems higher than in fact it is, hence errors of 10 to 20 percent *slower* than commanded speed occur. It takes several minutes to re-adapt. Hence speed adjustments on a corridor should not be modeled by a simple step function, but rather resemble an over damped first-order response with a time constant of two minutes or more.

3.15.2 Distractors On/Near Roadway

One of the problems that militates against smooth traffic flow on congested facilities is the "rubber neck" problem. Drivers passing by accident scenes, unusual businesses or activities on the road side, construction or maintenance work, or other occurrences irrelevant to the driving task tend to shift sufficient attention to degrade their driving performance. In *Positive Guidance* terms (Lunenfeld and Alexander 1990) such a situation on or near the roadway is a temporary violation of expectancy. How to model the driver response to such distractions? In the absence of specific driver performance data on distractors, the individual driver response could be estimated by injecting a sudden accelerator release with consequent deceleration from speed discussed in Section 3.11.2. This response begins as the distraction comes within a cone of 30 degrees centered around the straight-ahead direction on a tangent, and the outer delineation of the curve on a horizontal curve. A possible increase in the amplitude of lane excursions could also occur, similar to the task-loaded condition in the Dulas (1994) study discussed in Section 3.9.1.2.

3.15.3 Real-Time Driver Information Input

With the advent of Intelligent Transportation Systems (ITS), driver performance changes associated with increased information processing work load becomes a real possibility. ITS may place message screens, collision avoidance displays and much more in the vehicle of the future. Preliminary studies of the effects of using such technology in a traffic stream are just now appearing in the literature. There is clearly much more to come. For a review of some of the human factors implications of ITS, see Hancock and Pansuraman (1992) and any recent publications by Peter Hancock of The University of Minnesota.

Drivers as human beings have a very finite attentional resource capacity, as summarized in Dulas (1994). Resources can be allocated to additional information processing tasks only at the cost of decreasing the efficiency and accuracy of those tasks. When the competing tasks use the same sensory modality and similar resources in the brain, increases in errors becomes dramatic. To the extent that driving is primarily a psychomotor task at the skill-based level of behavior, it is relatively immune to higher-level information processing, if visual perception is not a dominant factor.

But as task complexity increases, say, under highly congested urban freeway conditions, any additional task becomes very disruptive of performance. This is especially true of older drivers. Even the use of cellular telephones in traffic has been found to be a potent disrupter of driving performance (McKnight and McKnight 1993). A study described by Dulas (1994) found that drivers using a touch screen CRT at speeds of 64 km tended to increase lane deviations such that the probability of lane excursion was 0.15. Early and very preliminary studies indicate that close attention to established human engineering principles for information display selection and design should result in real-time information systems that do not adversely affect driver performance.

References

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

Ang, A. H-S., and W. H. Tang (1975). *Probability Concepts in Engineering Planning and Design Vol. 1: Basic Principles.* New York: John Wiley and Sons.

Berman, A. H. (1994). *A Study of Factors Affecting Foot Movement Time in a Braking Maneuver.* Unpublished MS Thesis, Texas A&M University.

Blackwell, O. and H. Blackwell (1971). *Visual Performance Data for 156 Normal Observers of Various Ages.* Journal of Illumination Engineering Society 1, pp. 3-13.

Boff, K. R. and J. E. Lincoln (1988). *Engineering Data Compendium*. USAF, H. G. Armstrong Medical Research Laboratory, Wright-Patterson AFB, Ohio, (Section 8.1).

Brackett, R. Q. and R. J. Koppa (1988). *Preliminary Study of Brake Pedal Location Accuracy.* Proceedings of the Human Factors and Ergonomics Society 32nd Meeting.

Chang, M-S, C. J. Messer, and A. J. Santiago (1985). *Timing Traffic Signal Change Intervals Based on Driver Behavior.* Transportation Research Record 1027, Transportation Research Board, National Research Council, Washington, DC.

Drury, C. G. (1975). *Application of Fitt's Law to Foot Pedal Design.* Human Factors 12, pp. 559-561.

Dudek, C. L. (1990). *Guidelines on the Use of Changeable Message Signs.* FHWA-TS-90-043, Federal Highway Administration, Washington, DC.

Dulas, R. J. (1994). *Lane Deviation as a Result of Visual Workload and Manual Data Entry.* Unpublished doctoral dissertation, Texas A&M University.

Evans, L. and R. Rothery (1973). *Experimental Measurements of Perceptual Thresholds in Car Following.* Highway Research Record 454, Transportation Research Board, Washington, DC.

Fambro, D. et al. (1994). *Determination of Factors Affecting Stopping Sight Distance.* Working Paper I: Braking Studies (DRAFT) National Cooperative Highway Research Program, Project 3-42.

Federal Highway Administration (1983). *Traffic Control Device Handbook.*Washington, DC.

Fitts, P. M. (1954). *The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement.* Journal of Experimental Psychology, 47, pp. 381-391.

Fox, M. (1989). *Elderly Driver's Perceptions of Their Driving Abilities Compared to Their Functional Visual Perception Skills and Their Actual Driving Performance.* Taira, E. (Ed.). Assessing the Driving Ability of the Elderly New York: Hayworth Press.

Garner, K.C. (1967). *Evaluation of Human Operator Coupled Dynamic Systems.* In Singleton, W. T., Easterby, R. S., and Whitfield, D. C. (Eds.). The Human Operator in Complex Systems, London: Taylor and Francis.

Godthelp, H. (1986). *Vehicle Control During Curve Driving.* Human Factors, 28,2, pp. 211-221.

Godthelp, H. and J. Shumann (1994). *Use of Intelligent Accelerator as an Element of a Driver Support System.* Abstracted in Driver Performance Data Book Update: Older Drivers and IVHS Circular 419, Transportation Research Board, National Research Council, Washington, DC.

Greene, F. A. (1994). *A Study of Field and Laboratory Legibility Distances for Warning Symbol Signs.* Unpublished doctoral dissertation, Texas A&M University.

Gstalter, H. and C. G. Hoyos (1988). *Human Factors Aspects of Road Traffic Safety*. In Peters, G.A. and Peters, B.J., Automotive Engineering and Litigation Volume 2, New York: Garland Law Publishing Co.

Hancock, P. A. and R. Parasuraman (1992). *Human Factors and Safety in the Design of Intelligent Vehicle - Highway Systems (IVHS)*.

Hoffman, E. R. (1991). *Accelerator to Brake Movement Times and A Comparison of Hand and Foot Movement Times.* Ergonomics 34, pp. 277-287 and pp. 397-406.

Hooper, K. G. and H. W. McGee (1983). *Driver Perception-Reaction Time: Are Revisions to Current Specifications in Order?* Transportation Research Record 904, Transportation Research Board, National Research Council, Washington, DC, pp. 21-30.

Hostetter, R., H. McGee, K. Crowley, E. Seguin, and G. Dauber, (1986). *Improved Perception-Reaction Time Information for Intersection Sight Distance*. FHWA/RD-87/015, Federal Highway Administration, Washington, DC.

Hulbert, S. (1988). *The Driving Task.* In Peters, G.A. and Peters, B.J., Automotive Engineering and Litigation Volume 2, New York: Garland Law Publishing Co.

Institute of Transportation Engineers (1992). *Traffic Engineering Handbook*, Washington, DC.

IVHS America Publication B-571, Society of Automotive Engineers, (1992). *Intelligent Vehicle-Highway Systems (IVHS) Safety and Human Factors Considerations.* Warrendale, PA.

Jacobs, R. J., A. W. Johnston, and B. L. Cole, (1975). *The Visibility of Alphabetic and Symbolic Traffic Signs.* Australian Board Research, 5, pp. 68-86.

Johansson, G. and K. Rumar (1971). *Driver's Brake Reaction Times.* Human Factors, 12 (1), pp. 23-27.

Knight, J. L. (1987). *Manual Control of Tracking* in Salvendy, G. (ed) Handbook of Human Factors. New York: John Wiley and Sons.

Koppa, R. J. (1990). *State of the Art in Automotive Adaptive Equipment.* Human Factors, 32 (4), pp. 439-455.

Koppa, R. J., M. McDermott, et al., (1980). *Human Factors Analysis of Automotive Adaptive Equipment for Disabled Drivers.* Final Report DOT-HS-8-02049, National Highway Traffic Safety Administration, Washington, DC.

Kroemer, K. H. E., H. B. Kroemer, and K. E. Kroemer-Ebert (1994). *Ergonomics: How to Design for Ease and Efficiency* Englewood Cliffs, NJ: Prentice-Hall.

Lerner, Neil, et al. (1995). *Literature Review: Older Driver Perception-Reaction Time for Intersection Sight Distance and Object Detection,* Comsis Corp., Contract DTFH 61-90-00038 Federal Highway Administration, Washington, DC, Final Report.

Lunenfeld, H. and G. J. Alexander (1990). *A User's Guide to Positive Guidance (3rd Edition)* FHWA SA-90-017, Federal Highway Administration, Washington, DC.

Marmor, M. (1982). *Aging and the Retina.* In R. Sekuler, D. Kline, and K. Dismukes (eds.). Aging and Human Visual Function, New York: Alan R. Liss.

McGee, H. W. (1989). *Reevaluation of Usefulness and Application of Decision Sight Distance.* Transportation Research Record 1208, TRB, NRC, Washington, DC.

McGee, H., K. Hooper, W. Hughes, and W. Benson (1983). *Highway Design and Operational Standards Affected by Driver Characteristics.* FHWA RD 78-78, Federal Highway Administration, Washington, DC.

McKnight, A. J. and A. S. McKnight (1993). *The Effect of Cellular Phone Use Upon Driver Attention.* Accident Analysis and Prevention 25 (3), pp. 259-265.

McPherson, K., J. Michael, A. Ostrow, and P. Shaffron (1988). *Physical Fitness and the Aging Driver - Phase I.* AAA Foundation for Traffic Safety, Washington, DC.

McRuer, D. T. and R. H. Klein (1975). *Automobile Controllability - Driver/Vehicle Response for Steering Control.* Systems Technology Inc. TR-1040-1-I. DOT-HS-801-407, National Highway Traffic Safety Administration, Washington, DC.

Moir, A. and D. Jessel (1991). *Brain Sex.* New York: Carol Publishing Co.

Mortimer, R. G. (1988). *Rear End Crashes.* Chapter 9 of Peters, G.A. and Peters, B.J., *Automotive Engineering and Litigation,*Vol. 2, New York: Garland Law Publishing Co.

NHTSA Driver Performance Data Book (1987). R. L. Henderson (ed.) DOT HS 807 121, National Highway Traffic Safety Administration, Washington, DC.

NHTSA Driver Performance Data Book (1994). Transportation Reserach Circular 419, Transportation Research Board, National Research Council, Washington, DC.

Neuman, T. R. (1989). *New Approach to Design for Stopping Sight Distance.* Transportation Research Record, 1208, Transportation Research Board, National Research Council, Washington, DC.

O'Leary, A. and R. Atkins (1993). *Transportation Needs of the Older Driver.* Virginia Transportation Research Council MS-5232 (FHWA/VA-93-R14).

Odeh, R. E. (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening.* New York: arcel Dekker, Inc.

Ordy, J. , K. Brizzee, and H. Johnson (1982). *Cellular Alterations in Visual Pathways and the Limbic System: Implications for Vision and Short Term Memory.* ibid.

Picha, D. (1992). *Determination of Driver Capability in the Detection and Recognition of an Object.* In Kahl, K. B. *Investigation of Appropriate Objects for Safe Stopping Distance.* Unpublished Master's Thesis, Texas A&M University.

Price, D. (1988). *Effects of Alcohol and Drugs.* In Peters, G. A. and Peters, B. J., Automotive Engineering and Litigation Volume 2, New York: Garland Law Publishing Co.

Reason, J. (1990). *Human Error.* New York: Cambridge University Press.

Schiff, W. (1980). *Perception: An Applied Approach.* Boston: Houghton Mifflin.

Sekuler, R. and R. Blake (1990). *Perception.* 2nd Ed. New York: McGraw-Hill.

Sheridan, T. B. (1962). *The Human Operator in Control Instrumentation.* In MacMillan, R. H., Higgins, T. J., and Naslin, P. (Eds.)

Simmala, H. (1981). *Driver/Vehicle Steering Response Latencies.* Human Factors 23, pp. 683-692.

Smiley, A. M. (1974). *The Combined Effect of Alcohol and Common Psychoactive Drugs: Field Studies with an Instrumented Automobile.* Technical Bulletin ST-738, National Aeronautical Establishment, Ottawa, Canada.

Taoka, G. T. (1989). *Brake Reaction Times of Unalerted Drivers.* ITE Journal, Washington, DC.

Transportation Research Board (1985). *Highway Capacity Manual,* Special Report 209, Washington, DC.

Transportation Research Board (1987). *Zero Alcohol and Other Options - Limits for Truck and Bus Drivers*, Special Report 216, Transportation Research Board, National Research Council, Washington, DC.

Transportation Research Board, National Research Council (1988). *Transportation in an Aging Society.* Special Report 218, Washington, DC.

Transportation Research Board, National Research Council (1993). *Human Factors Research in Highway Safety*, TRB Circular 414, Washington, DC.

Triggs, T. J. (1988). Speed Estimation, Chapter 18, Vol. II, Peters, G. A. and Peters, B. *Automotive Engineering and Litigation*, op. cit.

Weir, David, (1976). Personal communication at a meeting, CA.

Wierwille, W. W., J. G. Casali, and B. S. Repa (1983). *Driver Steering Reaction Time to Abrupt-onset Crosswinds as Measured in a Moving Base Driving Simulator.* Human Factors 25, pp. 103-116.

Wortman, R. H. and J. S. Matthias (1983). *Evaluation of Driver Behavior at Signalized Intersections.* Transportation Research Record 904, TRB, NRC, Washington, DC.

CAR FOLLOWING MODELS

BY RICHARD W. ROTHERY[6]

[6] Senior Lecturer, Civil Engineering Department, The University of Texas, ECJ Building 6.204, Austin, TX 78712

CHAPTER 4 - Frequently used Symbols

α	=	Numerical coefficients
$a_{l,m}$	=	Generalized sensitivity coefficient
$a_f(t)$	=	Instantaneous acceleration of a following vehicle at time t
$a_l(t)$	=	Instantaneous acceleration of a lead vehicle at time t
β	=	Numerical coefficient
C	=	Single lane capacity (vehicle/hour)
τ	=	Rescaled time (in units of response time, T)
δ	=	Short finite time period
F	=	Amplitude factor
γ	=	Numerical coefficient
k	=	Traffic stream concentration in vehicles per kilometer
k_j	=	Jam concentration
k_m	=	Concentration at maximum flow
k_f	=	Concentration where vehicle to vehicle interactions begin
k_n	=	Normalized concentration
L	=	Effective vehicle length
L^{-1}	=	Inverse Laplace transform
λ	=	Proportionality factor
λ_i	=	Sensitivity coefficient, i = 1,2,3,...
$ln(x)$	=	Natural logarithm of x
q	=	Flow in vehicles per hour
q_n	=	Normalized flow
$<S>$	=	Average spacing rear bumper to rear bumper
S_i	=	Initial vehicle spacing
S_f	=	Final vehicle spacing
S_o	=	Vehicle spacing for stopped traffic
$S(t)$	=	Inter-vehicle spacing
ΔS	=	Inter-vehicle spacing change
$\overline{T}$	=	Average response time
T_o	=	Propagation time for a disturbance
t	=	Time
t_c	=	Collision time
T	=	Reaction time
U_l	=	Speed of a lead vehicle
U_f	=	Speed of a following vehicle
U_f	=	Final vehicle speed
U_f	=	Free mean speed, speed of traffic near zero concentration
U_i	=	Initial vehicle speed
U_{rel}	=	Relative speed between a lead and following vehicle
$u_l(t)$	=	Velocity profile of the lead vehicle of a platoon
V	=	Speed
V_f	=	Final vehicle speed
ω	=	Frequency of a monochromatic speed oscillation
$\ddot{x}_f(t)$	=	Instantaneous acceleration of a following vehicle at time t
$\ddot{x}_l(t)$	=	Instantaneous speed of a lead vehicle at time t
$\ddot{x}_f(t)$	=	Instantaneous speed of a following vehicle at time t
$\dot{x}_l(t)$	=	Instantaneous speed of a lead vehicle at time t
$\dot{x}_f(t)$	=	Instantaneous speed of a following vehicle at time t
$x_l(t)$	=	Instantaneous position of a lead vehicle at time t
$x_f(t)$	=	Instantaneous position of the following vehicle at time t
$x_i(t)$	=	Instantaneous position of the ith vehicle at time t
$z(t)$	=	Position in a moving coordinate system
$< x >$	=	Average of a variable x
Ω	=	Frequency factor

4.
CAR FOLLOWING MODELS

It has been estimated that mankind currently devotes over 10 million man-years each year to driving the automobile, which on demand provides a mobility unequaled by any other mode of transportation. And yet, even with the increased interest in traffic research, we understand relatively little of what is involved in the "driving task". Driving, apart from walking, talking, and eating, is the most widely executed skill in the world today and possibly the most challenging.

Cumming (1963) categorized the various subtasks that are involved in the overall driving task and paralleled the driver's role as an information processor (see Chapter 3). This chapter focuses on one of these subtasks, the task of one vehicle following another on a single lane of roadway (car following). This particular driving subtask is of interest because it is relatively simple compared to other driving tasks, has been successfully described by mathematical models, and is an important facet of driving. Thus, understanding car following contributes significantly to an understanding of traffic flow. Car following is a relatively simple task compared to the totality of tasks required for vehicle control. However, it is a task that is commonly practiced on dual or multiple lane roadways when passing becomes difficult or when traffic is restrained to a single lane. Car following is a task that has been of direct or indirect interest since the early development of the automobile.

One aspect of interest in car following is the average spacing, S, that one vehicle would follow another at a given speed, V. The interest in such speed-spacing relations is related to the fact that nearly all capacity estimates of a single lane of roadway were based on the equation:

$$C = (1000)\, V/S \qquad (4.1)$$

where

C = Capacity of a single lane (vehicles/hour)
V = Speed (km/hour)
S = Average spacing rear bumper to rear bumper in meters

The first *Highway Capacity Manual* (1950) lists 23 observational studies performed between 1924 and 1941 that were directed at identifying an operative speed-spacing relation so that capacity estimates could be established for single lanes of roadways. The speed-spacing relations that were obtained from these studies can be represented by the following equation:

$$S = \alpha + \beta V + \gamma V^2 \qquad (4.2)$$

where the numerical values for the coefficients, α, β, and γ take on various values. Physical interpretations of these coefficients are given below:

α = the effective vehicle length, L
β = the reaction time, T
γ = the reciprocal of twice the maximum average deceleration of a following vehicle

In this case, the additional term, γV^2, can provide sufficient spacing so that if a lead vehicle comes to a full stop instantaneously, the following vehicle has sufficient spacing to come to a complete stop without collision. A typical value empirically derived for γ would be ≈ 0.023 seconds ^{2}/ft . A less conservative interpretation for the non-linear term would be:

$$\gamma = 0.5(a_f^{-1} - a_\ell^{-1}) \qquad (4.3)$$

where a_f and a_ℓ are the average maximum decelerations of the following and lead vehicles, respectively. These terms attempt to allow for differences in braking performances between vehicles whether real or perceived (Harris 1964).

For $\gamma = 0$, many of the so-called "good driving" rules that have permeated safety organizations can be formed. In general, the speed-spacing Equation 4.2 attempts to take into account the physical length of vehicles; the human-factor element of perception, decision making, and execution times; and the net physics of braking performances of the vehicles themselves. It has been shown that embedded in these models are theoretical estimates of the speed at maximum flow, $(\alpha/\gamma)^{0.5}$; maximum flow, $[\beta + 2(\alpha\,\gamma)^{0.5}]^{-1}$; and the speed at which small changes in traffic stream speed propagate back through a traffic stream, $(\alpha/\gamma)^{0.5}$ (Rothery 1968).

The speed-spacing models noted above are applicable to cases where each vehicle in the traffic stream maintains the same or nearly the same constant speed and each vehicle is attempting to

maintain the same spacing (i.e., it describes a steady-state traffic stream).

Through the work of Reuschel (1950) and Pipes (1953), the dynamical elements of a line of vehicles were introduced. In these works, the focus was on the dynamical behavior of a stream of vehicles as they accelerate or decelerate and each driver-vehicle pair attempts to follow one another. These efforts were extended further through the efforts of Kometani and Sasaki (1958) in Japan and in a series of publications starting in 1958 by Herman and his associates at the General Motors Research Laboratories. These research efforts were microscopic approaches that focused on describing the detailed manner in which one vehicle followed another. With such a description, the macroscopic behavior of single lane traffic flow can be approximated. Hence, car following models form a bridge between individual "car following" behavior and the macroscopic world of a line of vehicles and their corresponding flow and stability properties.

4.1 Model Development

Car following models of single lane traffic assume that there is a correlation between vehicles in a range of inter-vehicle spacing, from zero to about 100 to 125 meters and provides an explicit form for this coupling. The modeling assumes that each driver in a following vehicle is an active and predictable control element in the driver-vehicle-road system. These tasks are termed psychomotor skills or perceptual-motor skills because they require a continued motor response to a continuous series of stimuli.

The relatively simple and common driving task of one vehicle following another on a straight roadway where there is no passing (neglecting all other subsidiary tasks such as steering, routing, etc.) can be categorized in three specific subtasks:

- Perception: The driver collects relevant information mainly through the visual channel. This information arises primarily from the motion of the lead vehicle and the driver's vehicle. Some of the more obvious information elements, only part of which a driver is sensitive to, are vehicle speeds, accelerations and higher derivatives (e.g., "jerk"), inter-vehicle spacing, relative speeds, rate of closure, and functions of these variables (e.g., a "collision time").

- Decision Making: A driver interprets the information obtained by sampling and integrates it over time in order to provide adequate updating of inputs. Interpreting the information is carried out within the framework of a knowledge of vehicle characteristics or class of characteristics and from the driver's vast repertoire of driving experience. The integration of current information and catalogued knowledge allows for the development of driving strategies which become "automatic" and from which evolve "driving skills".

- Control: The skilled driver can execute control commands with dexterity, smoothness, and coordination, constantly relying on feedback from his own responses which are superimposed on the dynamics of the system's counterparts (lead vehicle and roadway).

It is not clear how a driver carries out these functions in detail. The millions of miles that are driven each year attest to the fact that with little or no training, drivers successfully solve a multitude of complex driving tasks. Many of the fundamental questions related to driving tasks lie in the area of 'human factors' and in the study of how human skill is related to information processes.

The process of comparing the inputs of a human operator to that operator's outputs using operational analysis was pioneered by the work of Tustin (1947), Ellson (1949), and Taylor (1949). These attempts to determine mathematical expressions linking input and output have met with limited success. One of the primary difficulties is that the operator (in our case the driver) has no unique transfer function; the driver is a different 'mechanism' under different conditions. While such an approach has met with limited success, through the course of studies like

these a number of useful concepts have been developed. For example, reaction times were looked upon as characteristics of individuals rather than functional characteristics of the task itself. In addition, by introducing the concept of "information", it has proved possible to parallel reaction time with the rate of coping with information.

The early work by Tustin (1947) indicated maximum rates of the order of 22-24 bits/second (sec). Knowledge of human performance and the rates of handling information made it possible to design the response characteristics of the machine for maximum compatibility of what really is an operator-machine system.

The very concept of treating an operator as a transfer function implies, partly, that the operator acts in some continuous manner. There is some evidence that this is not completely correct and that an operator acts in a discontinuous way. There is a period of time during which the operator having made a "decision" to react is in an irreversible state and that the response must follow at an appropriate time, which later is consistent with the task.

The concept of a human behavior being discontinuous in carrying out tasks was first put forward by Uttley (1944) and has been strengthened by such studies as Telfor's (1931), who demonstrated that sequential responses are correlated in such a way that the response-time to a second stimulus is affected significantly by the separation of the two stimuli. Inertia, on the other hand, both in the operator and the machine, creates an appearance of smoothness and continuity to the control element.

In car following, inertia also provides direct feedback data to the operator which is proportional to the acceleration of the vehicle. Inertia also has a smoothing effect on the performance requirements of the operator since the large masses and limited output of drive-trains eliminate high frequency components of the task.

Car following models have not explicitly attempted to take all of these factors into account. The approach that is used assumes that a stimulus-response relationship exists that describes, at least phenomenologically, the control process of a driver-vehicle unit. The stimulus-response equation expresses the concept that a driver of a vehicle responds to a given stimulus according to a relation:

$$\text{Response} = \lambda \text{ Stimulus} \qquad (4.4)$$

where λ is a proportionality factor which equates the stimulus function to the response or control function. The stimulus function is composed of many factors: speed, relative speed, inter-vehicle spacing, accelerations, vehicle performance, driver thresholds, etc.

Do all of these factors come into play part of the time? The question is, which of these factors are the most significant from an explanatory viewpoint. Can any of them be neglected and still retain an approximate description of the situation being modeled?

What is generally assumed in car following modeling is that a driver attempts to: (a) keep up with the vehicle ahead and (b) avoid collisions.

These two elements can be accomplished if the driver maintains a small average relative speed, U_{rel} over short time periods, say δt, i.e.,

$$<U_l - U_f> = <U_{rel}> \frac{1}{\delta t}\int_{t-\delta t/2}^{t+\delta t/2} U_{rel}(t)dt \qquad (4.5)$$

is kept small. This ensures that 'collision' times:

$$t_c = \frac{S(t)}{U_{rel}} \qquad (4.6)$$

are kept large, and inter-vehicle spacings would not appreciably increase during the time period, δt. The duration of the δt will depend in part on alertness, ability to estimate quantities such as: spacing, relative speed, and the level of information required for the driver to assess the situation to a tolerable probability level (e.g., the probability of detecting the relative movement of an object, in this case a lead vehicle) and can be expressed as a function of the perception time.

Because of the role relative-speed plays in maintaining relatively large collision times and in preventing a lead vehicle from 'drifting' away, it is assumed as a first approximation that the argument of the stimulus function is the relative speed.

From the discussion above of driver characteristics, relative speed should be integrated over time to reflect the recent time history of events, i.e., the stimulus function should have the form like that of Equation 4.5 and be generalized so that the stimulus

at a given time, t, depends on the weighted sum of all earlier values of the relative speed, i.e.,

$$< U_\ell - U_f > = <U_{rel}> = \int_{t-\delta t/2}^{t+\delta t/2} \sigma(t-t')U_{rel}(t')dt \quad (4.7)$$

where $\sigma(t)$ is a weighing function which reflects a driver's estimation, evaluation, and processing of earlier information (Chandler et al. 1958). The driver weighs past and present information and responds at some future time. The consequence of using a number of specific weighing functions has been examined (Lee 1966), and a spectral analysis approach has been used to derive a weighing function directly from car following data (Darroch and Rothery 1969).

The general features of a weighting function are depicted in Figure 4.1. What has happened a number of seconds ($\approx$ 5 sec) in the past is not highly relevant to a driver now, and for a short time ($\approx$ 0.5 sec) a driver cannot readily evaluate the information available to him. One approach is to assume that

$$\sigma(t) = \delta(t-T) \quad (4.8)$$

where

$$\delta(t-T) = 0, \text{ for } t \neq T \quad (4.9)$$

$$\delta(t-T) = 1, \text{ for } t = T \quad (4.10)$$

and

$$\int_o^\infty \delta(t-T)dt = 1$$

For this case, our stimulus function becomes

$$\text{Stimulus}(t) = U_\ell(t - T) - U_f(t - T) \quad (4.11)$$

which corresponds to a simple constant response time, T, for a driver-vehicle unit. In the general case of $\sigma(t)$, there is an average response time, $\overline{T}$, given by

$$\overline{T(t)} = \int_0^t t'\sigma(t')dt' \quad (4.12)$$

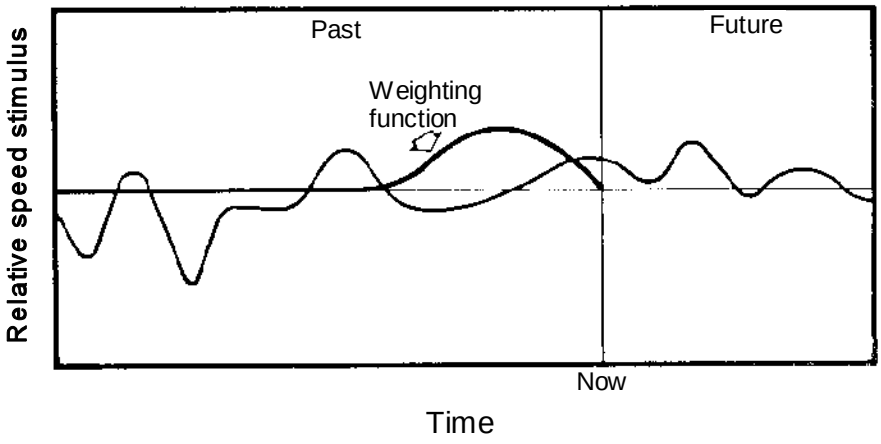


Figure 4.1
Schematic Diagram of Relative Speed Stimulus and a Weighting Function Versus Time (Darroch and Rothery 1972).

The main effect of such a response time or delay is that the driver is responding at all times to a stimulus. The driver is observing the stimulus and determining a response that will be made some time in the future. By delaying the response, the driver obtains "advanced" information.

For redundant stimuli there is little need to delay response, apart from the physical execution of the response. Redundancy alone can provide advance information and for such cases, response times are shorter.

The response function is taken as the acceleration of the following vehicle, because the driver has direct control of this quantity through the 'accelerator' and brake pedals and also because a driver obtains direct feedback of this variable through inertial forces, i.e.,

$$\text{Response } (t) = a_f(t) = \ddot{x}_f(t) \qquad (4.13)$$

where $x_i(t)$ denotes the longitudinal position along the roadway of the ith vehicle at time t. Combining Equations4.11 and 4.13 into Equation 4.4 the stimulus-response equation becomes (Chandler et al. 1958):

$$\ddot{x}_f(t) = \lambda[\dot{x}_i(t-T) - \dot{x}_f(t-T)] \qquad (4.14)$$

or equivalently

$$\ddot{x}_f(t+T) = \lambda[\dot{x}_i(t) - \dot{x}_f(t)] \qquad (4.15)$$

Equation 4.15 is a first approximation to the stimulus-response equation of car-following, and as such it is a grossly simplified description of a complex phenomenon. A generalization of car following in a conventional control theory block diagram is shown in Figure 4.1a. In this same format the linear car-following model presented in Equation 4.15 is shown in Figure 4.1b. In this figure the driver is represented by a time delay and a gain factor. Undoubtedly, a more complete representation of car following includes a set of equations that would model the dynamical properties of the vehicle and the roadway characteristics. It would also include the psychological and physiological properties of drivers, as well as couplings between vehicles, other than the forward nearest neighbors and other driving tasks such as lateral control, the state of traffic, and emergency conditions.

For example, vehicle performance undoubtedly alters driver behavior and plays an important role in real traffic where mixed traffic represents a wide performance distribution, and where appropriate responses cannot always be physically achieved by a subset of vehicles comprising the traffic stream. This is one area where research would contribute substantially to a better understanding of the growth, decay, and frequency of disturbances in traffic streams (see, e.g., Harris 1964; Herman and Rothery 1967; Lam and Rothery 1970).

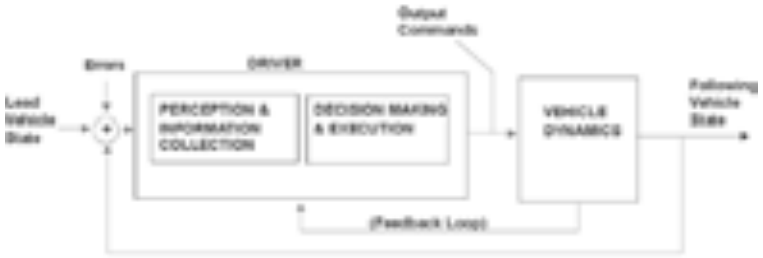


Figure 4.1a
Block Diagram of Car-Following.

Figure 4.1b
Block Diagram of the Linear Car-Following Model.

4.2 Stability Analysis

In this section we address the stability of the linear car following equation, Equation 4.15, with respect to disturbances. Two particular types of stabilities are examined: local stability and asymptotic stability.

Local Stability is concerned with the response of a following vehicle to a fluctuation in the motion of the vehicle directly in front of it; i.e., it is concerned with the localized behavior between pairs of vehicles.

Asymptotic Stability is concerned with the manner in which a fluctuation in the motion of any vehicle, say the lead vehicle of a platoon, is propagated through a line of vehicles.

The analysis develops criteria which characterize the types of possible motion allowed by the model. For a given range of model parameters, the analysis determines if the traffic stream (as described by the model) is stable or not, (i.e., whether disturbances are damped, bounded, or unbounded). This is an important determination with respect to understanding the applicability of the modeling. It identifies several characteristics with respect to single lane traffic flow, safety, and model validity. If the model is realistic, this range should be consistent with measured values of these parameters in any applicable situation where disturbances are known to be stable. It should also be consistent with the fact that following a vehicle is an extremely common experience, and is generally stable.

4.2.1 Local Stability

In this analysis, the linear car following equation, (Equation 4.15) is assumed. As before, the position of the lead vehicle and the following vehicle at a time, t, are denoted by $x_\ell(t)$ and $x_f(t)$, respectively. Rescaling time in units of the response time, T, using the transformation, $t = \tau T$, Equation 4.15 simplifies to

$$\ddot{x}_f(\tau+1) = C[(\dot{x}_\ell(\tau)-\dot{x}_f(\tau))] \qquad (4.16)$$

where $C = \lambda T$. The conditions for the local behavior of the following vehicle can be derived by solving Equation 4.16 by the method of Laplace transforms (Herman et al. 1959).

The evaluation of the inverse Laplace transform for Equation 4.16 has been performed (Chow 1958; Kometani and Sasaki 1958). For example, for the case where the lead and following vehicles are initially moving with a constant speed, u, the solution for the speed of the following vehicle was given by Chow where ν denotes the integral part of t/T. The complex form of Chow's solution makes it difficult to describe various physical properties (Chow 1958).

$$\dot{x}_n(t) = u + \sum_{\sigma=0}^{\nu-n} \int_{(n+\sigma)T}^{t} (-1)^\sigma \lambda^{n+\sigma} \frac{||\tau-(n+\sigma)T||^{n+\sigma-1}}{(n-1)!\sigma!} \cdot (u_0(t-\tau)-u)\,dt$$

However, the general behavior of the following vehicle's motion can be characterized by considering a specific set of initial conditions. Without any loss in generality, initial conditions are assumed so that both vehicles are moving with a constant speed, u. Then using a moving coordinate system $z(t)$ for both the lead and following vehicles the formal solution for the acceleration of the following vehicle is given more simply by:

$$L^{-1}[C(C + se^{s})^{-1}s \tag{4.16a}$$

where L^{-1} denotes the inverse Laplace transform. The character of the above inverse Laplace transform is determined by the singularities of the factor $(C + se^{s})^{-1}$ since $Cs^2Z_\ell(s)$ is a regular function. These singularities in the finite plane are the simple poles of the roots of the equation

$$C + se^{s} = 0 \tag{4.17}$$

Similarly, solutions for vehicle speed and inter-vehicle spacings can be obtained. Again, the behavior of the inter-vehicle spacing is dictated by the roots of Equation 4.17. Even for small t, the character of the solution depends on the pole with the largest real part, say , $s_0 = a_0 + ib_0$, since all other poles have considerably larger negative real parts so that their contributions are heavily damped.

Hence, the character of the inverse Laplace transform has the form $e^{a_0 t}\, e^{tb_0 t}$. For different values of C, the pole with the largest real part generates four distinct cases:

a) if $C \le e^{-1}(\approx 0.368)$, *then* $a_0 \le 0$, $b_0 = 0$, and the motion is non-oscillatory and exponentially damped.

b) if $e^{-1} < C < \pi / 2$, then $a_0 < 0$, $b_0 > 0$ and the motion is oscillatory with exponential damping.

c) if $C = \pi / 2$, then $a_0 = 0$, $b_0, > 0$ and the motion is oscillatory with constant amplitude.

d) if $C > \pi / 2$ then $a_0 > 0$, $b_0 > 0$ and the motion is oscillatory with increasing amplitude.

The above establishes criteria for the numerical values of C which characterize the motion of the following vehicle. In particular, it demonstrates that in order for the following vehicle not to over-compensate to a fluctuation, it is necessary that $C \le 1/e$. For values of C that are somewhat greater, oscillations occur but are heavily damped and thus insignificant. Damping occurs to some extent as long as $C < \pi/2$.

These results concerning the oscillatory and non-oscillatory behavior apply to the speed and acceleration of the following vehicle as well as to the inter-vehicle spacing. Thus, e.g., if $C \le e^{-1}$, the inter-vehicle spacing changes in a non-oscillatory manner by the amount ΔS, where

$$\Delta S = \frac{1}{\lambda}(V - U) \tag{4.18}$$

when the speeds of the vehicle pair changes from U to V. An important case is when the lead vehicle stops. Then, the final speed, V, is zero, and the total *change* in inter-vehicle spacing is $-U/\lambda$.

In order for a following vehicle to avoid a 'collision' from initiation of a fluctuation in a lead vehicle's speed the inter-vehicle spacing should be at least as large as U/λ. On the other hand, in the interests of traffic flow the inter-vehicle spacing should be small by having λ as large as possible and yet within the stable limit. Ideally, the best choice of λ is $(eT)^{-1}$.

The result expressed in Equation 4.18 follows directly from Chow's solution (or more simply by elementary considerations). Because the initial and final speeds for both vehicles are U and V, respectively, we have

$$\int_0^\infty \ddot{x}_f(t+T)dt = V - U \tag{4.19}$$

and from Equation 4.15 we have

$$\lambda \int_0^\infty [\dot{x}_l(t) - \dot{x}_f(t)]dt = \Delta S$$

or

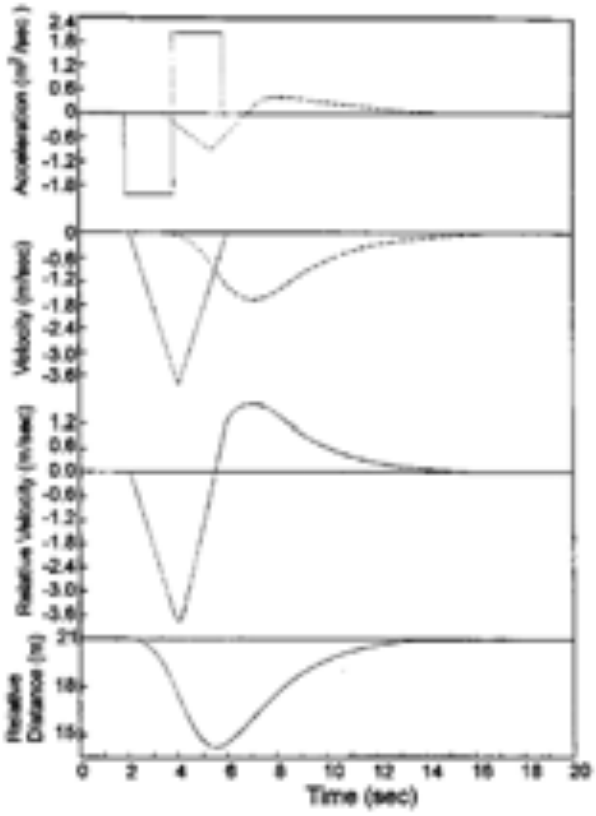
$$\Delta S = \int_0^\infty [\dot{x}_\ell(t) - \dot{x}_f(t)]dt = \frac{V - U}{\lambda} \tag{4.20}$$

as given earlier in Equation 4.18.

In order to illustrate the general theory of local stability, the results of several calculations using a Berkeley Ease analog computer and an IBM digital computer are described. It is interesting to note that in solving the linear car following equation for two vehicles, estimates for the local stability condition were first obtained using an analog computer for different values of C which differentiate the various type of motion.

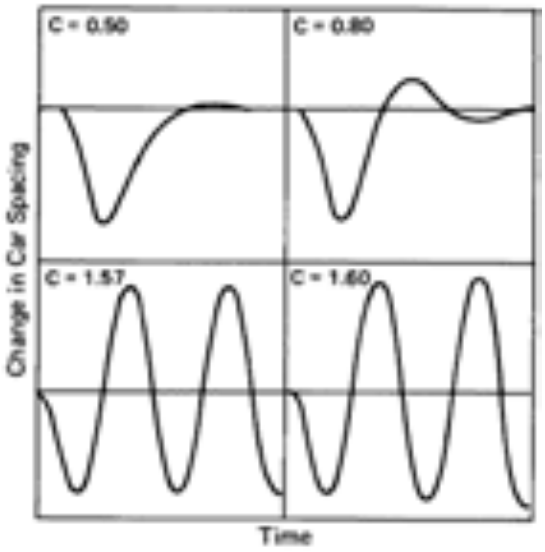
Figure 4.2 illustrates the solutions for $C = e^{-1}$, where the lead vehicle reduces its speed and then accelerates back to its original speed. Since C has a value for the locally stable limit, the acceleration and speed of the following vehicle, as well as the inter-vehicle spacing between the two vehicles are non-oscillatory.

In Figure 4.3, the inter-vehicle spacing is shown for four other values of C for the same fluctuation of the lead vehicle as shown in Figure 4.2. The values of C range over the cases of oscillatory



Note: Vehicle 2 follows Vehicle 1 (lead car) with a time lag T=1.5 sec and a value of $C=e^{-1}(\approx 0.368)$, the limiting value for local stability. The initial velocity of each vehicle is **u**

Figure 4.2
Detailed Motion of Two Cars Showing the
Effect of a Fluctuation in the Acceleration of the Lead Car (Herman et al. 1958).



Note: Changes in car spacings from an original constant spacing between two cars for the noted values of C. The acceleration profile of the lead car is the same as that shown in Figure 4.2.

Figure 4.3
Changes in Car Spacings from an
Original Constant Spacing Between Two Cars (Herman et al. 1958).

motion where the amplitude is damped, undamped, and of increasing amplitude.

For the values of $C = 0.5$ and 0.80, the spacing is oscillatory and heavily damped.

$$\text{For } C = 1.57 \left(\approx \frac{\pi}{2} \right),$$

the spacing oscillates with constant amplitude. For $C = 1.60$, the motion is oscillatory with increasing amplitude.

Local Stability with Other Controls. Qualitative arguments can be given of a driver's lack of sensitivity to variation in relative acceleration or higher derivatives of inter-vehicle spacings because of the inability to make estimates of such quantities. It is of interest to determine whether a control centered around such derivatives would be locally stable. Consider the car following equation of the form

$$\ddot{x}_f(\tau+1) = C\frac{d^m}{dt^m}[x_i(\tau)-x_f(\tau)] \qquad (4.21)$$

for m= 0,1,2,3..., i.e., a control where the acceleration of the following vehicle is proportional to the mth derivative of the inter-vehicle spacing. For $m = 1$, we obtain the linear car following equation.

Using the identical analysis for any m, the equation whose roots determine the character of the motion which results from Equation 4.21 is

$$C + s^m e^s = 0 \qquad (4.22)$$

None of these roots lie on the negative real axis when m is even, therefore, local stability is possible only for odd values of the mth derivative of spacing: relative speed, the first derivative of relative acceleration ($m = 3$), etc. Note that this result indicates that an acceleration response directly proportional to inter-vehicle spacing stimulus is unstable.

4.2.2 Asymptotic Stability

In the previous analysis, the behavior of one vehicle following another was considered. Here a platoon of vehicles (except for the platoon leader) follows the vehicle ahead according to the

linear car following equation, Equation 4.15. The criteria necessary for asymptotic stability or instability were first investigated by considering the Fourier components of the speed fluctuation of a platoon leader (Chandler et al. 1958).

The set of equations which attempts to describe a line of N identical car-driver units is:

$$\ddot{x}_{n+1}(t+T) = \lambda[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \qquad (4.23)$$

where $n = 0,1,2,3,...N$.

Any specific solution to these equations depends on the velocity profile of the lead vehicle of the platoon, $u_0(t)$, and the two parameters λ and T. For any inter-vehicle spacing, if a disturbance grows in amplitude then a 'collision' would eventually occur somewhere back in the line of vehicles.

While numerical solutions to Equation 4.23 can determine at what point such an event would occur, the interest is to determine criteria for the growth or decay of such a disturbance. Since an arbitrary speed pattern can be expressed as a linear combination of monochromatic components by Fourier analysis, the specific profile of a platoon leader can be simply represented by one component, i.e., by a constant together with a monochromatic oscillation with frequency, ω and amplitude, f_o, i.e.,

$$u_o(t) = a_o + f_o e^{i\omega t} \qquad (4.24)$$

and the speed profile of the nth vehicle by

$$u_n(t) = a_o + f_n e^{i\omega t} \qquad (4.25)$$

Substitution of Equations 4.24 and 4.25 into Equation 4.23 yields:

$$u_n(t) = a_o + F(\omega,\lambda,T,n)e^{i\Omega(\omega,\lambda,T,n)} \qquad (4.26)$$

where the amplitude factor $F(\omega, \lambda, T, n)$ is given by

$$[1+(\frac{\omega}{\lambda})^2+2(\frac{\omega}{\lambda})\sin(\omega T)]^{-n/2}$$

which decreases with increasing n if

$$1+(\frac{\omega}{\lambda})^2+2(\frac{\omega}{\lambda})\sin(\omega T) > 1$$

i.e. if

$$\frac{\omega}{\lambda} > 2\sin(\omega T)$$

The severest restriction on the parameter λ arises from the low frequency range, since in the limit as $\omega \to 0$, λ must satisfy the inequality

$$\lambda T < \frac{1}{2}[\lim_{\omega\to 0}(\omega T)/\sin(\omega T)] \qquad (4.27)$$

Accordingly, asymptotic stability is insured for all frequencies where this inequality is satisfied.

For those values of ω within the physically realizable frequency range of vehicular speed oscillations, the right hand side of the inequality of 4.27 has a short range of values of 0.50 to about 0.52. The asymptotic stability criteria divides the two parameter domain into stable and unstable regions, as graphically illustrated in Figure 4.4.

The criteria for local stability (namely that no local oscillations occur when $\lambda T \leq e^{-1}$) also insures asymptoticstability. It has also been shown (Chandler et al. 1958) that a speed fluctuation can be approximated by:

$$x_{n+1}(t) \approx u_0(t)\left[\frac{4\pi n}{\lambda}\left(\frac{1}{2\lambda}-T\right)\right]^{-1/2} \cdot \exp\left[\frac{[t-n/\lambda]}{4n/\lambda(1/2\lambda-T)}\right] \qquad (4.28)$$

Hence, the speed of propagation of the disturbance with respect to the moving traffic stream in number of inter-vehicle separations per second, n/t, is λ.

That is, the time required for the disturbance to propagate between pairs of vehicles is λ^{-1}, a constant, which is independent of the response time T. It is noted from the above equation that in the propagation of a speed fluctuation the amplitude of the

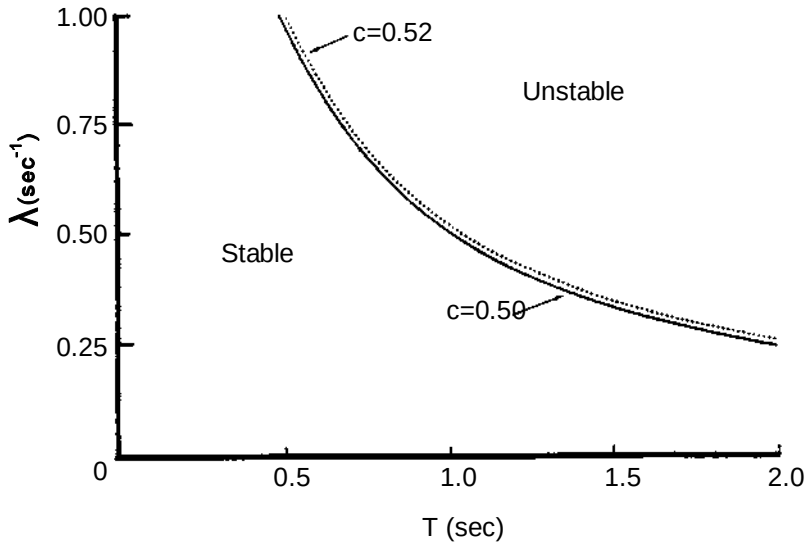


Figure 4.4
Regions of Asymptotic Stability (Rothery 1968).

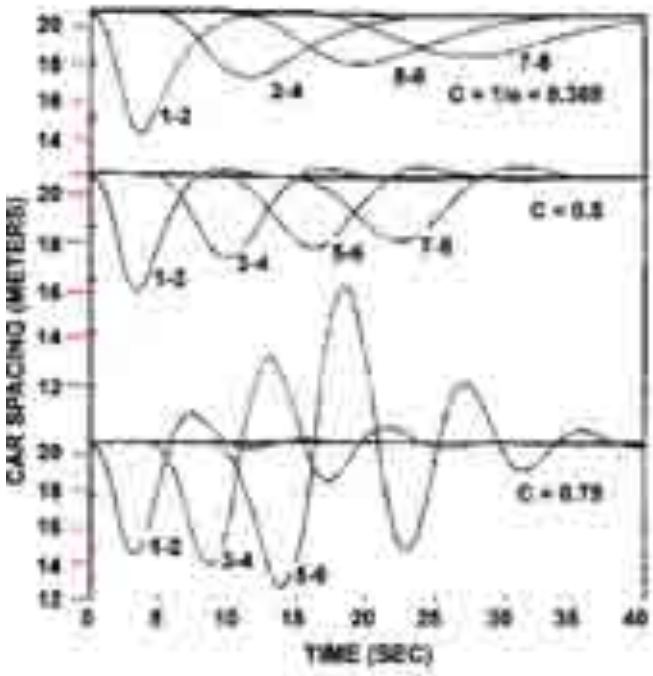
disturbance grows as the response time, T, approaches $1/(2\lambda)$ until instability is reached. Thus, while $\lambda T < 0.5$ ensures stability, short reaction times increase the range of the sensitivity coefficient, λ, that ensures stability. From a practical viewpoint, small reaction times also reduce relatively large responses to a given stimulus, or in contrast, larger response times require relatively large responses to a given stimulus. Acceleration fluctuations can be correspondingly analyzed (Chandler et al. 1958).

4.2.1.1 Numerical Examples

In order to illustrate the general theory of asymptotic stability as outlined above, the results of a number of numerical calculations are given. Figure 4.5 graphically exhibits the inter-vehicle spacings of successive pairs of vehicles versus time for a platoon of vehicles. Here, three values of C were used: $C = 0.368$, 0.5, and 0.75. The initial fluctuation of the lead vehicle, $n = 1$, was the same as that of the lead vehicle illustrated in Figure 4.2. This disturbance consists of a slowing down and then a speeding up to the original speed so that the integral of acceleration over time is zero. The particularly stable, non-oscillatory response is evident in the first case where $C = 0.368$ ($\approx 1/e$), the local stability limit. As analyzed, a heavily damped oscillation occurs in the second case where $C = 0.5$, the asymptotic limit. Note that the amplitude of the disturbance is damped as it propagates through the line of vehicles even though this case is at the asymptotic limit.

This results from the fact that the disturbance is not a single Fourier component with near zero frequency. However, instability is clearly exhibited in the third case of Figure 4.5 where $C = 0.75$ and in Figure 4.6 where $C = 0.8$. In the case shown in Figure 4.6, the trajectories of each vehicle in a platoon of nine are graphed with respect to a coordinate system moving with the initial platoon speed u. Asymptotic instability of a platoon of nine cars is illustrated for the linear car following equation, Equation 4.23, where $C = 0.80$. For $t = 0$, the vehicles are all moving with a velocity u and are separated by a distance of 12 meters. The propagation of the disturbance, which can be readily discerned, leads to "collision" between the 7th and 8th cars at about $t = 24$ sec. The lead vehicle at $t = 0$ decelerates for 2 seconds at 4 km/h/sec, so that its speed changes from u to u -8 km/h and then accelerates back to u. This fluctuation in the speed of the lead vehicle propagates through the platoon in an unstable manner with the inter-vehicle spacing between the seventh and eighth vehicles being reduced to zero at about 24.0 sec after the initial phase of the disturbance is generated by the lead vehicle of the platoon.

In Figure 4.7 the envelope of the minimum spacing that occurs between successive pairs of vehicles is graphed versus time



Note: Diagram uses Equation 4.23 for three values of C. The fluctuation in acceleration of the lead car, car number 1, is the same as that shown in Fig. 4.2 At t=0 the cars are separated by a spacing of 21 meters.

Figure 4.5
Inter-Vehicle Spacings of a Platoon of Vehicles
Versus Time for the Linear Car Following Model (Herman et al. 1958).

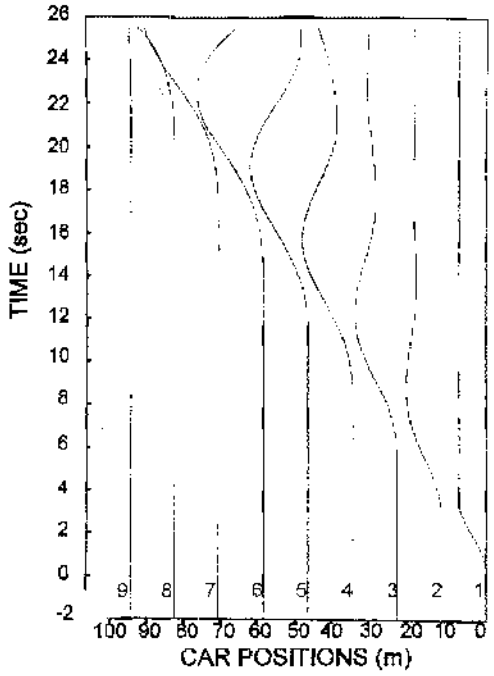
where the lead vehicle's speed varies sinusoidally with a frequency $\omega = 2\pi/10$ radian/sec. The envelope of minimum inter-vehicle spacing versus vehicle position is shown for three values of λ. The response time, *T*, equals 1 second. It has been shown that the frequency spectrum of relative speed and acceleration in car following experiments have essentially all their content below this frequency (Darroch and Rothery 1973).

The values for the parameter λ were 0.530, 0.5345, and 0.550/sec. The value for the time lag, *T*, was 1 sec in each case. The frequency used is that value of ω which just satisfies the stability equation, Equation 4.27, for the case where $\lambda = 0.5345$/sec. This latter figure serves to demonstrate not only the stability criteria as a function of frequency but the accuracy of the numerical results. A comparison between that which is predicted from the stability analysis and the numerical solution for the constant amplitude case (λ=0.5345/sec) serves as a check point. Here, the numerical solution yields a maximum and minimum amplitude that is constant to seven significant places.

4.2.1.2 Next-Nearest Vehicle Coupling

In the nearest neighbor vehicle following model, the motion of each vehicle in a platoon is determined solely by the motion of the vehicle directly in front. The effect of including the motion of the "next nearest neighbor" vehicle (i.e., the car which is two vehicles ahead in addition to the vehicle directly in front) can be ascertained. An approximation to this type of control, is the model

$$\ddot{x}_{n+2}(t+T) = \lambda_1[\dot{x}_{n+1}(t)-\dot{x}_{n+2(t)}]+\lambda_2[\dot{x}_n(t)-\dot{x}_{n+2}] \qquad (4.29)$$



Note: Diagram illustrates the linear car following equation, eq. 4.23, where C=080.

Figure 4.6
Asymptotic Instability of a Platoon of Nine Cars (Herman et al. 1958).

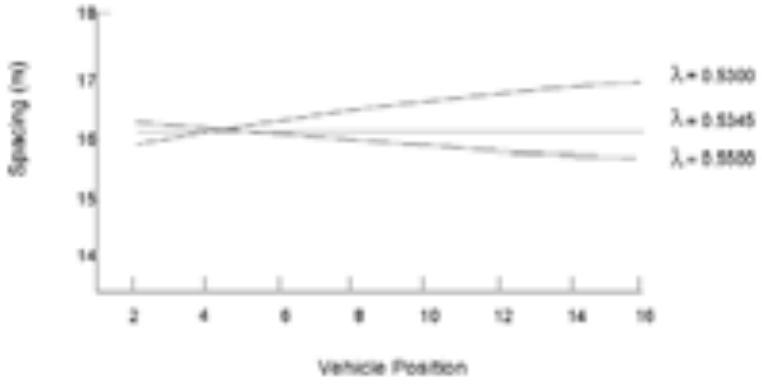


Figure 4.7
Envelope of Minimum Inter-Vehicle Spacing Versus Vehicle Position (Rothery 1968).

$$(\lambda_1+\lambda_2)T < \frac{1}{2}(\omega T)/\sin(\omega T)] \qquad (4.30)$$

which in the limit $\omega \rightarrow 0$ is

$$(\lambda_1+\lambda_2)T > \frac{1}{2} \qquad (4.31)$$

This equation states that the effect of adding next nearest neighbor coupling to the control element is, to the first order, to increase λ_1 to $(\lambda_1 + \lambda_2)$. This reduces the value that λ_1 can have and still maintain asymptotic stability.

4.3 Steady-State Flow

This section discusses the properties of steady-state traffic flow based on car following models of single-lane traffic flow. In particular, the associated speed-spacing or equivalent speed-concentration relationships, as well as the flow-concentration relationships for single lane traffic flow are developed.

The Linear Case. The equations of motion for a single lane of traffic described by the linear car following model are given by:

$$\ddot{x}_{n+1}(t+T) = \lambda[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \qquad (4.32)$$

where $n = 1, 2, 3, ...$

In order to interrelate one steady-state to another under this control, assume (up to a time t=0) each vehicle is traveling at a speed U_i and that the inter-vehicle spacing is S_i. Suppose that at t=0, the lead vehicle undergoes a speed change and increases or decreases its speed so that its final speed after some time, t, is U_f. A specific numerical solution of this type of transition is exhibited in Figure 4.8.

In this example $C = \lambda T$=0.47 so that the stream of traffic is stable, and speed fluctuations are damped. Any case where the asymptotic stability criteria is satisfied assures that each following vehicle comprising the traffic stream eventually reaches a state traveling at the speed U_f. In the transition from a speed U_i to a speed U_f, the inter-vehicle spacing S changes from S_i to S_f, where

$$S_f = S_i+\lambda^{-1}(U_f - U_i) \qquad (4.33)$$

This result follows directly from the solution to the car following equation, Equation 4.16a or from Chow (1958). Equation 4.33 also follows from elementary considerations by integration of Equation 4.32 as shown in the previous section (Gazis et al. 1959). This result is not directly dependent on the time lag, T, except that for this result to be valid the time lag, T, must allow the equation of motion to form a stable stream of traffic. Since vehicle spacing is the inverse of traffic stream concentration, k, the speed-concentration relation corresponding to Equation 4.33 is:

$$k_f^{-1} = k_i^{-1}+\lambda^{-1}(U_f-U_i) \qquad (4.34)$$

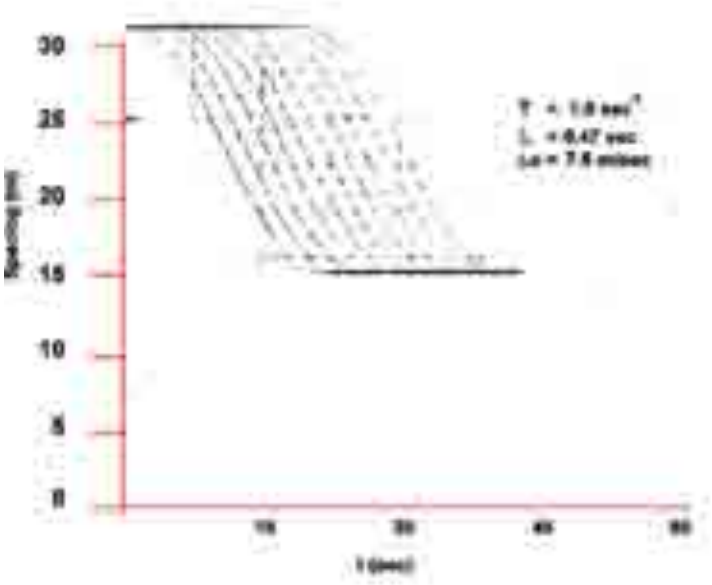
The significance of Equations 4.33 and 4.34 is that:

1) They link an initial steady-state to a second arbitrary steady-state, and

2) They establish relationships between macroscopic traffic stream variables involving a microscopic car following parameter, λ.

In this respect they can be used to test the applicability of the car following model in describing the overall properties of single lane traffic flow. For stopped traffic, $U_i = 0$, and the corresponding spacing, S_o, is composed of vehicle length and "bumper-to-bumper" inter-vehicle spacing. The concentration corresponding to a spacing, S_o, is denoted by k_j and is frequently referred to as the 'jam concentration'.

Given k_j, then Equation 4.34 for an arbitrary traffic state defined by a speed, U, and a concentration, k, becomes



Note: A numerical solution to Equation 4.32 for the inter-vehicle spacings of an 11- vehicle platoon going from one steady-state to another ($\lambda T = 0.47$). The lead vehicle's speed decreases by 7.5 meters per second.

Figure 4.8
Inter-Vehicle Spacings of an Eleven Vehicle Platoon (Rothery 1968).

$$U = \lambda(k^{-1} - k_j^{-1}) \qquad (4.35)$$

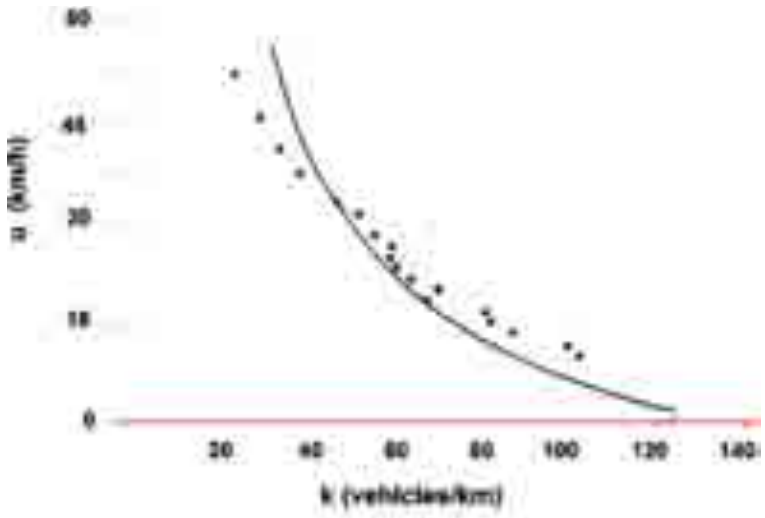
A comparison of this relationship was made (Gazis et al. 1959) with a specific set of reported observations (Greenberg 1959) for a case of single lane traffic flow (i.e., for the northbound traffic flowing through the Lincoln Tunnel which passes under the Hudson River between the States of New York and New Jersey). This comparison is reproduced in Figure 4.9 and leads to an estimate of 0.60 sec^{-1} for λ. This estimate of λ implies an upper bound for $T \approx 0.83$ sec for an asymptotic stable traffic stream using this facility.

While this fit and these values are not unreasonable, a fundamental problem is identified with this form of an equation for a speed-spacing relationship (Gazis et al. 1959). Because it is linear, this relationship does not lead to a reasonable description of traffic flow. This is illustrated in Figure 4.10 where the same data from the Lincoln Tunnel (in Figure 4.9) is regraphed. Here the graph is in the form of a normalized flow, versus a normalized concentration together with the corresponding theoretical steady-state result derived from Equation 4.35, i.e.,

$$q = Uk = \lambda(1 - \frac{k}{k_j}) \qquad (4.36)$$

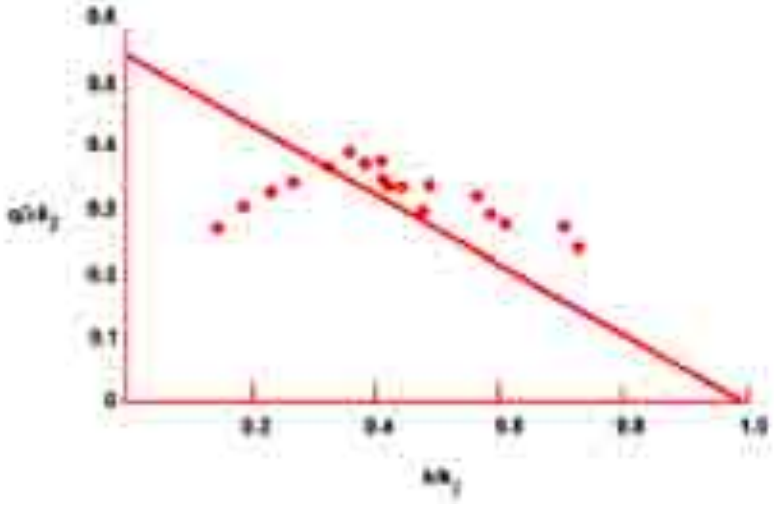
The inability of Equation 4.36 to exhibit the required qualitative relationship between flow and concentration (see Chapter 2) led to the modification of the linear car following equation (Gazis et al. 1959).

Non-Linear Models. The linear car following model specifies an acceleration response which is completely independent of vehicle spacing (i.e., for a given relative velocity, response is the same whether the vehicle following distance is small [e.g., of the order of 5 or 10 meters] or if the spacing is relatively large [i.e., of the order of hundreds of meters]). Qualitatively, we would expect that response to a given relative speed to increase with smaller spacings.



Note: The data are those of (Greenberg 1959) for the Lincoln Tunnel. The curve represents a "least squares fit" of Equation 4.35 to the data.

Figure 4.9
Speed (miles/hour) Versus Vehicle Concentration (vehicles/mile).(Gazis et al. 1959).



Note: The curve corresponds to Equation 4.36 where the parameters are those from the "fit" shown in Figure 4.9.

Figure 4.10
Normalized Flow Versus Normalized Concentration (Gazis et al. 1959).

In order to attempt to take this effect into account, the linear model is modified by supposing that the gain factor, λ, is not a constant but is inversely proportional to vehicle spacing, i.e.,

$$\lambda = \lambda_1 / S(t) = \lambda_1 / [x_n(t) - x_{n+1}(t)] \qquad (4.37)$$

where λ_1 is a new parameter - assumed to be a constant and which shall be referred to as the sensitivity coefficient. Using Equation 4.37 in Equation 4.32, our car following equation is:

$$\ddot{x}_{n+1}(t+\mathrm{T}) = \frac{\lambda_1}{[x_n(t) - x_{n+1}(t)]} [\dot{x}_n(t) - \dot{x}_{n+1}(t)] \qquad (4.38)$$

for n = 1,2,3,...

As before, by assuming the parameters are such that the traffic stream is stable, this equation can be integrated yielding the steady-state relation for speed and concentration:

$$u = \lambda_1 \ln (k_j / k) \qquad (4.39)$$

and for steady-state flow and concentration:

$$q = \lambda_1 k ln(k_j / k) \qquad (4.40)$$

where again it is assumed that for u=0, the spacing is equal to an effective vehicle length, $L = k^{-1}$. These relations for steady-state flow are identical to those obtained from considering the traffic stream to be approximated by a continuous compressible fluid (see Chapter 5) with the property that disturbances are propagated with a constant speed with respect to the moving medium (Greenberg 1959). For our non-linear car following equation, infinitesimal disturbances are propagated with speed λ_1 . This is consistent with the earlier discussion regarding the speed of propagation of a disturbance per vehicle pair.

It can be shown that if the propagation time, T_0, is directly proportional to spacing (i.e., $T_0 \approx S$), Equations 4.39 and 4.40 are obtained where the constant ratio S / T_o is identified as the constant λ_1.

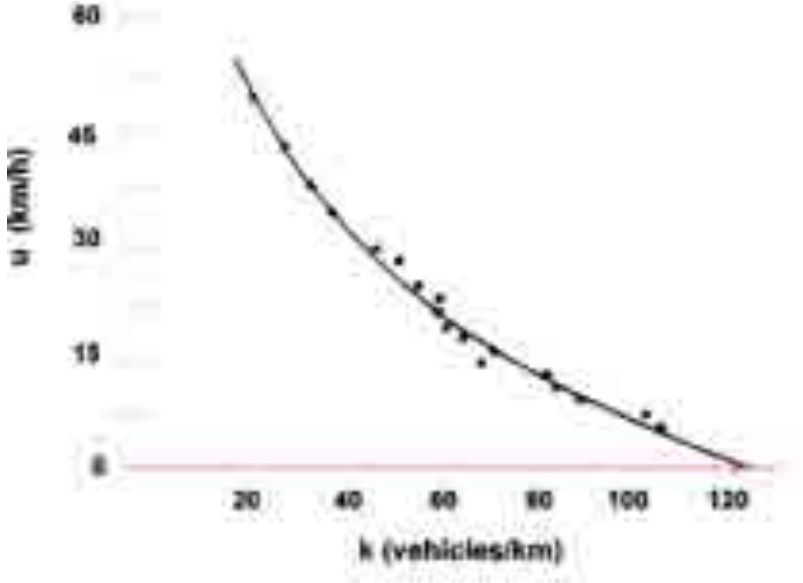
These two approaches are not analogous. In the fluid analogy case, the speed-spacing relationship is 'followed' at every instant before, during, and after a disturbance. In the case of car following during the transition phase, the speed-spacing, and therefore the flow-concentration relationship, does not describe the state of the traffic stream.

A solution to any particular set of equations for the motion of a traffic stream specifies departures from the steady-state. This is not the case for simple headway models or hydro-dynamical approaches to single-lane traffic flow because in these cases any small speed change, once the disturbance arrives, each vehicle instantaneously relaxes to the new speed, at the 'proper' spacing.

This emphasizes the shortcoming of these alternate approaches. They cannot take into account the behavioral and physical aspects of disturbances. In the case of car following models, the initial phase of a disturbance arrives at the nth vehicle downstream from the vehicle initiating the speed change at a time $(n\text{-}1)T$ seconds after the onset of the fluctuation. The time it takes vehicles to reach the changed speed depends on the parameter λ, for the linear model, and λ_1, for the non-linear model, subject to the restriction that $\lambda^{-1} > T$ or $\lambda_1 < S/T$, respectively.

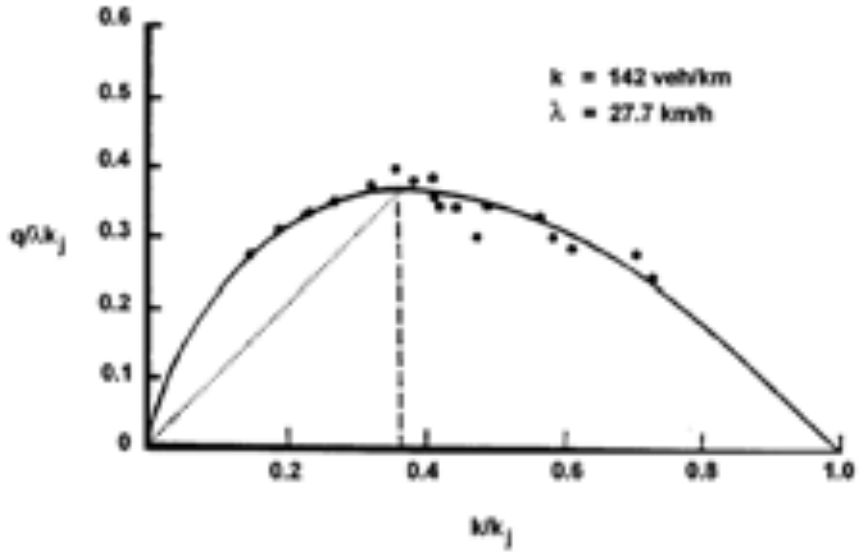
These restrictions assure that the signal speed can never precede the initial phase speed of a disturbance. For the linear case, the restriction is more than satisfied for an asymptotic stable traffic stream. For small speed changes, it is also satisfied for the non-linear model by assuming that the stability criteria results for the linear case yields a bound for the stability in the non-linear case. Hence, the inequality $\lambda T/S^* < 0.5$ provides a sufficient stability condition for the non-linear case, where S^* is the minimum spacing occurring during a transition from one steady-state to another.

Before discussing a more general form for the sensitivity coefficient (i.e., Equation 4.37), the same reported data (Greenberg 1959) plotted in Figures 4.9 and 4.10 are graphed in Figures 4.11 and 4.12 together with the steady-state relations (Equations 4.39 and 4.40 obtained from the non-linear model, Equation 4.38). The fit of the data to the steady-state relation via the method of "least squares" is good and the resulting values for λ_1 and k_j are 27.7 km/h and 142 veh/km, respectively. Assuming that this data is a representative sample of this facility's traffic, the value of 27.7 km/h is an estimate not only of the sensitivity coefficient for the non-linear car following model but it is the 'characteristic speed' for the roadway under consideration (i.e., the speed of the traffic stream which maximizes the flow).



Note: The curve corresponds to a "least squares" fit of Equation 4.39 to the data (Greenberg 1959).

Figure 4.11
Speed Versus Vehicle Concentration (Gazis et al. 1959).



Note: The curve corresponds to Equation 4.40 where parameters are those from the "fit" obtained in Figure 4.11.

Figure 4.12
Normalized Flow Versus Normalized Vehicle Concentration (Edie et al. 1963).

The corresponding vehicle concentration at maximum flow, i.e., when $u = \lambda_1$, is $e^{-1} k_j$. This predicts a roadway capacity of $\lambda_1 e^{-1} k_j$ of about ≈1400 veh/h for the Lincoln Tunnel. A noted undesirable property of Equation 4.40 is that the tangent *dq/dt* is infinite at $k = 0$, whereas a linear relation between flow and concentration would more accurately describe traffic near zero concentration. This is not a serious defect in the model since car following models are not applicable for low concentrations where spacings are large and the coupling between vehicles are weak. However, this property of the model did suggest the following alternative form (Edie 1961) for the gain factor,

$$\lambda = \lambda_2 \dot{x}_{n+1}(t+T)/[x_n(t)-x_{n+1}(t)]^2$$

This leads to the following expression for a car following model:

$$\ddot{x}_{n+1}(t+T) = \frac{\lambda_2 \dot{x}_{n+1}(t+T)}{[x_n(t)-x_{n+1}(t)]^2}[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \qquad (4.41)$$

As before, this can be integrated giving the following steady-state equations:

$$U = U_f e^{-k/k_m} \qquad (4.42)$$

and

$$q = U_f k e^{-k/k_m} \qquad (4.43)$$

where U_f is the "free mean speed", i.e., the speed of the traffic stream near zero concentration and k_m is the concentration when the flow is a maximum. In this case the sensitivity coefficient, λ_2 can be identified as k_m^{-1}. The speed at optimal flow is $e^{-1}U_f$ which, as before, corresponds to the speed of propagation of a disturbance with respect to the moving traffic stream. This model predicts a finite speed, U_f, near zero concentration.

Ideally, this speed concentration relation should be translated to the right in order to more completely take into account observations that the speed of the traffic stream is independent of vehicle concentration for low concentrations, .i.e.

$$U = U_f \quad for \quad 0 \le k \le k_f \qquad (4.44)$$

and

$$U = U_f \exp -\left[\frac{k - k_f}{k_m}\right] \qquad (4.45)$$

where k_f corresponds to a concentration where vehicle to vehicle interactions begin to take place so that the stream speed begins to decrease with increasing concentration. Assuming that interactions take place at a spacing of about 120 m, k_f would have a value of about 8 veh/km. A "kink" of this kind was introduced into a linear model for the speed concentration relationship (Greenshields 1935).

Greenshields' empirical model for a speed-concentration relation is given by

$$U = U_f(1-k/k_j) \qquad (4.46)$$

where U_f is a "free mean speed" and k_j is the jam concentration.

It is of interest to question what car following model would correspond to the above steady-state equations as expressed by Equation 4.46. The particular model can be derived in the following elementary way (Gazis et al. 1961). Equation 4.46 is rewritten as

$$U = U_f(1-L/S) \qquad (4.47)$$

Differentiating both sides with respect to time obtains

$$\dot{U} = (U_f L/S^2)\dot{S} \qquad (4.48)$$

which after introduction of a time lag is for the *(n+1)* vehicle:

$$\ddot{x}_{n+1}(t+T) = \frac{U_f L}{[x_n(t)-x_{n+1}(t)]^2}[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \qquad (4.49)$$

The gain factor is:

$$\frac{U_f L}{[x_n(t)-x_{n+1}(t)]^2} \tag{4.50}$$

The above procedure demonstrates an alternate technique at arriving at stimulus response equations from relatively elementary considerations. This method was used to develop early car following models (Reuschel 1950; Pipes 1951). The technique does pre-suppose that a speed-spacing relation reflects detailed psycho-physical aspects of how one vehicle follows another. To summarize the car-following equation considered, we have:

$$\ddot{x}_{n+1}(t+T) = \lambda[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \tag{4.51}$$

where the factor, λ, is assumed to be given by the following:

- A constant, $\lambda = \lambda_0$;
- A term inversely proportional to the spacing, $\lambda = \lambda_1/S$;
- A term proportional to the speed and inversely proportional to the spacing squared, $\lambda = \lambda_2 U/S^2$; and
- A term inversely proportional to the spacing squared, $\lambda = \lambda_3 / S^2$.

These models can be considered to be special cases of a more general expression for the gain factor, namely:

$$\lambda = a_{\ell,m}\dot{x}^m_{n+1}(t+T)/[x_n(t)-x_{n+1}(t)]^\ell \tag{4.52}$$

where $a_{\ell,m}$ is a constant to be determined experimentally. Model specification is to be determined on the basis of the degree to which it presents a consistent description of actual traffic phenomena. Equations 4.51 and 4.52 provide a relatively broad framework in so far as steady-state phenomena is concerned (Gazis et al. 1961).

Using these equations and integrating over time we have

$$f_m(U) = a \cdot f_l(S)+b \tag{4.53}$$

where, as before, U is the steady-state speed of the traffic stream, S is the steady-state spacing, and a and b are appropriate constants consistent with the physical restrictions and where $f_p(x)$, ($p = m$ or ℓ), is given by

$$f_p(x) = x^{1-p} \tag{4.54}$$

for $p \neq 1$ and

$$f_p(x) = \ell n x \tag{4.55}$$

for $p = 1$. The integration constant b is related to the "free mean speed" or the "jam concentration" depending on the specific values of m and ℓ. For $m > 1$, $\ell \neq 1$, or $m = 1$, $\ell > 1$

$$b = f_m(U_f) \tag{4.56}$$

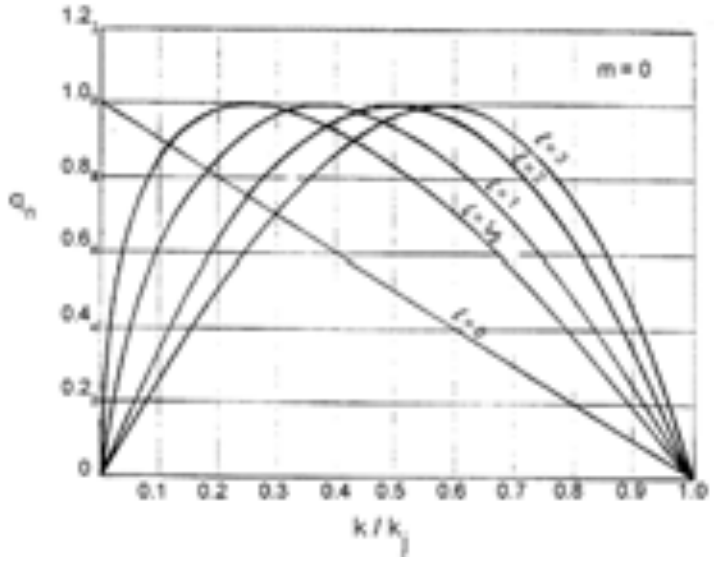
and

$$b = -af_l(L) \tag{4.57}$$

for all other combinations of m and ℓ, except $\ell < 1$ and $m = 1$.

For those cases where $\ell < 1$ and $m = 1$ it is not possible to satisfy either the boundary condition at $k = 0$ or k_j and the integration constant can be assigned arbitrarily, e.g., at k_m, the concentration at maximum flow or more appropriately at some 'critical' concentration near the boundary condition of a "free speed" determined by the "kink" in speed-concentration data for the particular facility being modeled. The relationship between k_m and k_j is a characteristic of the particular functional or model being used to describe traffic flow of the facility being studied and not the physical phenomenon involved. For example, for the two models given by $\ell = 1$, $m = 0$, and $\ell = 2$, m = 0, maximum flow occurs at a concentration of $e^{-1} k_j$ and $k_j / 2$, respectively. Such a result is not physically unrealistic. Physically the question is whether or not the measured value of q_{max} occurs at or near the numerical value of these terms, i.e., $k_m = e^{-1}k_j$ or $k_j/2$ for the two examples cited.

Using Equations 4.53, 4.54, 4.55, 4.56, 4.57, and the definition of steady-state flow, we can obtain the relationships between speed, concentration, and flow. Several examples have been given above. Figures 4.13 and 4.14 contain these and additional examples of flow versus concentration relations for various



Note: Normalized flow versus normalized concentration corresponding to the steady-state solution of Equations 4.51 and 4.52 for m=1 and various values of ℓ.

Figure 4.13
Normalized Flow Versus Normalized Concentration (Gazis et al. 1963).

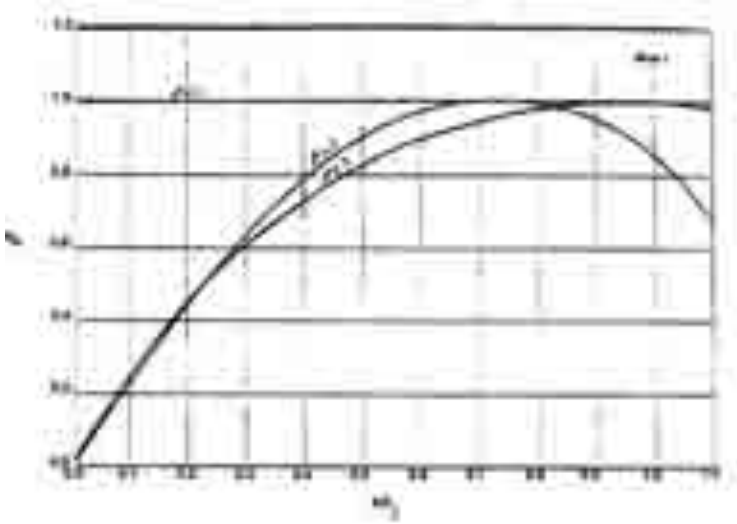


Figure 4.14
Normalized Flow versus Normalized Concentration Corresponding to the Steady-State Solution of Equations 4.51 and 4.52 for m=1 and Various Values of ℓ (Gazis 1963).

values of ℓ and m. These flow curves are normalized by letting $q_n = q/q_{max}$, and $k_n = k/k_j$.

It can be seen from these figures that most of the models shown here reflect the general type of flow diagram required to agree with the qualitative descriptions of steady-state flow. The spectrum of models provided are capable of fitting data like that shown in Figure 4.9 so long as a suitable choice of the parameters is made.

The generalized expression for car following models, Equations 4.51 and 4.52, has also been examined for non-integral values for m and ℓ (May and Keller 1967). Fittingdata obtained on the Eisenhower Expressway in Chicago they proposed a model with $m = 0.8$ and $\ell = 2.8$. Various values for m and ℓ can be identified in the early work on steady-state flow and car following .

The case $m = 0$, $\ell = 0$ equates to the "simple" linear car following model. The case $m = 0$, $\ell = 2$ can be identified with a model developed from photographic observations of traffic flow made in 1934 (Greenshields 1935). This model can also be developed considering the perceptual factors that are related to the car following task (Pipes and Wojcik 1968; Fox and Lehman 1967; Michaels 1963). As was discussed earlier, the case for $m = 0$, $\ell = 1$ generates a steady-state relation that can be developed bya fluid flow analogy to traffic (Greenberg 1959) and led to the reexamination of car following experiments and the hypothesis that drivers do not have a constant gain factor to a given relative-speed stimulus but rather that it varies inversely with the vehicle spacing, i.e., $m= 0$, $\ell =1$ (Herman et al. 1959). A generalized equation for steady-state flow (Drew 1965) and subsequently tested on the Gulf Freeway in Houston, Texas led to a model where $m = 0$ and $\ell = 3/2$.

As noted earlier, consideration of a "free-speed" near low concentrations led to the proposal and subsequent testing of the model $m = 1$, $\ell = 2$ (Edie 1961). Yet another model, $m = 1$, $\ell = 3$ resulted from analysis of data obtained on the Eisenhower Expressway in Chicago (Drake et al. 1967). Further analysis of this model together with observations suggest that the sensitivity coefficient may take on different values above a lane flow of about 1,800 vehicles/hr (May and Keller 1967).

4.4 Experiments And Observations

This section is devoted to the presentation and discussion of experiments that have been carried out in an effort to ascertain whether car following models approximate single lane traffic characteristics. These experiments are organized into two distinct categories.

The first of these is concerned with comparisons between car following models and detailed measurements of the variables involved in the driving situation where one vehicle follows another on an empty roadway. These comparisons lead to a quantitative measure of car following model estimates for the specific parameters involved for the traffic facility and vehicle type used.

The second category of experiments are those concerned with the measurement of macroscopic flow characteristics: the study of speed, concentration, flow and their inter-relationships for vehicle platoons and traffic environments where traffic is channeled in a single lane. In particular, the degree to which this type of data fits the analytical relationships that have been derived from car following models for steady-state flow are examined.

Finally, the degree to which any specific model of the type examined in the previous section is capable of representing a consistent framework from both the microscopic and macroscopic viewpoints is examined.

4.4.1 Car Following Experiments

The first experiments which attempted to make a preliminary evaluation of the linear car following model were performed a number of decades ago (Chandler et al. 1958; Kometani and Sasaki 1958). In subsequent years a number of different tests with varying objectives were performed using two vehicles, three vehicles, and buses. Most of these tests were conducted on test track facilities and in vehicular tunnels.

In these experiments, inter-vehicle spacing, relative speed, speed of the following vehicle, and acceleration of the following vehicles were recorded simultaneously together with a clock signal to assure synchronization of each variable with every other.

These car following experiments are divided into six specific categories as follows:

1) *Preliminary Test Track Experiments*. The first experiments in car following were performed by (Chandler et al. 1958) and were carried out in order to obtain estimates of the parameters in the linear car following model and to obtain a preliminary evaluation of this model. Eight male drivers participated in the study which was conducted on a one-mile test track facility.

2) *Vehicular Tunnel Experiments.* To further establish the validity of car following models and to establish estimates, the parameters involved in real operating environments where the traffic flow characteristics were well known, a series of experiments were carried out in the Lincoln, Holland, and Queens Mid-Town Tunnels of New York City. Ten different drivers were used in collecting 30 test runs.

3) *Bus Following Experiments.* A series of experiments were performed to determine whether the dynamical characteristics of a traffic stream changes when it is composed of vehicles whose performance measures are significantly different than those of an automobile. They were also performed to determine the validity and measure parameters of car following models when applied to heavy vehicles. Using a 4 kilometer test track facility and 53-passenger city buses, 22 drivers were studied.

4) *Three Car Experiments.* A series of experiments were performed to determine the effect on driver behavior when there is an opportunity for next-nearest following and of following the vehicle directly ahead. The degree to which a driver uses the information that might be obtained from a vehicle two ahead was also examined. The relative spacings and the relative speeds between the first and third vehicles and the second and third vehicles together with the speed and acceleration of the third vehicle were recorded.

5) *Miscellaneous Experiments.* Several additional car following experiments have been performed and reported on as follows:

a) *Kometani and Sasaki Experiments.* Kometani and Sasaki conducted and reported on a series of experiments that were performed to evaluate the effect of an additional term in the linear car following equation. This term is related to the acceleration of the lead vehicle. In particular, they investigated a model rewritten here in the following form:

$$\ddot{x}_{n+1}(t+T) = \lambda[\dot{x}_n(t) - \dot{x}_{n+1}(t)] + \gamma\ddot{x}_n(t) \qquad (4.58)$$

This equation attempts to take into account a particular driving phenomenon, where the driver in a particular state realizes that he should maintain a non-zero acceleration even though the relative speed has been reduced to zero or near zero. This situation was observed in several cases in tests carried out in the vehicular tunnels - particularly when vehicles were coming to a stop. Equation 4.58 above allows for a non-zero acceleration when the relative speed is zero. A value of γ near one would indicate an attempt to nearly match the acceleration of the lead driver for such cases. This does not imply that drivers are good estimators of relative acceleration. The conjecture here is that by pursuing the task where the lead driver is undergoing a constant or near constant acceleration maneuver, the driver becomes aware of this qualitatively after nullifying out relative speed - and thereby shifts the frame of reference. Such cases have been incorporated into models simulating the behavior of bottlenecks in tunnel traffic (Helly 1959).

b) *Experiments of Forbes et al.* Several experiments using three vehicle platoons were reported by Forbes et al. (1957). Here a lead vehicle was driven by one of the experimenters while the second and third vehicles were driven by subjects. At predetermined locations along the roadway relatively severe acceleration maneuvers were executed by the lead vehicle. Photographic equipment recorded these events with respect to this moving reference together with speed and time. From these recordings speeds and spacings were calculated as a function of time. These investigators did not fit this data to car following models. However, a partial set of this data was fitted to vehicle following models by another investigator (Helly 1959). This latter set consisted of six

tests in all, four in the Lincoln Tunnel and two on an open roadway.

c) *Ohio State Experiments.* Two different sets of experiments have been conducted at Ohio State University. In the first set a series of subjects have been studied using a car following simulator (Todosiev 1963). An integral part of the simulator is an analog computer which could program the lead vehicle for many different driving tasks. The computer could also simulate the performance characteristics of different following vehicles. These experiments were directed toward understanding the manner in which the following vehicle behaves when the lead vehicle moves with constant speed and the measurement of driver thresholds for changes in spacing, relative speed, and acceleration. The second set of experiments were conducted on a level two-lane state highway operating at low traffic concentrations (Hankin and Rockwell 1967). In these experiments the purpose was "to develop an empirically based model of car following which would predict a following car's acceleration and change in acceleration as a function of observed dynamic relationships with the lead car." As in the earlier car following experiments, spacing and relative speed were recorded as well as speed and acceleration of the following vehicle.

d) *Studies by Constantine and Young.* These studies were carried out using motorists in England and a photographic system to record the data (Constantine and Young 1967). The experiments are interesting from the vantage point that they also incorporated a second photographic system mounted in the following vehicle and directed to the rear so that two sets of car following data could be obtained simultaneously. The latter set collected information on an unsuspecting motorist. Although accuracy is not sufficient, such a system holds promise.

4.4.1.1 Analysis of Car Following Experiments

The analysis of recorded information from a car following experiment is generally made by reducing the data to numerical values at equal time intervals. Then, a correlation analysis is carried out using the linear car following model to obtain estimates of the two parameters, λ and T. With the data in discrete form, the time lag ,T, also takes on discrete values. The time lag or response time associated with a given driver is one for which the correlation coefficient is a maximum and typically falls in the range of 0.85 to 0.95.

The results from the preliminary experiments (Chandler et al. 1958) are summarized in Table 4.1 where the estimates are given for λ, their product; $C = \lambda T$, the boundary value for asymptotic stability; average spacing, $< S >$; and average speed, $< U >$. The average value of the gain factor is 0.368 sec^{-1}. The average value of λT is close to 0.5, the asymptotic stability boundary limit.

Table 4.1 Results from Car-Following Experiment

Driver	λ	$<U>$	$<S>$	λT
1	0.74 sec^{-1}	19.8 m/sec	36 m	1.04
2	0.44	16	36.7	0.44
3	0.34	20.5	38.1	1.52
4	0.32	22.2	34.8	0.48
5	0.38	16.8	26.7	0.65
6	0.17	18.1	61.1	0.19
7	0.32	18.1	55.7	0.72
8	0.23	18.7	43.1	0.47

Using the values for λ and the average spacing $<S>$ obtained for each subject a value of 12.1 m/sec (44.1 km/h) is obtained for an estimate of the constant $a_{I,\,0}$ (Herman and Potts 1959). This latter estimate compares the value λ for each driver with that driver's average spacing, $<S>$, since each driver is in somewhat different driving state. This is illustrated in Figure 4.15. This approach attempts to take into account the differences in the estimates for the gain factor λ or $a_{0,0}$, obtained for different drivers by attributing these differences to the differences in their respective average spacing. An alternate and more direct approach carries out the correlation analysis for this model using an equation which is the discrete form of Equation 4.38 to obtain a direct estimate of the dependence of the gain factor on spacing, $S(t)$.

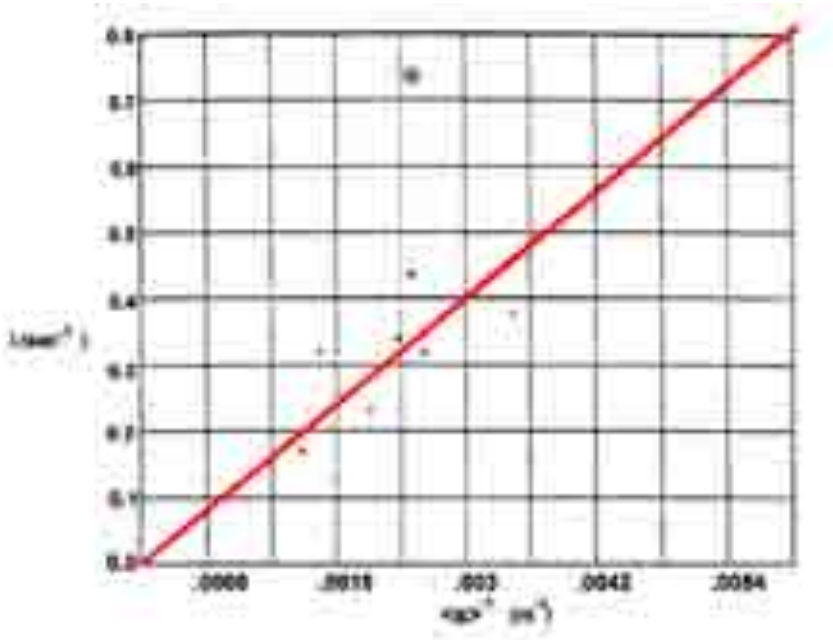


Figure 4.15
Sensitivity Coefficient Versus the Reciprocal of the Average Vehicle Spacing (Gazis et al. 1959).

Vehicular Tunnel Experiments. Vehicular tunnels usually have roadbeds that are limited to two lanes, one per direction. Accordingly, they consist of single-lane traffic where passing is prohibited. In order to investigate the reasonableness of the non-linear model a series of tunnel experiments were conducted. Thirty test runs in all were conducted: sixteen in the Lincoln Tunnel, ten in the Holland Tunnel and four in the Queens Mid-Town Tunnel. Initially, values of the parameters for the linear model were obtained, i.e., $\lambda = a_{0,0}$ and T. These results are shown in Figure 4.16 where the gain factor, $\lambda = a_{0,0}$ versus the time lag, T, for all of the test runs carried out in the vehicular tunnels. The solid curve divides the domain of this two parameter field into asymptotically stable and unstable regions.

It is of interest to note that in Figure 4.16 that many of the drivers fall into the unstable region and that there are drivers who have relatively large gain factors and time lags. Drivers with relatively slow responses tend to compensatingly have fast movement times and tend to apply larger brake pedal forces resulting in larger decelerations.

Such drivers have been identified, statistically, as being involved more frequently in "struck-from-behind accidents" (Babarik 1968; Brill 1972). Figures 4.17 and 4.18 graph the gain factor versus the reciprocal of the average vehicle spacing for the tests conducted in the Lincoln and Holland tunnels, respectively. Figure 4.17, the gain factor, λ, versus the reciprocal of the average spacing for the Holland Tunnel tests. The straight line is a "least-squares" fit through the origin. The slope, which is an estimate of $a_{1,0}$ and equals 29.21 km/h. Figure 4.18 graphs the gain factor, λ, versus the reciprocal of the average spacing for the Lincoln Tunnel tests. The straight line is a "least-squares" fit through the origin. These results yield characteristic speeds, $a_{1,0}$, which are within ± 3km/h for these two similar facilities. Yet these small numeric differences for the parameter $a_{1,0}$ properly reflect known differences in the macroscopic traffic flow characteristics of these facilities.

The analysis was also performed using these test data and the non-linear reciprocal spacing model. The results are not strikingly different (Rothery 1968). Spacing does not change significantly within any one test run to provide a sensitive measure of the dependency of the gain factor on inter-vehicular spacing for any given driver (See Table 4.2). Where the variation in spacings were relatively large (e.g., runs 3, 11, 13, and 14) the results tend to support the spacing dependent model. This time-dependent analysis has also been performed for seven additional functions for the gain factor for the same fourteen

Table 4.2
Comparison of the Maximum Correlations obtained for the Linear and Reciprocal Spacing Models for the Fourteen Lincoln Tunnel Test Runs

Number	$r_{0,0}$	$r_{l,0}$	<*S*> (m)	σ*S* (m)
1	0.686	0.459	13.4	4.2
2	0.878	0.843	15.5	3.9
3	0.77	0.778	20.6	5.9
4	0.793	0.748	10.6	2.9
5	0.831	0.862	12.3	3.9
6	0.72	0.709	13.5	2.1
7	0.64	0.678	5.5	3.2

Number	$r_{0,0}$	$r_{l,0}$	<*S*> (m)	σ*S* (m)
8	0.865	0.881	19.9	3.4
9	0.728	0.734	7.6	1.8
10	0.898	0.898	10.7	2.3
11	0.89	0.966	26.2	6.2
12	0.846	0.835	18.5	1.3
13	0.909	0.928	18.7	8.8
14	0.761	0.79	46.1	17.6

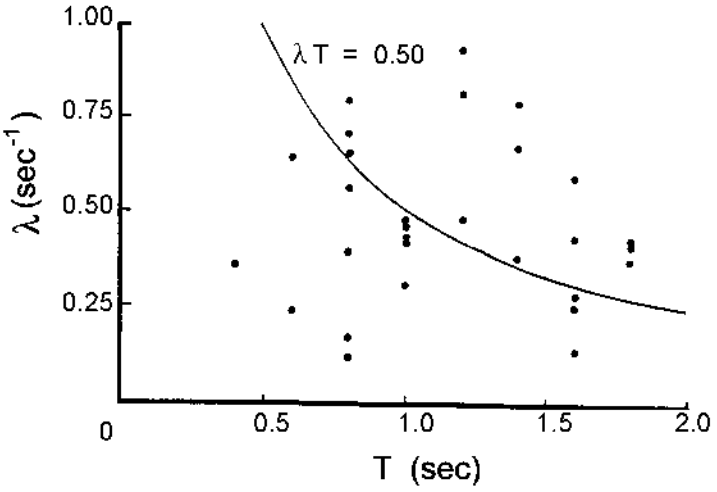
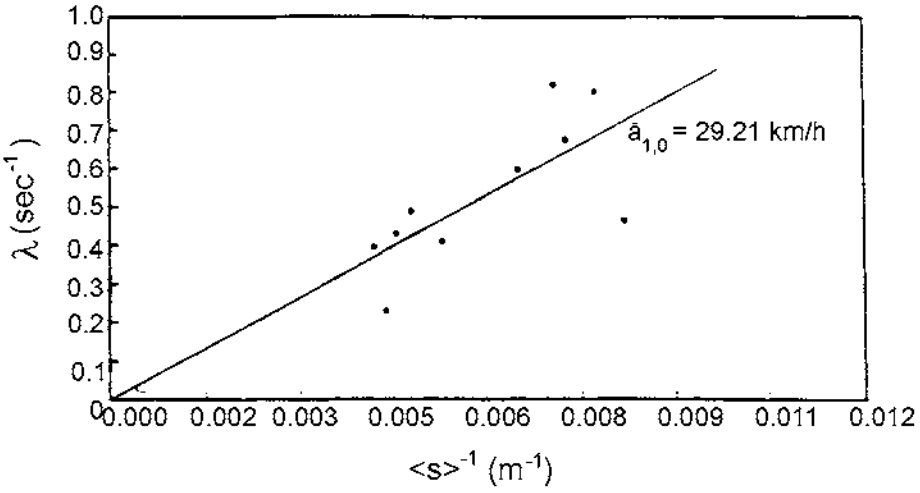
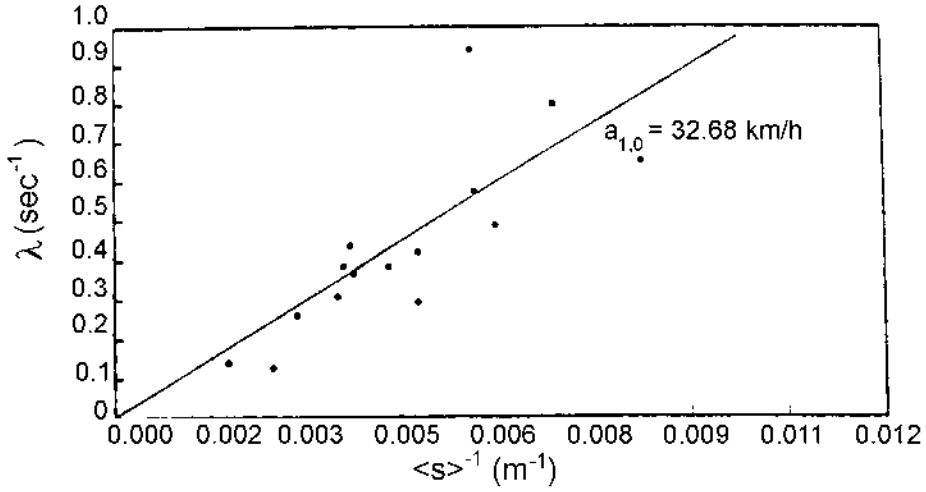


Figure 4.16
Gain Factor, λ, Versus the Time Lag, T, for All of the Test Runs (Rothery 1968).



Note: The straight line is a "least-squares" fit through the origin. The slope, which is an estimate of $a_{1,0}$, equals 29.21 km/h.

Figure 4.17
Gain Factor, λ, Versus the Reciprocal of the
Average Spacing for Holland Tunnel Tests (Herman and Potts 1959).



Note: The straight line is a "least-squares" fit through the origin. The slope, is an estimate of $a_{1,0}$, equals 32.68 km/h.

Figure 4.18
Gain Factor, λ ,Versus the Reciprocal of the
Average Spacing for Lincoln Tunnel Tests (Herman and Potts 1959).

little difference from one model to the other. There are definite trends however. If one graphs the correlation coefficient for a given ℓ, say ℓ=1 versus m; 13 of the cases indicate the best fits are with m = 0 or 1. Three models tend to indicate marginal superiority; they are those given by (ℓ=2; m=1), (ℓ=1; m=0) and (ℓ=2; m=0).

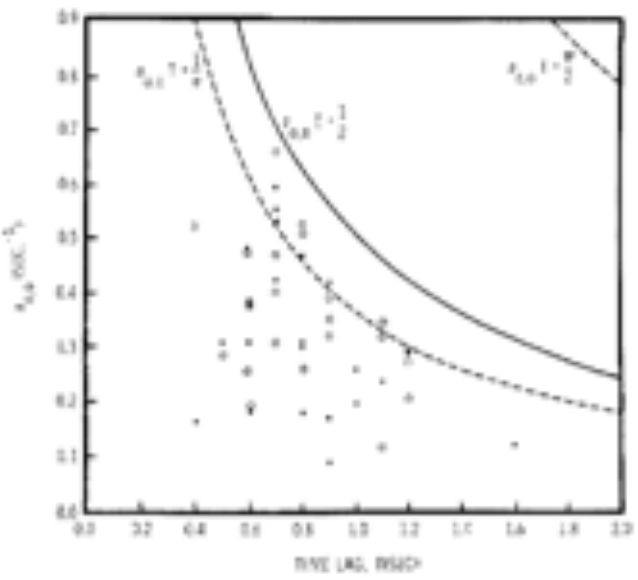
Bus Following Experiments. For each of the 22 drivers tested, the time dependent correlation analysis was carried out for the linear model (ℓ=0; m=0), the reciprocal spacing model (ℓ=1; m=0), and the speed, reciprocal-spacing-squared model (ℓ=2; m=1). Results similar to the Tunnel analysis were obtained: high correlations for almost all drivers and for each of the three models examined (Rothery et al. 1964).

The correlation analysis provided evidence for the reciprocal spacing effect with the correlation improved in about 75 percent of the cases when that factor is introduced and this model (ℓ=1; m=0) provided the best fit to the data. The principle results of the analysis are summarized in Figure 4.19 where the sensitivity coefficient $a_{0,0}$ versus the time lag, T, for the bus following experiments are shown. All of the data points obtained in these results fall in the asymptotically stable

region, whereas in the previous automobile experiments approximately half of the points fell into this region. In Figure 4.19, the sensitivity coefficient, $a_{0,0}$, versus the time lag, T, for the bus following experiments are shown. Some drivers are represented by more than one test. The circles are test runs by drivers who also participated in the ten bus platoon experiments. The solid curve divides the graph into regions of asymptotic stability and instability. The dashed lines are boundaries for the regions of local stability and instability.

Table 4.3
Maximum Correlation Comparison for Nine Models, $a_{\ell,m}$, for Fourteen Lincoln Tunnel Test Runs.

Driver	r(0,0)	r(1,-1)	r(1,0)	r(1,1)	r(1,2)	r(2,-1)	r(2,0)	r(2,1)	r(2,2)
1	0.686	0.408	0.459	0.693	0.721	0.310	0.693	0.584	0.690
2	0.878	0.770	0.843	0.847	0.746	0.719	0.847	0.827	0.766
3	0.770	0.757	0.778	0.786	0.784	0.726	0.786	0.784	0.797
4	0.793	0.730	0.748	0.803	0.801	0.685	0.801	0.786	0.808
5	0.831	0.826	0.862	0.727	0.577	0.805	0.728	0.784	0.624
6	0.720	0.665	0.709	0.721	0.709	0.660	0.720	0.713	0.712
7	0.640	0.470	0.678	0.742	0.691	0.455	0.745	0.774	0.718
8	0.865	0.845	0.881	0.899	0.862	0.818	0.890	0.903	0.907
9	0.728	0.642	0.734	0.773	0.752	0.641	0.773	0.769	0.759
10	0.898	0.890	0.898	0.893	0.866	0.881	0.892	0.778	0.865
11	0.890	0.952	0.966	0.921	0.854	0.883	0.921	0.971	0.940
12	0.846	0.823	0.835	0.835	0.823	0.793	0.835	0.821	0.821
13	0.909	0.906	0.923	0.935	0.927	0.860	0.935	0.928	0.936
14	0.761	0.790	0.790	0.771	0.731	0.737	0.772	0.783	0.775



Note: For bus following experiments - Some drivers are represented by more than one test. The circles are test runs by drivers who also participated in the ten bus platoons experiments. The solid curve divides the graph into regions of asymptotic stability and instability. The dashed lines are boundaries for the regions of local stability and instability.

Figure 4.19
Sensitivity Coefficient, $a_{0,0}$, Versus the Time Lag, T (Rothery et al. 1964).

The results of a limited amount of data taken in the rain suggest that drivers operate even more stably when confronted with wet road conditions. These results suggest that buses form a highly stable stream of traffic.

The time-independent analysis for the reciprocal-spacing model and the speed-reciprocal-spacing-squared model uses the time dependent sensitivity coefficient result, $a_{0,0}$, the average speed, $<U>$, and the average spacing, $<S>$, for eachof the car following test cases in order to form estimates of $a_{1,0}$ and $a_{2,1}$, i.e. by fitting

$$a_{0,0} = \frac{a_{1,0}}{<S>}$$

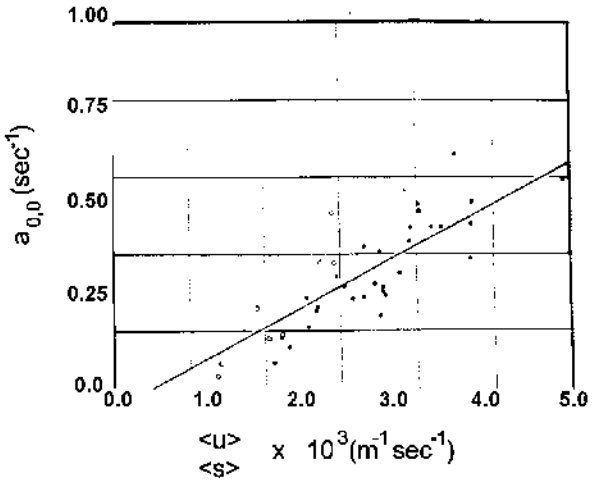
and

$$a_{0,0} = a_{2,1}\frac{<U>}{<S>^2}$$

respectively (Rothery et al. 1964).

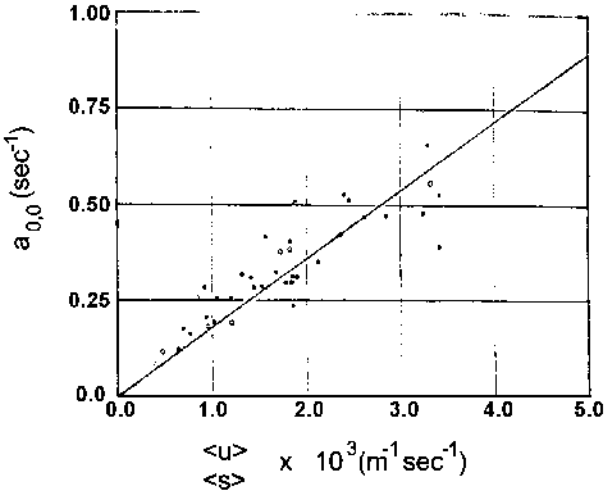
Figures 4.20 and 4.21 graph the values of $a_{0,0}$ for all test runs versus $<S>^{-1}$ and $<U>$ $<S>^{-2}$, respectively. In Figure 4.20, the sensitivity coefficient versus the reciprocal of the average spacing for each bus following experiment, and the "least-squares" straight line are shown. The slope of this regression is an estimate of the reciprocal spacing sensitivity coefficient. The solid dots and circles are points for two different test runs.

In Figure 4.21, the sensitivity coefficient versus the ratio of the average speed to the square of the average spacing for each bus following experiment and the "least-square" straight line are shown. The slope of this regression is an estimate of the speed-reciprocal spacing squared sensitivity coefficient. The solid dots and circles are data points for two different test runs. The slope of the straight line in each of these figures give an estimate of their respective sensitivity coefficient for the sample population. For the reciprocal spacing model the results indicate an estimate for $a_{1,0}$ = 52.8 ± .05 m/sec. (58 ± 1.61 km/h) and for the speed-reciprocal spacing squared model $a_{2,1}$ = 54.3 ± 1.86 m. The errors are one standard deviation.



Note: The sensitivity coefficient versus the reciprocal of the average spacing for each bus following experiment. The least squares straight line is shown. The slope of this regression is an estimate of the reciprocal spacing sensitivity coefficient. The solid dots and circles are data points for two different test runs.

Figure 4.20
Sensitivity Coefficient Versus the Reciprocal of the Average Spacing (Rothery et al. 1964).



Note: The sensitivity coefficient versus the ratio of the average speed to the square of the average spacing for each bus following experiment. The least squares straight line is shown. The slope of this regression is an estimate of the speed-reciprocal spacing squared sensitivity coefficient. The solid dots and circles are data points for two different test runs.

Figure 4.21
Sensitivity Coefficient Versus the Ratio of the Average Speed (Rothery et al. 1964).

Three Car Experiments. These experiments were carried out in an effort to determine, if possible, the degree to which a driver is influenced by the vehicle two ahead, i.e., next nearest interactions (Herman and Rothery 1965). The data collected in these experiments are fitted to the car following model:

$$\ddot{x}_{n+2}(t+T) = \lambda_1[\dot{x}_{n+1}(t)-\dot{x}_{n+2}(t)]+\lambda_2[\dot{x}_n(t)-\dot{x}_{n+2}(t)] \quad (4.61)$$

This equation is rewritten in the following form:

$$\ddot{x}_{n+2}(t+T) = \lambda_l\left\{[\dot{x}_{n+1}(t)-\dot{x}_{n+2}(t)]+\delta[\dot{x}_n(t)-\dot{x}_{n+2}(t)]\right\} \quad (4.62)$$

where

$$\delta = \lambda_2/\lambda_1$$

A linear regression analysis is then conducted for specific values of the parameter δ. For the case $\delta = 0$ there is nearest neighbor coupling only and for $\delta >> 1$ there is next nearest

neighbor coupling only. Using eight specific values of δ (0, 0.25, 0.50, 1, 5, 10, 100, and ∞) and a mean response time of 1.6 sec, a maximum correlation was obtained when δ = 0. However, if response times are allowed to vary, modest improvements can be achieved in the correlations.

While next nearest neighbor couplings cannot be ruled out entirely from this study, the results indicate that there are no significant changes in the parameters or the correlations when following two vehicles and that the stimulus provided by the nearest neighbor vehicle, i.e., the 'lead' vehicle, is the most significant. Models incorporating next nearest neighbor interactions have also been used in simulation models (Fox and Lehman 1967). The influence of including such interactions in simulations are discussed in detail by those authors.

Miscellaneous Car Following Experiments. A brief discussion of the results of three additional vehicle following experiments are included here for completeness.

The experiments of Kometani and Sasaki (1958) were car following experiments where the lead vehicle's task was closely approximated by: "accelerate at a constant rate from a speed *u* to a speed *u'* and then decelerate at a constant rate from the speed *u'* to a speed *u*." This type of task is essentially 'closed' since the external situation remains constant. The task does not change appreciably from cycle to cycle. Accordingly, response times can be reduced and even canceled because of the cyclic nature of the task.

By the driver recognizing the periodic nature of the task or that the motion is sustained over a period of time (≈ 13 sec for the acceleration phase and ≈ 3 sec for the deceleration phase) the driver obtains what is to him/her advanced information. Accordingly, the analysis of these experiments resulted in short response times ≈ 0.73 sec for low speed (20-40 km/h.) tests and ≈ 0.54 sec for high speed (40-80 km/h.) tests. The results also produced significantly large gain factors. All of the values obtained for each of the drivers for λT, exceeded the asymptotic stability limit. Significantly better fits of the data can be made using a model which includes the acceleration of the lead vehicle (See Equation 4.58) relative to the linear model which does not contain such a term. This is not surprising, given the task of following the lead vehicle's motion as described above.

A partial set of the experiments conducted by Forbes et al. (1958) were examined by Helly (1959), who fitted test runs to the linear vehicle model, Equation 4.41, by varying λ and T to minimize the quantity:

$$\sum_{j=1}^{N}[\dot{x}_n^{Exp.}(j.\delta t) - \dot{x}_n^{Theor.}(j.\delta t)]^2 \quad (4.63)$$

where the data has been quantitized at fixed increments of δt, $N\delta t$ is the test run duration, $\dot{x}_n su^{\text{Exp.}}(j.\delta t)$ is the experimentally measured values for the speed of the following vehicle at time $j\delta t$, and $\dot{x}_n^{\text{Theor.}}(j.\delta t)$ is the theoretical estimate for the speed of the following vehicle as determined from the experimentally measured values of the acceleration of the following vehicle and the speed of the lead vehicle *using the linear model*. These results are summarized in Table 4.4.

Ohio State Simulation Studies. From a series of experiments conducted on the Ohio State simulator, a relatively simple car following model has been proposed for *steady-state* car following (Barbosa 1961). The model is based on the concept of driver thresholds and can be most easily described by means of a 'typical' recording of relative speed versus spacing as

Table 4.4 Results from Car Following Experiments

Driver #	T	a_{00}	r_{max}
1	1.0	0.7	0.86
2	0.5	1.3	0.96
3	0.6	0.8	0.91
4	0.5	1.0	0.87
5	0.7	1.1	0.96
6	0.5	1.0	0.86

shown in Figure 4.22. At point "1," it is postulated that the driver becomes aware that he is moving at a higher speed than the lead vehicle and makes the decision to decelerate in order to avoid having either the negative relative speed becoming too large or the spacing becoming too small. At point "A," after a time lag, the driver initiates this deceleration and reduces the relative speed to zero. Since drivers have a threshold level below which relative speed cannot be estimated with accuracy, the driver continues to decelerate until he becomes aware of a positive relative speed because it either exceeds the threshold at this spacing or because the change in spacing has exceeded its threshold level. At point "2," the driver makes the decision to accelerate in order not to drift away from the lead vehicle. This decision is executed at point "B" until point "3" is reach and the cycle is more or less repeated. It was found that the arcs, e.g., AB, BC, etc. are "approximately parabolic" implying that accelerations can be considered roughly to be constant. These accelerations have been studied in detail in order to obtain estimates of relative speed thresholds and how they vary with respect to inter-vehicle spacing and observation times (Todosiev 1963). The results are summarized in Figure 4.23. This driving task, following a lead vehicle traveling at constant speed, was also studied using automobiles in a driving situation so that the pertinent data could be collected in a closer-to-reality situation and then analyzed (Rothery 1968).

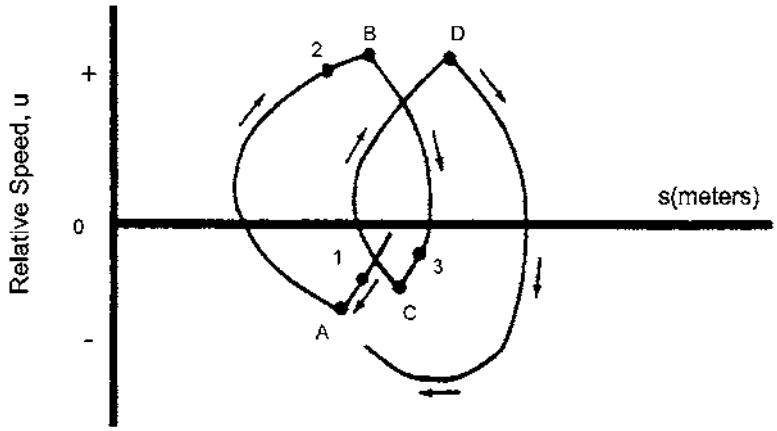


Figure 4.22
Relative Speed Versus Spacing (Rothery 1968).

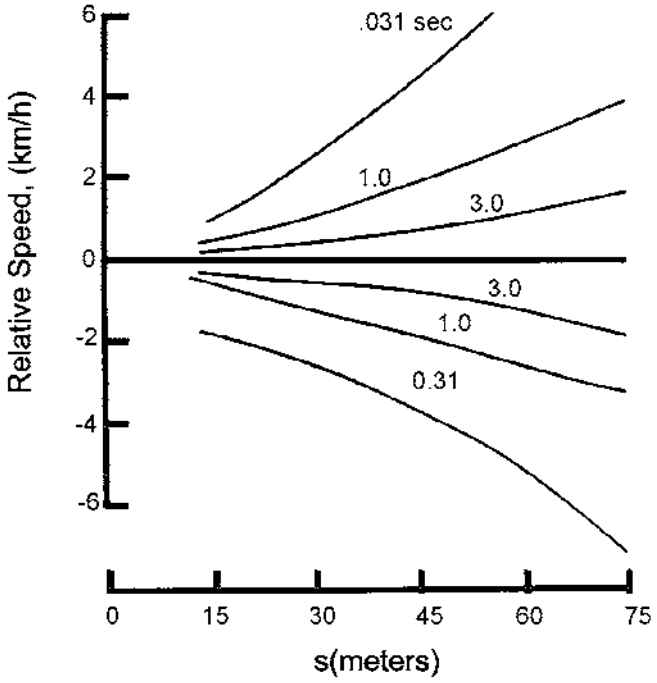


Figure 4.23
Relative Speed Thresholds Versus Inter-Vehicle Spacing for Various Values of the Observation Time. (Rothery 1968).

The interesting element in these latter results is that the character of the motion as exhibited in Figure 4.22 is much the same. However, the range of relative speeds at a given spacing that were recorded are much lower than those measured on the simulator. Of course the perceptual worlds in these two tests are considerably different. The three dimensional aspects of the test track experiment alone might provide sufficient additional cues to limit the subject variables in contrast to the two dimensional CRT screen presented to the 'driver' in the simulator. In any case, thresholds estimated in driving appear to be less than those measured via simulation.

Asymmetry Car Following Studies. One car following experiment was studied segment by segment using a model where the stimulus included terms proportional to deviations from the mean inter-vehicle spacing, deviations from the mean speed of the lead vehicle and deviations from the mean speed of the following car (Hankin and Rockwell 1967). An interesting result of the analysis of this model is that it implied an asymmetry in the response depending on whether the relative speed stimulus is positive or negative. This effect can be taken into account by rewriting our basic model as:

$$\dot{x}_{n+1}(t+\mathrm{T}) = \lambda_i[\dot{x}_n(t)-\dot{x}_{n+1}(t)] \qquad (4.64)$$

where $\lambda_i = \lambda_+$ or λ_- depending on whether the relative speed is greater or less than zero.

A reexamination of about forty vehicle following tests that were carried out on test tracks and in vehicular tunnels indicates, without exception, that such an asymmetry exists (Herman and Rothery 1965). The average value of λ_- is $\approx$10 percent greater than λ_+. The reason for this can partly be attributed to the fact that vehicles have considerably different capacities to accelerate and decelerate. Further, the degree of response is likely to be different for the situations where vehicles are separating compared to those where the spacing is decreasing. This effect creates a special difficulty with car following models as is discussed in the literature (Newell 1962; Newell 1965). One of

the principal difficulties is that in a cyclic change in the lead vehicle's speed - accelerating up to a higher speed and then returning to the initial speed, the asymmetry in acceleration and deceleration of the following car prevents return to the original spacing. With *n* such cycles, the spacing continues to increase thereby creating a drifting apart of the vehicles. A relaxation process needs to be added to the models that allows for this asymmetry and also allows for the return to the correct spacing.

4.4.2 Macroscopic Observations: Single Lane Traffic

Several data collections on single lane traffic have been carried out with the specific purpose of generating a large sample from which accurate estimates of the macroscopic flow characteristics could be obtained. With such a data base, direct comparisons can be made with microscopic, car following estimates - particularly when the car following results are obtained on the same facility as the macroscopic data is collected. One of these data collections was carried out in the Holland Tunnel (Edie et al. 1963). The resulting macroscopic flow data for this 24,000 vehicle sample is shown in Table 4.5.

The data of Table 4.5 is also shown in graphical form, Figures 4.24 and 4.25 where speed versus concentration and flow versus concentration are shown, respectively. In Figure 4.24, speed versus vehicle concentration for data collected in the Holland Tunnel is shown where each data point represents a speed class of vehicles moving with the plotted speed ± 1.61 m/sec. In Figure 4.25, flow versus vehicle concentration is shown; the solid points are the flow values derived from the speed classes assuming steady-state conditions. (See Table 4.5 and Figure 4.24.) Also included in Figure 4.25 are one-minute average flow values shown as encircled points. (See Edie et al. 1963). Using this macroscopic data set, estimates for three sensitivity coefficients are estimated for the particular car following models that appear to be of significance. These are: $a_{1,0}$, $a_{2,1}$, and $a_{2,0}$. These are sometimes referred to as the Reciprocal Spacing Model, Edie's Model, and Greenshields' Model, respectively. The numerical values obtained are shown and compared with the microscopic estimates from car following experiments for these same parameters.

The associated units for these estimates are ft/sec, ft^2/sec, and miles/car, respectively. As illustrated in this table, excellent agreement is obtained with the reciprocal spacing model. How well these models fit the macroscopic data is shown in Figure 4.26, where the speed versus vehicle concentration data is graphed together with the curves corresponding to the steady-state speed-concentration relations for the various indicated models. The data appears in Figure 4.24 and 4.25.

The curves are least square estimates. All three models provide a good estimate of the characteristic speed (i.e., the speed at optimum flow, namely 19, 24, and 23 mi/h for the reciprocal spacing, reciprocal spacing squared, and speed reciprocal spacing squared models, respectively).

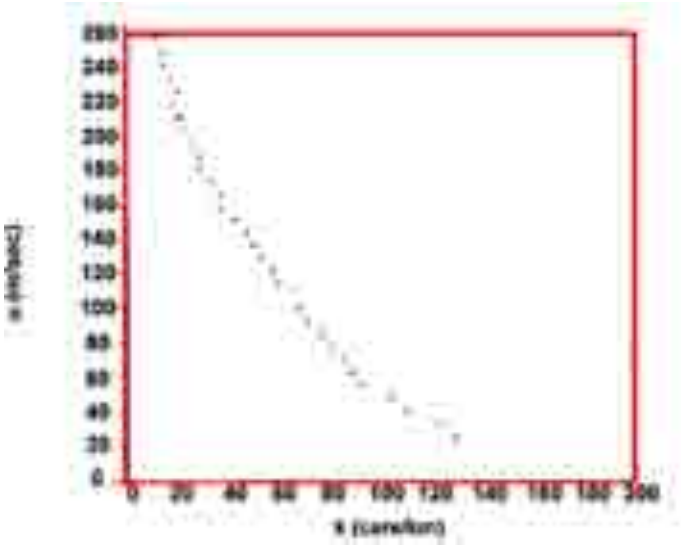
Edie's original motivation for suggesting the reciprocal spacing speed model was to attempt to describe low concentration, non-congested traffic. The key parameter in this model is the "mean free speed", i.e., the vehicular stream speed as the concentration goes to zero. The least squares estimate from the macroscopic data is 26.85 meters/second.

Edie also compared this model with the macroscopic data in the concentration range from zero to 56 vehicles/kilometer; the reciprocal spacing model was used for higher concentrations (Edie 1961). Of course, the two model fit is better than any one model fitted over the entire range, but marginally (Rothery 1968). Even though the improvement is marginal there is an apparent discontinuity in the derivative of the speed-concentration curve. This discontinuity is different than that which had previously been discussed in the literature. It had been suggested that there was an apparent break in the flow concentration curve near maximum flow where the flow drops suddenly (Edie and Foote 1958; 1960; 1961). That type of discontinuity suggests that the *u-k* curve is discontinuous.

However, the data shown in the above figures suggest that the curve is continuous and its derivative is not. If there is a discontinuity in the flow concentration relation near optimum flow it is considerably smaller for the Holland Tunnel than has been suggested for the Lincoln Tunnel. Nonetheless, the apparent discontinuity suggests that car following may be bimodal in character.

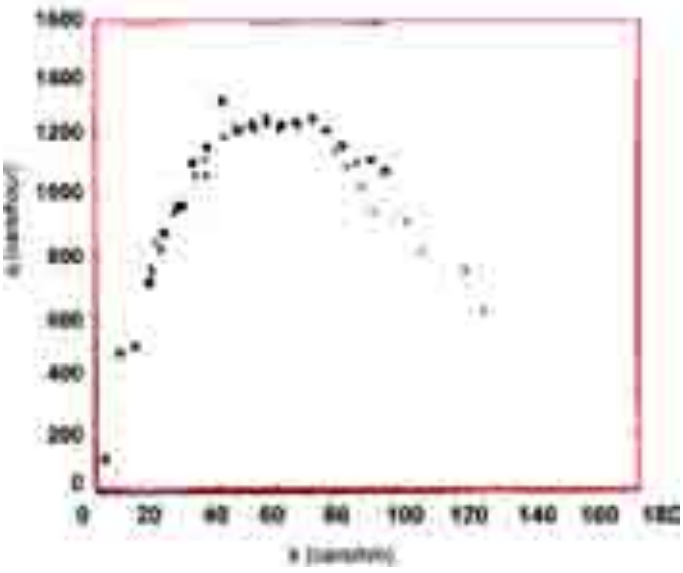
Table 4.5 Macroscopic Flow Data

Speed (m/sec)	Average Spacing (m)	Concentration (veh/km)	Number of Vehicles
2.1	12.3	80.1	22
2.7	12.9	76.5	58
3.3	14.6	67.6	98
3.9	15.3	64.3	125
4.5	17.1	57.6	196
5.1	17.8	55.2	293
5.7	18.8	52.6	436
6.3	19.7	50	656
6.9	20.5	48	865
7.5	22.5	43.8	1062
8.1	23.4	42	1267
8.7	25.4	38.8	1328
9.3	26.6	37	1273
9.9	27.7	35.5	1169
10.5	30	32.8	1096
11.1	32.2	30.6	1248
11.7	33.7	29.3	1280
12.3	33.8	26.8	1162
12.9`	43.2	22.8	1087
13.5	43	22.9	1252
14.1	47.4	20.8	1178
14.7	54.5	18.1	1218
15.3	56.2	17.5	1187
15.9	60.5	16.3	1135
16.5	71.5	13.8	837
17.1	75.1	13.1	569
17.7	84.7	11.6	478
18.3	77.3	12.7	291
18.9	88.4	11.1	231
19.5	100.4	9.8	169
20.1	102.7	9.6	55
20.7	120.5	8.1	56



Note: Each data point represents a speed class of vehicles moving with the plotted speed ± 1 ft/sec (See Table 4.4).

Figure 4.24
Speed Versus Vehicle Concentration (Edie et al. 1963).



Note: The solid points are the flow values derived from the speed classes assuming steady-state condition. Also included in Figure 4.25 are one minute average flow values shown as encircled points.

Figure 4.25
Flow Versus Vehicle Concentration (Edie et al. 1963).

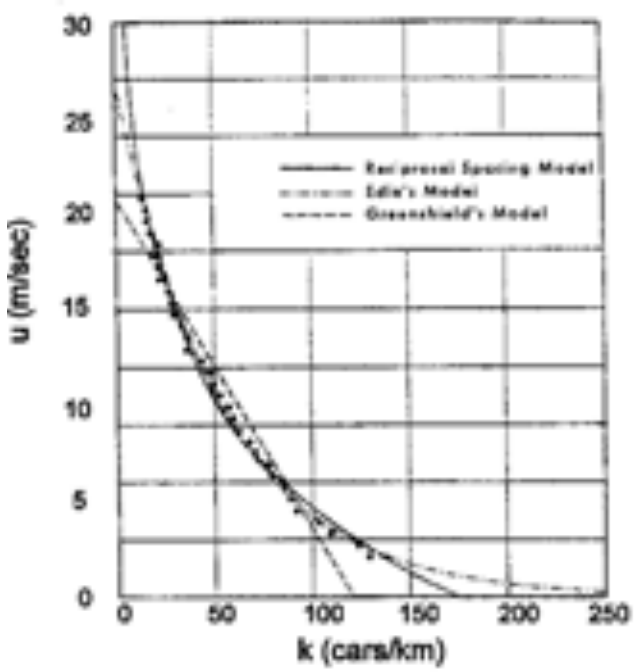


Figure 4.26
Speed Versus Vehicle Concentration (Rothery 1968).

Table 4.6
Parameter Comparison
(Holland Tunnel Data)

Parameters	Microscopic Estimates	Macroscopic Estimates
$a_{1,0}$	26.8	27.8
$a_{2,0}$	0.57	0.12
$a_{2,1}$	$(123)^{-1}$	$(54)^{-1}$

A totally different approach to modeling traffic flow variables which incorporates such discontinuities can be found in the literature. Navin (1986) and Hall (1987) have suggested that catastrophe theory (Thom 1975; Zeeman 1977) can be used as a vehicle for representing traffic relationships. Specifically, Navin followed the two regime approach proposed by Edie and cited above and first suggested that traffic relations can be represented using the cusp catastrophe. A serious attempt to apply such an approach to actual traffic data in order to represent flow variables without resorting to using two different expressions or two different sets of parameters to one expression has been made by Hall (1987). More recently, Acha-Daza and Hall (1994) have reported an analysis of freeway data using catastrophe theory which indicates that such an approach can effectively be applied to traffic flow. Macroscopic data has also been reported on single lane bus flow. Here platoons of ten buses were studied (Rothery et al. 1964).

Platoons of buses were used to quantify the steady-state stream properties and stability characteristics of single lane bus flow. Ideally, long chains of buses should be used in order to obtain the bulk properties of the traffic stream and minimize the end effects or eliminate this boundary effect by having the lead vehicle follow the last positioned vehicle in the platoon using a circular roadway. These latter type of experiments have been carried out at the Road Research Laboratory in England (Wardrop 1965; Franklin 1967).

In the platoon experiments, flow rates, vehicle concentration, and speed data were obtained. The average values for the speed and concentration data for the ten bus platoon are shown in Figure 4.28 together with the numerical value for the parameter

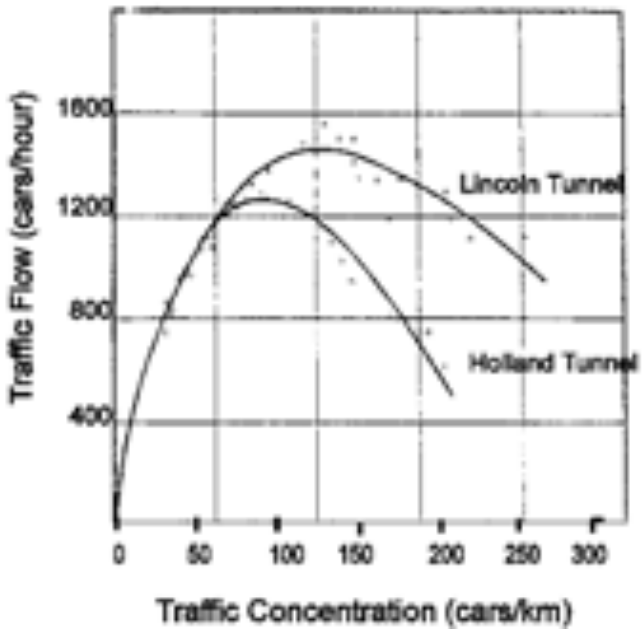


Figure 4.27
Flow Versus Concentration for the Lincoln and Holland Tunnels.

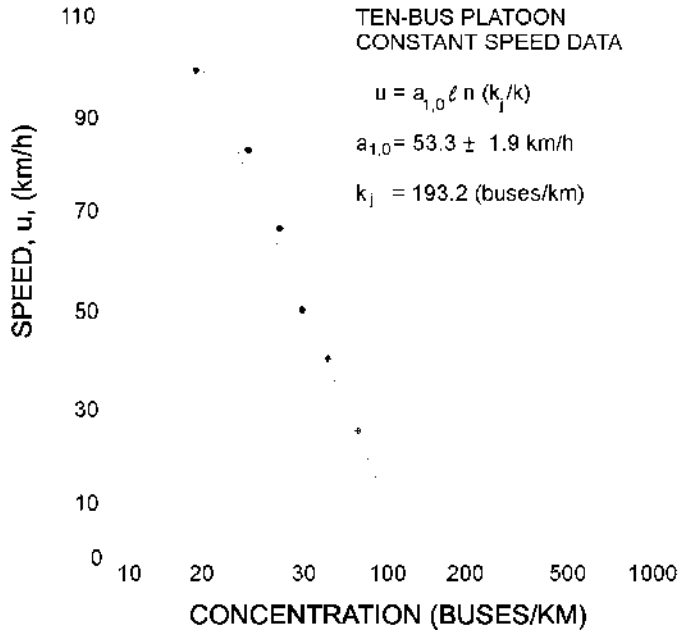


Figure 4.28
Average Speed Versus Concentration
for the Ten-Bus Platoon Steady-State Test Runs (Rothery 1968).

$a_{1,0}$ = 53 km/h which is to be compared to that obtained from the two bus following experiments discussed earlier namely, 58 km/h. Given these results, it is estimated that a single lane of standard size city buses is stable and has the capacity of over 65,000 seated passengers/hour. An independent check of this result has been reported (Hodgkins 1963). Headway times and speed of clusters of three or more buses on seven different highways distributed across the United States were measured and concluded that a maximum flow for buses would be approximately 1300 buses/hour and that this would occur at about 56 km/h.

4.5 Automated Car Following

All of the discussion in this chapter has been focused on manual car following, on what drivers do in following one or more other vehicles on a single lane of roadway. Paralleling these studies, research has also focused on developing controllers that would automatically mimic this task with specific target objectives in mind.

At the 1939 World's Fair, General Motors presented conceptually such a vision of automated highways where vehicles were controlled both longitudinally (car following) and laterally thereby freeing drivers to take on more leisurely activities as they moved at freeway speeds to their destinations. In the intervening years considerable effort has been extended towards the realization of this transportation concept. One prime motivation for such systems is that they are envisioned to provide more efficient utilization of facilities by increasing roadway capacity particularly in areas where constructing additional roadway lanes is undesirable and or impractical, and in addition, might improve safety levels. The concept of automated highways is one where vehicles would operate on both conventional roads under manual control and on specially instrumented guideways under automatic control. Here we are interested in automatic control of the car following task. Early research in this arena was conducted on both a theoretical and experimental basis and evaluated at General Motors Corporation (Gardels 1960; Morrison 1961; Flory et al. 1962), Ohio State University (Fenton 1968; Bender and Fenton 1969; Benton et al. 1971; Bender and Fenton 1970), Japan Governmental Mechanical Laboratory (Oshima et al. 1965), the Transportation Road Research Laboratory (Giles and Martin 1961; Cardew 1970), Ford Motor Corporation (Cro and Parker 1970) and the Japanese Automobile Research Institute (Ito 1973). During the past several decades three principal research studies in this arena stand out: a systems study of automated highway systems conducted at General Motors from 1971-1981, a long-range program on numerous aspects of automated highways conducted at The Ohio State University from 1964-1980, and the Program on Advanced Technology for the Highway (PATH) at the University of California, Berkeley from about 1976 to the present. Three overviews and detailed references to milestones of these programs can be found in the literature: Bender (1990), Fenton and Mayhan (1990), and Shladover et al. (1990), respectively.

The car following elements in these studies are focused on developing controllers that would replace driver behavior, carry out the car following task and would satisfy one or more performance and/or safety criteria. Since these studies have essentially been theoretical, they have by necessity required the difficult task of modeling vehicle dynamics. Given a controller for the driver element and a realistic model representation of vehicle dynamics a host of design strategies and issues have been addressed regarding inter-vehicular spacing control, platoon configurations, communication schemes, measurement and timing requirements, protocols, etc. Experimental verifications of these elements are underway at the present time by PATH.

4.6 Summary and Conclusions

Historically, the subject of car following has evolved over the past forty years from conceptual ideas to mathematical model descriptions, analysis, model refinements resulting from empirical testing and evaluation and finally extensions into advanced automatic vehicular control systems. These developments have been overlapping and interactive. There have been ebbs and flows in both the degree of activity and progress made by numerous researchers that have been involved in the contributions made to date.

The overall importance of the development of the subject of car following can be viewed from five vantage points, four of which this chapter has addressed in order. First, it provides a mathematical model of a relative common driving task and provides a scientific foundation for understanding this aspect of the driving task; it provides a means for analysis of local and asymptotic stability in a line of vehicles which carry implications with regard to safety and traffic disruptions and other dynamical characteristics of vehicular traffic flow; it provides a steady state description of single lane traffic flow with associated road capacity estimates; it provides a stepping stone for extension into advance automatic vehicle control systems; and finally, it has and will undoubtedly continue to provide stimulus and encouragement to scientists working in related areas of traffic theory.

References

Acha-Daza, J. A. and F. L. Hall (1994). *Application of Catastrophe Theory to Traffic Flow Variables.* Transportation Research - B, 28B(3). Elsevier Science Ltd., pp. 235-250.

Babarik, P. (1968). *Automobile Accidents and Driver Reaction Pattern*. Journal of Applied Psychology, 52(1), pp. 49-54.

Barbosa, L. (1961). *Studies on Traffic Flow Models.* Reports No. 202A-1. The Ohio State University Antenna Laboratory.

Bender, J. G. (1971). *An Experimental Study of Vehicle Automatic Longitudinal Control.* IEEE Transactions on Vehicular Technology, VT-20, pp. 114-123.

Bender, J. G. (1991). *An Overview of Systems Studies of Automated Highway Systems.* IEEE Transactions on Vehicular Technology 40(1). IEEE Vehicular Technology Society, pp. 82-99.

Bender, J. G. and R. E. Fenton (1969). *A Study of Automatic Car Following*. IEEE Transactions on Vehicular Technology, VT-18, pp. 134-140.

Cardew, K. H. F. (1970). *The Automatic Steering of Vehicles -An Experimental System Fitted to a Citroen Car.* RL 340, Road Research Laboratories.

Chandler, F. E., R. Herman, and E. W. Montroll, (1958). *Traffic Dynamics: Studies in Car Following*, Operations Research, 6, pp. 165-184.

Chow, T. S. (1958). *Operational Analysis of a Traffic Dynamics Problem*. Operations Research, 6(6), pp. 165-184.

Constantine, T. and A. P. Young (1967). *Traffic Dynamics: Car Following Studies.* Traffic Engineering and Control 8, pp. 551.

Cumming, R. W. (1963). *The Analysis of Skills in Driving*. Journal of the Australian Road Research Board 1, pp. 4.

Darroch, J. N. and R. W. Rothery (1973). *Car Following and Spectral Analysis.* Proceedings of the 5th International Symposium on the Theory of Traffic Flow and Transportation. Ed. Newell, G. F., American Elsevier Publishing Co., New York.

Drake, J. S., J. L. Schofer, and A. D. May, Jr. (1967). *A Statistical Analysis of Speed Density Hypotheses*. Highway Research Record 154, pp. 53-87.

Drew, D. R. (1965). *Deterministic Aspects of Freeway Operations and Control.* Highway Research Record, 99, pp. 48-58.

Edie, L. C. (1961). *Car-Following and Steady State Theory for Non-Congested Traffic.* Operations Research 9(1), pp. 66-76.

Edie, L. C. and E. Baverez (1967). *Generation and Propagation of Stop-Start Waves.* Vehicular Traffic Science Proceedings of the 3rd International Symposium on the Theory of Traffic Flow. L.C. Edie, R. Herman and R.W. Rothery (Eds.). American Elsevier, New York.

Edie, L. C. and R. S. Foote, (1958). *Traffic Flow in Tunnels.* Proceedings of the Highway Research Board, 37, pp. 334-344.

Edie, L. C. and R. S. Foote (1960). *Effect of Shock Waves on Tunnel Traffic Flow.* Proceedings of the Highway Research Board, 39, pp. 492-505.

Edie, L. C., and R. S. Foote (1961). *Experiments on Single Lane Flow in Tunnels.* Theory of Traffic Flow Proceedings of the Theory of Traffic Flow. Ed. R. Herman. Elsevier Pub. Co., Amsterdam, pp. 175-192.

Edie, L. C., R. S. Foote, R. Herman, and R. W. Rothery (1963). *Analysis of Single Lane Traffic Flow.* Traffic Engineering, 21, 27.

Ellson, D. G. (1949). *The Application of Operational Analysis to Human Motor Behavior*. Psychology Review 9, pp. 56.

Fenton, R. E. (1968). *One Approach to Highway Automation*. Proceedings IEEE 56; pp. 556-566.

Fenton, R. E. and R. J. Mayhan (1991). *Automated Highway Studies at The Ohio State University - An Overview.* IEEE Transaction on Vehicular Technology 40(1); IEEE Vehicular Technology Society, pp. 100-113.

Flory, L. E. (1962). *Electronic Techniques in a System of Highway Vehicle Control*. RCA Review 23, pp. 293-310.

Forbes, T. W., M. J. Zagorski, E. L. Holshouser, and W. A. Deterline (1959). *Measurement of Driver Reaction to Tunnel Conditions*. Proceedings of the Highway Research Board 37, pp. 345-357.

Gardels, K. (1960). *Automatic Car Controls for Electronic Highways*. GM Research Laboratories Report GMR-276, Warren, MI.

Gazis, D. C., R. Herman, and R. B. Potts (1959). *Car Following Theory of Steady State Traffic Flow*. Operations Research 7(4), pp. 499-505.

Gazis, D. C., R. Herman, and R. W. Rothery (1961). *Non-Linear Follow the Leader Models of Traffic Flow*. Operations Research, 9, pp. 545-567.

Gazis, D. C., R. Herman, and R. W. Rothery (1963). *Analytical Methods in Transportation: Mathematical Car-Following Theory of Traffic Flow*. Journal of the Engineering Mechanics Division, ASCE Proc. Paper 3724 89(Paper 372), pp. 29-46.

Giles, C. G. and J. A. Martin (1961). *Cable Installation for Vehicle Guidance Investigation in the New Research Track at Crowthorne.* Rep RN/4057/CGG, Road Research Laboratory, Crowthorne, England.

Greenburg, H. (1959). *An Analysis of Traffic Flow*. Operations Research 7(I), pp. 255-275.

Greenshields, B. D. (1935). *A Study of Traffic Capacity*. Proceedings of the Highway Research Board, 14, 468.

Hall, F. L. (1987). *An Interpretation of Speed-Flow Concentration Relationships Using Catastrophe Theory*. Transportation Research, 21A, pp. 335-344.

Hankin, A. and T. H. Rockwell (1967). *A Model of Car Following Derived Empirically by Piece-Wise Regression Analysis*. Vehicular Traffic Science Proceedings of the 3rd International Symposium on the Theory of Traffic Flow. L.C. Edie, R. Herman and R.W. Rothery (Eds.). American Elsevier, New York.

Harris, A. J. (1964). *Following Distances, Braking Capacity and the Probability of Danger of Collision Between Vehicles*. Australian Road Research Board, Proceedings 2, Part 1, pp. 496-412.

Helly, W. (1959). *Dynamics of Single-Lane Vehicular Traffic Flow.* Research Report No. 2, Center for Operations Research, MIT. Cambridge, Mass.

Herman, R., E. W. Montroll, R. B. Potts and R. W. Rothery (1958). *Traffic Dynamics: Analysis of Stability in Car Following.* Operations Research, E. 17, pp. 86-106.

Herman, R. and R. B. Potts (1959). *Single Lane Traffic Theory and Experiment.* Proceedings Symposium on Theory of Traffic Flow. Ed. R. Herman, Elsevier Publications Co., pp. 120-146.

Herman, R. and R. W. Rothery (1962). *Microscopic and Macroscopic Aspects of Single Lane Traffic Flow.* Operations Research, Japan, pp. 74.

Herman, R. and R. W. Rothery (1965). *Car Following and Steady-State Flow.* Proceedings of the 2nd International Symposium on the Theory of Traffic Flow. Ed J. Almond, O.E.C.D., Paris.

Herman, R. and R. W. Rothery (1969). *Frequency and Amplitude Dependence of Disturbances in a Traffic Stream*. Proceedings of 4th International Symposium on the Theory of Traffic Flow, Ed. W. Leutzbach and P. Baron. Bonn, Germany.

Highway Capacity Manual (1950). U.S. Government Printing Office, Washington, DC.

Hodgkins, E. A. (1963). *Effect of Buses on Freeway Capacity.* Presentation at the 43rd Annual Meeting of the Highway Research Board 33. Highway Research Abstracts, pp. 69.

Ito, T. (1973). *An Automatic Drivers System of Automobiles by Guidance Cables.* Society of Automotive Engineering (730127).

Kometani, E. and T. Suzaki (1958). *On the Stability of Traffic Flow.* J. Operations Research, Japan 2, pp. 11-26.

Lam, T. and R. W. Rothery (1970). *Spectral Analysis of Speed fluctuations on a Freeway.* Transportation Science 4(3).

Lee, G. (1966). *A Generalization of Linear Car-Following Theory.* Operations Research 14(4), pp. 595-606.

May, A. D. and H.E.M. Keller (1967). *Non-Integer Car Following Models.* Highway Research Record 199, pp. 19-32.

Michaels, R. M. (1963). P*erceptual Factors in Car Following.* Proceedings of the 2nd International Symposium on the Theory of Road Traffic Flow (London, England), OECD.

Navin, F. P. D. (1986). *Traffic Congestion Catastrophes.* Transportation Planning and Technology 11, pp. 19-25.

Newell, G. F. (1961). *Nonlinear Effects in the Dynamics of Car Following.* Operations Research 9(2), pp. 209-229.

Newell, G. F. (1962). *Theories of Instability in Dense Highway Traffic.* J. Operations Research Society of Japan 5(1), pp. 9-54.

Oshima, R. (1965). *Control System for Automobile Driving.* Proceedings of the Tokyo 1FAC Symposium, pp. 347-357.

Pipes, L. A. (1951). *A Proposed Dynamic Analogy of Traffic.* ITTE Report, Institute of Transportation and Traffic Engineering, University of California, Berkeley.

Pipes, L. A. (1953). *An Operational Analysis of Traffic Dynamics.* Journal of Applied Physics 24, pp. 271-281.

Reuschel, A. (1950). Fahrzeugbewegungen in der Kolonne Beigleichformig beschleunigtem oder vertzogerten Leitfahrzeub, Zeit. D. Oster. Ing. U. Architekt Vereines Ed. (Vehicle Movements in a Platoon with Uniform Acceleration or Deceleration of the Lead Vehicle), pp. 50-62 and 73-77.

Rothery, R. W., R. Silver, R. Herman and C. Torner (1964). *Analysis of Experiments on Single Lane Bus Flow.* Operations Research 12, pp. 913.

Shladover, S. E., C. A. Desoer, J. D. Hedrick, M. Tomizuka, J. Walrand, W. B. Zhang, D. H. McMahon, H. Peng, S. Shiekholeslam, and N. McKeown (1991). *Automated Vehicle Control Developments in the PATH Program.* IEEE Transactions on Vehicular Technology 40(1), pp. 114-130.

Taylor, F. V. (1949). *Certain Characteristics of the Human Serve*. Electrical Engineering 68, pp. 235.

Telfor, G. W. (1931). *The Refractory Phase of Voluntary and Associative Responses*. Journal of Experimental Psychology, 14, pp. 1.

Thorn, R. (1975). *Structural Stability and Morphogenesis*. An English Translation by D.H. Fowler. John Wiley & Sons, New York.

Todosiev, E. P. (1963). *The Action Point Model of the Driver Vehicle System*. Report No. 202A-3. Ohio State University, Engineering Experiment Station, Columbus, Ohio.

Tuck, E. (1961). *Stability of Following in Two Dimensions*. Operations Research, 9(4), pp. 479-495.

Tustin, A. (1947). *The Nature of the Operator Response in Manual Control and its Implication for Controller Design*. J.I.E.E. 92, pp. 56.

Unwin, E. A. and L. Duckstein (1967). *Stability of Reciprocal-Spacing Type Car Following Models*. Transportation Science, 1, pp. 95-108.

Uttley, A. (1941). *The Human Operator as an Intermittent Servo*. Report of the 5th Meeting of Manual Tracking Panel, GF/171.SR1A.

Wardrop, J. G. (1965). *Experimental Speed/Flow Relations in a Single Lane*. Proceedings of the 2nd International Symposium on the Theory of Road Traffic Flow. Ed. J. Almond O.E.C.D.

Zeeman, E. C. (1977). *Catastrophe Theory*. Selected Papers 1972-1977. Addison-Wesley, London.

CONTINUUM FLOW MODELS

REVISED BY **H. MICHAEL ZHANG***

ORIGINAL TEXT BY **REINHART KUHNE**[7]

PANOS MICHALOPOULOS[8]

* Professor, Department of Civil and Environmental Engineering, University of California Davis, One Shields Avenue, Davis, CA 95616

[7] Managing Director, Steierwald Schonharting und Partner, Hessbrühlstr. 21c 7 70565 Stuttgart, Germany

[8] Professor, Department of Civil Engineering, University of Minnesota Institute of Technology, 122 Civil Engineering Building, 500 Pillsbury Drive S.E., Minneapolis, MN 55455-0220

CHAPTER 5 - Frequently used Symbols

a	=	dimensionless traffic parameter
A	=	stop-start wave amplitude
α	=	sensitivity coefficient
b	=	net queue length at traffic signal
c	=	$g + r$ = cycle length
c_{0_2}	=	coefficient
c_0^2	=	constant, independent of density k
ds	=	infinitesimal time
Δt, Δx	=	the time and space increments respectively such that $\Delta x/\Delta t >$ free flow speed
$\delta\eta_i$, $\delta\eta_{i+1}$	=	deviations
η	=	state vector
η_i, η_{i+1}	=	state vector at position i, i+1
$f(x, v, t)$	=	vehicular speed distribution function
f	=	relative truck portion, $k_{pass} = k$
f_0	=	equilibrium speed distribution
Γ	=	fluctuating force as a stochastic quantity
g	=	effective green interval
g_j^n	=	is the generation (dissipation) rate at node j at $t = t_0 + n\Delta t$; if no sinks or sources exist g_j^n = 0 and the last term of Equation 5.28 vanishes
$g_{\min}$	=	minimum green time required for undersaturation
h	=	average space headway
i	=	station
j	=	node
k	=	density
k_-, k_+	=	density downstream, upstream shock
k_0	=	operating point
k_{10}	=	equilibrium density
K_A	=	constant value
k_a	=	density within L_2
k_{bumper}	=	density "bumper to bumper"
k_d, q_d	=	density, flow downstream
k_u, q_u	=	density, flow upstream
k_{hom}	=	vehicle density in homogeneous flow
k_j	=	jam density of the approach under consideration
k_j^n, q_j^n	=	density and flow rate on node j at $t = t_0 + n\Delta t$
k_m	=	density conditions
k_{pass}	=	density "bumper to bumper" for 100% passenger cars
k_{ref}	=	reference state
k_{truck}	=	density "bumper to bumper" for 100% trucks
L	=	distance
L	=	length of periodic interval
ld	=	logarithmus dualis
l_o	=	characteristic length
λ	=	wave length of stop-start waves
μ_0	=	dynamic viscosity
μ	=	viscosity term
n	=	current time step
N	=	normalization constant
n_i, n_2	=	exponents
N_i	=	number of cars (volume)
ω	=	eigenvalue
p	=	probability
q	=	actual traffic volume, flow
Q_0	=	net flow rate
q_a	=	average flow rate
$q_a k_a$	=	arrival flow and density conditions
q_{ni}	=	capacity flow
r	=	effective red interval
σ_0	=	quantity
T	=	oscillation time
t	=	time
t_0	=	the initial time
τ	=	relaxation time as interaction time lag
u	=	speed
$U_e(k)$	=	equilibrium speed-density relation
$u_e(k^n_{ij})$	=	equilibrium speed
u_f	=	free-flow speed of the approach under consideration
u_g	=	group velocity
$u_{max} - u_{min}$	=	speed range
u_w	=	shock wave speed
u_z	=	spatial derivative of profile speed
$v(k)$	=	viscosity
v_g	=	values of the group velocity
$W(q)$	=	distribution of the actual traffic volume values q
x	=	space
xh	=	estimated queue length
x_i, t_i, y_i	=	coordinates at point i
X_{ij}	=	length of any line ij
y	=	street width
$y(t)$	=	queue length at any time point t
y_{ij}	=	queue length from i to j assuming a positive direction opposite to x, i.e. from B to A
z	=	$x - U, t$, collective coordinate
$\dot{x}$	=	shockspeed

5. CONTINUUM FLOW MODELS

Original text by Reinhart Kühne and Panos Michalopoulos
Revised by H. M. Zhang

1 Conservation and traffic waves

There have been two kinds of traffic flow theories presented thus far in this monograph. Chapter 2 discussed the relations among aggregated quantities of flow rate, concentration and mean speeds for stationary vehicular traffic (these relations are also referred to as the equations of state in traffic flow), and Chapter 3 described the dynamic evolution of microscopic quantities of vehicle spacing, headway and speeds in single-lane traffic through the follow-the-leader type of models. In this Chapter we study another kind of traffic flow theory—continuum traffic flow theory—that describes the temporal-spatial evolution of macroscopic flow quantities as mentioned earlier. Continuum theories are natural extensions to the first two kinds of theories because, one one hand, they are closely related to the equations of state and car-following theories, and on the other hand, they overcome some of the drawbacks of the first (stationarity) and second (too much detail and limited observability) kinds of theories. The quantities and features that continuum theories describe, such as traffic concentration and shock waves, are mostly observable with current surveillance technology, which makes it easier to validate and calibrate these models.

Many theories exist in the continuum description of traffic flow. All of them share two fundamental relations: one is the conservation of vehicles and the other is the flow-concentration-speed relation

$$q = ku. \tag{1}$$

The $q - k - u$ relation is true by choice, i.e., one defines flow rate (q), concentration (k) and mean speed (u) in such a way that (1) always holds (see Chapter 2 for definitions). Conservation of vehicles, on the other hand, is true regardless of how q, k, u are defined, and can be expressed in different forms. Here we present three forms of the conservation law using variables in (1).

The first integral form of the conservation law. Consider a stretch of highway between x_1 and x_2 ($x_1 < x_2$). At time t the traffic concentration on this section is $k(x, t)$, and traffic flows into the section at a rate of $q(x_1, t)$, and out of the section at a rate of $q(x_2, t)$. The total number of

vehicles in this section at time t is then

$$\int_{x_1}^{x_2} k(x,t)dx.$$

Suppose no entries and exits exist between x_1 and x_2, then by conservation the rate of change in the number of vehicles

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} k(x,t)dx.$$

should equal to the net flow into this section

$$q(x_1,t) - q(x_2,t),$$

that is,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} k(x,t)dx = q(x_1,t) - q(x_2,t), \tag{2}$$

This is the first integral form of the conservation law.

The second integral form of the conservation law. The conservation of vehicles also means that, in the $x-t$ plane, the net number of vehicles passing through any closed curve is zero provided that no sources and sinks are there in the enclosed region (Fig. 2). Note that the number of vehicles across any segment dl of curve $\mathcal{C}$ is $-kdx + qdt$, the total number of vehicles across the enclosed region is therefore the line integral of the vector field $(-k, q)$, and we have the second integral form of the conservation law:

$$\int_{\mathcal{C}} -kdx + qdt = 0. \tag{3}$$

(3) implies that we can construct a scalar function $N(x,t)$ with the properties

$$q = \frac{\partial N}{\partial t},\ k = -\frac{\partial N}{\partial x}.$$

This function is the cumulative count of vehicles passing location x at time t, provided that $N(x, t = 0) = 0$. The fact that such a function exists is another consequence of the conservation law.

The differential form of the conservation law. By applying the divergence theorem to the second integral form we obtain

$$\int_{\mathcal{C}} -kdx + qdt = \int\int_D \left(\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x}\right) dxdt = 0,$$

which leads to the differential form of the conservation law:

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0. \tag{4}$$

The three different forms of the conservation law are not completely equivalent. Note that in the differential form, both k and q are required to be differentiable with respect to time t and space x. In contrast, it is perfectly fine in the integral forms that these variables are discontinuous in space and/or time. This is an important distinction because a special kind of traffic waves, called shock waves, are prevalent in vehicular traffic flow and play an essential role in the continuum

description of traffic flow. A shock wave is a drastic change in traffic concentration (and/or speed, flow rate) that propagates through a traffic stream. Examples of shock waves include the stoppage of traffic in front of a red light, and traffic slow-downs caused by an accident. In reality a shock always has a profile (that is, a transition region of non-zero width) that usually spans a few vehicle lengths, but this can be practically treated as a discontinuity when compared with the length of roads in consideration.

Because of the existence of shocks, it is necessary to expand the solution space of (4) to include the so-called *weak solutions*. In mathematical sense a weak solution is a function $(k, q)(x, t)$ that satisfies (4) everywhere except along a certain path $x(t)$. On $x(t)$ $(k, q)(x, t)$ are discontinuous but obey the integral forms of the conservation law. A shock wave solution to (4) is a weak solution. Moreover, its path $x(t)$ is governed by the following equation:

$$\dot{x}(t) = [q]/[k]. \tag{5}$$

where $[k] = k_r - k_l, [q] = q_r - q_l$, $\dot{x}(t)$ is the speed of the shock, also denoted as s, and $(k_{l,r}, q_{l,r})$ are the traffic states immediately to the right and left of the shock path $x(t)$, respectively. Expression (5) is also known as the Rankine-Hugoniot (R-H) condition (LeVeque 1992) , which is a consequence of the conservation law.

Expression (5) has a simple geometric interpretation in the (k, q) plane, also known as the $k-q$ phase plane. It is simply the slope of the segment that connects the two phase points (k_l, q_l) and (k_r, q_r) (Fig. 1(a)). Suppose that a unique curve connects $(k_{l,r}, q_{l,r})$, and $k_r \to k_l$ (i.e., the shock is weak), then s approaches the slope of the tangent of the connecting curve at (k_l, q_l) (Fig. 1(b)). We call this speed the speed of *traffic sound waves*, and give it a special notation $c(k)$. This is the speed that information is propagated in homogeneous traffic. It plays a central role in the basic kinematic wave model of traffic flow. It must be pointed out that a unique $q - k$ curve is not necessary to compute the shock speed. This can be done as long as both the densities and flow rates on both sides of the shock are known.

Next we provide two derivations of (5) from the conservation law. The first derivation is based on the first integral form of the conservation law. Suppose that traffic states on both sides of the shock are constant states $(k_{1,2}, q_{1,2})$, and suppose that two observers located at $x_1(t)$ and $x_2(t)$ on the two sides of the shock (i.e., $x_1 < x(t) < x_2$) travel at precisely the speed of the shock, s. We apply the conservation law (2) to the region between the two observers after doing a coordinate transformation $x^{'} = x - st$ to obtain:

$$\frac{\partial}{\partial t} \int_{x_1^{'}}^{x_2^{'}} k(x, t)dx = [q_1 - k_1 s] - [q_2 - k_2 s].$$

The integral, however, can be evaluated independently in $[x_1^{'}, x_l^{'})$ and $(x_r^{'}, x_2^{'}]$, where $x_{l,r}^{'}$ are points immediate to the right/left of shock path $x^{'}(t)$. This integral turns out to be

$$k_1(x_l^{'} - x_1^{'}) + k_2(x_2^{'} - x_r^{'}),$$

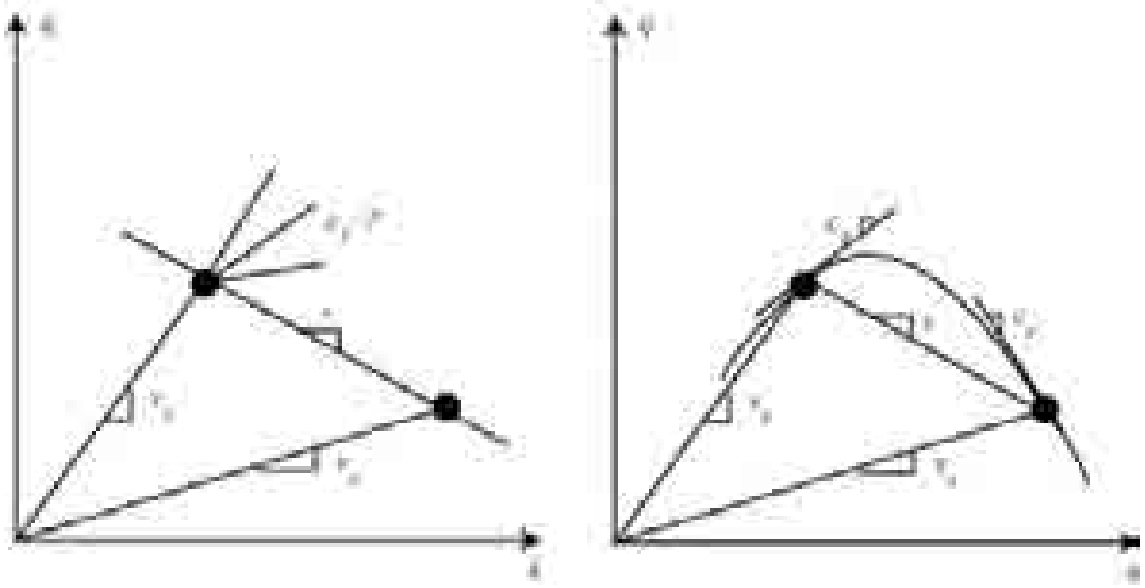


Figure 1: Geometric representation of shocks, sound waves, and traffic speeds in the $k - q$ phase plane

a constant, whose derivative is zero. Therefore we have

$$[q_1 - k_1 s] - [q_2 - k_2 s] = 0,$$

which leads to (5) after rearranging terms.

The second derivation of the Rankine-Hugoniot condition comes from the second integral form. As shown in Fig.2, a shock path $x(t)$ breaks the closed curve $\mathcal{C}$ into two parts $\mathcal{C}_l$ and $\mathcal{C}_r$. Let $\Gamma_{l,r}$ be the right/left edges of $x(t)$, respectively, then the conservation law applies to each of the three enclosures $\mathcal{C}$, $\mathcal{C}_l \cup \Gamma_l$, and $\mathcal{C}_r \cup \Gamma_r$:

$$\int_{\mathcal{C}} -kdx + qdt = 0. \tag{6}$$

$$\int_{\mathcal{C}_l \cup \Gamma_l} -kdx + qdt = 0. \tag{7}$$

$$\int_{\mathcal{C}_r \cup \Gamma_r} -kdx + qdt = 0. \tag{8}$$

Adding (7) and (8) together and rearranging terms we obtain

$$\int_{\mathcal{C}} -kdx + qdt + \int_{\Gamma_l} -kdx + qdt + \int_{\Gamma_r} -kdx + qdt = 0.$$

The first integral is zero, and the second and third integrand become

$$(-k_{l,r} s + q_{l,r})dt$$

on $\Gamma_{l,r}$. Taking account the direction of integration we obtain the following integral equation:

$$\int_{t_1}^{t_2} \{(-k_l s + q_l) - (-k_r s + q_r)\}dt = 0.$$

This implies that

$$(-k_l s + q_l) - (-k_r s + q_r) = 0,$$

which also leads to the Rankine-Hugoniot shock condition.

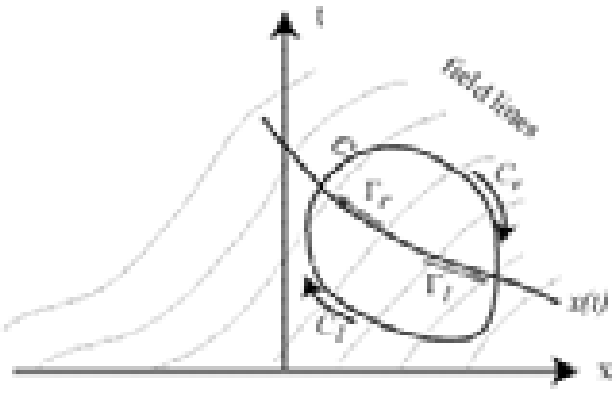


Figure 2: Field representation of shocks and conservation of flow

The differential conservation law (4), supplemented by the Rankine-Hugoniot condition (5), is the corner stone of all continuum vehicular traffic theories. Yet it is not a complete theory, nor is it unique to traffic flow. In fact, the differential conservation law is obeyed by many kinds of material fluids, such as gas flow in a pipe or water flow in a channel. As such, it does not capture the unique characters of vehicular traffic flow, a special kind of "fluid". It must be supplemented by additional relations to form a complete and useful traffic flow theory. Some of the relations between macroscopic traffic variables are already available and have been discussed in Chapter 2. These are the binary relations between flow rate, traffic concentration and mean travel speed. Although these relations are found among stationary traffic, they can also be used as a first approximation of dynamic flow-concentration-speed relations. This leads to the first, and the simplest continuum theory of traffic flow—the kinematic wave theory developed independently by Lighthill and Whitham (1955), and Richards (1956) (The LWR model). A more accurate relation between flow-concentration or speed-concentration is one that accounts for the dynamic behavior of drivers—anticipation and inertia. The usage of such a relation leads to the so-called higher-order models. We will study in this chapter these two kinds of continuum theories in great deal. In the sections to follow, we'll first study the properties of the LWR model and its Riemann problem, then study the properties of a special class of higher-order models and their respective Riemann problems. Next we develop numerical approximations of both types of models, and finally we provide some examples of the use of the numerical approximations.

2 The kinematic wave model of LWR

2.1 The LWR model and characteristics

The LWR model assumes that the relation observed in *stationary* flow, $q = f_*(k)$, also applies to dynamic traffic. With such a relation, the conservation law becomes:

$$k_t + f_*(k)_x = 0. \tag{9}$$

and the Rankine-Hugoniot condition becomes

$$s = [f_*]/[k].$$

This is the celebrated kinematic wave model of LWR. It is one of the simplest nonlinear scalar conservation laws in physics and engineering.

Experimental evidence indicates that $f_*(k)$ is usually smooth and concave (refer to Chapter 2), and satisfies the following boundary conditions:

$$f_*(0) = f_*(k_j) = 0,$$

$$f_*^{'}(0) = v_f > 0,\ f_*{}'(k_j) = c_j < 0.$$

where k_j is the jam concentration at which vehicles grind to a halt, usually ranging from 260 vpm to 330 vpm (in passenger car units), c_j is the speed of traffic sound waves at jam condition, taking values in the range of [-10 mph, -20 mph], and v_f is free-flow travel speed, ranging from $55mph$ to $75mph$ on freeways. All these parameter values are readily computable from conventional field measurements.

With the introduction of traffic wave speed (when no confusion arises we use traffic wave and traffic sound waves interchangably) $c(k) = f_*{}'(k)$, the LWR model also reads

$$k_t + c(k)k_x = 0. \tag{10}$$

Now let us look at what changes in concentration an observer sees when he travels at the speed of $c(k)$, that of traffic waves.

$$\frac{dk}{dt} = k_t + k_x\frac{dx}{dt} = k_t + c(k)k_x = 0.$$

That is to say, he sees no changes in density at all if he travels at the wave speed. As a result, if he knows the initial concentration at a point ξ, he'll know the concentration at any point on the path

$$\dot{x}(t) = c(k),\ x(0) = \xi. \tag{11}$$

This path is called the *characteristic curve* of the LWR model, and the wave speed $c(k)$ is also known as the *characteristic speed* of the LWR model. Because k is constant along the characteristic curve, the characteristic of (9) is therefore a straight line.

One must not confuse over traffic sound wave speeds, shock wave speeds, characteristic speeds, and vehicle speeds. Traffic sound wave speeds and characteristic speeds in the LWR model are

identical, which are the slopes of the tangential lines on the flow-concentration curve (also known as *the fundamental diagram* of traffic flow). Shock wave speeds are the slopes of the secant that connecting any two traffic states on $f_*(k)$, and vehicular speeds are the slopes of the rays from the origin to points on $f_*(k)$ (Fig. 1). Because of the concave shape of $f_*(k)$, we have the following relations among the various speeds:

$$v = f_*/k \equiv v_*(k) \geq c(k),\ s < v_*(k_{l,r}),$$

that is to say, all waves, including sound and shock waves, travel no faster than traffic. In other words, information in LWR traffic is propagated against the traffic stream. This property of information propagation in the LWR model is known as the *anisotropic property* of traffic flow. The anisotropic property may not hold if the fundamental diagram is not concave (Zhang 2000c).

2.2 The Riemann problem and entropy solutions

The LWR model is a well-posed hyperbolic partial differential equation (pde) [1] and can be solved with proper initial/boundary data. In fact, analytical solutions can be obtained for a special kind of problem called Riemann problem, whose initial data—the so-called Reimann data—are two constant states $k_{l,r} \geq 0$ separated by a single jump:

$$k(x, t=0) = \begin{cases} k_l, & x < 0, \\ k_r, & x > 0 \end{cases}$$

The analytical solution to the Riemann problem can either be a shock:

$$k(x,t) = \begin{cases} k_l, & x < st, \\ k_r, & x > st. \end{cases} \tag{12}$$

or a smooth expansion wave (also known as a rarefaction wave):

$$k(x,t) = \begin{cases} k_l, & x < c_l t, \\ (f_*')^{-1}\left(\frac{x}{t}\right), & c_l t \leq x \leq c_r t, \\ k_r, & x > c_r t \end{cases} \tag{13}$$

The condition for it to be a shock is governed by the so-called *entropy* condition, which states that

$$c_l > c_r. \tag{14}$$

Otherwise the solution will be an expansion wave. Because of the concavity condition $f_*''(k) < 0$, (14) implies that $k_l < k_r$, that is, shocks arising from the LWR theory are compressive. Moreover, they reach vehicles from upstream because $s < v_{l,r}$. Examples of the two kinds of solutions to a Riemann problem are shown in Figs. 3& 4.

For general initial conditions, it is usually tedious, if not difficult, to obtain analytical solutions of the LWR equation, and numerical approximations are often sought, where the Riemann problem

[1] A pde is hyperbolic if its characteristic speeds are real (as compared to imaginary).

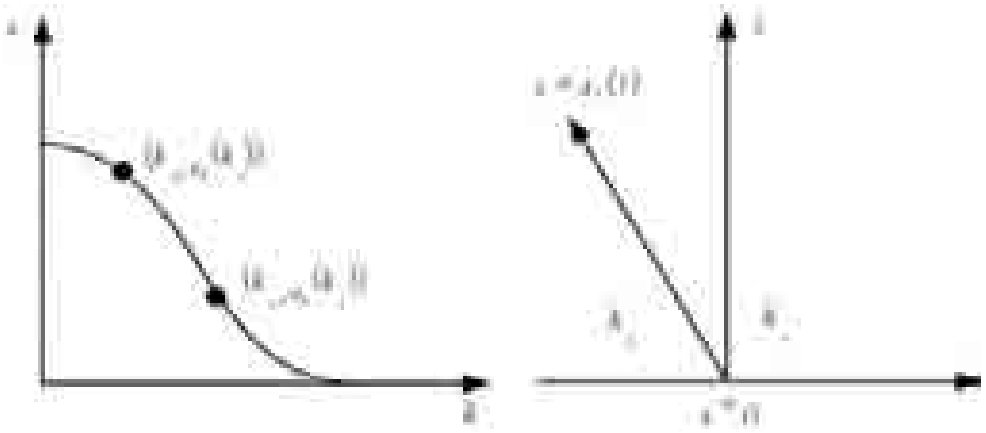


Figure 3: A shock solution

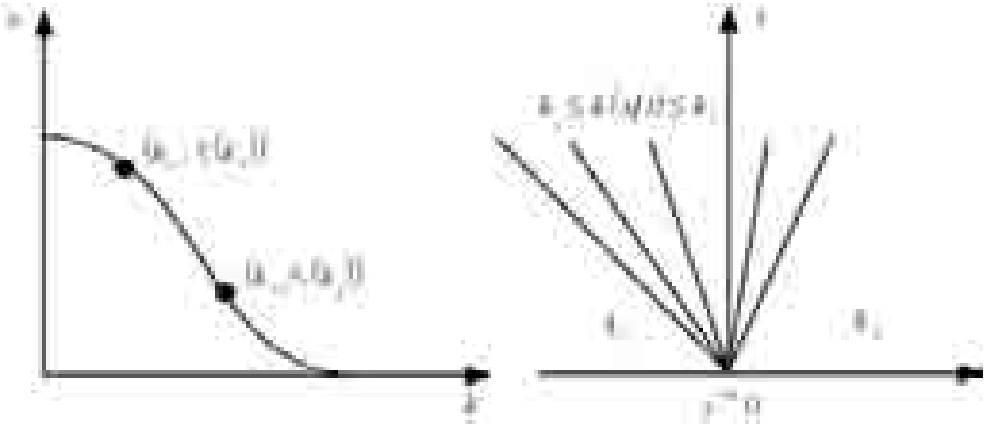


Figure 4: A rarefaction solution

plays a key role in developing some of the most efficient and accurate numerical schemes. This will be discussed in Section 4 together with the treatment of boundary conditions. There are special cases where an analytical solution is still straightforward to obtain. One case involves a special kind of initial condition—$k(x,0)$ is piece-wise constant and and increasing, and another case involves a special kind of fundamental diagram—the triangular shaped fundamental diagram that has only two wave speeds (Koshi et al 1983, Newell 1993). In the former case, a series of Riemann problems can be solved with consideration of wave interactions, and in the latter case, expansion waves are replaced by acceleration shocks, and it is much simpler to construct the solutions with shocks than expansion waves (see Newell 1993).

2.3 Applications

Same as Sections 5.1.3 & 5.1.4 in the original text.

2.4 Extensions to the LWR model

The LWR model we presented applies to traffic on a homogeneous highway with no entering and exiting traffic. In reality all highways have entries and exits and variable geometric features. Fortunately the LWR equation can be easily extended to model such inhomogeneities.

Suppose the net inflow to a road section is $s(x,t)dxdt$ from sources such as freeway ramps, then from conservation we have

$$k_t + f_*(k)_x = r(x,t), \tag{15}$$

This is the LWR model with sources. These source flows can be entering or exiting flows from ramps, or both.

When road geometries change, such as the drop of a lane at some locations, they often generate disturbances that cause traffic breakdowns. The effect of such inhomogeneities on traffic flow can be captured to some degree with a location-dependent fundamental diagram, $f_*(k,x)$. Because vehicle conservation still holds, applying the conservation principle to an inhomogeneous road leads to the following extended LWR model:

$$k_t + f_*(k,x)_x = 0. \tag{16}$$

If a road segment has both sources and geometric inhomogeneities, the following model applies:

$$k_t + f_*(k,x)_x = r(x,t). \tag{17}$$

Like the homogeneous LWR model, these extensions are also non-linear hyperbolic pdes, whose characteristics are

$$\dot{x} = f_{*k}(k,x). \tag{18}$$

These characteristics, however, are no longer straight lines as in the homogeneous model. This is because along the characteristics, we have

$$\dot{k} = k_t + k_x\dot{x} = k_t + f_{*k}k_x = \begin{cases} r, & \text{for (15)} \\ -f_{*x}, & \text{for (16)} \\ r - f_{*x}, & \text{for (17)} \end{cases} \tag{19}$$

(18) and (19) form a system of ordinary differential equations (ode), and can be solved iteratively using any ode solver such as the Runge-Kutta method. This way of solving the inhomogeneous LWR models is known as *the method of characteristics* (Courant and Hilbert 1962).

The basic LWR model and its various extensions we have covered treat traffic across lanes as homogeneous, which is somewhat restrictive because alternate motion of traffic in different lanes is often observed in heavily congested traffic. This limitation is easily overcome by modeling traffic evolution in each lane. Here the conservation principle still applies, but a continuum of sources exist for each lane. These sources are the exchange of flow between adjacent lanes. Without loss of generality, assume a two-lane highway with traffic $(k_1(x,t), q_1(x,t))$ and $(k_2(x,t), q_2(x,t))$. Let Q_1 be the net amount of traffic entering lane 1 from lane 2 in unit time and distance, and Q_1 be

the net amount of traffic entering lane 2 from lane 1 in unit time and distance, then the extended LWR model for a two-lane highway is (Dressler 1949)

$$k_{1t} + q_{1x} = Q_1, \tag{20}$$

$$k_{2t} + q_{2x} = Q_2. \tag{21}$$

where $q_1 = f_{1*}(k_1)$ and $q_2 = f_{2*}(k_2)$.

From conservation we know the source flows $Q_1,\ Q_2$ observe

$$Q_1 + Q_2 = 0$$

and experiences tell us that the change of flow between lanes is related to the differential of lane speeds/densities, for drivers always like to move into the faster lane in congested traffic. An example of $Q_1,\ Q_2$ was proposed by Gazis et al. (1962):

$$Q_{1,2} = \alpha[(k_{2,1} - k_{1,2}) - (k^0_{2,1} - k^0_{1,2})] \tag{22}$$

where $k^0_{1,2}$ are equilibrium densities of lanes 1 & 2, respectively (Gazis et al. 1962). These can be observed experimentally.

The above lane-specific LWR model is a system of quasi-linear PDEs and can be expressed in vector forms

$$\begin{pmatrix} \mathrm{k}_1 \\ \mathrm{k}_2 \end{pmatrix}_t + \begin{pmatrix} \mathrm{f}_{1*}(k_1) \\ \mathrm{f}_{2*}(k_2) \end{pmatrix}_x = \begin{pmatrix} \mathrm{Q}_1(k_1, k_2) \\ \mathrm{Q}_2(k_1, k_2) \end{pmatrix}, \tag{23}$$

or

$$\begin{pmatrix} \mathrm{k}_1 \\ \mathrm{k}_2 \end{pmatrix}_t + \begin{pmatrix} f_{1*,k_1} & 0 \\ 0 & f_{2*,k_2} \end{pmatrix} \begin{pmatrix} \mathrm{k}_1 \\ \mathrm{k}_2 \end{pmatrix}_x = \begin{pmatrix} \mathrm{Q}_1(k_1, k_2) \\ \mathrm{Q}_2(k_1, k_2) \end{pmatrix}. \tag{24}$$

For a system of quasi-linear PDEs, the characteristics are the eigenvalues of the Jacobian matrix $DF \equiv \begin{pmatrix} f_{1*,k_1} & 0 \\ 0 & f_{2*,k_2} \end{pmatrix}$, which contains the partial derivatives of the flow vector $F(U) \equiv \begin{pmatrix} \mathrm{f}_{1*}(k_1) \\ \mathrm{f}_{2*}(k_2) \end{pmatrix}$ with respect to the state vector $U \equiv \begin{pmatrix} \mathrm{k}_1 \\ \mathrm{k}_2 \end{pmatrix}$.

This system is said to be strictly hyperbolic if all its characteristics are real and distinct. It is easy to see that the characteristics of (23) are $f'_{1*}(k_1)$ and $f'_{2*}(k_2)$ because the dynamics of the two lanes are nearly decoupled (the only coupling comes from the source term). The system is strictly hyperbolic under the condition that $f_{1*}(k_1) \neq f_{2*}(k_2)$ and $f''_{1,2*} < 0$. Thus this system can be solved in the same way as other inhomogeneous models through the method of characteristics. Only this time one solves a system of four odes instead of two:

$$\dot{x}_1 = f'_{1*}(k_1),$$

$$\dot{x}_2 = f'_{2*}(k_2),$$

$$\dot{k}_1 = Q_1(k_1, k_2),$$

$$\dot{k}_2 = Q_2(k_1, k_2).$$

It should be noted that both the inhomogeneous models and the multiple-lane models can be solved numerically using the finite difference procedures developed in Section 5.

2.5 Limitations of the LWR model

The kinematic wave model of LWR is only a first approximation of the real traffic flow process and is deficient in describing a number of traffic features of potential importance. The include (i) driver differences, (ii) shock structure, (iii) forward moving waves in queued up traffic and (iv) traffic instability (Daganzo 1997).

The LWR model is inadequate for modeling light traffic because it does not recognize the segregation of fast and slow drivers in the traffic stream owing to passing. An adequate model for traffic where significant amount of passing takes place should track both fast and slow drivers and their interactions, not by lumping both types of drivers together through a single descriptor such as density $k(x,t)$. When drivers travel at roughly the same speed, however, the LWR model can still be used for light traffic, where density is independent of concentration and flow increases linearly with density.

The second deficiency of the LWR model lies in its treatment of shock waves. The shocks in the LWR model has no width, i.e., no transition region. A vehicle enters a shock thus dropping its speed in no time, implying an infinite deceleration. In reality, shocks always have a structure, i.e., a transition region of a few vehicles' length in which vehicles decelerate at finite rates. This deficiency can be addressed through introducing higher-order approximations of traffic dynamics (Kühne 1984,1989) or using a microscopic description (Newell 1961). It is one of the author's (Zhang) opinion that as long as the accurate knowledge of vehicle acceleration is not of main interest, the lack of a shock structure is not a major failing of the LWR theory, because in comparison with the space scale one is modeling, a few vehicle lengths of shock width can be practically considered as nil. To most applications, the important matter is whether the LWR theory gives a reasonable estimate of shock speeds. Empirical evidence indicates that it does (Chapter 4 in Daganzo 1997).

The LWR theory has only one family of waves which travels at a speed of $f_*^{'}(k)$. These waves always travel against the traffic stream because $f_*^{'}(k) < v_*(k)$. In particular, $f_*^{'}(k) < 0$ when concentration k exceeds a critical value k_* at which flow rate is maximal. An early study of tunnel traffic, however, revealed that a different type of waves exists in real traffic (Edie and Baverez 1967). In analyzing the tunnel traffic data, Edie and Baverez (1967) noted that " small changes in flow may not propagate at a speed equal to the slope of the tangent to a steady-state q-k curve as suggested by the hydrodynamic wave theories of traffic flow. Instead, they are carried along at about stream speed or only slightly less than stream speed right up to saturation flows, at which level they suddenly reverse directions." This observation indicates that apart from the family of waves that travels against the traffic stream, there's at least another family of traffic waves that travels with the traffic stream, even in congested traffic. The findings of Edie and Baverez (1967) prompted Newell (1965) to suggest a fundamental diagram with multiple branches—one branch for free flow, one for acceleration flow and another for deceleration flow. The extended LWR theory with this fundamental diagram was able to explain not only the forward wave motion in queued-up traffic, but also the instability found in tunnel traffic.

In comparison with other deficiencies of the LWR theory, the fourth deficiency, namely its

inability to model traffic instability, has more serious consequences, for traffic instability is at the very heart of the traffic congestion phenomenon. The LWR model is always stable in the sense that traffic disturbances, small or large, are always dampened. In other words, a driver who obeys the LWR driving law always responds to stimulus properly, i.e., he always manages to change his speed in the right amount of time and with the right magnitude that he simply absorbs the disturbance. In fact the reaction time of a LWR driver is zero and the rate of adjustment he makes is infinite. In reality, a driver responds to traffic events with a time delay, and not always precisely. As a result, some disturbances in real traffic may get magnified as they propagate through the traffic stream, causing traffic break-downs (stop-and-go) that could last for several hours. Such stop-and-go traffic patterns exhibit, in physical space (i.e., $x-t$ domain), periodic oscillations with amplitude-dependent oscillation time, and in phase space (i.e., $q-k$ or $k-u$ domain), hysteresis loops and wide scatter of data points.

Evidences of traffic instability and the resulting stop-and-go flow pattern are found on highways around the globe. Perhaps the most impressive measurements of transients and stop-start wave formation are gained from European freeways. Due to space restrictions, there are numerous freeways with two lanes per highway in Europe. These freeways, often equipped with a dense measurement grid not only for volume and occupancy but also for speed detection, show stable stop-start waves lasting in some cases for more than three hours. Measurement data exists for Germany (Leutzbach 1991), the Netherlands (Verweij 1985), and Italy (Ferrari 1989). First we examine the measurements from German highways. The German data were collected from the Autobahn A5 Karlsruhe-Basel at 617 km by the Institute of Transport Studies at Karlsruhe University (Kühne 1987). Each measurement point is a mean value of a two-minute ensemble actuated every 30 sec. These data were collected during a holiday when no trucks used that stretch of road.

All the data sets have traffic densities over the critical density and show signs of instability (i.e., stop-start waves with more or less regular shape and of long duration - in some series up to 12 traffic breakdowns). These data can characterized in the time domain by their oscillation times T and magnitudes A. These characteristic features derived from the data shown in **Figures 5.9a,b and 5.9c,d (in the original text)** (Kühne 1987,Michalopoulos and Pisharody 1980) are listed in the following table:

Table 1: Oscillation time and magnitudes of stop-and-go traffic from German measurement

oscillation time T	16 min	15 min	7.5 min	5 min
amplitude A	70km/h	70 km/h	40 km/h	25 km/h
measurement figure	2a	2b	3a	3b

These characteristic values show a proportionally between amplitude and oscillation time. This strong dependence is a sign for the non-linear and anharmonic character of stop-start waves. In the case of harmonic oscillations, the amplitude is independent of the oscillation time as the linear

pendulum shows. Obviously, the proportionally holds only for the range between traffic flow at a critical lane speed of about 80 km/h (= speed corresponding to the critical density kc . 25 veh/km) and creeping with jam speed of about 10 km/h. For oscillations covering the whole range between free-flow speed and complete gridlock, saturation effects will reduce the proportionally.

Examples of stop-start waves from other locations, such as the Netherlands (Verweij 1985), Japan (Koshi et al. 1983), Italy (Ferrari 1989) and the U.S.A. also show similar characters as those observed on German highways. Such oscillations, when viewed from the $q - k$ phase plane, show wider scatter of data points in the congested regime. Embedded in the scatter are sharp drops (often referred to as *the capacity drop*) and hysteresis loops (e.g., Treiterer and Myers 1974) of irregular shapes that differ from the equilibrium phase curve $q = f_*(k)$ and cannot be simply explained away by stochastic arguments. All but the first deficiencies of the LWR model can be addressed, to various degrees of success, through the introduction of higher-order approximations or "dynamic" fundamental diagrams. This leads to two classes of traffic flow models—higher-order and lower-order continuum models. Higher-order models introduce a dynamic speed-concentration or flow concentration relation that accounts for driver reaction time and anticipation of traffic conditions ahead, i.e.,

$$v(x, t + \tau) = V_*(k(x + \Delta x, t)),$$

$$q(x, t + \tau) = kV_*(k(x + \Delta x, t)).$$

The approximations of these dynamic relations lead to evolution equations for travel speed or flow rate (Payne 1971, Whitham 1974, Zhang 1998). Consequently the traffic model becomes a system of partial differential equations and its solutions, when shown in the phase plane, deviate from the equilibrium fundamental diagram, producing the scatter and forward waves in the congested region. Lower-order models, on the other hand, models different traffic motions—acceleration, deceleration and coasting—explicitly on the fundamental diagram. Their fundamental diagrams $f_*(k)$ has multiple branches and connecting curves, each describes a particular kind of motion (Newell 1965, Zhang 2001). For one reason or another, higher-order models are more widely studied and used than lower-order models. Therefore we will focus our presentation on higher-order models in the remaining text. Readers who are interested in lower-order traffic flow models can refer to Newell (1965), Daganzo (1999a,b) and Zhang (2001).

3 Higher-order continuum models

The development of higher-order traffic flow models again originated from the seminal work of Lighthill and Whitham (1955), in which they suggested the following higher-order extension of their first-order model:

$$\text{The LW model:} \qquad q_t + Cq_x + Tq_{tt} - Dq_{xx} = 0, \tag{25}$$

where C is convection speed, T is a reaction time constant and D is the diffusion coefficient. Because of a lack of strong experimental evidence in support of such an extension, higher-order

approximations of traffic flow were not pursued further till 1971, when Payne (1971) and later Whitham (1974) derived a so-called 'momentum equation' from a car-following argument:

$$kv_t + vv_x + \frac{c_0^2}{k}k_x = \frac{v_*(k) - v}{\tau}, \tag{26}$$

where $v_*(k)$ is the equilibrium speed-concentration relation, $c_0 < 0$ is the 'sound' speed, and is given by $c_0^2 = \frac{\mu}{\tau}$, where μ is often referred to as the anticipation coefficient and τ relaxation time [2]. Recently, Zhang (1998) proposed a new addition to the existing momentum equations:

$$v_t + vv_x + kv_*'(k)^2 k_x = \frac{v_*(k) - v}{\tau}, \tag{27}$$

which is structurally similar to the momentum equation of Payne (1971) and Whitham (1974).

On the left hand side of the momentum equations, the second term is the change of speed due to convection, and the third term captures drivers' adjustment to travel speeds owing to anticipation. The term on the right hand side captures drivers' affinity to equilibrium travel speeds. In the momentum equations, the acceleration of a vehicle, expressed by the material derivative $v_t + vv_x$, responds negatively to the increase of concentration downstream, and positively (negatively) to travel speeds that are lower (higher) than the corresponding equilibrium speeds for the same concentration. As a result, travel speed v in both momentum equations usually differs from the equilibrium speed $v_*(k)$ under the same traffic condition, but this difference is reduced over time because of relaxation effects. The parameter τ decides the strength of relaxation. In literature τ is often interpreted as driver reaction time, whose value ranges from 1 sec to 1.8 sec.

Through the definition of a concentration dependent sound speed $c(k)$ ($= kv_*'(k)$, or c_0, or $-\sqrt{\frac{\mu}{\tau}}$), both momentum equations can be expressed in a general form:

$$v_t + vv_x + \frac{c^2(k)}{k}k_x = \frac{v_*(k) - v}{\tau}. \tag{28}$$

This evolution equation of travel speed, coupled with the continuity equation

$$k_t + (kv)_x = 0, \tag{29}$$

forms what we call a generalized PW higher-order model. The general PW model comprises a system of partial differential equations that can be compactly expressed using vector notation

$$\begin{pmatrix} k \\ v \end{pmatrix}_t + \begin{pmatrix} v & k \\ \frac{c^2(k)}{k} & v \end{pmatrix} \begin{pmatrix} k \\ v \end{pmatrix}_x = \begin{pmatrix} 0 \\ \frac{v_* - v}{\tau} \end{pmatrix}. \tag{30}$$

or

$$U_t + A(U)U_x = R(U), \tag{31}$$

where $U = \begin{pmatrix} k \\ v \end{pmatrix}$, $A(U) = \begin{pmatrix} v & k \\ \frac{c^2(k)}{k} & v \end{pmatrix}$ and $R(U) = \begin{pmatrix} 0 \\ \frac{v_* - v}{\tau} \end{pmatrix}$.

[2] There are other versions of (26)that differ in values taken by μ. Payne (1971), for example, gives $\mu = -\frac{v_*'}{2}$

This system is strictly hyperbolic because its characteristics, the eigenvalues of the Jacobian matrix $A(U)$, are real and distinctive

$$\lambda_{1,2} = v \pm c(k),\ \lambda_1 < \lambda_2. \tag{32}$$

The same is true for the LW model, because, through introduction of two auxiliary variables $w = q_t$ and $z = q_x$ the LW model can be transformed into a system of PDEs

$$\begin{pmatrix} w \\ z \end{pmatrix}_t + \begin{pmatrix} 0 & 1 \\ -\frac{D}{T} & 0 \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix}_x = - \begin{pmatrix} 0 \\ w + Cz \end{pmatrix}, \tag{33}$$

whose characteristics are

$$\lambda_{1,2} = \mp\sqrt{\frac{D}{T}},\ \lambda_1 < \lambda_2.$$

The two higher-order models differ, however, in that the general PW model is genuinely nonlinear while the LW model is linearly degenerate.

Although both the LWR model and the general PW model are hyperbolic, the latter has two characteristics, one is always slower than traffic and the other always faster than traffic, owing to $c(k) < 0$. This is constrasted to the single characteristic of the LWR model, $\lambda_* = v_*(k) + kv_*'(k)$ that is always slower than traffic. These differences have profound consequences on the behavior of these models, which we shall discuss in the context of various kinds traffic waves.

3.1 Propagation of traffic sound waves in higher-order models

We first note that traffic sound waves—the propagation of small disturbances in homogeneous traffic—travel at characteristic speeds in the LWR model. In higher-order models, there are two families of characteristics, thus two characteristic speeds. If a disturbance is still propagated at characteristic speeds, then it travels in both directions of traffic with different speeds, reaching drivers from front and behind. This can be checked through writing the higher-order model in question, here the LW model, in a special form

$$\left\{ (\partial_t + C\partial_x) + \left(\partial_t + \sqrt{\frac{D}{T}}\partial_x \right) \left(\partial_t - \sqrt{\frac{D}{T}}\partial_x \right) \right\} q = 0.$$

where $\partial_t + () * \partial_x$ is called a wave operator.

From this special form, one can clearly see that the LW model possesses three families of waves: the first order wave traveling at the convection speed $c_* = C$ (which is also the characteristic speed of the corresponding first-order model $q_t + Cq_x = 0$), the slower second order wave traveling at the first characteristic speed $c_1 = -\sqrt{\frac{D}{T}}$ and the faster second order wave traveling at the speed of the second characteristic $c_2 = \sqrt{\frac{D}{T}}$. For certain parameter values of D and T, the second characteristic speed can be greater than the convection speed, indicating that fast waves reach traffic from behind (*the rear view mirror effect*). In fact, this is also required for LW traffic to be stable. Otherwise

the first-order signals would violate the second-order signals, which leads to traffic instability. The general stability condition for the above higher-order model is

$$c_1 \leq c_* \leq c_2, \tag{34}$$

that is, the first order waves are sandwiched between the two second order waves. This stability condition can be derived from Fourier stability analysis (e.g., Whitham 1974, Kühne 1984, Zhang 1999).

For the general PW model, small perturbations around equilibrium points (k_0, v_0) would propagate in the same way as the waves in the LW model, only with different wave speeds. Instead of $c_* = C,\ c_1 = -\sqrt{\frac{D}{T}},\ c_2 = \sqrt{\frac{D}{T}}$, here we have $c_* = \lambda_*(k_0),\ c_1 = v_0 + c(k_0)$, and $c_2 = v_0 - c(k_0)$. Because of $c(k) < 0$, fast waves in the general PW model also reaches vehicles from behind.

There are subtle differences among special cases of the general PW model. In the PW model, where $c(k) = c_0$, the stability condition (34) can be violated and waves grow in magnitude and eventually become shocks in the form of roll waves (Whitham 1974, Kuhne 1984) while in Zhang's model where $c(k) = kv_*^{'}(k)$ the stability condition is always satisfied and waves always damp in magnitude (Zhang 1999). Thus, like the LWR model, Zhang's model is also inherently stable. Moreover, the asymptotic behavior of small perturbations of the kind $k = k_0 + \xi(x,t),\ v = v_0 + w(x,t)$ near equilibrium point (k_0, v_0) ($v_0 = v_*(k_0)$ and ξ, w are small perturbations) in the PW and Zhang models also differ. Such perturbations to Zhang's model can be accurately approximated by a convection equation

$$\xi_t + \lambda_*(k_0)\xi_x = 0, \tag{35}$$

while those to the PW model can be accurately approximated by a diffusion equation

$$\xi_t + \lambda_*(k_0)\xi_x = D_*\xi_{xx}, \quad \text{or} \tag{36}$$

$$\xi_{t'} = D_*\xi_{x'x'} \tag{37}$$

if a moving coordinate $x^{'} = x - \lambda_*(k_0)t,\ t^{'} = t$ is used, where D_* is the diffusion coefficient, and is given by $-(\lambda_1(k_0, v_0) - \lambda_*(k_0))(\lambda_2(k_0, v_0) - \lambda_*(k_0))$, a positive constant (Whitham 1974, del Castillo et al. 1994, and Zhang 1999). Driven by relaxation, any non-equilibrium state (k, v) in the PW model will become closer and closer to its corresponding equilibrium state $(k, v_*(k))$ with the increase of time. This latter behavior, together with the properties of Eq. 37 (i.e., its solutions are smooth), implies that at the end of a queue the PW model does backward smoothing to a sharp density/speed profile, thus predicting possibly negative travel speeds (Daganzo 1995a). Zhang's model, however, is absent of this problem because its solutions are not diffusive.

3.2 Propagation of shock and expansion waves

When traffic conditions undergo sharp transitions, shocks or expansion waves arise. The LWR theory possesses both types of waves and whether a particular kind of wave arises is determined by the entropy condition. A LWR shocks has zero width and travels at a particular speed given by

the Rankine-Hugoniot condition $s = [f_*]/[k]$, which is derived from the integral conservation law. The general PW model also has similar properties. It has both kinds of waves, actually two for each kind—one associated with the first characteristic and the other with the second characteristic. The first is referred to as 1-shock or 1-rarefaction waves, and the second 2-shock or 2- rarefaction waves. Whether a shock or expansion wave arises in a general PW solution is also dependent on entropy conditions, and they are as follows (Zhang 2000a):

$$\text{1-Shock, (E-H1):} \quad \lambda_1(U_r) < s < \lambda_1(U_l),\ s < \lambda_2(U_r), \tag{38}$$

$$\text{2-Shock, (E-H2):} \quad \lambda_2(U_r) < s < \lambda_2(U_l),\ s > \lambda_1(U_l), \tag{39}$$

$$\text{1-Rarefaction, (E-R1):} \quad \lambda_1(U_l) < \lambda_1(U_r), \tag{40}$$

$$\text{2-Rarefaction, (E-R2):} \quad \lambda_2(U_l) < \lambda_2(U_r), \tag{41}$$

And the speeds of shocks are also given by the Rankine-Hugoniot conditions derived from integral conservation laws. A difficulty arises, however, from the fact that the momentum equation is not a conservation law, it is simply an evolution equation for travel speeds. As a result, we do not have a *unique* corresponding integral conservation law for the momentum equation, that is, there are many different integral conservation forms that lead to the momentum equation. Two such examples are:

$$\begin{aligned} \frac{\partial}{\partial t}\int_{x_1}^{x_2} v(x,t)dx &= \left(\frac{v^2}{2}+\phi(k)\right)(x_1,t) - \left(\frac{v^2}{2}+\phi(k)\right)(x_2,t) \\ &+ \int_{x_1}^{x_2}\frac{v_*-v}{\tau}dx, \quad \phi^{'}(k) = \frac{c^2(k)}{k}, \end{aligned} \tag{42}$$

$$\begin{aligned} \frac{\partial}{\partial t}\int_{x_1}^{x_2} (kv)(x,t)dx &= \left(\frac{(kv)^2}{k}+\psi(k)\right)(x_1,t) - \left(\frac{(kv)^2}{k}+\psi(k)\right)(x_2,t) \\ &+ \int_{x_1}^{x_2}\frac{kv_*-kv}{\tau}dx, \quad \psi^{'}(k) = c^2(k). \end{aligned} \tag{43}$$

We cannot tell, from physical principles, which integral form is "correct". The selection of a particular integral form should therefore be guided by field observations, i.e., choosing the one that produces closest shock speeds to field measurements. For illustration purposes, we shall take the simplest, i.e., the first integral form in our subsequent presentations. The same arguments can be applied directly to other integral forms. The first integral momentum equation leads to following *conservative differential form*:

$$v_t + \left(\frac{v^2}{2}+\phi(k)\right)_x = \frac{v_*-v}{\tau}.$$

together with the conservation of mass

$$k_t + (kv)_x = 0,$$

we obtain the system of "conservation " laws with a source term:

$$U_t + F(U)_x = R(U), \tag{44}$$

where $F(U) = \left(kv, \frac{v^2}{2} + \phi(k)\right)^t$, and U, R are as defined before.

The shock speeds of the general PW model is not affected by the presence of the source term, and is still given by the Rankine-Hugoniot condition:

$$s[U] = [F(U)], \tag{45}$$

or

$$s[k] = [kv],\ s[v] = \left[\frac{v^2}{2} + \phi(k)\right].$$

These equations imply a certain relation between k and v. This relation leads to the shock curves (1-shock & 2-shock) in the $k - v$ phase plane (Zhang 2000a).

Expansion (or rarefaction) waves of the general PW model, however, are influenced by the relaxation source term. The strength of this influence is determined by the relaxation time. In a short time (compared to the relaxation time), one can neglect the effect of the source term and obtain certain relations between k, v in an expansion wave solution. These are the 1-rarefaction and 2-rarefaction curves in the $k - v$ phase plane (Zhang 2000a). Over time, however, the cumulative effects of the source term, relaxation, build up into rarefaction wave solutions, and one addresses this problem by modifying the solutions obtained without the source term. Also because of relaxation, the general PW model either approaches to the LWR model or a viscous LWR model $k_t + (kv_*)_x = D_* k_{xx}$, depending on the choice of $c(k)$.

Clearly one cannot ignore the effects of relaxation. Nevertheless, the study of the general PW model without relaxation

$$U_t + F(U)_x = 0 \tag{46}$$

reveals much information about the properties of the general PW model with relaxation. And computations of numerical solutions to the general PW model often rely on the Riemann solutions of (46). Therefore we provide here the solutions of Riemann problems to the general PW model without relaxation (46) (see Zhang 2000a for details).

The solution to the Riemann problem

$$U_t + F(U)_x = 0, \tag{47}$$

$$U(x, t=0) = \begin{cases} U_l, & x < 0 \\ U_r, & x > 0 \end{cases} \tag{48}$$

are of 8 kinds:

1. 1-shock:

$$\text{H1: } v_r - v_l = -\sqrt{\frac{2(k_l - k_r)(\phi(k_l) - \phi(k_r))}{k_l + k_r}},\ k_l < k_r. \tag{49}$$

2. 2-shock:

$$\text{H2: } v_r - v_l = -\sqrt{\frac{2(k_l - k_r)(\phi(k_l) - \phi(k_r))}{k_l + k_r}},\ k_l > k_r. \tag{50}$$

3. 1-rarefaction:

$$\text{R1: } v_r = \int_{k_l} \frac{c(k)}{k} dk,\ k_l > k_r \tag{51}$$

4. 2-rarefaction:

$$\text{R2: } v_r = \int_{k_l} \frac{c(k)}{k} dk,\ k_l < k_r \tag{52}$$

5. 1-shock + 2-shock:

$$\text{H1: } v_m - v_l = -\sqrt{\frac{2(k_l - k_m)(\phi(k_l) - \phi(k_m))}{k_l + k_m}},\ k_l < k_m. \tag{53}$$

$$\text{H2: } v_r - v_m = -\sqrt{\frac{2(k_m - k_r)(\phi(k_m) - \phi(k_r))}{k_m + k_r}},\ k_m > k_r. \tag{54}$$

6. 1-rarefaction + 2-rarefaction:

$$\text{R1: } v_m = \int_{k_l} \frac{c(k)}{k} dk,\ k_l > k_m \tag{55}$$

$$\text{R2: } v_r = \int_{k_m} \frac{c(k)}{k} dk,\ k_m < k_r \tag{56}$$

7: 1-rarefaction + 2-shock:

$$\text{H1: } v_m - v_l = -\sqrt{\frac{2(k_l - k_m)(\phi(k_l) - \phi(k_m))}{k_l + k_m}},\ k_l < k_m. \tag{57}$$

$$\text{H2: } v_r - v_m = -\sqrt{\frac{2(k_m - k_r)(\phi(k_m) - \phi(k_r))}{k_m + k_r}},\ k_m > k_r. \tag{58}$$

8: 1-shock + 2-rarefaction:

$$\text{H1: } v_m - v_l = -\sqrt{\frac{2(k_l - k_m)(\phi(k_l) - \phi(k_m))}{k_l + k_m}},\ k_l < k_m. \tag{59}$$

$$\text{R2: } v_r = \int_{k_m} \frac{c(k)}{k} dk,\ k_m < k_r \tag{60}$$

where $U_m = (k_m, v_m)^t$ is an intermediate state that provides the transition from a 1-wave to a 2-wave.

The transitions are most clearly seen from the phase diagram (Fig. 5). Note that for a state U_l, there are four special curves (H1, R1, H2, R2) emanating from that point, dividing the quarter plane of $k - v$ into four regions: I, II, III, IV. If the downstream U_r in the Riemann problem falls in region I, the solution would be of Type 6 (R1+ R2); in region II, Type 7 (R1+ H2), in region III, Type 5 (H1+H2), and in region IV, Type 8 (H1 + R2). Because of the entropy conditions, the intermediate states always fall on a 1-wave curve (i.e., R1, H1). When U_r falls on a particular wave curve, then the transition involves only one kind of wave and no intermediate state is produced. These are the types 1-4 solutions.

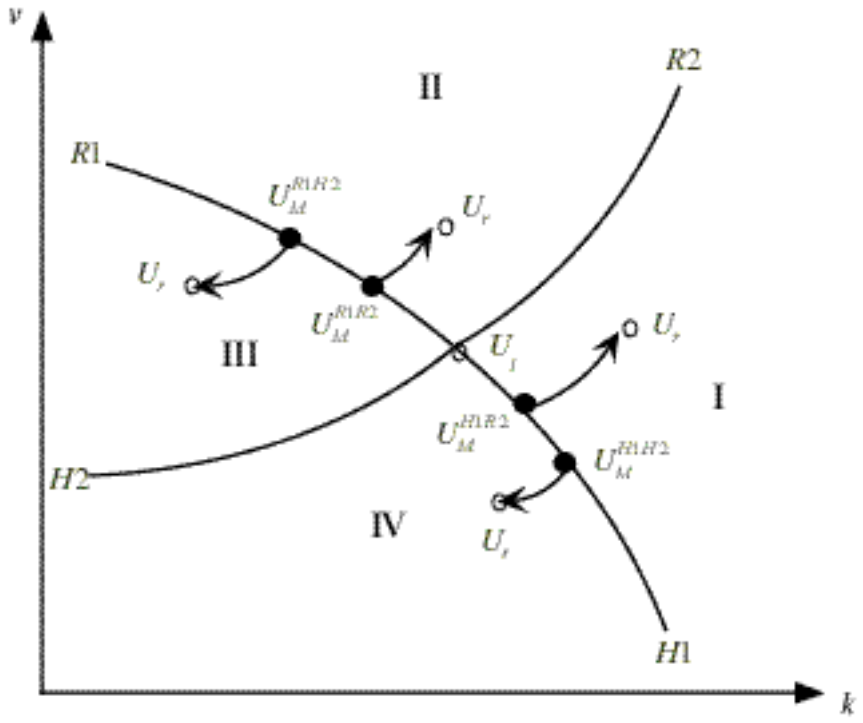


Figure 5: Phase transition diagram in the solution of Riemann problems (adopted from Zhang (2000a))

3.3 Traveling waves, instability and roll waves

Besides traffic sound waves, shock waves and expansion waves, there is another kind of waves in the general PW theory. This is the so-called traveling waves (Whitham 1974, Zhang 1999). It is a wave that has a smooth profile and travels at a constant speed, and takes the following form: $(k, v)(x, t) = (k, v)(\chi),\ \chi = x - st$ (Fig. 7(a)). For the traveling wave solution to arise, a certain stability condition has to be met. When this condition is violated, another kind of waves, the roll waves, may arise. A roll wave is a series of smooth, monotonic profiles separated by jumps (Fig. 6). The existence of roll waves is a direct consequence of traffic instability.

In this section we first obtain the traveling waves, then construct roll waves based on the traveling wave solution. Recall that a traveling wave is a steady profile with a translation speed s:

$$(k, v)(x, t) = (k, v)(\chi),\ \chi = x - st. \tag{61}$$

Substitute the traveling wave solution (61) into the general PW model, one obtains

$$-sk_\chi + (kv)_\chi = 0, \tag{62}$$

$$-sv_\chi + vv_\chi + \frac{c^2(k)}{k}k_\chi = \frac{v_* - v}{\tau}, \tag{63}$$

which after integration and further substitution of terms become

$$\left(c^2(k) - \frac{Q^2}{k^2}\right) k_\chi = f_*(k) - Q - ks, \tag{64}$$

where Q is an integration constant and

$$Q = k(v - s). \tag{65}$$

Because the two states at $x = \pm\infty$, $(k, v) = (k_{1,2}, v_{1,2})$ both satisfy (65), one can compute the speed of the traveling wave

$$s = \frac{k_1 v_1 - k_2 v_2}{k_1 - k_2},$$

which is the same as the shock speed given by the Rankine-Hugonoit condition.

Let $h(k)$ denote the right hand side of (64), then $h^{"}(k) = f^{"}(k) < 0$. Moreover, $h(k)$ crosses the $k-$axis at most twice at $k_a,\ k_b,\ k_a \leq k_b$. These crossing points corresponding to equilibrium concentration values and in between them $h(k) > 0$ (Zhang 1999). Since both $(k, v) = (k_{1,2}, v_{1,2})$ are equilibrium points, $k_{1,2}$ are also roots of $h(k)$. Moreover, we have $k_1 < k_2$, or $k_\chi > 0$ from the fact that $f_*(k)$ is strictly concave. Thus, for a smooth profile of $k(\chi)$ to exist, we must have

$$c^2(k) - \frac{Q^2}{k^2} > 0,$$

which yields the following stability condition

$$v + c(k) < s < v - c(k). \tag{66}$$

That is, the traveling wave must travel slower than the fast characteristic and faster than the slow characteristic.

When the stability condition is met, the smooth traveling wave profile can be obtained by integrating

$$\frac{d\chi}{dk} = \frac{c^2(k) - \frac{Q^2}{k^2}}{f_*(k) - Q - ks}, \tag{67}$$

which is given by:

$$\chi = \int^k \frac{c^2(\eta) - \frac{Q^2}{\eta^2}}{f_*(\eta) - Q - \eta s} d\eta. \tag{68}$$

When the stability condition (66) is violated, that is, the traveling wave travels either faster than the fast characteristic or slower than the slow characteristic, a smooth profile connecting k_1 and k_2 is no longer possible and a discontinuity (i.e., shock) must be inserted at the location where the wave turns back on itself. This time one still obtains a monotonic profile but with a shock separates two smooth pieces (Fig. 7(b)), and the speed of the shock is determined by the R-H shock condition.

There is a special case when the stability condition is not met. This is the case where both the numerator and denominator of (67) vanishes, which leads to non-smooth and periodic solutions to

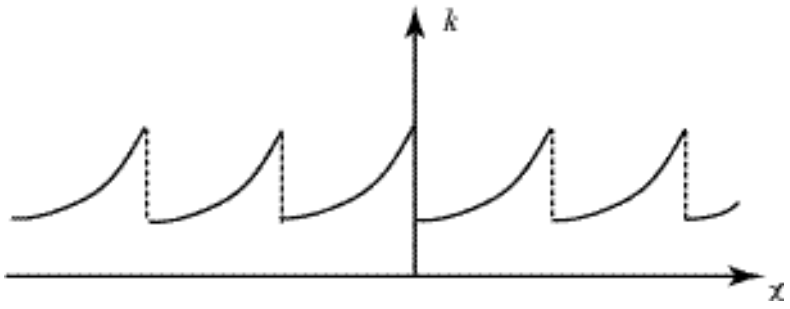


Figure 6: Roll waves in the moving coordinate χ

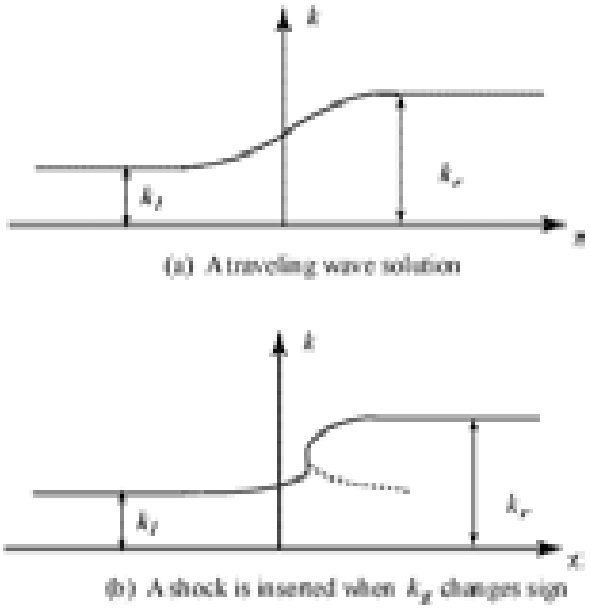


Figure 7: Traveling waves and shocks in the PW model

the general PW model, often referred to as *roll waves* in hydraulic literature. Note that the physical solution of these two algebraic equations is $Q = -k_0 c(k_0)$, where Q is the constant flux measured relative to the moving coordinate χ and k_0 is the critical density that makes both the numerator and denominator of (67) vanish. It is necessary that (k_0, v_0) is an equilibrium state, which fixes the shock speed $s = v_*(k_0) + c(k_0)$. The critical state also fixes the integration constant in the profile equation (68). The roll waves can then be constructed, in the same manner as shown in Dressler (1949), by piecing together the smooth profiles with shocks at appropriate locations, which in turn are determined by the R-H shock conditions. Although the existence of roll wave solutions in the PW model were known to transportation researchers not long after the development of the PW model (e.g., Leutzbach 1985), no one to date has obtained specific roll wave solutions to compare with real world observations of stop-start waves. With the development of Section 3.2, however, this can now be done.

3.4 Summary and Discussions

The PW-like higher-order traffic flow models extend the basic kinematic wave traffic flow model in two ways: they allow for non-equilibrium phase transitions and introduce instability. These are achieved through the addition of a momentum equation that describes speed evolution. The resulting flow obtained from these higher-order models, however, are not all that different than those obtained from the basic kinematic wave model—shocks, for example, exist in both types of models. Moreover, owning to relaxation, similar types of solutions (e.g. shocks or expansion waves) of these two families of models become much similar in long time (typically 10τ).

On the other hand, some significant differences also exist between the two family of models. For one, higher-order models have two family of characteristics and waves as compared to one family of the kinematic wave model. While the first family of waves in the higher-order models behave much like those in the kinematic wave model, the second family behaves quite differently—they travel faster than traffic, and reaches vehicles from behind. This property of the second characteristic has led to doubts about the validity of higher-order traffic models and interesting discussions over the pros and cons of higher-order approximations of traffic flow in general (e.g., Daganzo 1995a, Papageorgiou 1998, Lebacque 1999, Zhang 2001). Because following vehicles usually cannot force leading ones to speed up or slow down. fast-than-traffic waves are quite unrealistic when traffic is on a single-lane highway. When traffic is flowing on a multi-lane highway and passing is allowed, such faster-than-traffic waves do arise as a result of 1) lane-changing, or 2) averaging, or both (see Zhang 2000c for details). Another difference between the two types of models is that higher-order models contain traveling wave and roll wave solutions while the KW model does not. Roll wave solutions are particularly interesting because of their similarity to observed stop-start waves.

Like the kinematic wave model, higher-order models can also be extended to model inhomogeneous roads. This is particularly straightforward with the PW model— one simply replace the homogeneous equilibrium speed-density relation $v_*(k)$ with a space-dependent one $v_*(k, x)$. The extension of higher-order models to model road networks, however, is not as straightforward and is

a worthwhile research topic for traffic flow researchers.

One of the motivations for developing the PW model was to remove shocks from the model solutions. This, however, is not fulfilled by the higher-order models that we have covered up to now, although the PW model does admit smooth traveling wave solutions. Higher-order space derivatives, in the form of k_{xx} or v_{xx} must be introduced to obtain shocks with a structure. We call these models diffusive or viscous (higher-order) models. In contrast, PW like models are called inviscid higher-order models. The next section briefly introduces some of the popular viscous traffic models, and a stochastic extension to a particular viscous higher-order model.

4 Diffusive, viscous and stochastic traffic flow models

4.1 Diffusive and viscous traffic flow models

The first diffusive traffic flow model, which was mentioned in the classical book of Whitham (1974) was obtained by considering a flux that is dependent not only on vehicle concentration, but also on the concentration gradient:

$$q = f_*(k) - \nu k_x,$$

which, after substitution into the conservation equation, leads to

$$k_t + f_*(k)_x = \nu k_{xx}, \tag{69}$$

where ν is a positive parameter.

This diffusive model, when $f_*(k)$ is quadratic, can be manipulated into the following form

$$c_t + cc_x = \nu c_{xx}, \quad c = f_*^{'}(k), \tag{70}$$

which is the well-known Burger's equation.

Analytical solutions to the Burger's equation with given initial data can be obtained through the so-called Cole-Hopf transformation. We refer the reader to Whitham (1974, pp 96-112) for the detailed solution formulas and only discuss the qualitative properties of various kinds of solutions to the Burger's equation. These are:

1. The solution to the initial value problem $c_t + cc_x = \nu c_{xx},\ c(x, t = 0) = F(x)$ always exists, and is smooth after $t > 0$.

2. When $\nu \to 0$, the solution approaches that of $c_t + cc_x = 0$.

3. For initial data $F(x) = \begin{cases} c_l & x < 0 \\ c_r & x > 0, \end{cases}$ $c_l > c_r$ (A step function), the solution is a traveling wave $c(x - st)$, with $s = \frac{c_1+c_2}{2}$. The width of the traveling wave, as measured by the range where 90% of the change in $c_l - c_r$ occurs, is proportional to $\frac{\nu}{c_l - c_r}$. As $\nu \to 0$, the width of the traveling wave becomes nil and the traveling wave becomes a shock of $c_t + cc_x = 0$. Thus the viscous model provides a shock structure to the LWR model.

4. For initial data $F(x) = A\delta(x)$ (A single hump), the solution is a nonlinear diffusion wave, similar to those of the heat equation $c_t = \nu c_{xx}$ when $R = \frac{A}{2\nu} \ll 1$.

5. Also, N-waves and periodic waves can be found in the solution with proper initial data.

Clearly, the addition of the second order derivative k_{xx} to the LWR conservation law does two things to it because of diffusion: 1) it smoothes the shocks of the LWR conservation law, thereby providing a shock structure, and 2) it guarantees the uniqueness of solution for small ν, thereby providing a way to pick solutions from the LWR conservation law. This latter property is often exploited in numerical computations of shock wave solutions.

The diffusion corrected conservation law of (69), when $f_*(k)$ is not quadratic, can be approximated by the Burger's equation. In fact, the aforementioned properties of the Burger's equation (except Property #2, which must be modified) are shared by all known diffusion corrected or viscous traffic flow models, including the following popular viscosity-corrected PW model:

$$\begin{pmatrix} k \\ v \end{pmatrix}_t + \begin{pmatrix} v & k \\ \frac{c^2(k)}{k} & v \end{pmatrix} \begin{pmatrix} k \\ v \end{pmatrix}_x = \begin{pmatrix} 0 \\ \frac{v_*-v}{\tau} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \nu \end{pmatrix} \begin{pmatrix} k \\ v \end{pmatrix}_{xx}. \tag{71}$$

The viscosity-corrected PW model, however, can become unstable in certain ranges of traffic and therefore has additional properties. These properties, including the collapse of homogeneous traffic under local and global perturbations and the formation of vehicle-clusters in stop-and-go traffic, are well documented in (Kerner &Konhaüser 1993, 1994; Kerner, Konhaüser & Schilke 1995; Kerner and Rehborn 1999; Kühne & Beckschulte 1993). Interested readers are referred to the aforementioned literature for detailed discussions of these properties.

4.2 Acceleration noise and a stochastic flow model

Same as section 5.3 in the original text.

5 Numerical approximations of continuum models

All of our continuum models are described by partial differential equations, some (i.e, LWR, PW and Zhang) are hyperbolic while others (e.g., the viscous model of (69)) parabolic. Proper initial/boundary conditions must be prescribed to these equations to form a well posed problem. The solutions of continuum models involving general initial/boundary conditions are tedious, if not difficult, to obtain analytically. Numerical procedures are often employed to solve such problems.

Typically, finite difference methods are applied to solve hyperbolic traffic flow models while finite difference or finite element methods are used to solve parabolic traffic flow models. Both methods start with a discretization of the time-space domain ($t \geq 0, -L < x < L$), with the following grid mesh being the most common:

$$\begin{aligned} x_i &= ih, \quad i = 0, \pm 1, \pm 2, \cdots, L/h, \\ t_j &= jk, \quad j = 0, \ \ 1, \ \ 2, \ \cdots, T/k. \end{aligned}$$

where $h \equiv \Delta x$ and $k \equiv \Delta t$, and (x_i, t_j) are the grid points of this mesh.

Let the values of $U(x,t)$ on those grid points denoted by U_i^j (see Fig. 8). Then the time-space derivatives in a continuum model can be approximated using values at these grid points, and we obtain a set of finite difference equation(s) (FDE). For example, the space derivative U_x can be approximated in a number of ways:

$$[U_x]_i^j = \begin{cases} \frac{U_i^j - U_{i-1}^j}{h}, & \text{Forward Difference} \\ \frac{U_{i+1}^j - U_{i-1}^j}{2h}, & \text{Center Difference} \\ \frac{U_{i+1}^j - U_i^j}{h}, & \text{Backward Difference} \end{cases} \tag{72}$$

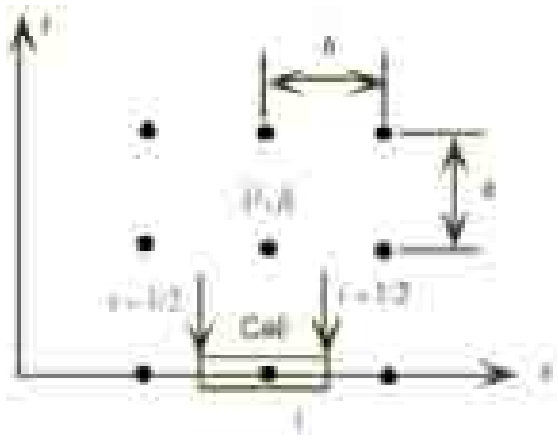


Figure 8: Time-space grid

However, the numerical approximation of continuum models is not a simple exercise of replacing continuous variables and their derivatives with discrete ones and their differences. This is particularly true for hyperbolic traffic models, because in these models discontinuities or shock waves can develop spontaneously even from smooth initial data. The existence of shocks presents a major challenge to the development of numerical approximations that are *consistent* with and *convergent* to the model equations when the mesh size is further and further refined. A valid approximation must meet the following three conditions:

1. it is consistent with the original PDE, i.e., the FDE converges to the PDE when h and κ approach 0 (*consistency*),

2. numerical errors introduced by the FDE do not increase over time (*stability*), and

3. its solutions converge to the right solutions of the original PDE when h and κ approach 0 (*convergence*).

In this section we present two consistent, stable and convergent finite difference approximations of the system (31) with initial/boundary data of (73) and a finite element approximation of the

viscous model (33) with the same initial and boundary data. It should noted that the LWR, PW and Zhang models can all be expressed in the form of (31). Therefore the finite difference approximations presented here applies to all three models. We begin our presentation with finite difference approximations.

5.1 Finite difference methods for solving inviscid models

In this section we present two finite difference schemes to solve the system of (31) with the following initial/boundary data:

$$\begin{array}{ll} \text{I.C.:} & U(x,0) = U_0(x),\ -L \le x \le L \\ \text{B.C.:} & U(-L,t) = U_-(t),\ U(L,t) = U_+(t),\ t \ge 0. \end{array} \tag{73}$$

where U_0 and $U_\pm$ are vector valued functions of space and time, respectively, and may contain a countable number of jumps. This system of equations include the Kinematic wave model of LWR, the higher-order models of Payne-Whitham and of Zhang.

It turns out that a particular form of hyperbolic PDEs, called the conservative form, is specially suited for developing finite difference schemes that ensure the aforementioned three conditions. A conservative form of (31), for example, is given by (44), i.e.,

$$U_t + F(U)_x = R(U).$$

(44) is called a conservative form because it arises from certain conservative phenomena in convective transport:

$$\frac{\partial}{\partial t}\int_L U(x,t)dx + \int_{\partial L} F(U)dx = \int_L R(U)dx. \tag{74}$$

For example, the conservation of vehicles on a finite road segment $[x_1, x_2]$ leads to a specific case of (74):

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} k(x,t)dx + (kv)(x_2,t) - (kv)(x_1,t) = \int_{x_1}^{x_2} r(k,v)dx. \tag{75}$$

Using the conservative form, we can develop a *conservative finite difference approximation* of (44). The advantage of a conservative finite difference approximation is that it ensures the correct computation of shock speeds. A finite difference approximation of (44) is conservative if it can be written as

$$\frac{U_i^{j+1} - U_i^j}{\kappa} + \frac{\tilde{F}(U_{i+1}^j, U_i^j) - \tilde{F}(U_i^j, U_{i-1}^j)}{h} = \tilde{R}(U_{i+1}^j, U_i^j, U_{i-1}^j), \tag{76}$$

where $\tilde{F}$ in (76) is called *numerical flux* (explained later) [3]. When a finite difference scheme is in conservative form, the condition for consistency is particularly simple. It requires that the numerical flux function satisfies

$$\tilde{F}(U,U) = F(U).$$

[3]The argument list of the numerical flux function can involve more than two nodal points, depending on the required accuracy of the finite difference approximation. (76) uses two nodal values to compute its numerical flux and is first order accurate.

Examples of consistent conservative finite difference schemes in traffic flow are the finite difference schemes of (Michalopoulos, Beskos & Lin (1984)), (Daganzo 1994), and (Lebacque, 1996) in the scalar case (i.e., the LWR model), the finite difference scheme of Leo and Pretty (Leo and Pretty 1992) and Zhang (2000b) in the system case (e.g., the PW model).

A consistent, conservative finite difference approximation, if linear, is guaranteed to converge to the correct solution if it meets a stability condition (LeVeque 1992). This condition, generally referred to as the CFL (Courant-Friedrichs-Lewy) condition, says that the cell advance speed $\frac{h}{\kappa}$ cannot be greater than the maximum absolute characteristic velocity, i.e.,

$$\max \left| \frac{\kappa}{h} \lambda_i \right| \leq 1, \ i = 1, \cdots, n.$$

Although this theorem has not been proven for most nonlinear systems, computational experiences indicate that the CFL condition is sufficient to ensure convergence for a large number of nonlinear systems. Thus we require the CFL condition in all of our finite difference approximations.

The remaining task in our finite difference approximations is to obtain the numerical flux function that ensures the consistency, stability and convergence properties. Before explaining what a numerical flux function is, we first make clear what U_i^j represents in our scheme of things. Suppose that $u(x,t)$ is a weak solution of the integral conservation law (74). We can write (74) as

$$\begin{aligned} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_{j+1})dx &= \int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_j)dx + \int_{t_j}^{t_{j+1}} \int_{x_{i-1/2}}^{x_{i+1/2}} R(u(x,t))dxdt \\ &- \left[\int_{t_j}^{t_{j+1}} F(u(x_{i+1/2},t)dt - \int_{t_j}^{t_{j+1}} F(u(x_{i-1/2},t)dt \right] \end{aligned} \tag{77}$$

where $i-1/2$ and $i+1/2$ denotes the left and right boundary of cell i (see Fig. 8), respectively. If we interpret U_i^j as the cell average

$$U_i^j = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_j)dx \tag{79}$$

then

$$\tilde{F}(U_{i+1}^j, U_i^j) = \frac{1}{\kappa} \int_{t_j}^{t_{j+1}} F(u(x_{i+1/2},t)dt \tag{80}$$

is the average flux passing through the cell boundary $x_{i+1/2}$ in the time interval (t_j, t_{j+1}), and

$$h\tilde{R} = \frac{1}{\kappa} \int_{t_j}^{t_{j+1}} \int_{x_{i-1/2}}^{x_{i+1/2}} R(u(x,t))dxdt$$

is the average inflow into cell i from the source during time interval (t_j, t_{j+1}). With these definitions, the integral conservation law (74) reduces to (76), and this is why (76) is called a conservative approximation.

There are a number of ways to construct a numerical flux function, perhaps the most intuitive and illustrative is the one obtained using Godunov's finite difference method. The Godunov method solves locally a Riemann problem at each cell boundary for the time interval (t_j, t_{j+1}), using the

cell averages U_i^j as initial data. It then pieces together these Riemann solutions at time t_{j+1} and average them using (79) to obtain new initial data for t_{j+2}, and this process is repeated until T/k is reached. Recall that the Riemann problem for the homogeneous equation

$$U_t + F(U)_x = 0 \tag{81}$$

of our continuum traffic models have been solved in Sections 2 and 3. We can then apply the principle of superposition to solve the traffic models with source terms:

$$U_t + F(U)_x = R(U). \tag{82}$$

Using the aforementioned procedure, we obtain a Godunov type of difference equation for (44):

$$\frac{U_i^{j+1} - U_i^j}{k} + \frac{F\left(U_{i+1/2}^{*j}\right) - F\left(U_{i-1/2}^{*j}\right)}{h} = \tilde{R}. \tag{83}$$

in which the numerical flux function reads

$$\tilde{F}(U_{i+1}^j, U_i^j) = F\left(U_{i+1/2}^{*j}\right), \tag{84}$$

and the source flux is computed by

$$\tilde{R} = R\left(\frac{U_{i+1} + U_{i-1}}{2}\right).$$

The variables $U_{i+1/2}^{*j}, i = 0, \pm 1, \pm 2, \cdots, \pm L/h$ are obtained from solving a series of Riemann problems to the homogeneous equation at cell boundaries. Owing to space limitations, we cannot in this monograph fully describe how this is done for higher-order models (interested readers are referred to Zhang 2000b). However, the computation of $U_{i+1/2}^{*j}$ is particularly simple when the equilibrium relation $v = v_*(k)$ is assumed. It leads to the following finite difference equation

$$\frac{k_i^{j+1} - k_i^j}{k} + \frac{f_*\left(k_{i+1/2}^{*j}\right) - f_*\left(k_{i-1/2}^{*j}\right)}{h} = 0, \; v_i^{j+1} = v_*(k_i^{j+1}), \tag{85}$$

where the cell boundary flow $f_*\left(k_{i+1/2}^{*j}\right)$ is given by a simple formula (LeVeque 1992, Bui, et. al. 1992):

$$f_*\left(\rho_{i+1/2}^{*j}\right) = \begin{cases} \min_{k_i^j \le k \le \rho_{i+1}^j} f_*(k), & \text{if } k_i^j < k_{i+1}^j \\ \max_{k_i^j \ge k \ge \rho_{i+1}^j} f_*(k), & \text{if } k_i^j > k_{i+1}^j \end{cases}.$$

When $f_*(k)$ is concave, this formula can be further streamlined through the introduction of a supply and demand function for each cell (Daganzo 1994, 1995b; Lebacque 1996):

$$\text{Demand: } D(i,j) = \begin{cases} f_*(k_i^j), & \text{if } \frac{j}{i} < k* \\ f(k^*), & \text{if } \frac{j}{i} \ge k* \end{cases}, \tag{86}$$

$$\text{Supply: } S(i,j) = \begin{cases} f_*(k_i^j), & \text{if } k_i^j > k* \\ f(k^*), & \text{if } k_i^j \le k* \end{cases}, \tag{87}$$

where k^* is the critical density at which $f_*^{'}(k^*) = 0$. And we have

$$f_* \left(k_{i+1/2}^{*j} \right) = \min\{D_i, S_{i+1}\}, \tag{88}$$

which also applies to boundary cells and bottlenecks.

Another difference approximation of (31) uses the idea of Lax-Friedrichs center differencing. It leads to the following numerical flux:

$$\tilde{F}(U_{i+1}^j, U_i^j) = \frac{F(U_{i+1}^j) + F(U_i^j)}{2} - \frac{h}{\kappa} \frac{U_{i+1}^j + U_i^j}{2}. \tag{89}$$

It is easy to check that this flux function meets the consistency requirement and the resulting finite difference approximation

$$U_i^{j+1} = \frac{U_{i+1}^j + U_i^j}{2} - \frac{\kappa}{2h} \left[F(U_{i+1}^j) - F(U_{i-1}^j) \right] + h\kappa\tilde{R}$$

is also conservative. This center difference scheme remained till recent years a popular choice of approximation of the kinematic wave model (e.g., Michalopolous 1988, Michalopolous, Beskos & Yamauchi 1984, Michalopolous, Kwon, & Khang 1991, Michalopolous, Lin, & Beskos 1987) and being used lately to approximate higher-order models (Zhang 2000d). Zhang and Wu (1999) investigated the convergence properties of this scheme and found that in comparison with Godunov-type of schemes, the center difference scheme has faster convergence rate with respect to expansion wave solutions and slower convergence rate with respect to shock solutions. The reason is that this difference scheme has built-in numerical viscosity which smoothes shocks.

5.2 Finite element methods for solving viscous models

Apart from finite difference methods, the method of finite element is also employed to solve continuum traffic flow equations. In this section we show how the latter is used to solve the viscosity-corrected PW equations.

First we introduce an auxiliary variable w : $w = v_x$, and normalize all the state and time-space variables in the following way:

$$k' = \frac{k}{k_{jam}} \qquad v' = \frac{v}{v_f} \qquad w' = \frac{w}{v_f} \qquad x' = \frac{x}{v_f \tau} \qquad t' = \frac{t}{\tau}. \tag{90}$$

Then the unknown variables in the viscosity-corrected PW model can be expressed by a vector

$$\eta = \begin{pmatrix} k' \\ v' \\ w' \end{pmatrix} \tag{91}$$

and the model itself by the following vector-valued quasi-linear partial differential equation:

$$A\eta_t + B\eta_x = C \tag{92}$$

with (note that for convenience the $'$'s are dropped from the notations)

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} v & 0 & 0 \\ \frac{1}{k}\frac{1}{Fr} & 0 & \frac{1}{R_e} \\ 0 & 1 & 0 \end{pmatrix} \qquad C = \begin{pmatrix} & -kw & \\ -vw & +v_* & -v \\ & w & \end{pmatrix}, \tag{93}$$

$$\begin{aligned} Fr \equiv \text{Froude number} &= \frac{\text{kinetic energy (inertia influence)}}{\text{potential energy(pressure)}} \\ &= \frac{\left(\frac{1}{2}\right) k v_f^2}{c_0^2 k} = \left(\frac{v_f}{c_0}\right)^2 \\ R \equiv \text{Reynolds number} &= \frac{\text{length velocity}}{\text{kinem. viscosity}} = \frac{v_f^2 \tau}{\nu_0}. \end{aligned} \tag{94}$$

The problem also comes with possibly two initial conditions and six boundary conditions. Because of the hyperbolic nature of the viscosity-corrected PW model, however, only certain combinations of these initial/boundary data are allowed.

In the finite element method, we replace the continuous functions

$$\eta(x,t) = \begin{pmatrix} k(x,t) \\ u(x,t) \\ w(x,t) \end{pmatrix} \tag{95}$$

by functions defined as a lattice:

$$\eta(x_0 + i\Delta x, t_0 + j\Delta t) \equiv \eta_{i,j} \tag{96}$$

and all derivatives by center difference quotients:

$$\eta_x \to \frac{1}{2\Delta x}(\eta_{i+1,j+1} - \eta_{i,j+1} + \eta_{i+1,j} - \eta_{i,j}) \tag{97}$$

and the function values by the midpoint values:

$$\eta_t \to \frac{1}{2\Delta t}(\eta_{i+1,j+1} + \eta_{i,j+1} - \eta_{i+1,j} - \eta_{i,j}). \tag{98}$$

$$\eta \to \frac{1}{4}(\eta_{i+1,j+1} + \eta_{i,j+1} + \eta_{i+1,j} + \eta_{i,j}) \tag{99}$$

We then do step-wise integration, starting from time step $j = 0$ and ending at time step $j = J$, as shown in Fig. (fig 5.16 from original text). To ensure the stability of the numerical procedure, an implicit integration scheme is used to compute the unknown variables $\eta_{i,j},\ \eta_{i+1,j}$ (for notational simplicity we'll drop subscript j in the remaining text of this section). It turns out that the Newtonian iteration procedure is perfectly suited for this purpose. In this procedure the variables $\eta_i,\ \eta_{i+1}$ are replaced by an approximation $\tilde{\eta}_i,\ \tilde{\eta_{i+1}}$ and the deviations $\delta\eta_i,\ \delta\eta_{i+1}$ are computed by linearizing the original equations. Denoting the deviation vector by

$$\delta_i \equiv \begin{pmatrix} \delta & k_i \\ \delta & u_i \\ \delta & w_i \end{pmatrix} \tag{100}$$

then the basic equations can be written in the form of

$$A_i\delta_{i+1} + B_i\delta_i = R_i \tag{101}$$

with

$$\left(\alpha = \frac{2}{\Delta x}, \beta = \frac{2}{\Delta t}, \kappa = \frac{1}{Fr} = \frac{c_0^2}{v_f^2}\right) \tag{102}$$

$$A_i = \begin{pmatrix} \beta + \alpha U + W & K_x & K \\ -U_e'(K) - \kappa\frac{K_x}{K^2} + \alpha\kappa\frac{1}{K} & \beta + W + 1 & U - \alpha\nu \\ 0 & \alpha & -1 \end{pmatrix} \tag{103}$$

$$B_i = \begin{pmatrix} \beta - \alpha U + W & K_x & K \\ -U_e'(K) - \kappa\frac{K_x}{K^2} - \alpha\kappa\frac{1}{K} & \beta + W + 1 & U + \alpha\nu \\ 0 & \alpha & -1 \end{pmatrix} \tag{104}$$

$$R_i = -4\begin{pmatrix} K_t + K_x U + K \\ U_t + UW - U_e(K) + U + \kappa\frac{K_x}{K} - \nu W_x \\ U_x - W \end{pmatrix} \tag{105}$$

where abbreviations

$$K = \frac{1}{4}(\overline{k}_{i+1} + \overline{k}_i + k_{i+1,j} + k_{i,j}) \tag{106}$$

$$K_t = \frac{1}{2\Delta t}(\overline{k}_{i+1} + \overline{k}_i - k_{i+1,j} - k_{i,j}) \tag{107}$$

are used.

Starting with the initial condition as the lowest approximation

$$\overline{\eta}_i = \eta_{i,j=0} \quad \eta_{i-1} = 0 \tag{108}$$

and using the left boundary condition

$$k_{i=0,j}, V_{i=0,j} \Longrightarrow \delta_0 = \begin{pmatrix} 0 \\ 0 \\ \delta w_0 \end{pmatrix} \tag{109}$$

the δ_i is computed recursively by

$$\delta_{i+1} = A_i^{-1}(R_i - B_i\delta_i) \tag{110}$$

as a function of δw_0, which in turn is determined by the right boundary condition

$$v_{i=I,j} \Longrightarrow \delta_I = \begin{pmatrix} \delta k_I \\ 0 \\ \delta w_I \end{pmatrix}. \tag{111}$$

An alternative rearrangement of the deviations δ_i is possible in order to produce a tridiagonal form which facilitates the fit of the boundary conditions (Kerner and Konhaüser, 1993).

5.3 Applications

5.3.1 Calibration of model parameters with field measurements

All the continuum models discussed in this monograph contain certain static relations and parameters. These relations include, for the LWR model, the fundamental diagram $f_*(k)$, and for higher-order models, $v_*(k)$. Since $f_*(k) = kv_*(k)$, knowing one would know the other. The parameters include, but not limited to, free flow speed v_f, jam wave speed c_j, sound speed $c_0(< 0)$ (PW model), jam density k_{jam}, critical density k_c, capacity q_c, and relaxation time τ (PW model). These relations and parameters capture certain fundamental characteristics of the *local* driving environment and driver population, and have to be calibrated/obtained locally before application of the corresponding models. The attainment of the parameters are achieved in two ways: direct measurement and data fitting. The former, when its cost is acceptable, is always preferred if there's a a choice of the two.

The interpretations of the parameters, in most cases, are straightforward and intuitive, which also suggest ways to measure them directly from field data. Among the various speeds, for example, one can easily measure free flow speed and jam wave speed, but not traffic sound speed c_0. For those parameters that can be directly measured, to obtain them is a simple matter of data gathering and processing and we will not elaborate on them here. Rather, we focus on the calibration of those parameters that have confusing interpretations in literature and are difficult to measure directly. These include sound speed c_0 and relaxation time τ.

Recall that the definition of traffic sound speed is the speed of sound waves minus the speed of traffic that carries these sound waves. In the LWR model, the sound wave speeds are $f_*^{'}(k)$, and the traffic speed is $v_*(k)$, therefore $c_0 = f_*^{'}(k) - v_*(k) = kv_*^{'}(k)$ is variable. That is, sound speed in the LWR (and Zhang's model for that matter) is not a fundamental parameter. The PW model, however, fixes the sound speed c_0 as a fundamental parameter and assumes that it is a constant. The latter assumption is questionable because it is unlikely that drivers respond to stimuli with the same intensity under free-flow and jam traffic conditions. With that being said, we turn our attention to the possible ways of measuring c_0. Recall that $\lambda_{1,2} = v \pm c_0$ in the PW model, which suggests that if we can measure the speed of the slower wave λ_1 and traffic speed v, we can then compute the sound speed c_0. This can be done with instrumented vehicles on a single lane-highway where one can measure the acceleration, speed and position of each vehicle in the traffic stream. Clearly, this is a costly way of obtaining sound speed and is rarely done in practice. In reality, c_0, together with another parameter τ, is obtained through data fitting.

Before describing the data fitting procedure, we want to reexamine the interpretation of τ, such that we have a sense of its range and would know roughly if the value obtained from our data fitting exercise makes sense. Relaxation in traffic flow refers to the process where non-equilibrium traffic approaches equilibrium traffic overtime[4]. The pace of this process is controlled by relaxation time τ. Clearly, τ takes the human reaction time as its lower limit, which is about $1 - 1.8$ seconds. Its

[4]one should not confuse the latter with free-flow traffic, although free-flow traffic is also equilibrium traffic, so is jam traffic!

upper limit, in theory, can be infinite. In practice, one never observes congestion that lasts longer than a day, not to mention infinite. The upper limit of relaxation, one speculates, would be in the order of a few minutes, the period of a stop-start wave. This speculation tends to be supported by existing calibration exercises (del Castillo and Benitez 1995, Cremer et al 1993, Kühne 1991, 1984, Papageorgiou et al 1990) that reported τ values ranging from1.8s to 108s.

The calibration of model parameters through data fitting usually involves the following steps:

1. Data collection: collection of road data in the forms of number of lanes, locations of ramps, and so forth, and traffic data, in the forms of time series data of flux, occupancy and spot speeds, at various locations,

2. Numerical approximations of the traffic flow model involved,

3. Calibration: obtain the fundamental diagram for each location from measured data, which in turn determines parameters such as free-flow speed, jam density, capacity flux, and jam wave speed, and obtain other parameters in the model through data fitting.

The process of data fitting involves minimizing some pre-defined performance measures. One common measure is the sum of square errors between model outputs and measured data:

$$PI = \gamma_1 \int dt \left(v_{cal.}(d,t) - v_{meas.}(d,t)\right)^2 + \gamma_2 \int dt \left(k_{cal.}(d,t) - k_{meas.}(d,t)\right)^2 \tag{112}$$

which is a function of model parameters to be calibrated, e.g.,

$$PI = PI(c_0, \tau) \tag{113}$$

for the PW model.

Because of the hyperbolic nature of the continuum traffic flow models, it is crucial to make sure that the finite difference or finite element approximations are correct and accurate. For this reason it is advocated that another step be added to the calibration or validation procedure: the step of checking the finite difference approximation (Zhang 2001). The best way to check the correctness of a numerical approximation, apart from theoretical considerations, is to run through benchmark problems, such as Riemann problems. In this way one can rid of transient and boundary conditions that may hide the inadequacies of the approximation (Zhang 2001).

For the specific examples of calibrating the model parameters, the readers are referred to the following literature:

- del Castillo and Benitez 1995 (A2 Amsterdam-Utrecht, the Netherlands).
- Kühne and Langbein-Euchner 1993 (A3 Fürth-Erlangen near Nuremberg, Germany)
- Papageorgiou et al 1990 (Boulevard Peripherique, Paris, France), and
- Sailer 1996 (Interstate 35W in Minneapolis, Minnesota, U.S.A.).

5.3.2 Multilane traffic flow dynamics

Numerical examples of a two-lane ring road to be provided by Panos?

5.3.3 Traffic flow on a ring road with a bottleneck[5]

In this section we use the finite difference approximations of the LWR and PW models developed earlier to simulate traffic on a ring road. The length of the ring road is $L = 800l = 22.4$ km. The simulation time is $T = 500\tau = 2500$ s $= 41.7$ min. We partition the road $[0, L]$ into $N = 100$ cells and the time interval $[0, T]$ into $K = 500$ steps. Hence, the length of each cell is $\Delta x = 0.224$ km and the length of each time step is $\Delta t = 5$ s. Since $\lambda_* \leq v_f = 5l/\tau$, we find the CFL condition number

$$\lambda_* \frac{\Delta t}{\Delta x} \quad \leq \quad 0.625 < 1.$$

Moreover, we adopt in this simulation the fundamental diagram used in (Kerner and Konhäuser, 1994; Herrman and Kerner, 1998) with the following parameters: the relaxation time $\tau = 5$ s; the unit length $l = 0.028$ km; the free flow speed $v_f = 5.0l/\tau = 0.028$ km/s $= 100.8$ km/h; the jam density of a single lane $\rho_j = 180$ veh/km/lane; $c_0 = 2.48445l/\tau = 0.014$ km/s $= 50.0865$ km/h; The equilibrium speed-density relationship is therefore

$$v_*(\rho, a(x)) = 5.0461 \left[\left(1 + \exp\{[\frac{\rho}{a(x)\rho_j} - 0.25]/0.06\} \right)^{-1} - 3.72 \times 10^{-6} \right] l/\tau,$$

where $a(x)$ is the number of lanes at location x. The equilibrium functions $v_*(\rho, a(x))$ and $f_*(\rho, a(x))$ are given in **Figure** 9.

The first simulation is about the homogeneous LWR model. Here we assume that the ring road has single lane everywhere; i.e., $a(x) = 1$, for $x \in [0, L]$, and use a global perturbation as the initial condition

$$\begin{array}{rcll} \rho(x,0) & = & \rho_h + \Delta\rho_0 \sin \frac{2\pi x}{L}, & x \in [0, L], \\ v(x,0) & = & v_*(\rho(x,0), 1), & x \in [0, L], \end{array} \tag{114}$$

with $\rho_h = 28$ veh/km and $\Delta\rho_0 = 3$ veh/km; and the corresponding initial condition (114) is depicted in **Figure** 10.

The results are shown as *contour plots* in **Figure** 11, from which we observe that initially wave interactions are strong but gradually the bulge sharpens from behind and expands from front to form a so-called N-wave that travels around the ring with a nearly fixed profile.

In the second simulation we created a bottleneck on the ring road with the following lane configuration:

$$a(x) \quad = \quad \begin{cases} 1, & x \in [320l, 400l), \\ 2, & \text{elsewhere .} \end{cases} \tag{115}$$

[5]The results in this section are from unpublished material of Jin and Zhang (2000a,b)

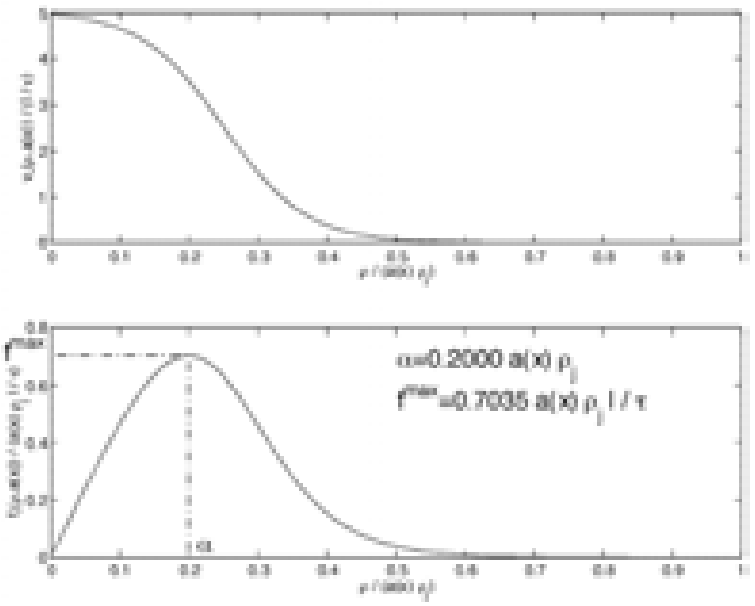


Figure 9: The Kerner-Konhäuser model of speed-density and flow-density relations

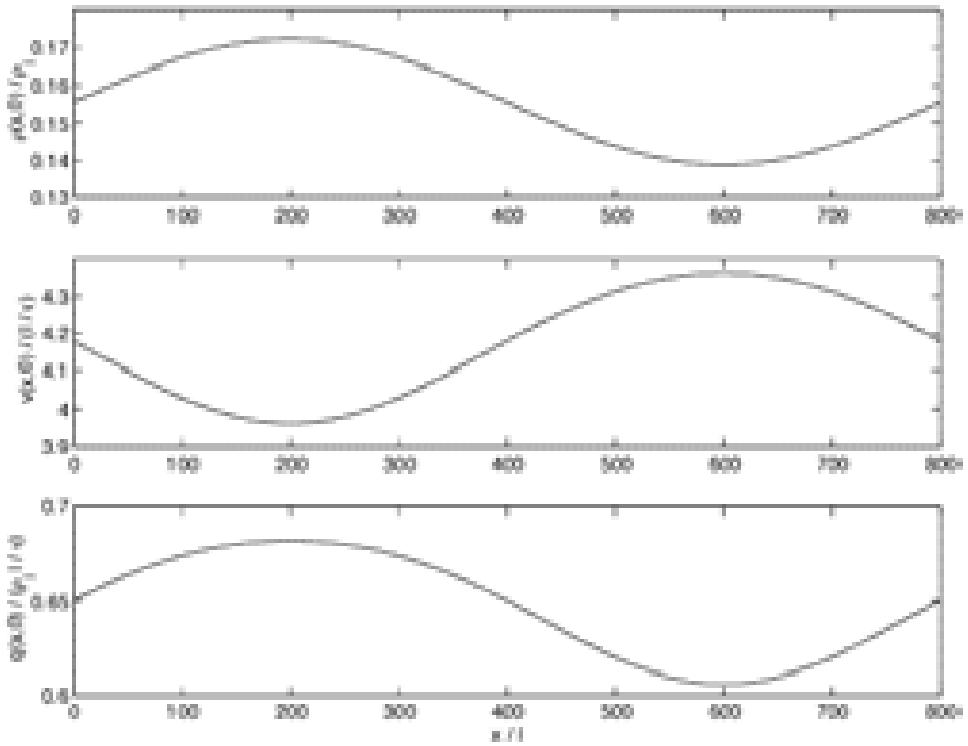


Figure 10: Initial condition (114) with $\rho_h = 28$ veh/km and $\Delta\rho_0 = 3$ veh/km

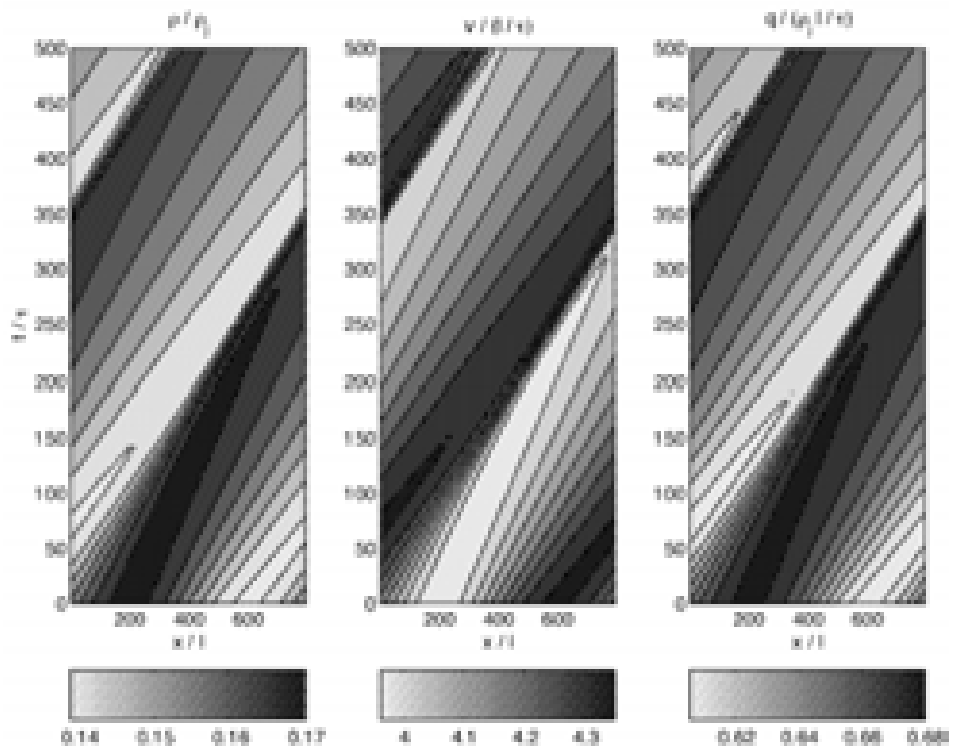


Figure 11: Solutions of the homogeneous LWR model with initial condition in **Figure** 10

As before, we also use a global perturbation as the initial condition

$$\begin{aligned} \rho(x,0) &= a(x)(\rho_h + \Delta\rho_0 \sin \tfrac{2\pi x}{L}), \quad x \in [0,L], \\ v(x,0) &= v_*(\rho(x,0), a(x)), \qquad\qquad x \in [0,L], \end{aligned} \tag{116}$$

with $\rho_h = 28$ veh/km/lane and $\Delta\rho_0 = 3$ veh/km/lane (the corresponding initial condition (116) is depicted in **Figure** 12).

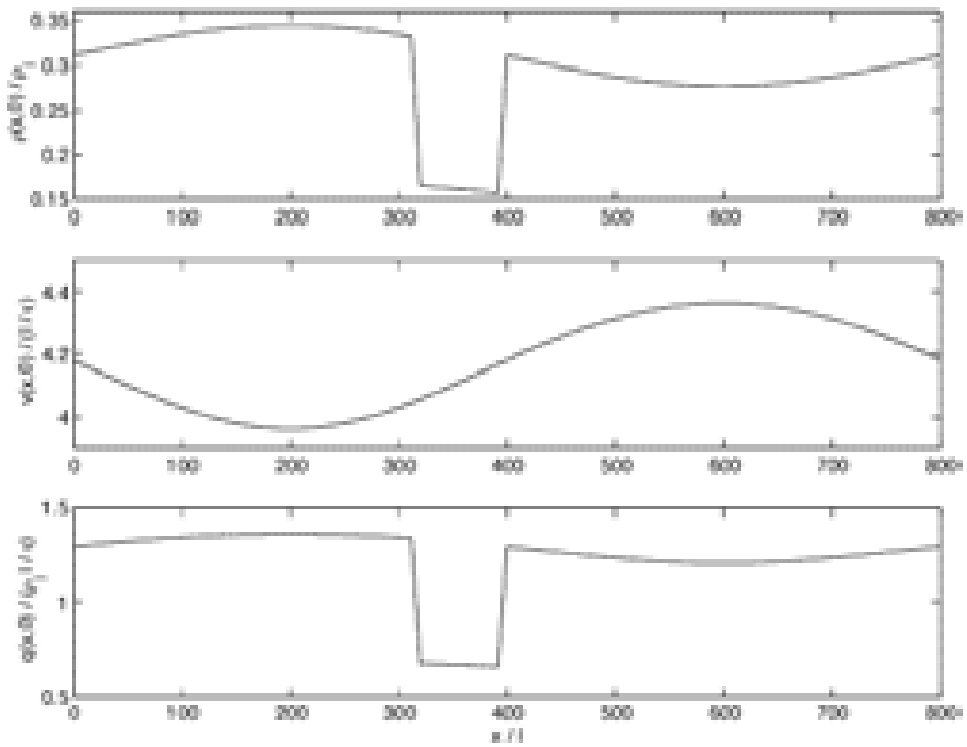


Figure 12: Initial condition (116) with $\rho_h = 28$ veh/km/lane and $\Delta\rho_0 = 3$ veh/km/lane

The results for this simulation are shown in **Figure** 13, and are more interesting. We observe from this figure that at first flow increases in the bottleneck to make the bottleneck saturated, then a queue forms upstream of the bottleneck, whose tail propagates upstream as a shock. In the same time, traffic emerges from the bottleneck accelerates in an expansion wave. After a while, all the commotion settles and an equilibrium state is reached, where a stationary queue forms upstream of the bottleneck, whose in/out flow rate equals the capacity of the bottleneck.

The third and fourth simulation runs are for the PW model, where we used the same initial conditions for density as in the first and second simulations, respectively, but different initial conditions for traffic speeds. These initial conditions (I.C.) are

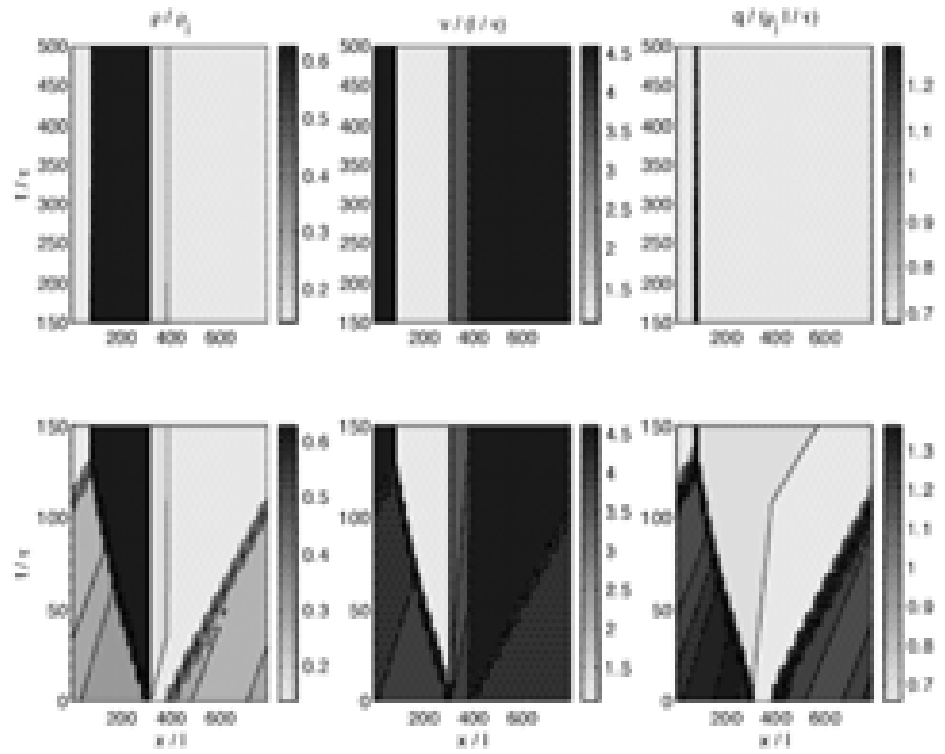


Figure 13: Solutions of the inhomogeneous LWR model with initial condition (116)

I.C. for the third simulation

$$\begin{aligned} a(x) &= 1, \quad x \in [0, L] \\ \rho(x,0) &= \rho_h + \Delta\rho_0 \sin\tfrac{2\pi x}{L}, \qquad x \in [0, L], \\ v(x,0) &= v_*(\rho_h, 1) + \Delta v_0 \sin\tfrac{2\pi x}{L}, \quad x \in [0, L]. \end{aligned} \tag{117}$$

I.C. for the fourth simulation

$$\begin{aligned} a(x) &= \begin{cases} 1, & x \in [320l, 400l) \\ 2, & \text{elsewhere} \end{cases} \\ \rho(x,0) &= a(x)(\rho_h + \Delta\rho_0 \sin\tfrac{2\pi x}{L}), \qquad x \in [0, L], \\ v(x,0) &= v_*(\rho_h, a(x)) + \Delta v_0 \sin\tfrac{2\pi x}{L}, \quad x \in [0, L]. \end{aligned} \tag{118}$$

The parameters are $\rho_h = 28$ veh/km, $\Delta\rho_0 = 3$ veh/km and $\Delta v_0 = 0.002$ km/s.

Again we use the same time step and cell size, which yields a CFL number of

$$\lambda_2 \frac{\Delta t}{\Delta x} \leq 0.9375 < 1$$

that ensures numerical stability of our finite difference approximation. The results of these simulation runs are shown in **Figure** 14 and **Figure** 15 respectively. Note that the PW model solution for the homogeneous road is slightly different that the corresponding LWR solution due to non-equilibrium initial speed, but the PW solution soon (about 10τ) looks very much like the LWR solution. This can be seen more clearly from time-slice plots of vehicle density, speed and flow rate shown in **Figure** 16. As can been seen from that figure, the solutions are nearly indistinguishable after $t = 140\tau$. In contrast, the PW solution for the inhomogeneous road, although shows similar patterns as the the corresponding LWR solution, does not converge to the LWR solution in long time (see **Figure** 15 & **Figure** 17). In long time, both solutions predict the same location of the tail anb head of the queue, but different discharge rate from the queue—traffic leaving the queue at capacity flow rate in the LWR solution, but below capacity flow rate in the PW solution (see **Figure** 17). This result highlights not only the differences between the two models, but also the importance and need for careful experimental validations of these models[6].

[6]The pikes in density that exceed jam density are possibly caused by traffic being unstable near the tail of the queue.

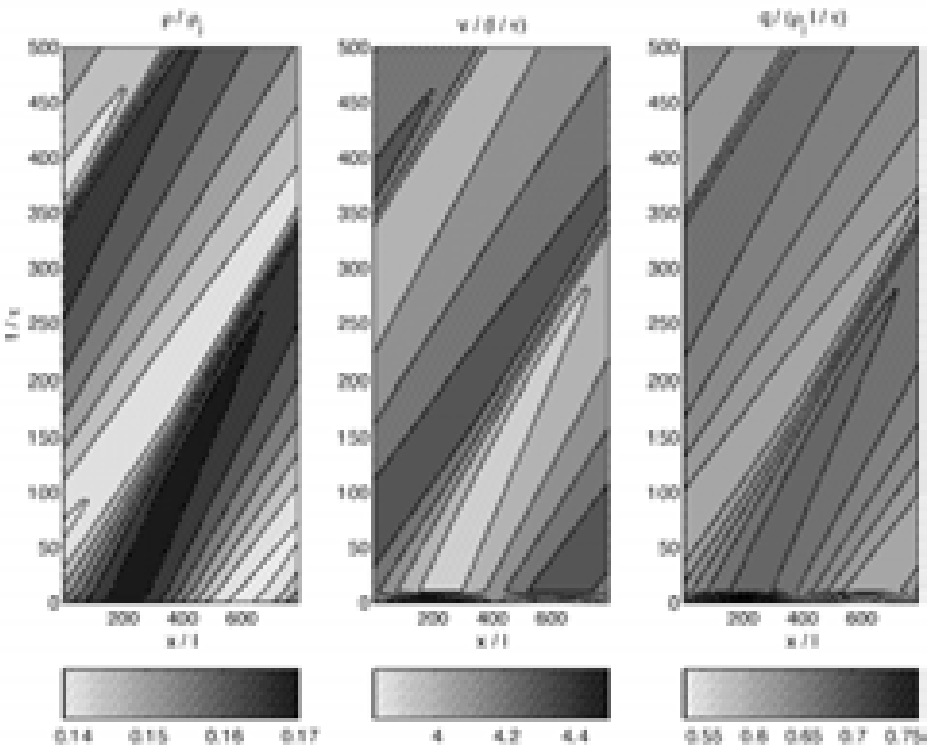


Figure 14: Solutions of the PW model with initial condition (117)

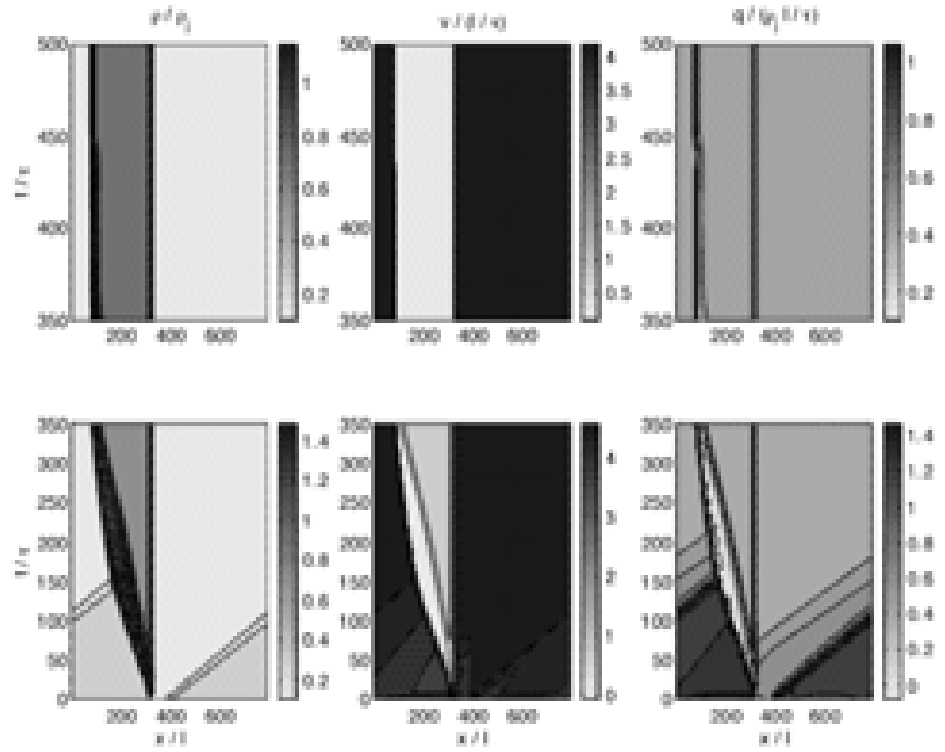


Figure 15: Solutions of the PW model with initial condition (118)

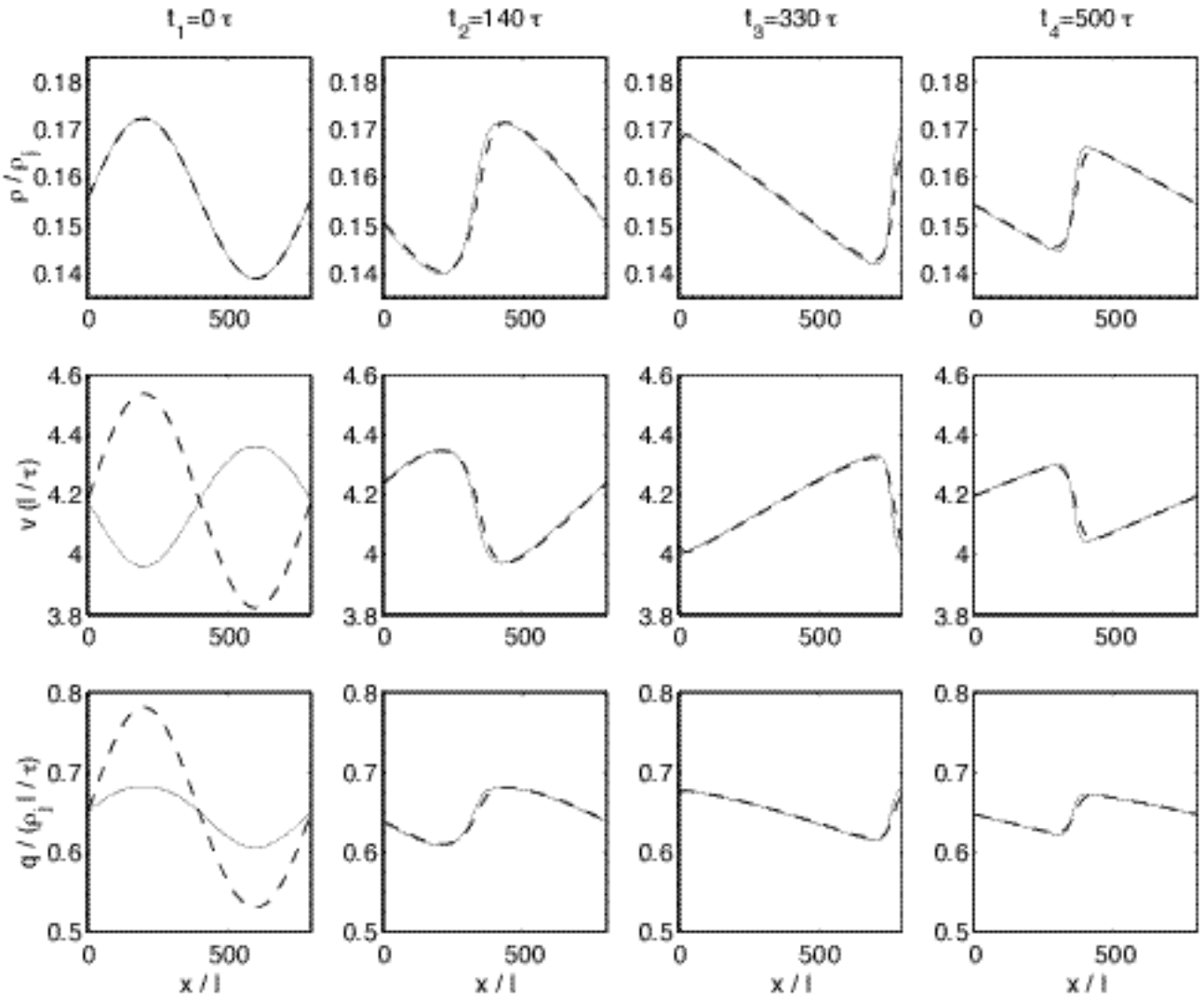


Figure 16: Comparison of the LWR model and the PW model on a homogeneous ring road: Solid line is used for the LWR model, and dashed line for the PW model

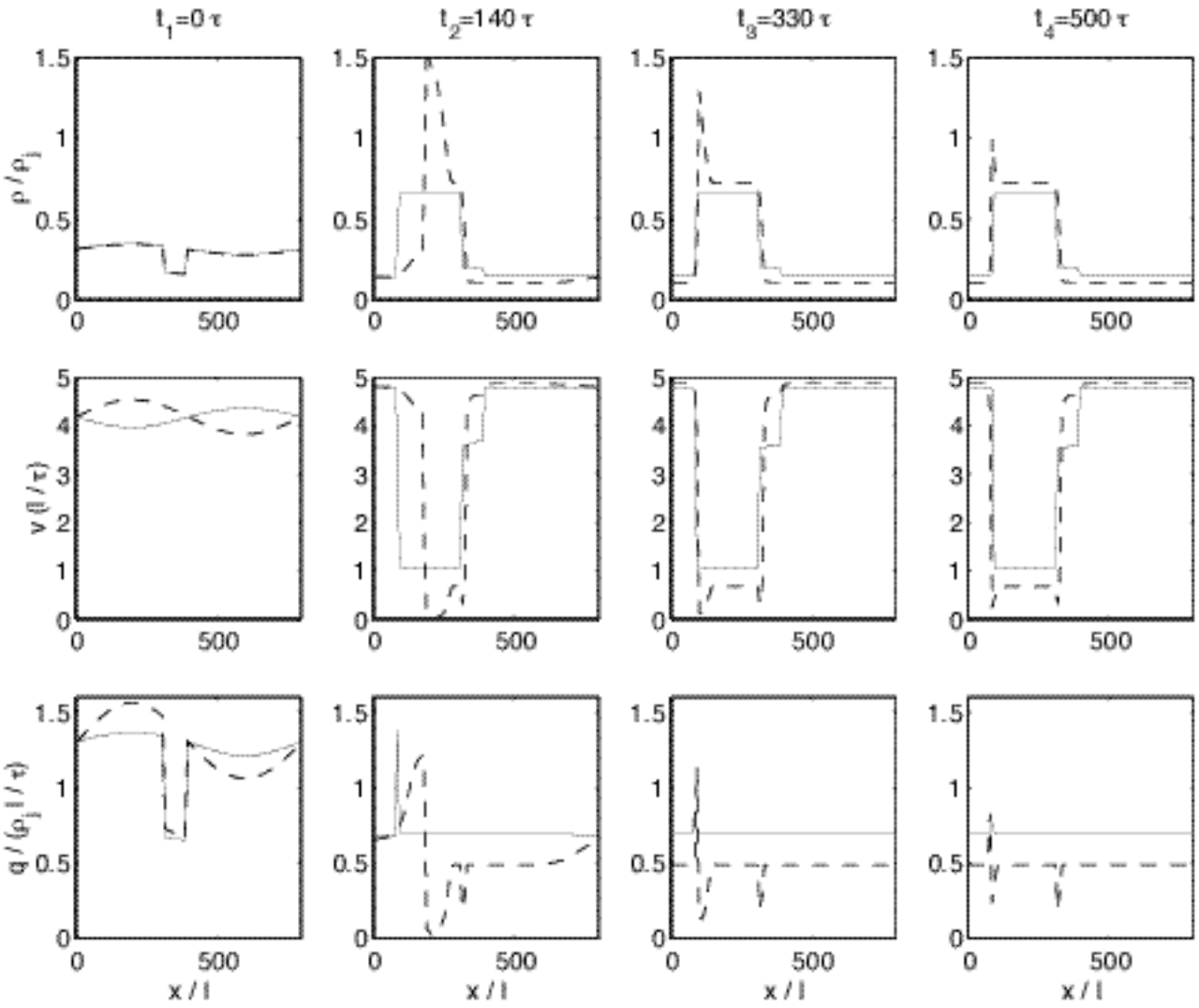


Figure 17: Comparison of the LWR model and the PW model on an inhomogeneous ring road: Solid line is used for the LWR model, and dashed line for the PW model

References

Bui, D., Nelson, P. & Narasimhan, S. L. (1992). Computational realizations of the entropy condition in modeling congested traffic flow, Technical Report, Texas Transportation Research Institute.

Courant, R. & Friedrichs, K. O. (1948) , *Supersonic flow and shock waves*, Interscience Publishers, Inc., New York.

Courant, R. and Hilbert, D. (1962). *Methods of mathematical physics*, Interscience Publishers, Inc., New York.

Cremer, M., F. Meiner, and S. Schrieber (1993). On Predictive Schemes in Dynamic Rerouting Strategies. In C. Daganzo (ed). *Theory of Transportation and Traffic Flow*, pp. 407-462.

Daganzo, C. F. (1994). The cell transmission model: a dynamic representation of highway traffic consistent with hydrodynamic theory, *Transpn. Res. B*, Vol. 28B, No.4, 269-287.

Daganzo, C. F. (1995a). Requiem for second-order approximations of traffic flow. *Transpn. Res. -B*, vol. 29B, No.4, 277-286.

Daganzo, C. F. (1995b). The cell transmission model, Part II: network traffic. *Transpn. Res. -B*, vol. 29B, No. 2, 79-93.

Daganzo, C. F. (1995c) , 'A finite difference approximation of the kinematic wave model of traffic flow', *Transportation Research, B* **29B**(4), 261–276.

Daganzo, C. F. (1997). *Fundamentals of Transportation and Traffic Operations*. Pergamon, Oxford.

Daganzo, C. F. (1999a). A Behavioral Theory of Multi-Lane Traffic Flow Part I: Long Homogeneous Freeway Sections, ITS Working Paper, UCB-ITS-RR-99-5.

Daganzo, C. F. (1999b). A Behavioral Theory of Multi-Lane Traffic Flow Part II: Merges and the Onset of Congestion", ITS Working Paper, UCB-ITS-RR-99-6.

del Castillo, J. M. Pintado, P. and F. G. Benitez (1994). The reaction time of drivers and the stability of traffic flow. *Transpn. Res.-B.* Vol. 28B, No. 1, 35-60.

del Castillo, J. M. and F. G. Benitez (1995). On functional form of the speed-density relationship—I: general theory, II: empirical investigation. *Transp. Res.* 29B, 373-406.

Dressler, R. F. (1949). Mathematical Solution of the Problem of Roll-Waves in Inclined Open Channels. *Commun. Pure and Appl. Math*, 2, 149-194.

Edie, L. C. and E. Bavarez (1967). Generation and propagation of stop-start traffic waves. In L. C. Edie, R. Herman and R. Rothery, eds., *Vehicular Traffic Science*, Elsevier, 26-37.

Gazis, D. C., R. Herman, and G. H. Weiss (1962). Density Oscillations Between Lanes of a Multilane Highway. *Oper. Res.*,10, 658-667.

Ferrari, P. (1989). The Effect of Driver Behaviour on Motorway Reliability. *Transportation Research, 23B*, 139-150.

Herrmann, M. and B.S. Kerner (1998). Local cluster effect in different traffic flow models, {em Physica A 255, 163-198.

Jin, W. L. and H. M. Zhang (2000a). The inhomogeneous kinematic wave traffic flow model as a resonant nonlinear system, submitted to *Transpn. Sci.*.

Jin, W. L. and H. M. Zhang (2000b). Numerical studies on the PW model, Working Paper, Institute of Transportation Studies, University of California at Davis.

Kerner, B. S. and Konhaüser, P. (1993). Cluster effect in initially homogeneous trafficflow, *Physical Review E* **48**(4), 2335–2338.

Kerner,B.S. and P. Konhäuser (1994). Structure and parameters of clusters in traffic flow, *Physical Review E* Volume 50, Number 1, 54-83.

Kerner, B. S., Konhaüser, P. and M. Shilke (1996). A new approach to problems of traffic flow theory. in: *Proceedings of the 13th Int. Symp. on Transportation and Traffic Theory*, J. B. Lesort, editor. 79-102, Pergamon, Oxford, U.K..

Kerner, B. S. and H. Rehborn (1999). Theory of congested traffic flow: self organization without bottlenecks. *Proceedings of the 14th Int. Symp. on Transportation and Traffic Theory*, Cedar, A., editor. 147-171, Pergamon, New York, NY.

Koshi, M., Iwasaki, M. and I. Ohkura (1983). Some findings and an overview on vehicular flow characteristics. *Proceedings of the 8th Int. Symp. on Transportation and Traffic Theory*, Hurdle, V., Hauer, E. and G. Stuart, editors. 403-451, University of Toronto Press, Toronto, Canada.

Kuhne, R. D. (1984). Macroscopic freeway model for dense traffic— stop-start waves and incident detection, *Ninth International Symposium on Transportation and Traffic Theory*, VNU Science Press, 20-42.

Kuhne, R. D. (1987). Freeway Speed Distribution and Acceleration Noise in N. H. Gartner, N. H. M. Wilson (eds.). *Proceedings of the Tenth International Symposium on Transportation and Traffic Theory*, Elsevier, 119-137.

Kuhne, R. D. (1989). Freeway control and incident detection using a stochastic continuum theory of traffic flow. In *Proc. 1st int. conf. on applied advanced technology in transportation engineering*, pages 287–292, San Diego, CA.

Kühne, R. D. and Beckschulte, R. (1993). Non-linearity stochastics of unstable traffic flow, *in* C. F. Daganzo, ed., 'Transportation and Traffic Theory', Elsevier Science Publishers., pp. 367–386.

Kühne, R. D. and K. Langbein-Euchner (1995). Parameter Validation. Report as an Order of Daimler-Benz Research, Stuttgart. Available by Steierwald Schnharting und Partner, Stuttgart.

Lax, P. D. (1972). *Hyperbolic systems of conservations laws and the mathematical theory of shock waves*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Lebacque, J. P. (1996). The Godunov scheme and what it means for first order traffic flow models. *Proceedings of the 13th Int. Symp. on Transportation and Traffic Theory*, J-P Lesort, editor. 647-677, Pergamon, New York, NY.

Lebacque, J. P. (1999). Macroscopic traffic flow models: a question of order. *Proceedings of the 14th Int. Symp. on Transportation and Traffic Theory*, Cedar, A., editor. 147-171, Pergamon, New York, NY.

Leo, C. J., and R. L. Pretty (1992). Numerical simulations of macroscopic continuum traffic models. *Transpn. Res.-B.* 26B, No.3, 207-220.

Leutzbach, W. (1985). *Introduction to the Theory of Traffic Flow*. Springer Pub., 184-193.

Leutzbach, W. (1991) Measurements of Mean Speed Time Series Autobahn A5 near Karlsruhe, Germany. Institute of Transport Studies, University of Karlsruhe, Germany.

LeVeque, R. (1992) *Numerical methods for conservation laws*, Birkhäuser Verlag.

Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves: II. a theory of traffic flow on long crowded roads. In *Proc. Royal Society*, volume 229(1178) of *A*, 317–345.

Michalopoulos, P. G., Beskos, D. E. & Lin, J. K. (1984) , 'Analysis of interrupted traffic flow by finite difference methods', *Transportation Research, B* **18B**, 377–396.

Michalopoulos, P. G., D. E. Beskos, and Y. Yamauchi, (1984). Multilane Traffic Flow Dynamics: Some Macroscopic Considerations. *Transportation Research*, 18B, No. 4/5, pp. 377-395.

Michalopoulos, P. G., Jaw Kuan Lin, and D. E. Beskos, (1987). Integrated Modelling and Numerical Treatment of Freeway Flow. *Appl. Mathem. Model*, Vol. 11, No. 401, 447-458.

Michalopoulos, P. G., E. Kwon, and E. G. Khang, (1991). Enhancements and Field Testing of a Dynamic Simulation Program. *Transportation Research Record*, 1320, pp. 203-215.

Newell, G. F. (1961). Nonlinear Effects in the Dynamics of Car Following, *Operations Research, Vol 9*, pp.209-229.

Newell, G. F. (1965). Instability in dense highway traffic, a review. In Almond, P., editor, *Proceedings of the Second International Symposium on the Theory of Traffic Flow*, pages 73–85.

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. *Transpn. Res. B*, Vol. 27, 281-313.

Papageorgiou, M., Blosseville, J., and Hadj-Salem, H.(1990) Modeling and Real-Time Control of Traffic Flow on The Southern Part of Boulevard Peripherique in Paris: Part I: Modeling, *Transportation Research, A*, 24 (5) 345-359.

Papageorgiou, M. (1998). Some remarks on macroscopic flow modeling. *Transportation Research, A*, 32 (5) 323-329.

Payne, H. J. (1971). Models of freeway traffic and control. In Bekey, G. A., editor, *Mathematical Models of Public Systems*, volume 1 of *Simulation Councils Proc. Ser.*, 51–60.

Payne, H. J. (1979). FREFLO: A Macroscopic Simulation Model of Freeway Traffic. *Transportation Research Record 722*, pp. 68-77.

Richards, P. I. (1956). Shock waves on the highway. *Operations Research*, 4:42–51.

Sailer, H. (1996). Fuid Dynamical Modelling of Traffic Flow on Highways–Physical Basis and Numerical Examples; Dissertation. (submitted for a Diploma) at the Institute for Theoretical Physics, University Innsbruck.

Treiterer, J. and J. A. Myers (1974). The hysteresis phenomena in traffic flow. *Proceedings of the sixth Int. Symp. on Transportation and Traffic Theory*, D. J. Buckley, editor. 13-38.

Verweij, H. D. (1985). Filewaarschuwing en Verkeer-safwikkeling. Ministry of Traffic and Water-Ways, Delft, The Netherlands.

Whitham, G. B. (1974). *Linear and nonlinear waves.* John Wiley & Sons, New York.

Zhang, H. M. (1998). A theory of nonequilibrium traffic flow. *Transportation Research, B*, 32(7):485–498.

Zhang, H. M. (1999). Analyses of the stability and wave properties of a new continuum traffic theory. *Transportation Research, B*, 33 (6) 399-415.

Zhang, H. M. and T. Wu (1999) Numerical simulation and analysis of traffic flow. *Transportation Research Record 1678*, 251-260.

Zhang, H. M. (2000a). Structural properties of solutions arising from a nonequilibrium traffic flow theory. *Transportation Research, B.* 34, 583-603 .

Zhang, H. M. (2000b). A finite difference approximation of a non-equilibrium traffic flow model. *Transportation Research, B.* (in press).

Zhang, H. M. (2000c). Anisotropic property revisited—when is it violated in traffic flow? Working paper, Institute of Transportation Studies, University of California at Davis (accepted for publication in *Trans. Res., -B*).

Zhang, H. M. (2000d). Phase transitions in a non-equilibrium traffic theory, " *ASCE Journal of Transportation Engineering*. Vol. 126, No. 1, 1-12.

Zhang, H. M. and T. Kim (2000). Effects of relaxation and anticipation on Riemann solutions of the PW model-a numerical investigation. *Transportation Research Record* (in press).

Zhang, H. M. (2001). New perspectives on continuum traffic flow models. *Special issue on traffic flow theory*, H. M. Zhang, ed., *Journal of Networks and Spatial Economics*, Vol. 1, Issue 1. 2001.

MACROSCOPIC FLOW MODELS

BY JAMES C. WILLIAMS[9]

[9] Associate Professor, Department of Civil Engineering, University of Texas at Arlington, Box 19308, Arlington, TX 76019-0308.

CHAPTER 6 - Frequently used Symbols

Note to reader: The symbols used in Chapter 6 are the same as those used in the original sources. Therefore, the reader is cautioned that the same symbol may be used for different quantities in different sections of this chapter. The symbol definitions below include the sections in which the symbols are used if the particular symbol definition changes within the chapter or is a definition particular to this chapter. In each case, the symbols are defined as they are introduced within the text of the chapter. Symbol units are given only where they help define the quantity; in most cases, the units may be in either English or metric units as necessary to be consistent with other units in a relation.

A	=	area of town (Section 6.2.1)
c	=	capacity (vehicles per unit time per unit width of road) (Section 6.2.1)
D	=	delay per intersection (Section 6.2.2)
f	=	fraction of area devoted to major roads (Section 6.1.1)
f	=	fraction of area devoted to roads (Section 6.2.1)
f	=	number of signalized intersections per mile (Section 6.2.2)
f_r	=	fraction of moving vehicles in a designated network (Section 6.3)
f_s	=	fraction of stopped vehicles in a designated network (Section 6.3)
$f_{s,min}$	=	minimum fraction of vehicles stopped in a network (Section 6.4)
I	=	total distance traveled per unit area, or traffic intensity (pcu/hour/km) (Sections 6.1.1 and 6.2.3)
J	=	fraction of roadways used for traffic movement (Section 6.2.1)
K	=	average network concentration (ratio of the number of vehicles in a network and the network length, Section 6.4)
K_j	=	jam network concentration (Section 6.4)
N	=	number of vehicles per unit time that can enter the CBD (Section 6.2.1)
n	=	quality of traffic indicator (two-fluid model parameter, Section 6.3)
Q	=	capacity (pcu/hr) (Section 6.2.2)
Q	=	average network flow, weighted average over all links in a designated network (Section 6.4)
q	=	average flow (pcu/hr)
R	=	road density, i.e., length or area of roads per unit area (Section 6.2.3)
r	=	distance from CBD
T	=	average travel time per unit distance, averaged over all vehicles in a designated network (Section 6.3)
T_m	=	average minimum trip time per unit distance (two-fluid model parameter, Section 6.3)
T_r	=	average moving (running) time per unit distance, averaged over all vehicles in a designated network (Section 6.3)
T_s	=	average stopped time per unit distance, averaged over all vehicles in a designated network (Section 6.3)
V	=	average network speed, averaged over all vehicles in a designated network (Section 6.4)
V_f	=	network free flow speed (Section 6.4)
V_m	=	average maximum running speed (Section 6.2.3)
V_r	=	average speed of moving (running) vehicles, averaged over all in a designated network (Section 6.3)
v	=	average speed
v	=	weighted space mean speed (Section 6.2.3)
v_r	=	average running speed, i.e., average speed while moving (Section 6.2.2)
w	=	average street width
α	=	Zahavi's network parameter (Section 6.2.3)
λ	=	g/c time, i.e. ration of effective green to cycle length

6.
MACROSCOPIC FLOW MODELS

Mobility within an urban area is a major component of that area's quality of life and an important issue facing many cities as they grow and their transportation facilities become congested. There is no shortage of techniques to improve traffic flow, ranging from traffic signal timing optimization (with elaborate, computer-based routines as well as simpler, manual, heuristic methods) to minor physical changes, such as adding a lane by the elimination of parking. However, the difficulty lies in evaluating the effectiveness of these techniques. A number of methods currently in use, reflecting progress in traffic flow theory and practice of the last thirty years, can effectively evaluate changes in the performance of an intersection or an arterial. But a dilemma is created when these individual components, connected to form the traffic network, are dealt with collectively.

The need, then, is for a consistent, reliable means to evaluate traffic performance in a network under various traffic and geometric configurations. The development of such performance models extends traffic flow theory into the network level and provides traffic engineers with a means to evaluate system-wide control strategies in urban areas. In addition, the quality of service provided to the motorists could be monitored to evaluate a city's ability to manage growth. For instance, network performance models could also be used by a state agency to compare traffic conditions between cities in order to more equitably allocate funds for transportation system improvements.

The performance of a traffic system is the response of that system to given travel demand levels. The traffic system consists of the network topology (street width and configuration) and the traffic control system (e.g., traffic signals, designation of one- and two-way streets, and lane configuration). The number of trips between origin and destination points, along with the desired arrival and/or departure times comprise the travel demand levels. The system response, i.e., the resulting flow pattern, can be measured in terms of the level of service provided to the motorists. Traffic flow theory at the intersection and arterial level provides this measurement in terms of the three basic variables of traffic flow: speed, flow (or volume), and concentration. These three variables, appropriately defined, can also be used to describe traffic at the network level. This description must be one that can overcome the intractabilities of existing flow theories when network component interactions are taken into account.

The work in this chapter views traffic in a network from a macroscopic point of view. Microscopic analyses run into two major difficulties when applied to a street network:

1) Each street block (link) and intersection are modeled individually. A proper accounting of the interactions between adjacent network components (particularly in the case of closely spaced traffic signals) quickly leads to intractable problems.

2) Since the analysis is performed for each network component, it is difficult to summarize the results in a meaningful fashion so that the overall network performance can be evaluated.

Simulation can be used to resolve the first difficulty, but the second remains; traffic simulation is discussed in Chapter 10.

The Highway Capacity Manual (Transportation Research Board 1994) is the basic reference used to evaluate the quality of traffic service, yet does not address the problem at the network level. While some material is devoted to assessing the level of service on arterials, it is largely a summation of effects at individual intersections. Several travel time models, beginning with the travel time contour map, are briefly reviewed in the next section, followed by a description of general network models in Section 6.2. The two-fluid model of town traffic, also a general network model, is discussed separately in Section 6.3 due to the extent of the model's development through analytical, field, and simulation studies. Extensions of the two-fluid model into general network models are examined in Section 6.4, and the chapter references are in the final section.

6.1 Travel Time Models

Travel time contour maps provide an overview of how well a street network is operating at a specific time. Vehicles can be dispatched away from a specified location in the network, and each vehicle's time and position noted at desired intervals.

Contours of equal travel time can be established, providing information on the average travel times and mean speeds over the network. However, the information is limited in that the travel times are related to a single point, and the study would likely have to be repeated for other locations. Also, substantial resources are required to establish statistical significance. Most importantly, though, is that it is difficult to capture network performance with only one variable (travel time or speed in this case), as the network can be offering quite different levels of service at the same speed.

This type of model has be generalized by several authors to estimate average network travel times (per unit distance) or speeds as a function of the distance from the central business district (CBD) of a city, unlike travel time contour maps which consider only travel times away from a specific point.

6.1.1 General Traffic Characteristics as a Function of the Distance from the CBD

Vaughan, Ioannou, and Phylactou (1972) hypothesized several general models using data from four cities in England. In each case, general model forms providing the best fit to the data were selected. Traffic intensity (I, defined as the total distance traveled per unit area, with units of pcu/hour/km) tends to decrease with increasing distance from the CBD,

$$I = A \exp\left(-\sqrt{r/a}\right) \tag{6.1}$$

where r is the distance from the CBD, and A and a are parameters. Each of the four cities had unique values of A and a, while A was also found to vary between peak and off-peak periods. The data from the four cities is shown in Figure 6.1. A similar relation was found between the fraction of the area which is major road (f) and the distance from the CBD,

$$f = B \exp\left(-\sqrt{r/b}\right) \quad , \tag{6.2}$$

where b and B are parameters for each town. Traffic intensity and fraction of area which is major road were found to be linearly related, as was average speed and distance from the CBD. Since only traffic on major streets is considered, these

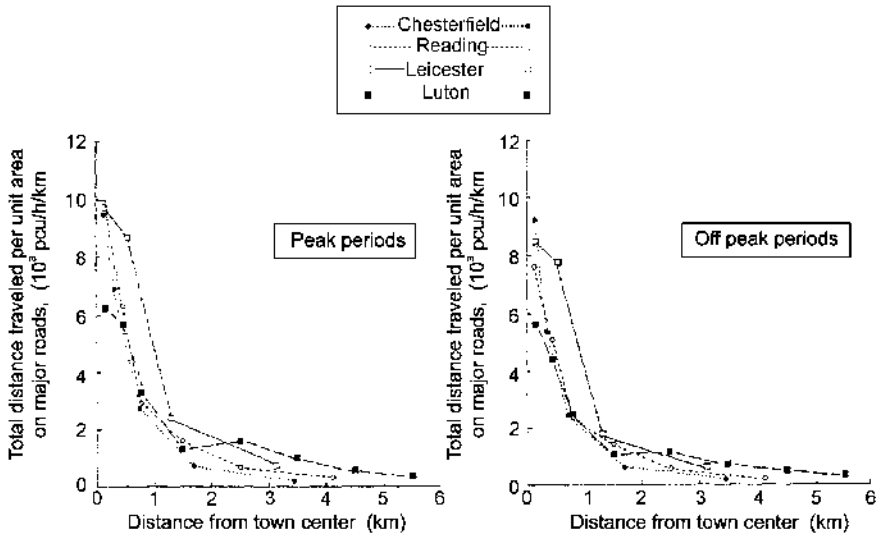


Figure 6.1
Total Vehicle Distance Traveled Per Unit Area on Major Roads as a Function of the Distance from the Town Center (Vaughan et al. 1972).

results are somewhat arbitrary, depending on the streets selected as major.

6.1.2 Average Speed as a Function of Distance from the CBD

Branston (1974) investigated five functions relating average speed (v) to the distance from the CBD (r) using data collected by the Road Research Laboratory (RRL) in 1963 for six cities in England. The data was fitted to each function using least-squares regression for each city separately and for the aggregated data from all six cities combined. City centers were defined as the point where the radial streets intersected, and the journey speed in the CBD was that found within 0.3 km of the selected center. Average speed for each route section was found by dividing the section length by the actual travel time (miles/minute). The five selected functions are described below, where a, b, and c are constants estimated for the data. A power curve,

$$v = a r^{b} \qquad (6.3)$$

was drawn from Wardrop's work (1969), but predicts a zero speed in the city center (at $r = 0$). Accordingly, Branston also fitted a more general form,

$$v = c + a r^{b} \quad , \qquad (6.4)$$

where c represents the speed at the city center.

Earlier work by Beimborn (1970) suggested a strictly linear form, up to some maximum speed at the city edge, which was defined as the point where the average speed reached its maximum (i.e., stopped increasing with increasing distance from the center). None of the cities in Branston's data set had a clear maximum limit to average speed, so a strict linear function alone was tested:

$$v = a + b r \quad . \qquad (6.5)$$

A negative exponential function,

$$v = a - b e^{-cr} \quad , \qquad (6.6)$$

had been fitted to data from a single city (Angel and Hyman 1970). The negative exponential asymptotically approaches some maximum average speed.

The fifth function, suggested by Lyman and Everall (1971),

$$v = \frac{1 + b^2 r^2}{a + c b^2 r^2} \qquad (6.7)$$

also suggested a finite maximum average speed at the city outskirts. It had originally be applied to data for radial and ring roads separately, but was used for all roads here.

Two of the functions were quickly discarded: The linear model (Equation 6.5) overestimated the average speed in the CBDs by 3 to 4 km/h, reflecting an inability to predict the rapid rise in average speed with increasing distance from the city center. The modified power curve (Equation 6.4) estimated negative speeds in the city centers for two of the cities, and a zero speed for the aggregated data. While obtaining the second smallest sum of squares (negative exponential, Equation 6.6, had the smallest), the original aim of using this model (to avoid the estimation of a zero journey speed in the city center) was not achieved.

The fitted curves for the remaining three functions (negative exponential, Equation 6.6; power curve, Equation 6.3; and Lyman and Everall, Equation 6.7) are shown for the data from Nottingham in Figure 6.2. All three functions realistically predict a leveling off of average speed at the city outskirts, but only the Lyman-Everall function indicates a leveling off in the CBD. However, the power curve showed an overall better fit than the Lyman-Everall model, and was preferred.

While the negative exponential function showed a somewhat better fit than the power curve, it was also rejected because of its greater complexity in estimation (a feature shared with the Lyman-Everall function). Truncating the power function at measured downtown speeds was suggested to overcome its drawback of estimating zero speeds in the city center. The complete data set for Nottingham is shown in Figure 6.3, showing the fitted power function and the truncation at $r = 0.3$ km.

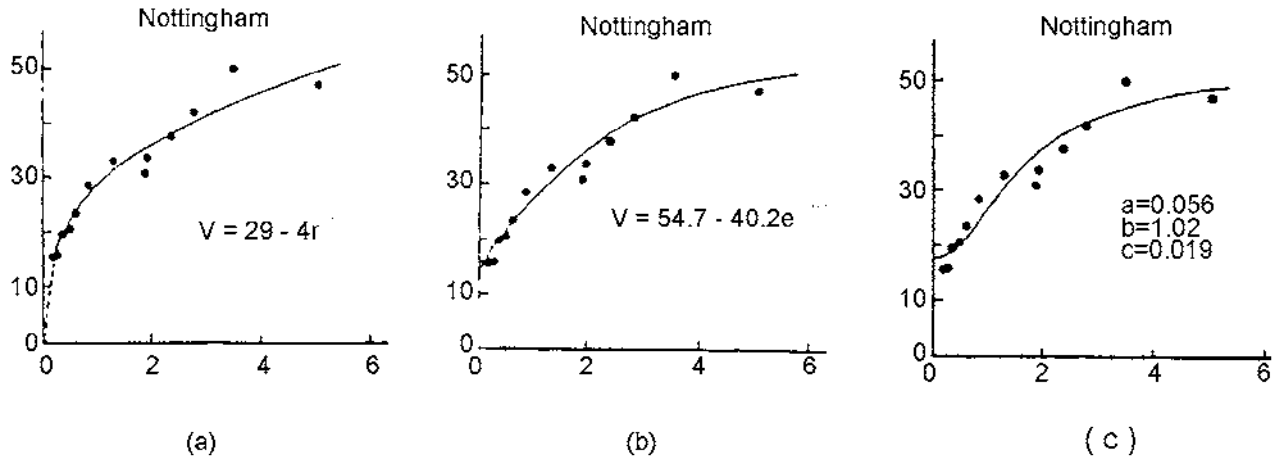


Figure 6.2
Grouped Data for Nottingham Showing Fitted a) Power Curve, b) Negative Exponential Curve, and c) Lyman-Everall Curve (Branston 1974, Portions of Figures 1A, 1B, and 1C).

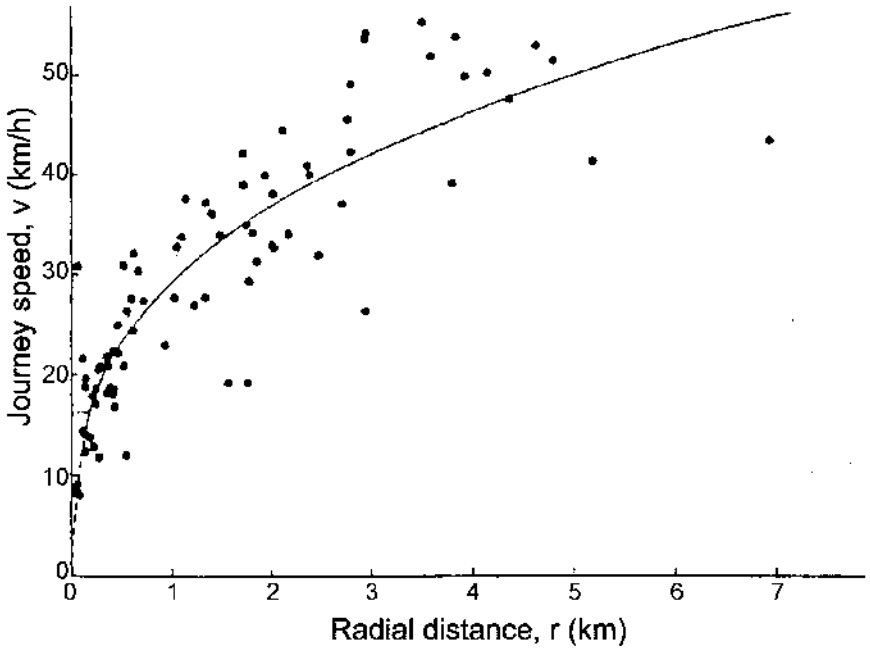


Figure 6.3
Complete Data Plot for Nottingham; Power Curve Fitted to the Grouped Data (Branston 1974, Figure 3).

If the data is broken down by individual radial routes, as shown in Figure 6.4, the relation between speed and distance from the city center is stronger than when the aggregated data is examined.

Hutchinson (1974) used RRL data collected in 1967 from eight cities in England to reexamine Equations 6.3 and 6.6 (power curve and negative exponential) with an eye towards simplifying them.

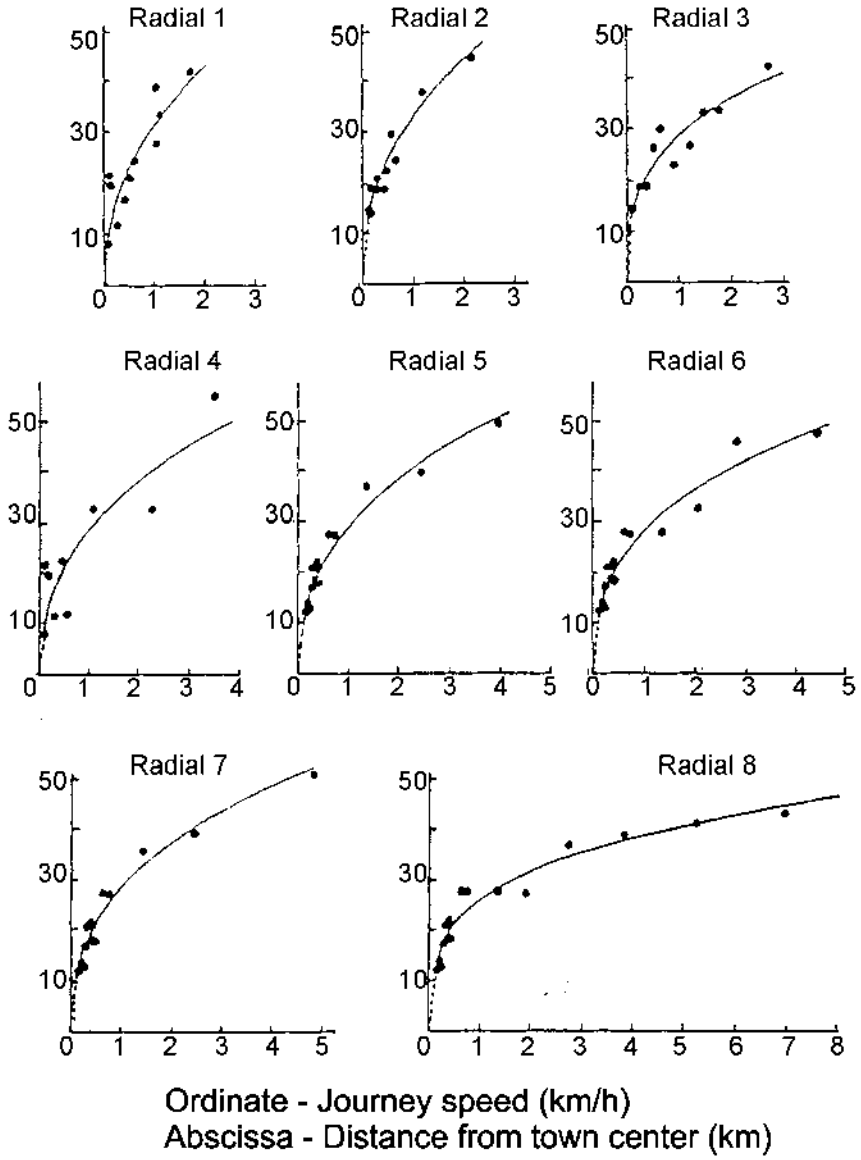


Figure 6.4
Data from Individual Radial Routes in Nottingham,
Best Fit Curve for Each Route is Shown (Branston 1974, Figure 4).

The exponents of the power functions fitted by Branston (1974) fell in the range 0.27 to 0.36, suggesting the following simplification

$$v = k r^{1/3} \qquad (6.8)$$

When fitted to Branston's data, there was an average of 18 percent increase in the sum of squares. The other parameter, k, was found to be significantly correlated with the city population, with different values for peak and off-peak conditions. The parameter k was found to increase with increasing population, and was 9 percent smaller in the peak than in the off-peak.

In considering the negative exponential model (Equation 6.6), Hutchinson reasoned that average speed becomes less characteristic of a city with increasing r, and, as such, it would be reasonable to select a single maximum limit for v for every city. Assuming that any speed between 50 and 75 km/h would make little difference, Hutchinson selected 60 km/h, and

$$v = 60 - a e^{-r/R} \quad . \qquad (6.9)$$

Hutchinson found that this model raised the sum of squares by 30 percent (on the average) over the general form used by Branston. R was found to be strongly correlated with the city population, as well as showing different averages with peak and off-peak conditions, while a was correlated with neither the city population nor the peak vs. off-peak conditions. The difference in the Rs between peak and off-peak conditions (30 percent higher during peaks) implies that low speeds spread out over more of the network during the peak, but that conditions in the city center are not significantly different. Hutchinson (1974) used RRL data collected in 1967 from eight cities in England to reexamine Equations 6.3 and 6.6 (power curve and negative exponential) with an eye towards simplifying them.

6.2 General Network Models

A number of models incorporating performance measures other than speed have been proposed. Early work by Wardrop and Smeed (Wardrop 1952; Smeed 1968) dealt largely with the development of macroscopic models for arterials, which were later extended to general network models.

6.2.1 Network Capacity

Smeed (1966) considered the number of vehicles which can "usefully" enter the central area of a city, and defined N as the number of vehicles per unit time that can enter the city center. In general, N depends on the general design of the road network, width of roads, type of intersection control, distribution of destinations, and vehicle mix. The principle variables for towns with similar networks, shapes, types of control, and vehicles are: A, the area of the town; f, the fraction of area devoted to roads; and c, the capacity, expressed in vehicles per unit time per unit width of road (assumed to be the same for all roads). These are related as follows:

$$N = \alpha f c \sqrt{A}, \qquad (6.10)$$

where α is a constant. General relationships between f and $(N/c\sqrt{A})$ for three general network types (Smeed 1965) are shown in Figure 6.5. Smeed estimated a value of c (capacity per unit width of road) by using one of Wardrop's speed-flow equations for central London (Smeed and Wardrop 1964),

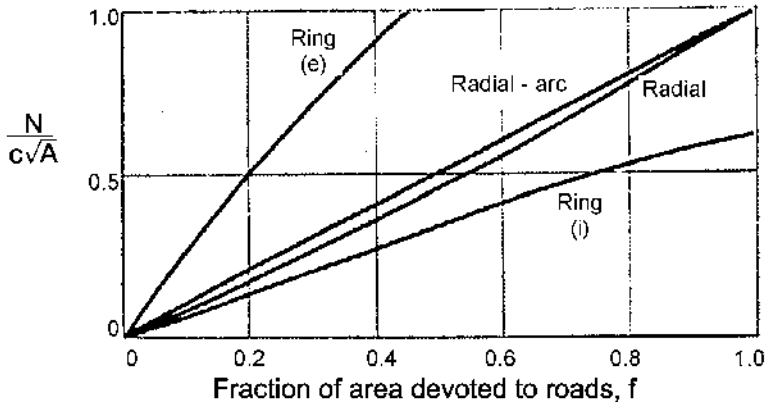
$$q = 2440 - 0.220\, v^3 \quad , \qquad (6.11)$$

where v is the speed in kilometers/hour, and q the average flow in pcus/hour, and divided by the average road width, 12.6 meters,

$$c = 58.2 - 0.00524\, v^3 \ . \qquad (6.12)$$

A different speed-flow relation which provided a better fit for speeds below 16 km/h resulted in $c = 68 - 0.13\, v^2$ (Smeed 1963).

Equation 6.12 is shown in Figure 6.6 for radial-arc, radial, and ring type networks for speeds of 16 and 32 km/h. Data from several cities, also plotted in Figure 6.6, suggests that αc=30,



Note: (e=excluding area of ring road, I=including area of ring road)

Figure 6.5
Theoretical Capacity of Urban Street Systems (Smeed 1966, Figure 2).

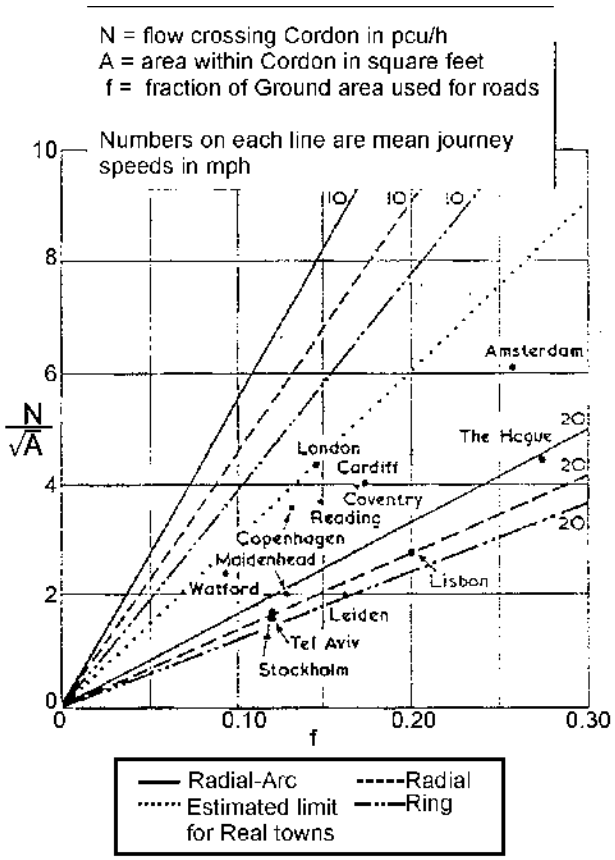


Figure 6.6
Vehicles Entering the CBDs of Towns Compared with the Corresponding Theoretical Capacities of the Road Systems (Smeed 1966, Figure 4).

and using the peak period speed of 16 km/h in central London, Equation 6.10 becomes

$$N = \left(33 - 0.003\, v^3\right) f\sqrt{A}\ , \qquad (6.13)$$

where v is in miles/hour and A in square feet. It should be noted that f represents the fraction of total area usefully devoted to roads. An alternate formulation (Smeed 1968) is

$$N = \left(33 - 0.003\, v^3\right) Jf\sqrt{A} \qquad (6.14)$$

where f is the fraction of area actually devoted to roads, while J is the fraction of roadways used for traffic movement. J was found to range between 0.22 and 0.46 in several cities in England. The large fraction of unused roadway is mostly due to the uneven distribution of traffic on all streets. The number of vehicles which can circulate in a town depends strongly on their average speed, and is directly proportional to the area of usable roadway. For a given area devoted to roads, the larger the central city, the smaller the number of vehicles which can circulate in the network, suggesting that a widely dispersed town is not necessarily the most economical design.

6.2.2 Speed and Flow Relations

Thomson (1967b) used data from central London to develop a linear speed-flow model. The data had been collected once every two years over a 14-year period by the RRL and the Greater London Council. The data consisted of a network-wide average speed and flow each year it was collected. The average speed was found by vehicles circulating through central London on predetermined routes. Average flows were found by first converting measured link flows into equivalent passenger carunits, then averaging the link flows weighted by their respective link lengths. Two data points (each consisting of an average speed and flow) were found for each of the eight years the data was collected: peak and off-peak.

Plotting the two points for each year, Figure 6.7, resulted in a series of negatively sloped trends. Also, the speed-flow capacity (defined as the flow that can be moved at a given speed) gradually increased over the years, likely due to geometric and traffic control improvements and "more efficient vehicles." This indicated that the speed-flow curve had been gradually changing, indicating that each year's speed and flow fell on different curves. Two data points were inadequate to determine the shape of the curve, so all sixteen data points were used by accounting for the

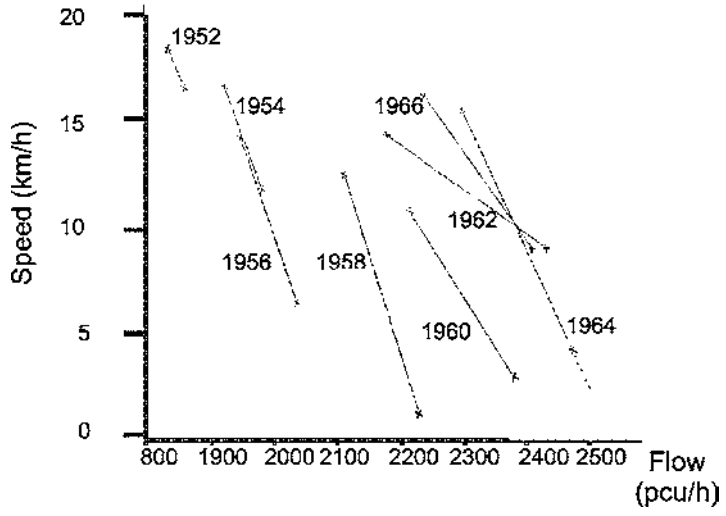


Figure 6.7
Speeds and Flows in Central London, 1952-1966, Peak and Off-Peak (Thomson 1967b, Figure 11)

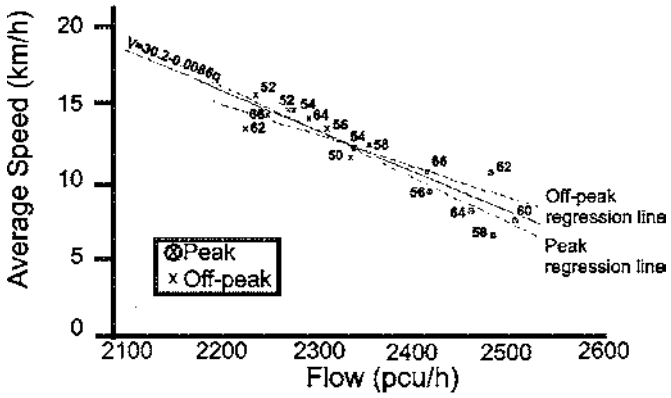
changing capacity of the network, and scaling each year's flow measurement to a selected base year. Using linear regression, the following equation was found:

$$v = 30.2 - 0.0086\,q \qquad (6.15)$$

where v is the average speed in kilometers/hour and q is the average flow in pcu/hour. This relation is plotted in Figure 6.8.

The equation implies a free-flow speed of about 48.3 km/h however, there were no flows less than 2200 pcu/hour in the historical data.

Thomson used data collected on several subsequent Sundays (Thomson 1967a) to get low flow data points. These are reflected in the trend shown in Figure 6.9.Also shown is a curve developed by Smeed and Wardrop using data from a single year only.



Note: Scaled to 1964 equivalent flows.

Figure 6.8
Speeds and Scaled Flows, 1952-1966 (Thomson 1967b, Figure 2).

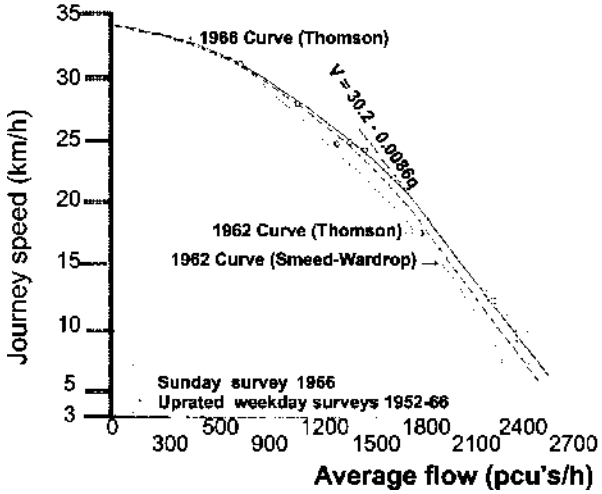


Figure 6.9
Estimated Speed-Flow Relations in Central London (Main Road Network) (Thomson 1967b, Figure 4).

The selected area of central London could be broken into inner and outer zones, distinguished principally by traffic signal densities, respectively 7.5 and 3.6 traffic signals per route-mile. Speed and flow conditions were found to be significantly different between the zones, as shown in Figure 6.10, and for the inner zone,

$$v = 24.3 - 0.0075\,q \quad , \tag{6.16}$$

and for the outer zone,

$$v = 34.0 - 0.0092\,q \quad . \tag{6.17}$$

Wardrop (1968) directly incorporated average street width and average signal spacing into a relation between average speed and flow, where the average speed includes the stopped time. In order to obtain average speeds, the delay at signalized intersections must be considered along with the running speed between the controlled intersections, where running speed is defined as the average speed while moving. Since speed is the inverse of travel time, this relation can be expressed as:

$$\frac{1}{v} = \frac{1}{v_r} + f d \quad , \tag{6.18}$$

where v is the average speed in mi/h, v_r, the running speed in mi/h, d the delay per intersection in hours, and f the number of signalized intersections per mile. Assuming $v_r = a(1-q/Q)$ and $d = b/(1-q/\lambda s)$, where q is the flow in pcu/hr, Q is the capacity in pcu/hr, λ is the g/c time, and s is the saturation flow in pcu/hr, and combining into Equation 6.18,

$$\frac{1}{v} = \frac{1}{a\,(1 - q/Q)} + \frac{f\,b}{1 - q/\lambda s} \tag{6.19}$$

Using an expression for running speed found for central London (Smeed and Wardrop 1964; RRL 1965),

$$v_r = 31 - \frac{0.70q + 430}{3w} \tag{6.20}$$

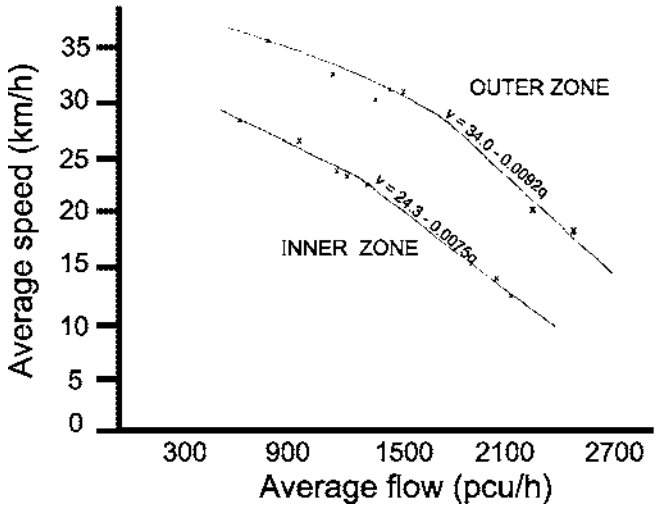


Figure 6.10
Speed-Flow Relations in Inner and Outer Zones of Central Area (Thomson 1967a, Figure 5).

where w is the average roadway width in feet, and an average street width of 42 feet (in central London), Equation 6.20 becomes $v_r = 28 - 0.0056\,q$, or 24 mi/h, whichever is less. The coefficient of q was modified to 0.0058 to better fit the observed running speed.

Using observed values of 0.038 hours/mile stopped time, 2180 pcu/hr flow, and 2610 pcu/hr capacity, the numerator of the second term of Equation 6.19 (fb) was found to be 0.0057. Substituting the observed values into Equation 6.19,

$$\frac{1}{v} = \frac{1}{28 - 0.0058\,q} + \frac{0.0057}{1 - \frac{q}{2610}} \quad .$$

Simplifying,

$$\frac{1}{v} = \frac{1}{28 - 0.0058\,q} + \frac{1}{197 - 0.0775\,q} \qquad (6.21)$$

Revising the capacity to 2770 pcu/hour (to reflect 1966 data), thus changing the coefficient of q in the second term of Equation 6.21 to 0.071, this equation provided a better fit than Thomson's linear relation (Thomson 1967b) and recognizes the known information on the ultimate capacity of the intersections.

Generalizing this equation for urban areas other than London, and knowing that the average street width in central London was 12.6 meters, the running speed can be written

$$v_r = 31 - \frac{430}{3w} - \frac{a\,q}{w}$$
$$= 31 - \frac{140}{w} - \frac{a\,q}{w} \quad .$$

Since $a/w = 0.0058$ when $w = 42$ by Equation 6.21, a=0.0244, then

$$v_r = 31 - \frac{140}{w} - 0.0244\,\frac{q}{w} \quad . \qquad (6.22)$$

For the delay term, five controlled intersections per mile and a g/c of 0.45 were found for central London. Additionally, the intersection capacity was assumed to be proportional to the average stop line width, given that it is more than 5 meters wide (RRL 1965), which was assumed to be proportional to the roadway width. The general form for the delay equation (second term of Equation 6.21) is

$$fd = \frac{f\,b}{1 - q/k\lambda w} \qquad (6.23)$$

where k is a constant. For central London, $w = 42$, $\lambda = 0.45$, and $k\lambda w = Q = 2770$, thus $k = 147$, yielding

$$fd = \frac{f\,b}{1 - q/147\,\lambda w} \qquad (6.24)$$

Given that $f = 5$ signals/mile and $fb = 0.00507$ for central London, $b = 0.00101$, yielding

$$fd = \frac{f}{1000 - 6.8\,q/\lambda w} \quad . \qquad (6.25)$$

Combining, then, for the general equation for average speed:

$$\frac{1}{v} = \frac{1}{31 - \frac{140}{w} - 0.0244\,\frac{q}{w}} + \frac{f}{1000 - 6.8\frac{q}{\lambda w}} \quad . \qquad (6.26)$$

The sensitivity of Equation 6.26 to flow, average street width, number of signalized intersections per mile, and the fraction of green time are shown in Figures 6.11, 6.12, and 6.13. By calibrating this relation on geometric and traffic control features in the network, Wardrop extended the usefulness of earlier speed flow relations. While fitting nicely for central London, the applicability of this relation to other cities in its generalized format (Equation 6.26) is not shown, due to a lack of available data.

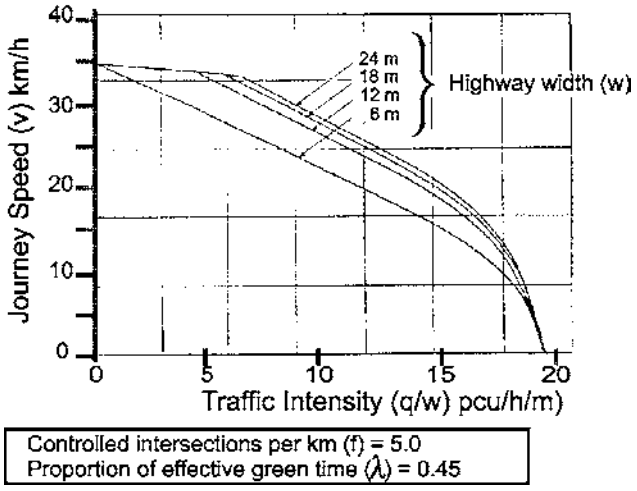


Figure 6.11
Effect of Roadway Width on Relation Between Average (Journey) Speed and Flow in Typical Case (Wardrop 1968, Figure 5).

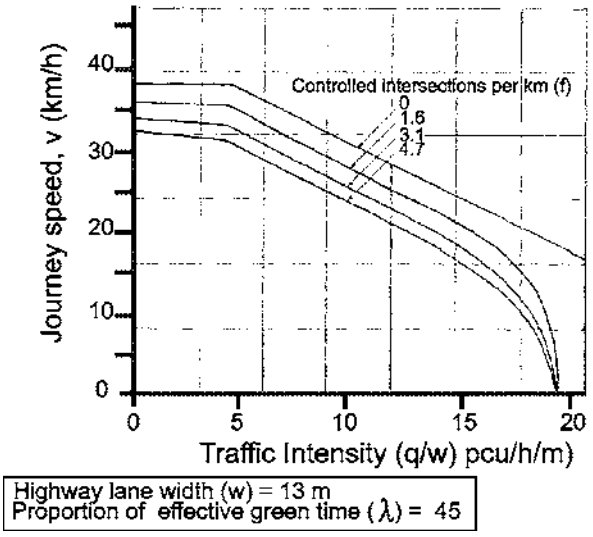


Figure 6.12
Effect of Number of Intersections Per Mile on Relation Between Average (Journey) Speed and Flow in Typical Case (Wardrop 1968, Figure 6).

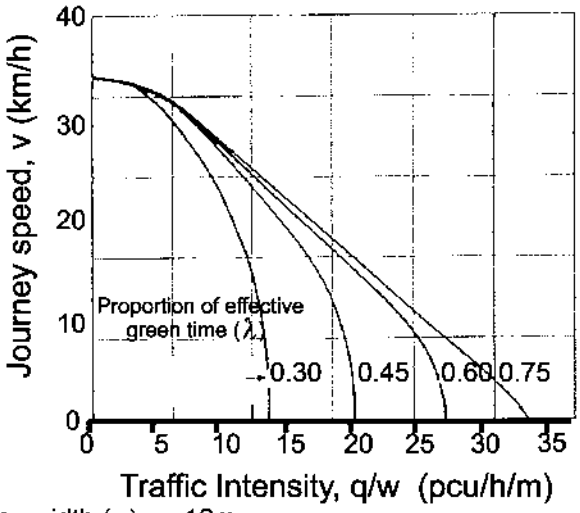


Figure 6.13
Effect of Capacity of Intersections on Relation Between Average (Journey) Speed and Flow in Typical Case (Wardrop 1968, Figure 7).

Godfrey (1969) examined the relations between the average speed and the concentration (defined as the number of vehicles in the network), shown in Figure 6.14, and between average speed and the vehicle miles traveled in the network in one hour, shown in Figure 6.15. Floating vehicles on circuits within the network were used to estimate average speed and aerial photographs were used to estimate concentration.

There is a certain concentration that results in the maximum flow (or the maximum number of miles traveled, see Figure 6.15), which occurs around 10 miles/hour. As traffic builds up past this optimum, average speeds show little deterioration, but there is excessive queuing to get into the network (either from car parking lots within the network or on streets leading into the designated network). Godfrey also notes that expanding an intersection to accommodate more traffic will move the queue to another location within the network, unless the bottlenecks downstream are cleared.

6.2.3 General Network Models Incorporating Network Parameters

Some models have defined specific parameters which intend to quantify the quality of traffic service provided to the users in the network. Two principal models are discussed in this chapter, the α-relationship, below, and the two-fluid theory of town traffic. The two-fluid theory has been developed and applied to a greater extent than the other models discussed in this section, and is described in Section 6.3.

Zahavi (1972a; 1972b) selected three principal variables, I, the traffic intensity (here defined as the distance traveled per unit area), R, the road density (the length or area of roads per unit area), and v, the weighted space mean speed. Using data from England and the United States, values of I, v, and R were found for different regions in different cities. In investigating various relationships between I and v/R, a linear fit was found between the logarithms of the variables:

$$I = \alpha (v/R)^m \quad , \qquad (6.27)$$

where α and m are parameters. Trends for London and Pittsburgh are shown in Figure 6.16. The slope (m) was found to be close to -1 for all six cities examined, reducing Equation 6.27 to

$$I = \alpha R/v \quad , \qquad (6.28)$$

where α is different for each city. Relative values of the variables were calculated by finding the ratio between observed

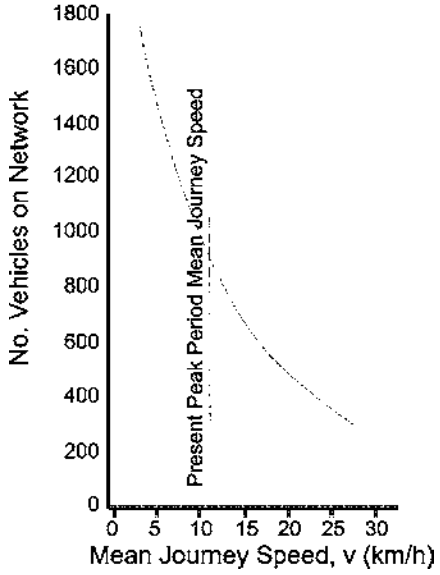


Figure 6.14
Relationship Between Average (Journey) Speed and Number of Vehicles on Town Center Network (Godfrey 1969, Figure 1).

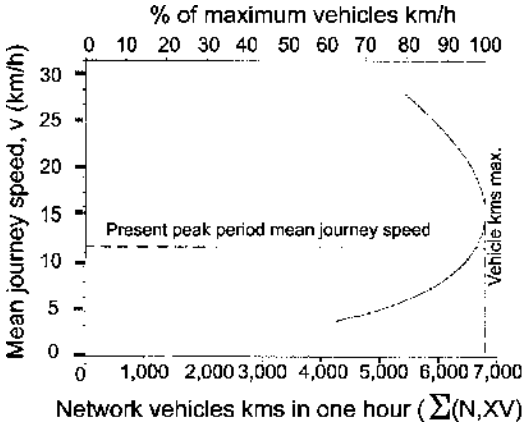


Figure 6.15
Relationship Between Average (Journey) Speed of Vehicles and Total Vehicle Mileage on Network (Godfrey 1969, Figure 2).

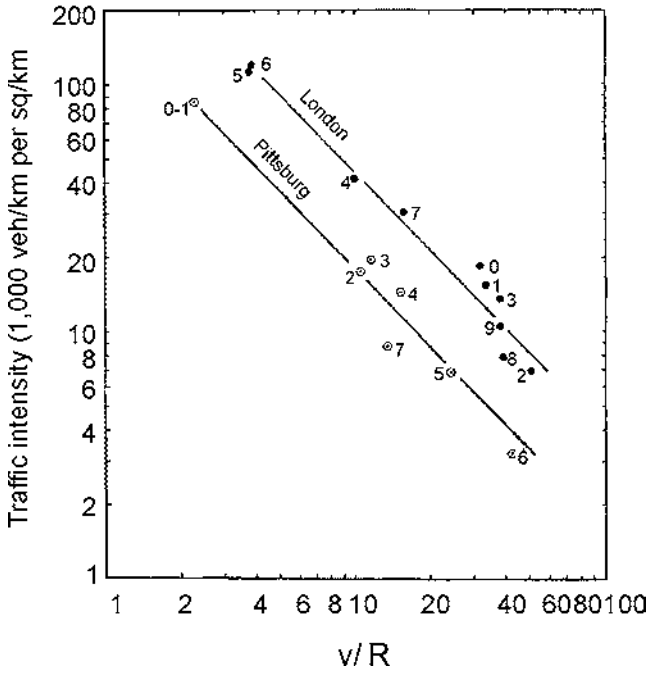


Figure 6.16
The α-Relationship for the Arterial Networks of London and Pittsburgh, in Absolute Values (Zahavi 1972a, Figure 1).

values of *I* and *v/R* for each sector and the average value for the entire city. The relationship between the relative values is shown in Figure 6.17, where the observations for London and Pittsburgh fall along the same line.

The physical characteristics of the road network, such as street widths and intersection density, were found to have a strong effect on the value of α for each zone in a city. Thus, α may serve as a measure of the combined effects of the network characteristics and traffic performance, and can possibly be used as an indicator for the level of service. The α map of London is shown in Figure 6.18. Zones are shown by the dashed lines, with dotted circles indicating zone centroids. Values of α were calculated for each zone and contour lines of equal α were drawn, showing areas of (relatively) good and poor traffic flow conditions. (The quality of traffic service improves with increasing α.)

Unfortunately, Buckley and Wardrop (1980) have shown that α is strongly related to the space mean speed, and Ardekani (1984), through the use of aerial photographs, has shown that α has a high positive correlation with the network concentration.

The two-fluid model also uses parameters to evaluate the level of service in a network and is described in Section 6.3.

6.2.4 Continuum Models

Models have been developed which assume an arbitrarily fine grid of streets, i.e., infinitely many streets, to circumvent the errors created on the relatively sparse networks typically used during the trip or network assignment phase in transportation planning (Newell 1980). A basic street pattern is superimposed over this continuum of streets to restrict travel to appropriate directions. Thus, if a square grid were used, travel on the street network would be limited to the two available directions (the x and y directions in a Cartesian plot), but origins and destinations could be located anywhere in the network.

Individual street characteristics do not have to be specifically modeled, but network-wide travel time averages and capacities (per unit area) must be used for traffic on the local streets. Other street patterns include radial-ring and other grids (triangular, for example).

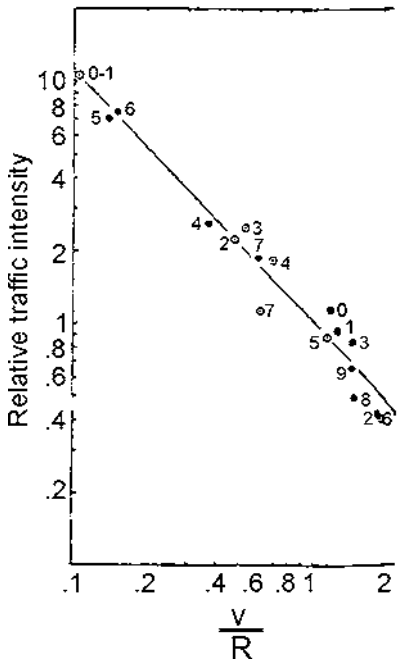


Figure 6.17
The α-Relationship for the Arterial Networks of London and Pittsburgh, in Relative Values (Zahavi 1972a, Figure 2).

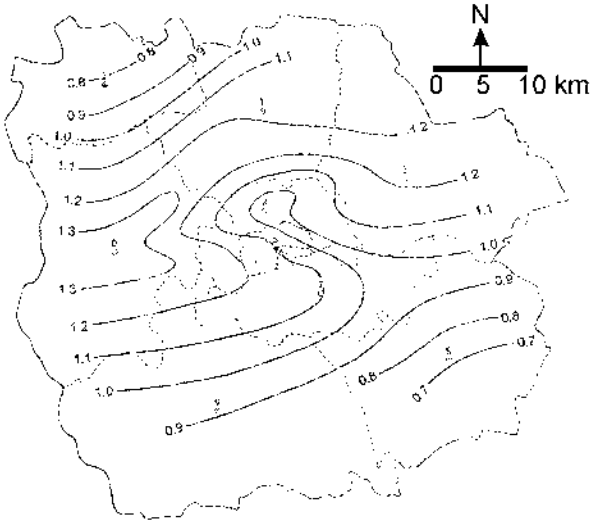


Figure 6.18
The α-Map for London, in Relative Values (Zahavi 1972b, Figure 1).

While the continuum comprises the local streets, the major streets (such as arterials and freeways) are modeled directly. Thus, the continuum of local streets provides direct access (within the constraints provided by the superimposed grid) to the network of major streets.

6.3 Two-Fluid Theory

An important result from Prigogine and Herman's (1971) kinetic theory of traffic flow is that two distinct flow regimes can be shown. These are individual and collective flows and are a function of the vehicle concentration. When the concentration rises so that the traffic is in the collective flow regime, the flow pattern becomes largely independent of the will of individual drivers.

Because the kinetic theory deals with multi-lane traffic, the two-fluid theory of town traffic was proposed by Herman and Prigogine (Herman and Prigogine 1979; Herman and Ardekani 1984) as a description of traffic in the collective flow regime in an urban street network. Vehicles in the traffic stream are divided into two classes (thus, two fluid): moving and stopped vehicles. Those in the latter class include vehicles stopped in the traffic stream, i.e., stopped for traffic signals and stop signs, stopped for vehicles loading and unloading which are blocking a moving lane, stopped for normal congestion, etc., but excludes those out of the traffic stream (e.g., parked cars).

The two-fluid model provides a macroscopic measure of the quality of traffic service in a street network which is independent of concentration. The model is based on two assumptions:

(1) The average running speed in a street network is proportional to the fraction of vehicles that are moving, and
(2) The fractional stop time of a test vehicle circulating in a network is equal to the average fraction of the vehicles stopped during the same period.

The variables used in the two-fluid model represent network-wide averages taken over a given period of time.

The first assumption of the two-fluid theory relates the average speed of the moving (running) vehicles, V_r, to the fraction of moving vehicles, f_r, in the following manner:

$$V_r = V_m f_r^n \ , \tag{6.29}$$

where V_m and n are parameters. V_m is the average maximum running speed, and n is an indicator of the quality of traffic service in the network; both are discussed below. The average speed, V, can be defined as $V_r f_r$, and combining with Equation 6.29,

$$V = V_m f_r^{n+1} \ . \tag{6.30}$$

Since $f_r + f_s = 1$, where f_s is the fraction of vehicles stopped, Equation 6.30 can be rewritten

$$V = V_m (1 - f_s)^{n+1} \ . \tag{6.31}$$

Boundary conditions are satisfied with this relation: when f_s=0, V=V_m, and when f_s=1, V=0.

This relation can also be expressed in average travel times rather than average speeds. Note that T represents the average travel time, T_r the running (moving) time, and T_s the stop time, all per unit distance, and that $T=1/V$, $T_r=1/V_r$, and $T_m=1/V_m$, where T_m is the average minimum trip time per unit distance.

The second assumption of the two-fluid model relates the fraction of time a test vehicle circulating in a network is stopped to the average fraction of vehicles stopped during the same period, or

$$f_s = \frac{T_s}{T} \ . \tag{6.32}$$

This relation has been proven analytically (Ardekani and Herman 1987), and represents the ergodic principle embedded in the model, i.e., that the network conditions can be represented by a single vehicle appropriately sampling the network.

Restating Equation 6.31 in terms of travel time,

$$T = T_m \, (1 - f_s)^{-(n+1)} \, . \qquad (6.33)$$

Incorporating Equation 6.32,

$$T = T_m \left[1 - (T_s / T) \right]^{-(n+1)} \, , \qquad (6.34)$$

realizing that $T = T_r + T_s$, and solving for T_r,

$$T_r = T_m^{\frac{1}{n+1}} \; T^{\frac{n}{n+1}} \, . \qquad (6.35)$$

The formal two-fluid model formulation, then, is

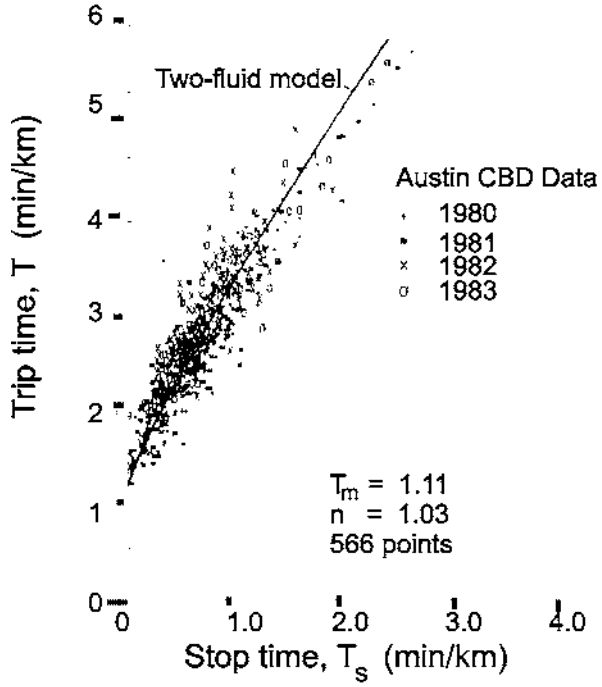
$$T_s = T - T_m^{\frac{1}{n+1}} \; T^{\frac{n}{n+1}} \, . \qquad (6.36)$$

A number of field studies have borne out the two-fluid model (Herman and Ardekani 1984; Ardekani and Herman 1987; Ardekani et al. 1985); and have indicated that urban street networks can be characterized by the two model parameters, n and T_m. These parameters have been estimated using observations of stopped and moving times gathered in each network. The log transform of Equation 6.35,

$$ln \, T_r = \frac{1}{n+1} ln \, T_m + \frac{n}{n+1} ln \, T \qquad (6.37)$$

provides a linear expression for the use of least squares analysis.

Empirical information has been collected with chase cars following randomly selected cars in designated networks. Runs have been broken into one- or two-mile trips, and the running time (T_r) and total trip time (T) for each one- or two-mile trip from the observations for the parameter estimation. Results tend to form a nearly linear relationship when trip time is plotted against stop time (Equation 6.36) as shown in Figure 6.19 for data collected in Austin, Texas. The value of T_m is reflected by the y-intercept (i.e., T at T_s=0), and n by the slope of the curve. Data points representing higher concentration levels lie higher along the curve.



Note: Each point represents one test run approximately 1 or 2 miles long.

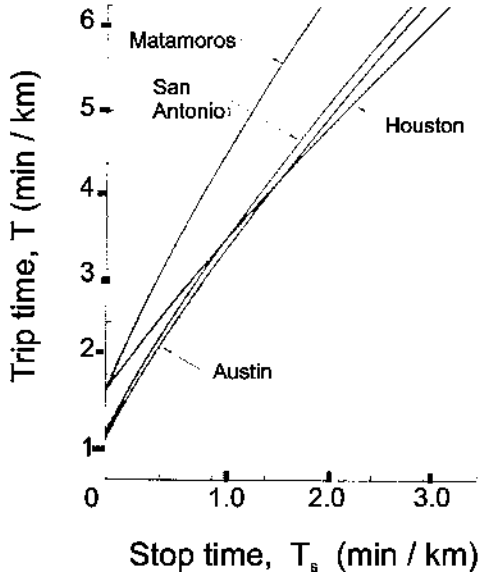
Figure 6.19
Trip Time vs. Stop Time for the Non-Freeway Street Network of the Austin CBD (Herman and Ardekani 1984, Figure 3).

6.3.1 Two-Fluid Parameters

The parameter T_m is the average minimum trip time per unit distance, and it represents the trip time that might be experienced by an individual vehicle alone in the network with no stops. This parameter is unlikely to be measured directly, since a lone vehicle driving though the network very late at night is likely to have to stop at a red traffic signal or a stop sign. T_m, then, is a measure of the uncongested speed, and a higher value would indicate a lower speed, typically resulting in poorer operation. T_m has been found to range from 1.5 to 3.0 minutes/mile, with smaller values typically representing better operating conditions in the network.

As stop time per unit distance (T_s) increases for a single value of n, the total trip time also increases. Because $T=T_r+T_s$, the total trip time must increase at least as fast as the stop time. If n=0, T_r is constant (by Equation 6.35), and trip time would increase at the same rate as the stop time. If n>0, trip time increases at a faster rate than the stop time, meaning that running time is also increasing. Intuitively, n must be greater than zero, since the usual cause for increased stop time is increased congestion, and when congestion is high, vehicles when moving, travel at a lower speed (or higher running time per unit distance) than they do when congestion is low. In fact, field studies have shown that n varies from 0.8 to 3.0, with a smaller value typically indicating better operating conditions in the network. In other words, n is a measure of the resistance of the network to degraded operation with increased demand. Higher values of n indicate networks that degrade faster as demand increases. Because the two-fluid parameters reflect how the network responds to changes in demand, they must be measured and evaluated in a network over the entire range of demand conditions.

While lower n and T_m values represent, in general, better traffic operations in a network, often there is a tradeoff. For example, two-fluid trends for four cities are shown in Figure 6.20. In comparing Houston (T_m=2.70 min/mile, n=0.80) and Austin (T_m=1.78 min/mile, n=1.65), one finds that traffic in Austin moves at significantly higher average speeds during off-peak conditions (lower concentration); at higher concentrations, the curves essentially overlap, indicating similar operating conditions. Thus, despite a higher value of n, traffic conditions



Note: Trip Time vs. Stop Time Two-Fluid Model Trends for CBD Data From the Cities of Austin, Houston, and San Antonio, Texas, and Matamoros, Mexico.

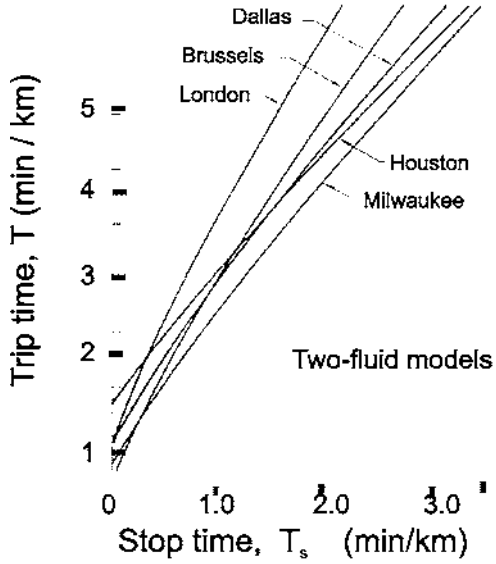
Figure 6.20
Trip Time vs. Stop Time Two-Fluid Model Trends
(Herman and Ardekani 1984, Figure 6).

are better in Austin than Houston, at least at lower concentrations. Different values of the two-fluid parameters are found for different city street networks, as was shown above and in Figure 6.21. The identification of specific features which have the greatest effect on these parameters has been approached through extensive field studies and computer simulation.

6.3.2 Two-Fluid Parameters: Influence of Driver Behavior

Data for the estimation of the two-fluid parameters is collected through chase car studies, where the driver is instructed to follow a randomly selected vehicle until it either parks or leaves the designated network, after which a nearby vehicle is selected and followed. The chase car driver is instructed to follow the vehicle being chased imitating the other driver's actions so as to reflect, as closely as possible, the fraction of time the other driver spends stopped. The objective is to sample the behavior of the drivers in the network as well as the commonly used routes in the street network. The chase car's trip history is then broken into one-mile (typically) segments, and T_r and T calculated for each mile. The (T_r, T) observations are then used in the estimation of the two-fluid parameters.

One important aspect of the chase car study is driver behavior, both that of the test car driver and the drivers sampled in the network. One study addressed the question of extreme driver behaviors, and found that a test car driver instructed to drive aggressively established a significantly different two-fluid trend than one instructed to drive conservatively in the same network at the same time (Herman et al. 1988).

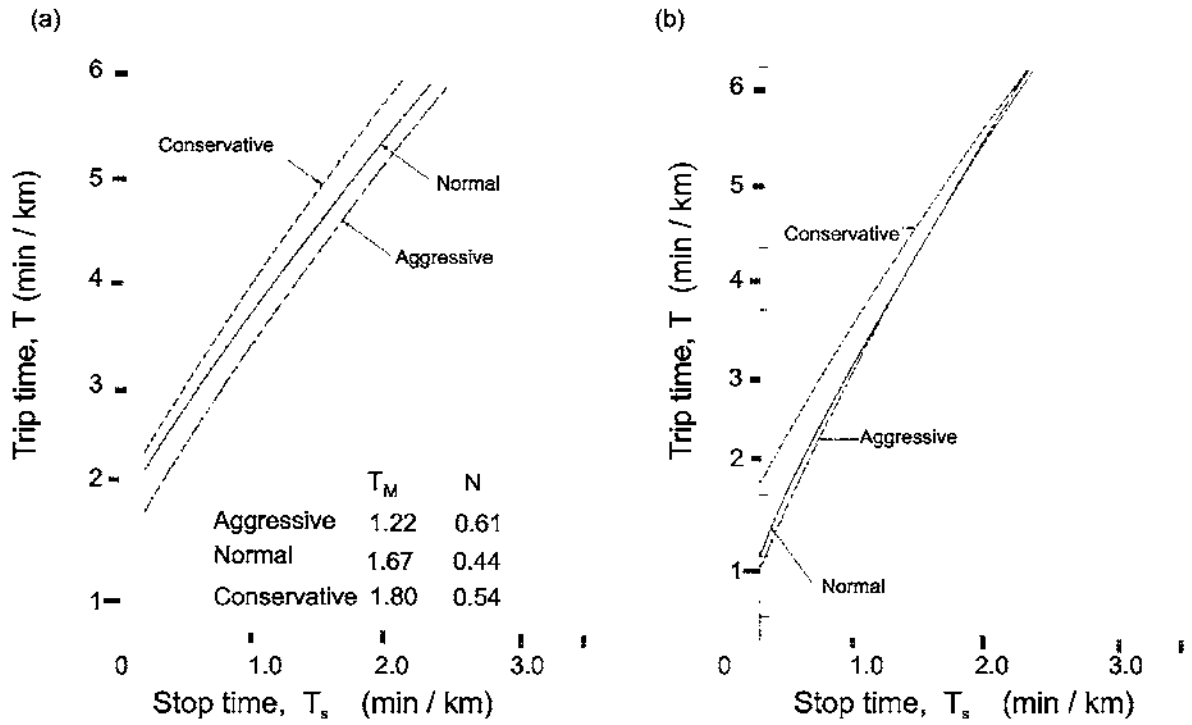


Note: Trip Time vs. Stop Time Two-Fluid Model Trends for Dallas and Houston, Texas, compared to the trends in Milwaukee, Wisconsin, and in London and Brussels.

Figure 6.21
Trip Time vs. Stop Time Two-Fluid Model Trends Comparison
(Herman and Ardekani 1984, Figure 7).

The two-fluid trends resulting from the these studies in two cities are shown in Figure 6.22. In both cases, the normal trend was found through a standard chase car study, conducted at the same time as the aggressive and conservative test drivers were in the network. In both cases, the two-fluid trends established by the aggressive and conservative driver are significantly different. In Roanoke (Figure 6.22a), the normal trend lies between the aggressive and conservative trends, as expected. However, the aggressive trend approaches the normal trend at high demand levels, reflecting the inability of the aggressive driver to reduce his trip and stop times during peak periods. On the other hand, at lower network concentrations, the aggressive driver can take advantage of the less crowded streets and significantly lower his trip times.

As shown in Figure 6.22b, aggressive driving behavior more closely reflects normal driving habits in Austin, suggesting more aggressive driving overall. Also, all three trends converge at high demand (concentration) levels, indicating that, perhaps, the Austin network would suffer congestion to a greater extent than Roanoke, reducing all drivers to conservative behavior (at least as represented in the two-fluid parameters).



Note: The two-fluid trends for aggressive, normal, and conservative drivers in (a) Roanoke, Virginia, and (b) Austin, Texas

Figure 6.22
Two-Fluid Trends for Aggressive, Normal, and Conservative Drivers
(Herman et al. 1988, Figures 5 and 8).

The results of this study reveal the importance of the behavior of the chase car driver in standard two-fluid studies. While the effects on the two-fluid parameters of using two different chase car drivers in the same network at the same time has not been investigated, there is thought to be little difference between two well-trained drivers. To the extent possible, however, the same driver has been used in different studies that are directly compared.

6.3.3 Two-Fluid Parameters: Influence of Network Features (Field Studies)

Geometric and traffic control features of a street network also play an important role in the quality of service provided by a network. If relationships between specific features and the two-fluid parameters can be established, the information could be used to identify specific measures to improve traffic flow and provide a means to compare the relative improvements.

Ayadh (1986) selected seven network features: lane miles per square mile, number of intersections per square mile, fraction of one-way streets, average signal cycle length, average block length, average number of lanes per street, and average block length to block width ratio. The area of the street network under consideration is used with the first two variables to allow a direct comparison between cities. Data for the seven variables were collected for four cities from maps and in the field. Through a regression analysis, the following models were selected:

$$T_m = 3.59 - 0.54\,C_6 \quad \text{and}$$
$$n = -0.21 + 2.97\,C_3 + 0.22\,C_7 \qquad (6.38)$$

where C_3 is the fraction of one-way streets, C_6 the average number of lanes per street, and C_7 the average block length to block width ratio. Of these network features, only one (the fraction of one-way streets) is relatively inexpensive to implement. One feature, the block length to block width ratio, is a topological feature which would be considered fixed for any established street network.

Ardekani et al. (1992), selected ten network features: average block length, fraction of one-way streets, average number of lanes per street, intersection density, signal density, average speed limit, average cycle length, fraction of curb miles with parking allowed, fraction of signals actuated, and fraction of approaches with signal progression. Of these, only two features (average block length and intersection density) can be considered fixed, and, as such, not useful in formulating network improvements. In addition, one feature (average number of lanes per street), also used in the previous study (Ayadh 1986), can typically be increased only by eliminating parking (if present), yielding only limited opportunities for improvement of traffic flow. Data was collected in ten cities; in seven of the cities, more than one study was conducted as major geometric changes or revised signal timings were implemented, yielding nineteen networks for this study. As before, the two-fluid parameters in each network were estimated from chase car data and the network features were determined from maps, field studies, and local traffic engineers. Regression analysis yielded the following models:

$$T_m = 3.93 + 0.0035\,X_5 - 0.047\,X_6 - 0.433\,X_{10} \qquad (6.39)$$
$$\text{and } n = 1.73 + 1.124\,X_2 - 0.180\,X_3 - 0.0042\,X_5 - 0.271\,X_9$$

where X_2 is the fraction of one-way streets, X_3 the average number of lanes per street, X_5 the signal density, X_6 the average speed limit, X_9 the fraction of actuated signals, and X_{10} the fraction of approaches with good progression. The R^2 for these equations, 0.72 and 0.75 (respectively), are lower than those for Equation 6.38 (both very close to 1), reflecting the larger data size. The only feature in common with the previous model (Equation 6.38) is the appearance of the fraction of one-way streets in the model for n. Since all features selected can be changed through operational practices (signal density can be changed by placing signals on flash), the models have potential practical application. Computer simulation has also been used to investigate these relationships, and is discussed in Section 6.3.4.

6.3.4 Two-Fluid Parameters: Estimation by Computer Simulation

Computer simulation has many advantages over field data in the study of network models. Conditions not found in the field can be evaluated and new control strategies can be easily tested. In the case of the two-fluid model, the entire vehicle population in the network can be used in the estimation of the model parameters, rather than the small sample used in the chase car studies. TRAF-NETSIM (Mahmassani et al. 1984), a

microscopic traffic simulation model, has been used successfully with the two-fluid model.

Most of the simulation work to-date has used a generic grid network in order to isolate the effects of specific network features on the two-fluid parameters (FHWA 1993). Typically, the simulated network has been a 5 x 5 intersection grid made up entire of two-way streets. Traffic signals are at each intersection and uniform turning movements are applied throughout. The network is closed, i.e., vehicles are not allowed to leave the network, thus maintaining constant concentration during the simulation run. The trip histories of all the vehicles circulating in the network are aggregated to form a single (T_r , T) observation for use in the two-fluid parameter estimation. A series of five to ten runs over a range of network concentrations (nearly zero to 60 or 80 vehicles/lane-mile) are required to estimate the two-fluid parameters.

Initial simulation runs in the test network showed both T and T_s increasing with concentration, but T_r remaining nearly constant, indicating a very low value of n (Mahmassani et al. 1984). In its default condition, NETSIM generates few of the vehicle interaction of the type found in most urban street networks, resulting in flow which is much more idealized than in the field. The short-term event feature of NETSIM was used to increase the inter-vehicular interaction (Williams et al. 1985). With this feature, NETSIM blocks the right lane of the specified link at mid-block; the user specifies the average time for each blockage and the number of blockages per hour, which are stochastically applied by NETSIM. In effect, this represents a vehicle stopping for a short time (e.g., a commercial vehicle unloading goods), blocking the right lane, and requiring vehicles to change lanes to go around it. The two-fluid parameters (and n in particular) were very sensitive to the duration and frequency of the short-term events. For example, using an average 45-second event every two minutes, n rose from 0.076 to 0.845 and T_m fell from 2.238 to 2.135. With the use of the short-term events, the values of both parameters were within the ranges found in the field studies. Further simulation studies found both block length (here, distance between signalized intersections) and the use of progression to have significant effects on the two-fluid parameters (Williams et al. 1985).

Simulation has also provided the means to investigate the use of the chase car technique in estimating the two-fluid parameters (Williams et al. 1995). The network-wide averages in a simulation model can be directly computed; and chase car data can be simulated by recording the trip history of a single vehicle for one mile, then randomly selecting another vehicle in the network. Because the two-fluid model is non-linear (specifically, Equation 6.35, the log transform of which is used to estimate the parameters), estimations performed at the network level and at the individual vehicle level result in different values of the parameters, and are not directly comparable. The sampling strategy, which was found to provide the best parameter estimates, required a single vehicle circulating in the network for at least 15 minutes. However, due to the wide variance of the estimate (due to the possibility of a relatively small number of "chased" cars dominating the sample estimation), the estimate using a single vehicle was often far from the parameter estimated at the network level. On the other hand, using 20 vehicles to sample the network resulted in estimates much closer to those at the network level. The much smaller variance of the estimates made with twenty vehicles, however, resulted in the estimate being significantly different from the network-level estimate. The implication of this study is that, while estimates at the network and individual vehicle levels can not be directly compared, as long as the same sampling strategy is used, the resulting two-fluid parameters, although biased from the "true" value, can be used in making direct comparisons.

6.3.5 Two-Fluid Parameters: Influence of Network Features (Simulation Studies)

The question in Section 6.3.3, above, regarding the influence of geometric and control features of a network on the two-fluid parameters was revisited with an extensive simulation study (Bhat 1994). The network features selected were: average block length, fraction of one-way streets, average number of lanes per street, signals per intersection, average speed limit, average signal cycle length, fraction of curb miles with parking, and fraction of signalized approaches in progression. A uniform-precision central composite design was selected as the experimental design, resulting in 164 combination of the eight network variables. The simulated network was increased to 11 by 11 intersections; again, vehicles were not allowed to leave the network, but traffic data was collected only on the interior 9 by 9 intersection grid, thus eliminating the edge effects caused by the necessarily different turning movements at the boundaries. Ten simulation runs were made for each combination of

variables over a range of concentrations from near zero to about 35 vehicles/lane-mile.

Regression analysis yielded the following models:

$$T_m = 1.049 + 1.453\,X_2 + 0.684\,X_3 - 0.024\,X_6 \quad \text{and}$$
$$n = 4.468 - 1.391\,X_3 - 0.048\,X_5 + 0.042\,X_6 \qquad (6.40)$$

where X_2 is the fraction of one-way streets, X_3 the number of lanes per street, X_5 average speed limit, and X_6 average cycle length. The R^2 (0.26 and 0.16 for Equation 6.40) was considerably lower than that for the models estimated with data from field studies (Equation 6.39). Additionally, the only variable in common between Equations 6.39 and 6.40 is the number of lanes per street in the equation for n. Additional work is required to clarify these relationships.

6.3.6 Two-Fluid Model: A Practical Application

When the traffic signals in downtown San Antonio were retimed, TRAF-NETSIM was selected to quantify the improvements in the network. In order to assure that the results reported by NETSIM reflected traffic conditions in San Antonio, NETSIM was calibrated with the two-fluid model.

Turning movement counts used in the development of the new signal timing plans were available for coding NETSIM. Simulation runs were made for 31 periods throughout the day, and the two-fluid parameters were estimated and compared with those found in the field. By a trial and error process, NETSIM was calibrated by

- Increasing the sluggishness of drivers, by increasing headways during queue discharge at traffic signals and reducing maximum acceleration,
- Adding vehicle/driver types to increase the range of sluggishness represented in the network, and
- Reducing the desired speed on all links to 32.2 km/h during peaks and 40.25 km/h otherwise (Denney 1993).

Three measures of effectiveness (MOEs) were used in the evaluation: total delay, number of stops, and fuel consumption. The changes noted for all three MOEs were greater between calibrated and uncalibrated NETSIM results than between before and after results. Reported relative improvements were also affected. The errors in the reported improvements without calibration ranged from 16 percent to 132 percent (Denney 1994).

6.4 Two-Fluid Model and Traffic Network Flow Models

Computer simulation provides an opportunity to investigate network-level relationships between the three fundamental variables of traffic flow, speed *(V)*, flow *(Q)*, and concentration *(K)*, defined as average quantities taken over all vehicles in the network over some observation period (Mahmassani et al. 1984). While the existence of "nice" relations between these variables could not be expected, given the complexity of network interconnections, simulation results indicate relationships similar to those developed for arterials may be appropriate (Mahmassani et al. 1984; Williams et al. 1985). A series of simulation runs, as described in Section 6.3.4, above, was made at concentration levels between 10 and 100 vehicles/lane-mile. The results are shown in Figure 6.23, and bear a close resemblance to their counterparts for individual road sections. The fourth plot shows the relation of f_s, the fraction of vehicles stopped from the two-fluid model, to the concentration. In addition, using values of flow, speed, and concentration independently computed from the simulations, the network-level version of the fundamental relation $Q=KV$ was numerically verified (Mahmassani et al. 1984; Williams et al. 1987).

Three model systems were derived and tested against simulation results (Williams et al. 1987; Mahmassani et al. 1987); each model system assumed $Q=KV$ and the two-fluid model, and consisted of three relations:

$$V = f(K)\ , \qquad (6.41)$$

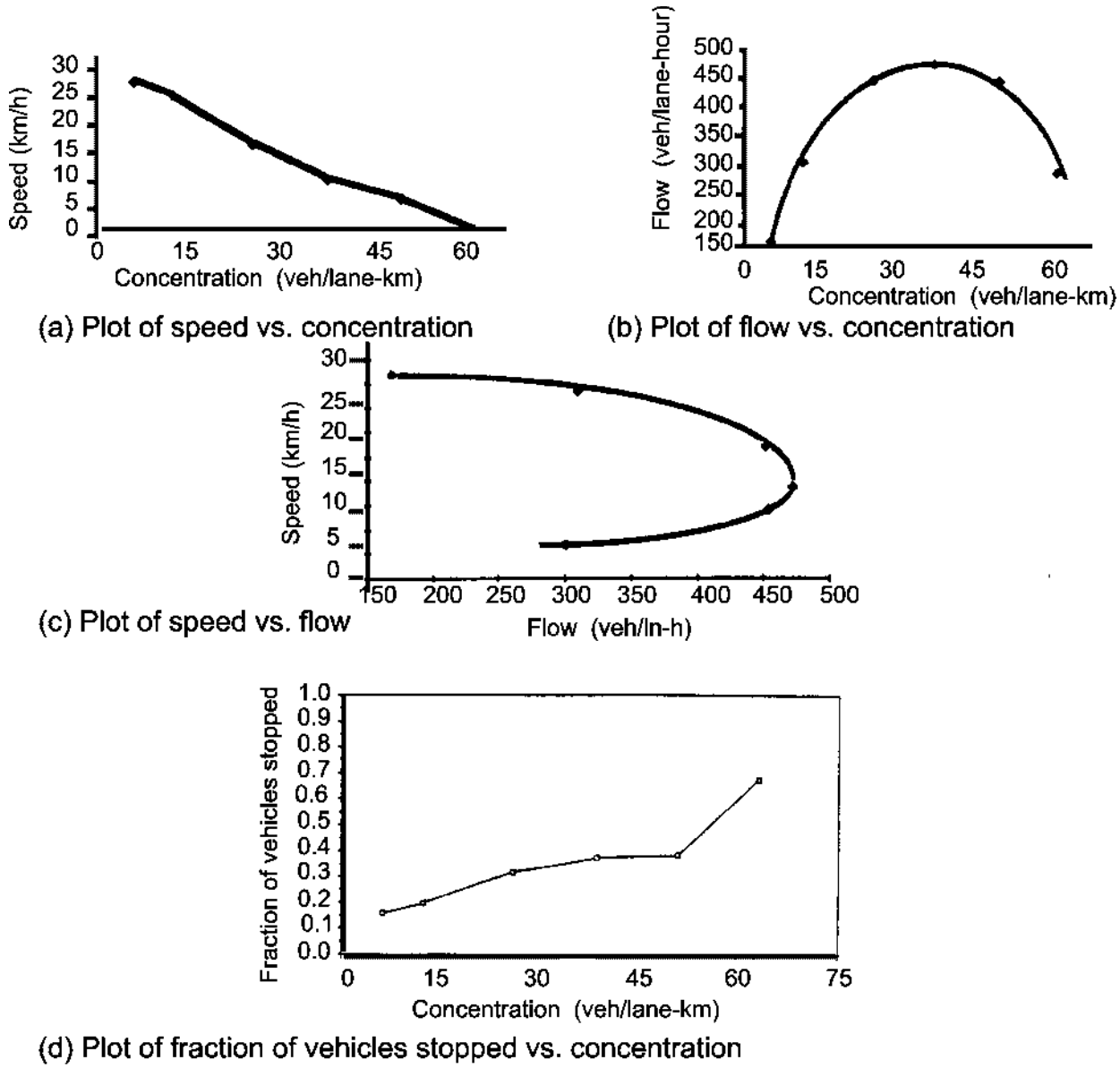


Figure 6.23
Simulation Results in a Closed CBD-Type Street Network.
(Williams et al. 1987, Figures 1-4).

A model system is defined by specifying one of the above relationships; the other two can then be analytically derived. (A relation between Q and V could also be derived.)

$$Q = g(K) \quad , \quad \text{and} \tag{6.42}$$

$$f_s = h(K) \quad . \tag{6.43}$$

Model System 1 is based on a postulated relationship between the average fraction of vehicles stopped and the network concentration from the two-fluid theory (Herman and Prigogine

1979), later modified to reflect that the minimum $f_s > 0$ (Ardekani and Herman 1987):

$$f_s = f_{s,\min} + (1 - f_{s,\min})(K/K_j)^{\pi} \quad , \qquad (6.44)$$

where $f_{s,min}$ is the minimum fraction of vehicles stopped in a network, K_j is the jam concentration (at which the network is effectively saturated), and π is a parameter which reflects the quality of service in a network. The other two relations can be readily found, first by substituting f_s from Equation 6.44 into Equation 6.31:

$$V = V_m (1 - f_{s,\min})^{n+1} [1 - (K/K_j)^{\pi}]^{n+1} \quad , \qquad (6.45)$$

then by using $Q=KV$,

$$Q = K V_m (1 - f_{s,\min})^{n+1} [1 - (K/K_j)^{\pi}]^{n+1} \quad . \qquad (6.46)$$

Equations 6.44 through 6.46 were fitted to the simulated data and are shown in Figure 6.24. Because the point representing the highest concentration (about 100 vehicles/lane-mile) did not lie in the same linear $\ln T_r$ - $\ln T$ trend as the other points, the two-fluid parameters n and T_m were estimated with and without the highest concentration point, resulting in the Method 1 and Method 2 curves, respectively, in the *V-K* and *Q-K* curves in Figure 6.24.

Model System 2 adopts Greenshields' linear speed-concentration relationship (Gerlough and Huber 1975),

$$V = V_f (1 - K/K_j) \quad , \qquad (6.47)$$

where V_f is the free flow speed (and is distinct from V_m; $V_f \le V_m$ always, and typically $V_f < V_m$). The f_s-K relation can be found by substituting Equation 6.47 into Equation 6.31 and solving for f_s:

$$f_s = 1 - [(V_f/V_m)(1 - K/K_j)]^{1/(n+1)} \quad , \qquad (6.48)$$

then by using $Q=KV$,

$$Q = V_f (K - K^2/K_j) \quad . \qquad (6.49)$$

Equations 6.47 through 6.49 were fitted to the simulation data and are shown in Figure 6.25. The difference between the Method 1 and Method 2 curves in the f_s-K plot (Figure 6.25) is described above. Model System 3 uses a non-linear bell-shaped function for the *V-K* model, originally proposed by Drake, et al., for arterials (Gerlough and Huber 1975):

$$V = V_f \exp[-\alpha (K/K_m)^d] \quad , \qquad (6.50)$$

where K_m is the concentration at maximum flow, and α and d are parameters. The f_s-K and *Q-K* relations can be derived as shown for Model System 2:

$$f_s = 1 - \{(V_f/V_m) \exp[-\alpha (K/K_m)^d]\}^{1/(n+1)} \quad \text{and} \qquad (6.51)$$

$$Q = K V_f \exp[-\alpha (K/K_m)^d] \quad . \qquad (6.52)$$

Equations 6.50 through 6.52 were fitted to the simulation data and are shown in Figure 6.26.

Two important conclusions can be drawn from this work. First, that relatively simple macroscopic relations between network-level variables appear to work. Further, two of the models shown are similar to those established at the individual facility level. Second, the two-fluid model serves well as the theoretical link between the postulated and derived functions, providing another demonstration of the model's validity. In the second and third model systems particularly, the derived f_s-K function performed remarkable well against the simulated data, even though it was not directly calibrated using that data.

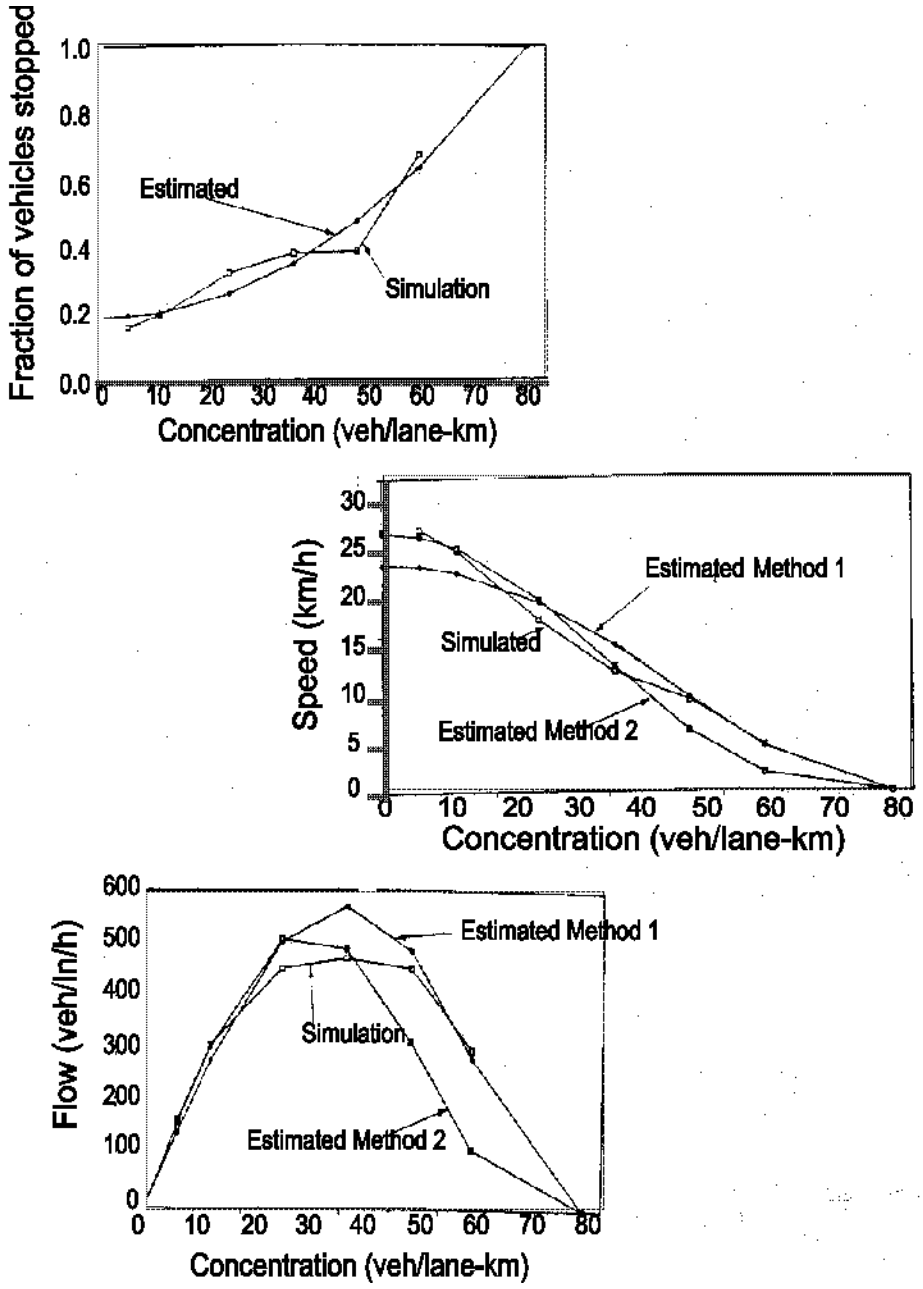


Figure 6.24
Comparison of Model System 1 with Observed Simulation Results
(Williams et al. 1987, Figure 5, 7, and 8).

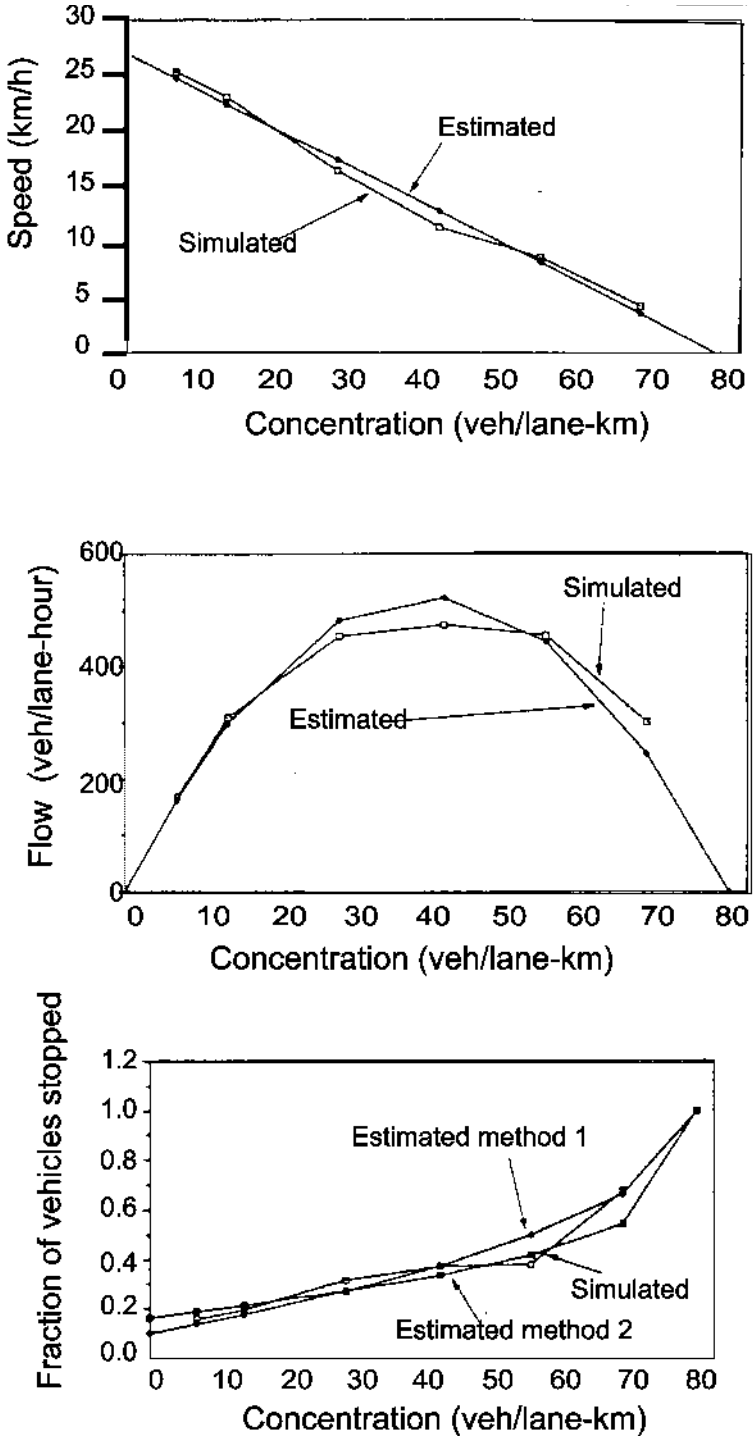


Figure 6.25
Comparison of Model System 2 with Observed Simulation Results (Williams et al. 1987, Figures 9-11).

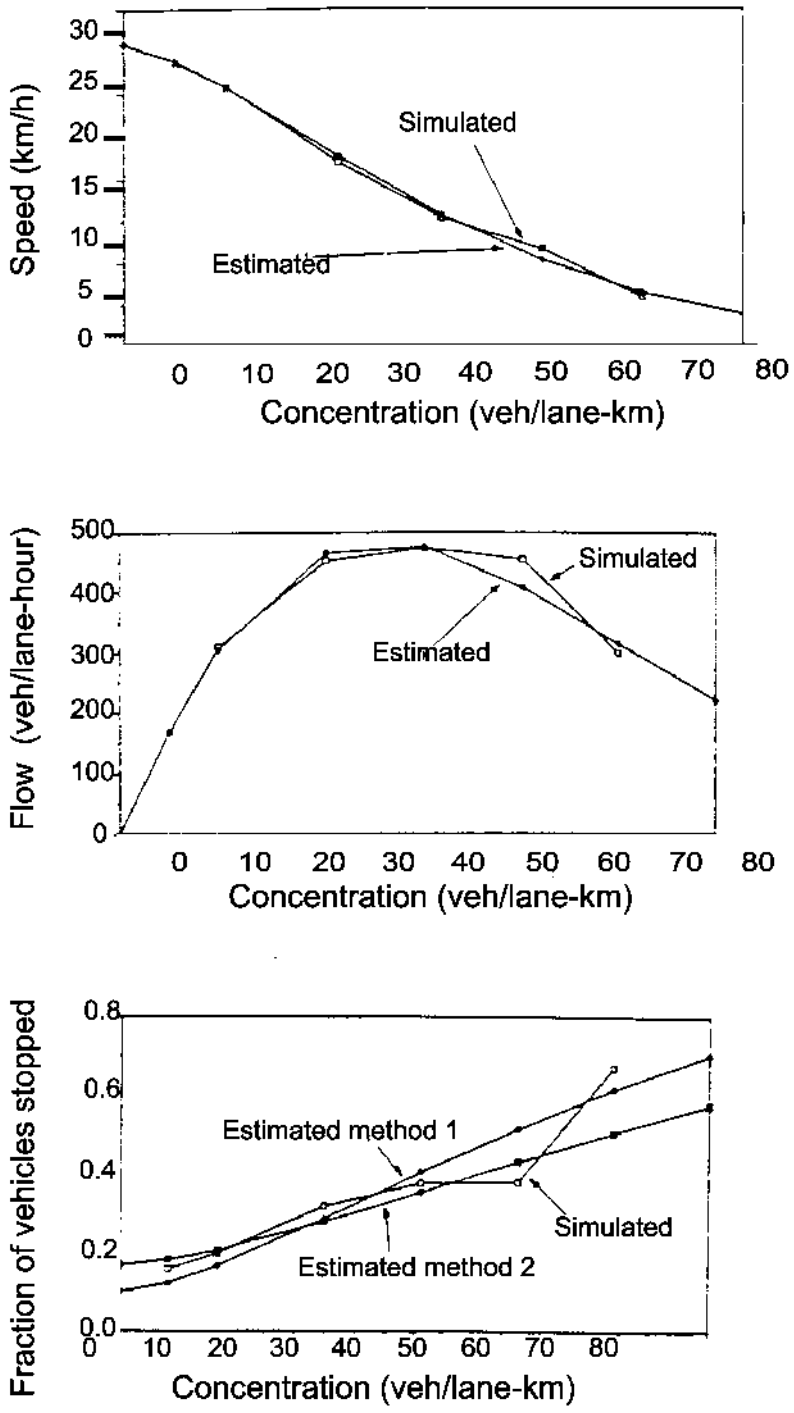


Figure 6.26
Comparison of Model System 3 with Observed Simulation Results (Williams et al. 1987, Figures 12-14).

6.5 Concluding Remarks

As the scope of traffic control possibilities widens with the development of ITS (Intelligent Transportation Systems) applications, the need for a comprehensive, network-wide evaluation tool (as well as one that would assist in the optimization of the control system) becomes clear. While the models discussed in this chapter are not ready for easy implementation, they do have promise, as in the application of the two-fluid model in San Antonio (Denney et al. 1993; 1994).

References

Angel, S. and G. M. Hyman (1970). *Urban Velocity Fields.* Environment and Planning, Vol. 2.

Ardekani, S. A. (1984). *The Two-Fluid Characterization of Urban Traffic: Theory, Observation, and Experiment.* Ph.D. Dissertation, University of Texas at Austin.

Ardekani, S. A., V. Torres-Verdin, and R. Herman (1985). *The Two-Fluid Model and Traffic Quality in Mexico City (El Modelo Bifluido y la Calidad del Tránsito en la Ciudad de México).* Revista Ingeniería Civil.

Ardekani, S. A. and R. Herman, (1987). *Urban Network-Wide Variables and Their Relations.* Transportation Science, Vol. 21, No. 1.

Ardekani, S. A., J. C. Williams,, and S. Bhat, (1992). *Influence of Urban Network Features on Quality of Traffic Service.* Transportation Research Record 1358, Transportation Research Board.

Ayadh, M. T. (1986). *Influence of the City Geometric Features on the Two Fluid Model Parameters*, M.S. Thesis, Virginia Polytechnic Institute and State University.

Beimborn, E. A. (1970). *A Grid Travel Time Model*.
Transportation Science, Vol. 4.

Bhat, S .C. S. (1994). *Effects of Geometric and Control Features on Network Traffic: A Simulation Study*. Ph.D. Dissertation, University of Texas at Arlington.

Branston, D. M. (1974). *Urban Traffic Speeds—I: A Comparison of Proposed Expressions Relating Journey Speed to Distance from a Town Center.* Transportation Science, Vol. 8, No. 1.

Buckley, D. G. and J. G. Wardrop, (1980). *Some General Properties of a Traffic Network.* Australian Road Research, Vol. 10, No. 1.

Denney, Jr., R. W., J. C. Williams, S. C. S. Bhat, and S. A. Ardekani, (1993). *Calibrating NETSIM for a CBD Using the Two-Fluid Model.* Large Urban Systems (Proceedings of the Advanced Traffic Management Conference), S. Yagar, A. Santiago, editors, Federal Highway Administration, U.S. Department of Transportation.

Denney, Jr., R. W., J. C. Williams, and S. C. S. Bhat, (1994). *Calibrating NETSIM Using the Two-Fluid Model.* Compendium of Technical Papers (Proceedings of the 64th ITE Annual Meeting, Dallas), Institute of Transportation Engineers.

Federal Highway Administration, U. S. Department of Transportation (1993). *TRAF User Reference Guide*, Version 4.0.

Gerlough, D. L. and M. J. Huber, (1975). *Traffic Flow Theory: A Monograph*, Special Report 165, Transportation Research Board.

Godfrey, J. W. (1969). *The Mechanism of a Road Network.* Traffic Engineering and Control, Vol. 11, No. 7.

Herman, R. and S. A. Ardekani, (1984). *Characterizing Traffic Conditions in Urban Areas.* Transportation Science, Vol. 18, No. 2.

Herman, R. and I. Prigogine, (1979). *A Two-Fluid Approach to Town Traffic.* Science, Vol. 204, pp. 148-151.

Hutchinson, T. P. (1974). *Urban Traffic Speeds—II: Relation of the Parameters of Two Simpler Models to Size of the City and Time of Day*. Transportation Science, Vol. 8, No. 1.

Lyman, D. A. and P. F. Everall, (1971). *Car Journey Times in London—1970*. RRL Report LR 416, Road Research Laboratory.

Mahmassani, H., J. C. Williams, and R. Herman, (1984). *Investigation of Network-Level Traffic Flow Relationships: Some Simulation Results.* Transportation Research Record 971, Transportation Research Board.

Mahmassani, H. S., J. C. Williams, and R. Herman, (1987). *Performance of Urban Traffic Networks*. Transportation and Traffic Theory (Proceedings of the Tenth International on Transportation and Traffic Theory Symposium, Cambridge, Massachusetts), N.H. Gartner, N.H.M. Wilson, editors, Elsevier.

Newell, G. F. (1980). *Traffic Flow on Transportation Networks*, Chapter 4, MIT Press.

Prigogine, I. and R. Herman, (1971). *Kinetic Theory of Vehicular Traffic*, American Elsevier.

Road Research Laboratory (1965). *Research on Road Traffic.*

Smeed, R. J. (1963). *International Road Safety and TrafficReview*, Vol. 11, No. 1.

Smeed, R. J. and J. G. Wardrop, (1964). *An Exploratory Comparison of the Advantages of Cars and Buses for Travel in Urban Areas*. Journal of the Institute of Transportation, Vol. 30, No. 9.

Smeed, R. J. (1965). *Journal of the Institute of Mathematics and its Applications*, Vol. 1.

Smeed, R. J. (1966). *Road Capacity of City Centers.* Traffic Engineering and Control, Vol. 8, No. 7.

Smeed, R. J. (1968). *Traffic Studies and Urban Congestion.* Journal of Transport Economics and Policy, Vol. 2, No. 1.

Thomson, J. M. (1967a). *Speeds and Flows of Traffic in Central London: 1. Sunday Traffic Survey*. Traffic Engineering and Control, Vol. 8, No. 11.

Thomson, J. M. (1967b). *Speeds and Flows of Traffic in Central London: 2. Speed-Flow Relations*. Traffic Engineering and Control, Vol. 8, No. 12.

Transportation Research Board (1994). *Highway Capacity Manual*, Special Report 209, (with revisions).

Vaughan, R., A. Ioannou, and R. Phylactou (1972). *Traffic Characteristics as a Function of the Distance to the Town Centre*. Traffic Engineering and Control, Vol. 14, No. 5.

Wardrop, J. G. (1952). *Some Theoretical Aspects of Road Traffic Research*. Proceedings of the Institution of Civil Engineers, Vol. 1, Part 2.

Wardrop, J. G. (1968). *Journey Speed and Flow in Central Urban Areas*. Traffic Engineering and Control, Vol. 9, No. 11.

Wardrop, J. G. (1969). *Minimum Cost Paths in Urban Areas.* Beiträge zur Theorie des Verkehrsflusses, Universitat (TH) Karlsruhe.

Williams, J. C., H. S. Mahmassani, and R. Herman (1995). *Sampling Strategies for Two-Fluid Model Parameter Estimation in Urban Networks*. Transportation Research - Part A, Vol. 29A, No. 3.

Williams, J. C., H. S. Mahmassani, and R. Herman, (1985). *Analysis of Traffic Network Flow Relations and Two-Fluid Model Parameter Sensitivity*. Transportation Research Record 1005, Transportation Research Board.

Williams, J. C., H. S. Mahmassani, and R. Herman, (1987). *Urban Traffic Network Flow Models*. Transportation Research Record 1112, Transportation Research Board.

Zahavi, Y. (1972a). *Traffic Performance Evaluation of Road Networks by the α-Relationship.* Traffic Engineering and Control, Vol. 14, No. 5.

Zahavi, Y. (1972b). *Traffic Performance Evaluation of Road Networks by the α-Relationship, Part 2*. Traffic Engineering and Control, Vol. 14, No. 6.

TRAFFIC IMPACT MODELS

BY SIA ARDEKANI[10]
EZRA HAUER[11]
BAHRAM JAMEI[12]

[10] Associate Professor, Civil Engineering Department, University of Texas at Arlington, Box 19308, Arlington, TX 76019-0308

[11] Professor, Department of Civil Engineering, University of Toronto, Toronto, Ontario Cananda M5S 1A4

[12] Transportation Planning Engineer, Virginia Department of Transportation, Fairfax, Va.

Chapter 7 - Frequently used Symbols

Φ	=	fuel consumption per unit distance
ω	=	parameter
a	=	instantaneous acceleration rate (a>0, km/hr/sec)
A	=	acceleration rate
$AADT$	=	average annual daily traffic (vehicles/day)
$ACCT$	=	acceleration time
ADT	=	average daily traffic (vehicles/day)
$C1$	=	average peak hour concentration
$C8$	=	annual maximum 8-hour concentration
CLT	=	Deceleration time
d_s	=	average stopped delay per vehicle (secs)
E	=	Engine size (cm^3)
EFI	=	idle emission factor
EFL	=	intersection link composite emission factor
f	=	instantaneous fuel consumption (mL/sec)
F	=	fuel consumed (lit/km) [Watson, et al model]
F	=	average fuel consumption per roadway section (mL) [Akcelik model]
f_1	=	fuel consumption rate while cruising (mL/km)
f_2	=	fuel consumption rate while idling (mL/sec)
f_3	=	excess fuel consumption per vehicle stop (mL)
f_c	=	steady-state fuel consumption rate at cruising speed (mL/km)
h	=	average number of stops per vehicle
K_1	=	parameter representing idle flow rate (mL/sec)
K_2	=	parameter representing fuel consumption to overcome rolling resistance
K_4,K_5	=	parameters related to fuel consumption due to positive acceleration
L	=	payload (Kg)
LQU	=	queue length
m	=	expected number of single-vehicle accidents per unit time
$NDLA$	=	vehicles delayed per cycle per lane
P	=	probability of a single-vehicle accident
PKE	=	positive kinetic energy
q	=	flow (vph)
q_i	=	flow rate of type i vehicles (vph)
S	=	speed
SPD	=	cruise speed
T	=	average travel time per unit distance
u	=	model parameter related to driving conditions
V	=	average speed (km/hr) [Elemental model]
V	=	instantaneous speed (km/hr) [Akcelik and Bayley model]
V_c	=	steady-state cruising speed (km/hr)
V_f	=	final speed (km/hr)
V_i	=	initial speed (km/hr)
VMT	=	vehicle miles of travel
V_s	=	space mean speed (km/hr)
VSP	=	at-rest vehicle spacing

7.
TRAFFIC IMPACT MODELS

7.1 Traffic and Safety

7.1.1 Introduction

This section ought to be about how traffic flow, speed and the like are related to accident frequency and severity. However, due to limitation of space, only the relationship between accident frequency and traffic flow will be discussed. The terminology that pertains to characteristics of the traffic stream has already been established and only a few definitions need to be added. The '**safety**' of an entity is defined as '**the number of accidents by type, expected to occur on the entity in a certain period, per unit of time**'. In this definition, 'accident types' are categories such as rear-end, sideswipe, single-vehicle, multi-vehicle, injury, property damage only, etc. The word 'expected' is as in probability theory: what would be the average-in-the-long-run if it was possible to freeze all the relevant circumstances of the period at their average, and then repeat it over and over again. The word 'entity' may mean a specific road section or intersection, a group of horizontal curves with the same radius, the set of all signalized intersections in Philadelphia, etc. Since the safety of every entity changes in time, one must be specific about the period. Furthermore, to facilitate communication, safety is usually expressed as a frequency. Thus, eg., one might speak about the expected number of fatal accidents/year in 1972-1976 for a certain road section. To standardize further one often divides by the section length; now the units may be, say, accidents/(year × km).

As defined, the safety of an entity is a string of expected frequencies, $m_1, m_2, \ldots, m_i, \ldots,$ one for each accident type chosen. However, for the purpose of this discussion it will suffice to speak about one (unspecified) accident type, the expected accident frequency of which is m_i.

7.1.2 Flow and Safety

The functional relationship between m_i and the traffic flow which the entity serves, is a **'safety performance function'**. A safety performance function is depicted schematically in Figure 7.1. For the moment its shape is immaterial. It tells how for some entity the expected frequency of accidents of some type would be changing if traffic flow on the entity changed while all other conditions affecting accident occurrence remained fixed. While the flow may be in any units, it is usually understood that it pertains to the same period of time which the accident frequency represents. Thus, eg., if the ordinate shows the expected number of fatal accidents/year in 1972-1976 for a certain road section, then the AADT is the average for the period 1972-1976.

Naturally, m_i can be a function of more than one traffic flow. Thus, eg., head-on collisions may depend on the two opposing flows; collisions between pedestrians and left-turning traffic depend on the flow of pedestrians, the flow of straight-through vehicles, and the flow of left-turning vehicles etc.. In short, the arguments of the safety performance function can be several flows.

In practice it is common to use the term '**accident rate**'. The accident rate is proportional to the slope of the line joining the origin and a point of the safety performance function. Thus, at point A of Figure 7.1, where AADT is 3000 vehicles per day and where the expected number of accidents for this road section is 1.05 accidents per year, the accident rate is $1.05/(3000\times365)=0.96\times10^{-6}$ accidents/vehicle. At point B the accident rate is $1.2/(4000\times365)=0.82\times10^{-6}$ accidents/vehicle. If the road section was, say, 1.7 km long, the same accident rates could be written as $1.05/(3000\times365\times1.7)=0.56\times10^{-6}$ accidents/vehicle-km and $1.2/(4000\times365\times1.7)=0.48\times10^{-6}$ accidents/vehicle-km.

The safety performance function of an entity is seldom a straight line. If so, the accident rate is not constant but varies with traffic flow. As a consequence, if one wishes to compare the safety of two or more entities serving different flows, one can not use the accident rate for this purpose. The widespread habit of using accident rates to judge the relative safety of different entities or to assess changes in safety of the same entity is inappropriate and often harmful. To illustrate, suppose that the AADT on the road section in Figure 7.1 increased from 3000 'before pavement resurfacing' to 4000 'after pavement resurfacing' and that the average accident frequency increased from 1.05 'before' to 1.3 'after'. Note that 1.2 accidents/year would be expected at AADT=4000 had the road surface remained unchanged (see Figure 7.1). Since $1.3 > 1.2$ one must conclude that following resurfacing there was a **deterioration** of 0.1 accidents/year. But

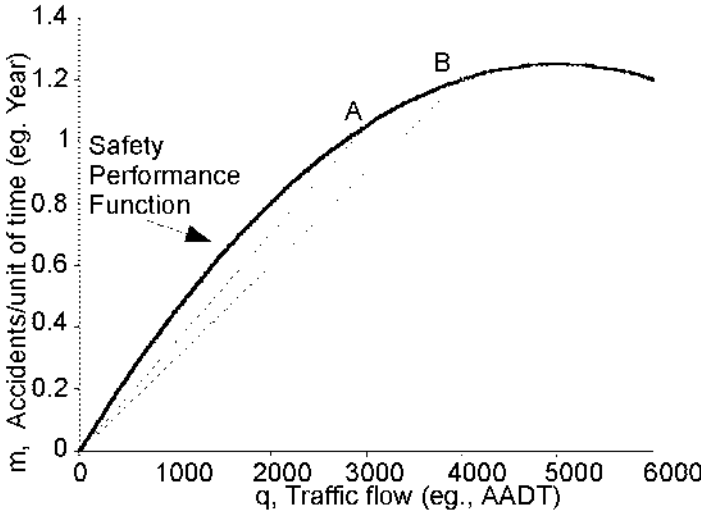


Figure 7.1
Safety Performance Function and Accident Rate.

less than the accident rate 'before resurfacing' ($1.05/3000\times365 = 0.96\times10^{-6}$) which erroneously suggests that there has been an improvement. Similar arguments against the use of accident rates can be found in Pfundt (1969), Hakkert et al. (1976), Mahalel (1986), Brundell-Freij & Ekman (1991), Andreassen (1991).

To avoid such errors, the simple rule is that safety comparisons are legitimate only when the entities or periods are compared **as if they served the same amount of traffic**. To accomplish such equalization, one needs to know the safety performance function. Only in the special case when the performance function happens to be a straight line, may one divide by traffic flow and then compare accident rates. However, to judge whether the safety performance function is a straight line, one must know its shape, and when the shape of the safety performance function is known, the computation of an accident rate is superfluous. It is therefore best not to make use of accident rates. For this reason, the rest of the discussion is about expected accident frequencies, not rates.

Knowledge of safety performance functions is an important element of rational road safety management. The nature and shape of this function is subject to some logical considerations. However, much of the inquiry must be empirical.

7.1.3 Logical Considerations

It stands to reason that there is some kind of relationship between traffic flow and safety. For one, without traffic there are no traffic accidents. So, the safety performance function must go through the origin. Also, the three interrelated characteristics of the traffic stream - flow, speed and density - all influence the three interrelated aspects of safety - the frequency of opportunities for accidents to occur, the chance of accident occurrence given an opportunity, and the severity of the outcome given an accident. However, while a relationship may be presumed to exist, it is rather difficult to learn much about its mathematical form by purely deductive reasoning.

Using logic only, one could argue as follows: "If, as in probability theory, the passage of a vehicle through a road section or an intersection is a 'trial' the 'outcome' of which can be 'accident' or 'no-accident' with some fixed probability. Assume further that vehicle passages are so infrequent that this probability is not influenced by the frequency at which the 'trials' occur. Under such conditions the expected number of single-vehicle accidents in a fixed time period must be proportional to the number of trials in that time period - that is to flow." In symbols, $m_{\text{single-vehicle}}=qp$, where q is flow and p is the probability of a single-vehicle accident in one passage of a

vehicle. In this, p is a constant that does not depend on q. Thus, one may argue, that the number of single-vehicle accidents ought to be proportional to flow, but only at very low flows.

As flow and density increase to a point where a driver can see the vehicle ahead, the correspondence between the mental picture of independent trials and between reality becomes strained. The probability of a 'trial' to result in a single-vehicle accident now depends on how close other vehicles are; that is $p = p(q)$. Should $p(q)$ be an increasing function of q, then $m_{\text{single-vehicle}}$ would increase more than in proportion with flow. Conversely, if an increase in flow diminishes the probability that a vehicle passing the road section will be in a single-vehicle accident, then $m_{\text{single-vehicle}}$ would increase less than in proportion to traffic flow; indeed, $m_{\text{single-vehicle}} = qp(q)$ can even decrease as traffic flow increases beyond a certain point. Thus, by logical reasoning one can only conclude that **near the origin**, the safety performance function for single-vehicle accidents ought to be a straight line.

If the safety performance function depends on two conflicting flows (car-train collisions at rail-highway grade crossings, car-truck collisions on roads, car-pedestrian collision at intersections etc.) then, near the origin, m_i should be proportional to the **product** of the two flows. One could also use the paradigm of probability theory to speculate that (at very low flows) the expected number of collisions with vehicles parked on shoulders is proportional to the **square** of the flows: in the language of 'trials', 'outcomes', the number of vehicles parked on the shoulder ought to be proportional to the passing flow and the number of vehicles colliding with the parked cars ought to be proportional to the same flow. From here there is only a small step to argue that, say, the number of rear-end collisions should also be proportional to q^2. Again, this reasoning applies only to very low flows. How m depends on q when speed choice, alertness and other aspects of behavior are also a function of flow, cannot be anticipated by speculation alone.

This is as far as logical reasoning seems to go at present. It only tells us what the shape of the safety performance function should be near the origin. Further from the origin, when p changes with q, not much is gained thinking of m as the product $qp(q)$. Since the familiar paradigm of 'trials' and 'outcomes' ceased to fit reality, and the notion of 'opportunity to have an accident' is vague, it might be better to focus directly on the function $m = m(q)$ instead of its decomposition into the product $qp(q)$.

Most theoretical inquiry into the relationship between flow and safety seems to lack detail. Thus, eg., most researchers try to relate the frequency of right-angle collisions at signalized intersections to the two conflicting flows. However, on reflection, the second and subsequent vehicles of a platoon may have a much lesser chance to be involved in such a collision than the first vehicle. Therefore it might make only a slight difference whether 2 or 20 vehicles have to stop for the same red signal. For this reason, the total flow is likely to be only weakly and circuitously related to the number of situations which generate right angle collisions at signalized intersections. There seems to be scope and promise for more detailed, elaborate and realistic theorizing. In addition, most theorizing to date attempted to relate safety to flow only. However, since flow, speed and density are connected, safety models could be richer if they contained all relevant characteristics of the traffic stream. Thus, eg., a close correspondence has been established between the number of potential overtakings derived from flow and speed distribution and accident involvement as a function of speed (Hauer 1971). Welbourne (1979) extends the ideas to crossing traffic and collisions with fixed objects. Ceder (1982) attempts to link accidents to headway distributions (that are a function of flow) through probabilistic modeling.

There is an additional aspect of the safety performance function which may benefit from logical examination. The claim was that it is reasonable to postulate the existence of a relationship between the traffic triad 'flow, speed and density' and between the safety triad 'frequency of opportunities, chance of accident given opportunity, and severity of the outcome given accident'. However, if there is a cause-effect relationship, it must be between accidents and the traffic characteristics near the time of their occurrence. One must ask whether there still is some meaningful safety performance function between accidents and traffic flow when flow is averaged over, say, a year. Whether the habit of relating accidents to AADTs (that is, averages over a year) materially distorts the estimated safety performance function is at present unclear. In a study by Quaye et al. (1993) three separate models were estimated from 15 minute flows, which then were aggregated into 1 hour flows and then into 7 hour flows. The three models differed but little. Persaud and Dzbik (1993) call models that relate hourly flows to accidents "microscopic" and models that relate AADT to yearly accident counts "macroscopic".

As will become evident shortly, empirical inquiries about safety performance functions display a disconcerting variety of results.

A part of this variety could be explained by the fact that the most ubiquitous data for such inquiries consist of flow estimates which are averages pertaining to periods of one year or longer.

7.1.4 Empirical Studies

Empirical studies about the association of traffic flow and accident frequency seldom involve experimentation; their nature is that of fitting functions to data. What is known about safety performance functions comes from studies in which the researcher assembles a set of data of traffic flows and accident counts, chooses a certain function that is thought to represent the relationship between the two, and then uses statistical techniques to estimate the parameters of the chosen function. Accordingly, discussion here can be divided into sections dealing with: (1) the kinds of study and data, (2) functional forms or models, (3) parameter estimates.

7.1.4.1 Kinds Of Study And Data

Data for the empirical investigations are accident counts and traffic flow estimates. A number-pair consisting of the accident count for a certain period and the estimated flow for the same period is a 'data point' in a Cartesian representation such as Figure 7.1. To examine the relationship between traffic flow and accident frequency, many such points covering an adequate range of traffic flows are required.

There are two study prototypes (for a discussion see eg., Jovanis and Chang 1987). The most common way to obtain data points for a range of flows is to choose many road sections or intersections that are similar, except that they serve different flows. In this case we have a '**cross-section'** study. In such a study the accident counts will reflect not only the influence of traffic flow but also of all else that goes with traffic flow. In particular, facilities which carry larger flows tend to be built and maintained to higher standards and tend to have better markings and traffic control features. This introduces a systematic bias into 'cross-section' models. If a road that is built and maintained to a higher standard is safer, then accident counts on high-flow roads will tend to be smaller than had the roads been built, maintained and operated in the same way as the lower flow roads. Thus, in a cross-section study, it is difficult to separate what is due to traffic flow, and what is due to all other factors which depend on traffic flow. It is therefore questionable whether the result of a cross section study enables one to anticipate how safety of a certain facility would change as a result of a change in traffic flow.

Less common are studies that relate different traffic flows on the same facility to the corresponding accident counts. In this case we have a '**time-sequence**' study. In such a study one obtains the flow that prevailed at the time of each accident and the number of hours in the study period when various flows prevailed. The number of accidents in a flow group divided by the number of hours when such flows prevailed is the accident frequency (see, eg., Leutzbach et al. 1970, Ceder and Livneh 1978, Hall and Pendelton 1991). This approach might obviate some of the problems that beset the cross-section study. However, the time-sequence study comes with its own difficulties. If data points are AADTs and annual accident counts over a period of many years, then the range of the AADTs is usually too small to say much about any relationship. In addition, over the many years, driver demography, norms of behavior, vehicle fleet, weather, and many other factors also change. It is therefore difficult to distinguish between what is due to changes in traffic flow and what is due to the many other factors that have also changed. If the data points are traffic flows and accidents over a day, different difficulties arise. For one, the count of accidents (on one road and when traffic flow is in a specified range) is bound to be small. Also, low traffic flows occur mostly during the night, and can not be used to estimate the safety performance function for the day. Also, peak hour drivers are safer en-route to work than on their way home in the afternoon, and off-peak drivers tend to be a different lot altogether.

7.1.4.2 Models

The first step of an empirical study of the relationship between traffic flow and accident frequency is to assemble, plot and examine the data. The next step is to select the candidate model equation(s) which might fit the data and serve as the safety performance function. Satterthwaite (1981, section 3) reviews the most commonly used models. Only those that are plausible and depend on traffic flow only are listed below. Traffic flow, while important, is but one of the many variables on which the expected accident frequency depends. However, since the monograph is devoted to traffic flow, the dependence on other variables will not be pursued here. The Greek letters are the unknown parameters to be estimated from data.

When only one traffic stream is relevant the power function and polynomial models have been used:

$$m = \alpha q^{\beta} \quad (7.1)$$

$$m = \alpha q + \beta q^2 + \ldots \quad (7.2)$$

At times the more complex power form

$$m = \alpha q^{\beta + \gamma \log(q)} \quad (7.1a)$$

is used which is also akin to the polynomial model 7.2 when written in logarithms

$$\log(m) = \log(\alpha) + \beta \log(q) + \gamma [\log(q)]^2 \quad (7.2a)$$

When two or more traffic streams or kinds of vehicles are relevant, the product of power functions seems common:

$$m = \alpha q_1^{\beta} q_2^{\gamma} \ldots \quad (7.3)$$

The common feature of the models most often used is that they are linear or can be made so by logarithmization. This simplifies statistical parameter estimation. The shapes of these functions are shown in Figure 7.2.

The power function (Equation 7.1) is simple and can well satisfy the logical requirements near the origin (namely, that when q=0 m=0, and that β=1 when one flow is involved or β=2 when two flows are involved). However, its simplicity is also its downfall. If logic dictates, eg., that near the origin β=1 (say, for single-vehicle accidents), then the safety performance function has to be a straight line even for the higher flows where $p(q)$ is not constant any more. Similarly, if logic says that β=2, the quadratic growth applies for all q. In short, if β is selected to meet requirements of logic, the model may not fit the data further from the origin. Conversely, if β is selected to fit the data best, the logical requirements will not be met.

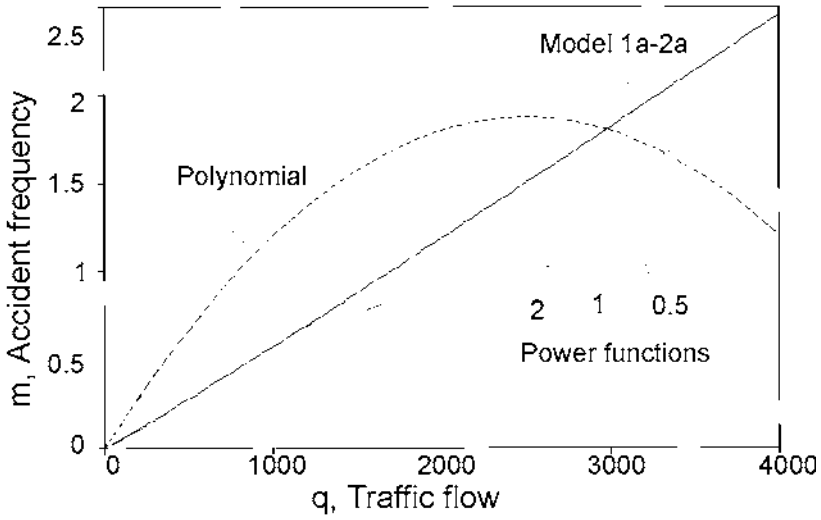


Figure 7.2
Shapes of Selected Model Equations.

The popularity of the power function in empirical research derives less from its suitability than from it being 'lazy user friendly'; most software for statistical parameter estimation can accommodate the power function with little effort. The polynomial model (Equation 7.2) never genuinely satisfies the near-origin requirements. Its advantage is that by using more terms (and more parameters) the curve can be bent and shaped almost at will. This is achieved at the expense of parsimony in parameters.

If the data suggest that as flow increases beyond a certain level, the slope of $m(q)$ is diminishing, perhaps even grows negative, an additional model that is parsimonious in parameters might deserve consideration:

$$m = \alpha q^k e^{\beta q} \qquad (7.4)$$

where k = 1 or 2 in accord with the near-origin requirements. When $\beta<0$ the function has a maximum at $q = -k/\beta$. Its form when k=1 and k=2 is shown in Figure 7.3. The advantage of this model is that it can meet the near-origin requirements and still can follow the shape of the data.

A word of caution is in order. In the present context the focus is on how accident frequency depends on traffic flow. Accordingly, the models were written with flow (q) as the principal independent variable. However, traffic flow is but one of the many causal factors which affect accident frequency. Road geometry, time of day, vehicle fleet, norms of behaviour and the like all play a part. Therefore, what is at times lumped into a single parameter (α in equation 7.1, 7.1a, 7.3 and 7.4) really represents a complex multivariate expression. In short, the modeling of accident frequency is multivariate in nature.

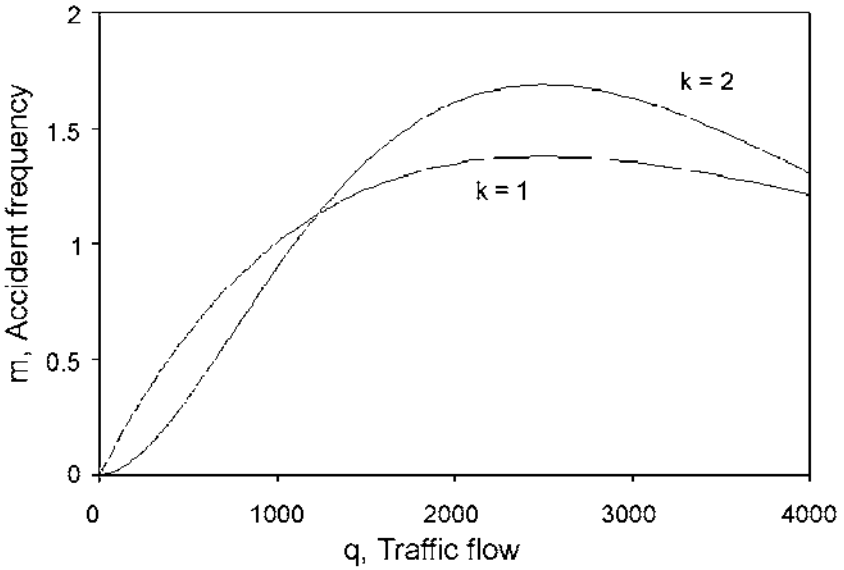


Figure 7.3
Two Forms of the Model in Equation 7.4.

7.1.4.3 Parameter Estimates

With the data in hand and the functional form selected, the next step is to estimate the parameters (α, β,...). In earlier work, estimation was often by minimization of squared deviations. This practice now seems deficient. Recognizing the discrete nature of accident counts, the fact that their variance increases with their mean, and the possible existence of over-dispersion, it now seems that more appropriate statistical techniques are called for (see eg., Hauer 1992, Miaou and Lum 1993).

Results of past research are diverse. Part of the diversity is due to the problems which beset both the cross-section and the time sequence of studies; another part is due to the use of AADT and similar long-period averages that have a less than direct tie to accident occurrence; some of the diversity may come from various methodological shortcomings (focus on accident rates, choice of simplistic model equations, use of inappropriate statistical techniques); and much is due to the diversity between jurisdictions in what counts as a reportable accident and in the proportion of reportable accidents that get reported. Hauer and Persaud (1996) provide a comprehensive review of safety performance functions and their parameter values (for two-lane roads, multi-lane roads without access control, freeways, intersections and interchanges) based on North American data. A brief summary of this information and some international results are given below.

A. Road Sections. In a cross-section study of Danish *rural roads* Thorson (1967) estimated the exponent of ADT to be 0.7. In a similar study of German rural roads Pfundt (1968) estimated the exponent of ADT to be 0.85. Kihlberg and Tharp (1968) conducted an extensive cross-section study using data from several states. For sections that are 0.5 miles long, they estimate the parameters for a series of road types and geometric features. The model used is an elaborated power function $m=\alpha(ADT)^{\beta}(ADT)^{\gamma\log(ADT)}$. The report contains a rich set of results but creates little order in the otherwise bewildering variety. Ceder and Livneh (1982) used both cross-sectional and time-sequence data for interurban road sections in Israel, using the simple power function model (Equation. 1). The diverse results are difficult to summarize. Cleveland et al. (1985) divide low-volume rural two-lane roads into 'bundles' by geometry and find the ADT exponents to range from 0.49 to 0.93 for off-road accidents Recent studies in the UK show that on urban road sections, for single-vehicle accidents the exponent of AADT in model 7.1 is 0.58; for rear-end accidents the exponent of AADT is 1.43. In a time-sequence study, Hall and Pendelton (1990) use ten mile 2 and 4-lane road segments surrounding permanent counting stations in New Mexico and provide a wealth of information about accident rates in relations to hourly flows and time of day. In an extensive cross-section study, Zegeer et al. (1986) find that the exponent of ADT is 0.88 for the total number of accidents on rural two-lane roads. Ng and Hauer (1989) use the same data as Zegeer and show that the parameters differ from state to state and also by lane width. For non-intersection accidents on rural two lane roads in New York State, Hauer et al. (1994) found that when m is measured in [accidents/(mile-year)] and AADT is used for q, then in model 1, in 13 years α varies from 0.0024-0.0028 and β=0.78. Persaud (1992) using data on rural roads in Ontario finds the exponent of AADT to vary between 0.73 and 0.89, depending on lane and shoulder width. For urban two-lane roads in Ontario the exponent is 0.72 For urban multi-lane roads (divided or undivided) β=1.14, for rural multi-lane divided roads it is 0.62 but for undivided roads it is again 1.13.

For California freeways Lundy (1965) shows that the accidents per million vehicle miles increase roughly linearly with ADT. This implies the quadratic relationship of model 2. Based on the figures in Slatterly and Cleveland (1969), with m measured in [accidents/day], $m = (5.8\times10^{-7})ADT + (2.4\times10^{-11})ADT^2$ for four-lane freeways, $m = (6.6\times10^{-7})ADT + (.94 \times 10^{-11})ADT^2$ for six-lane freeways and $m = (5.4 \times 10^{-7}) \times ADT + (.78 \times 10^{-11}) \times ADT^2$ for eight-lane freeways. Leutzbach (1970) examines daytime accidents on a stretch of an autobahn. Fitting a power function to his Figure (1c) and with m measured in accidents per day, $m =(3\times10^{-11})\times(\text{hourly flow})^3$. However, there is an indication in this and other data that as flow increases, the accident rate initially diminishes and then increases again. If so a third degree polynomial might be a better choice. Jovanis and Chang (1987) fit model 7.3 to the Indiana Toll Road and find the exponents to be 0.25 and 0.23 for car and truck-miles. Persaud and Dzbik (1993) find that when yearly accidents are related to AADT, $m=0.147\times(AADT/1000)^{1.135}$ for 4-lane freeways but, when hourly flows are related to accidents/hour, $m = 0.00145 \times (\text{hourly flow}/1000)^{0.717}$. Huang et al. (1992) report for California that Number of accidents = 0.65 + 0.666 × million-vehicle-miles.

B. Intersections. Tanner (1953) finds for rural T intersections in the UK the exponents to be 0.56 and 0.62 for left turning traffic and main road traffic respectively. Roosmark (1966) finds for similar intersections in Sweden the corresponding exponents to be 0.42 and 0.71. For intersections on divided highways McDonald (1953) gives m in accidents per year as $m = 0.000783 \times (\text{major road ADT})^{0.455} \times (\text{cross road ADT})^{0.633}$. For signalized intersections in California, Webb (1955) gives m in accidents per year, as $m = 0.00019 \times (\text{major road ADT})^{0.55} \times (\text{cross road ADT})^{0.55}$ in urban areas with speeds below 40 km/h ; $= 0.0056 \times (\text{major road ADT})^{0.45} \times (\text{cross road ADT})^{0.38}$ in semi-urban areas with speeds in the 40-70 km/h range; and $= 0.007 \times (\text{major road ADT})^{0.51} \times (\text{cross road ADT})^{0.29}$ in rural areas with speeds above 70 km/h. For rural stop-controlled intersections in Minnesota, Bonneson and McCoy (1993) give m [in accidents/year] $= 0.692 \times (\text{major road ADT}/1000)^{0.256} \times (\text{cross road ADT}/1000)^{0.831}$. Recent studies done in the UK show that at signalized intersections for single vehicle accidents the exponent of AADT in model 1 is 0.89; for right angle accidents (model 7.3) the exponents of AADT are 0.36 and 0.60 and for accident to left-turning traffic (their right turn) the exponents are 0.57 and 0.46. Using data from Quebec, Belangér finds that the expected annual number of accidents at unsignalized rural intersections is $0.002 \times (\text{Major road ADT})^{0.42} \times (\text{Cross road ADT})^{0.51}$.

C. Pedestrians. Studies in the UK show that for nearside pedestrian accidents on urban road sections the exponents for vehicle and pedestrian AADTs in model 7.3 are 0.73 and 0.42. With m measured in [pedestrian accidents/year] Brüde and Larsson (1993) find that, at intersections, $m = (7.3 \times 10^{-6})(\text{incoming traffic/day})^{0.50}$ (crossing pedestrians/day)$^{0.7}$. With m measured in (pedestrian accidents/hour), Quaye et al. (1993) find that if left-turning vehicles at signalized intersections do not face opposing vehicular traffic then $m = 1.82 \times 10^{-8}$(hourly flow of left-turning cars)$^{1.32} \times$ (hourly pedestrian flow)$^{0.34}$; when the left turning vehicles have to find gaps in the opposing traffic, $m = 1.29 \times 10^{-7}$ (hourly flow of left-turning cars)$^{0.36} \times$ (hourly pedestrian flow)$^{0.86}$.

7.1.5 Closure

Many aspects of the traffic stream are related to the frequency and severity of accidents; only the relationship with flow has been discussed here. How safety depends on flow is important to know. The relationship of traffic flow to accident frequency is called the 'safety performance function'. Only when the safety performance function is known, can one judge whether one kind of design is safer than another, or whether an intervention has affected the safety of a facility. Simple division by flow to compute accident rates is insufficient because the typical safety performance function is not linear.

Past research about safety performance functions has led to diverse results. This is partly due to the use of flow data which are an average over a long time period (such as AADT), partly due to the difficulties which are inherent in the cross-sectional and the time-sequence studies, and partly because accident reporting and roadway definitions vary among jurisdictions. However, a large part of the diversity is due to the fact that accident frequency depends on many factors in addition to traffic flow and that the dependence is complex.Today, some of the past difficulties can be overcome. Better information about traffic flows is now available (eg. from freeway-traffic-management-systems, permanent counting stations, or weight-in-motion devices); also better methods for the multivariate analysis of accident counts now exist. However, in addition to progress in statistical modeling, significant advances seem possible through the infusion of detailed theoretical modeling which makes use of all relevant characteristics of the traffic stream such as speed, flow, density, headways and shock waves.

7.2 Fuel Consumption Models

Substantial energy savings can be achieved through urban traffic management strategies aimed at improving mobility and reducing delay. It is conservatively estimated, for example, that if all the nearly 250,000 traffic signals in the U.S. were optimally timed, over 19 million liters (5 million gallons) of fuel would be saved each day. It is further estimated that 45 percent of the total energy consumption in the U.S. is by vehicles on roads. This amounts to some 240 million liters (63 million gallons) of petroleum per day, of which nearly one-half is used by vehicles under urban driving conditions.

Fuel consumption and emissions have thus become increasingly important measures of effectiveness in evaluating traffic management strategies. Substantial research on vehicular energy consumption has been conducted since the 1970's, resulting in an array of fuel consumption models. In this section, a number of such models which have been widely adopted are reviewed.

7.2.1 Factors Influencing Vehicular Fuel Consumption

Many factors affect the rate of fuel consumption. These factors can be broadly categorized into four groups: vehicle, environment, driver, and traffic conditions. The main variables in the traffic category include speed, number of stops, speed noise and acceleration noise. Speed noise and acceleration noise measure the amount of variability in speed and acceleration in terms of the variance of these variables. The degree of driver aggressiveness also manifests itself in speed and acceleration rates and influences the fuel consumption rate.

Factors related to the driving environment which could affect fuel consumption include roadway gradient, wind conditions, ambient temperature, altitude, and pavement type (for example, AC/PCC/gravel) and surface conditions (roughness, wet/dry). Vehicle characteristics influencing energy consumption include total vehicle mass, engine size, engine type (for example, gasoline, diesel, electric, CNG), transmission type, tire type and size, tire pressure, wheel alignment, the status of brake and carburetion systems, engine temperature, oil viscosity, gasoline type (regular, unleaded, etc.), vehicle shape, and the degree of use of auxiliary electric devices such as air-conditioning, radio, wipers, etc. A discussion of the degree of influence of most of the above variables on vehicle fuel efficiency is documented by the Ontario Ministry of Transportation and Communications (TEMP 1982).

7.2.2 Model Specifications

Fuel consumption models are generally used to evaluate the effectiveness and impact of various traffic management strategies. As such these models are developed using data collected under a given set of vehicle fleet and performance characteristics such as weight, engine size, transmission type, tire size and pressure, engine tune-up and temperature conditions, etc. The variation in fuel consumption due to other factors such as driver characteristics, ambient temperature, pavement roughness, grades, wind, altitude, etc. is hoped to be small due to collective effects if data points represent aggregate values over sufficiently long observation periods. Alternatively, model estimates may be adjusted for known effects of roadway grade, ambient temperature, altitude, wind conditions, payload, etc.

Given a fixed set of vehicle and driver characteristics and environmental conditions, the influence of traffic-related factors on fuel consumption can be modeled. A number of studies in Great Britain (Everall 1968), Australia (Pelensky et al. 1968), and the United States (Chang et al. 1976; Evans and Herman 1978; Evans et al. 1976) all indicated that the fuel consumption per unit distance in urban driving can be approximated by a linear function of the reciprocal of the average speed. One such model was proposed by Evans, Herman, and Lam (1976), who studied the effect of sixteen traffic variables on fuel consumption. They concluded that speed alone accounts for over 70 percent of the variability in fuel consumption for a given vehicle. Furthermore, they showed that at speeds greater than about 55 km/h, fuel consumption rate is progressively influenced by the aerodynamic conditions. They classified traffic conditions as urban (V < 55 km/h) versus highway (V > 55 km/h) traffic showing that unlike the highway regime, in urban driving fuel efficiency improves with higher average speeds (Figure 7.4).

7.2.3 Urban Fuel Consumption Models

Based on the aforementioned observations, Herman and co-workers (Chang and Herman 1981; Chang et al. 1976; Evans and Herman 1978; Evans et al. 1976) proposed a simple theoretically-based model expressing fuel consumption in urban conditions as a linear function of the average trip time per unit distance (reciprocal of average speed). This model, known as the Elemental Model, is expressed as:

$$\Phi = K_1 + K_2 T, \quad V < 55 \text{ km/hr} \tag{7.5}$$

where,

Φ : fuel consumption per unit distance

T : average travel time per unit distance

and

$V(=1/T)$: average speed

K_1 and K_2 are the model parameters. K_1 (in mL/km) represents fuel used to overcome the rolling friction and is closely related

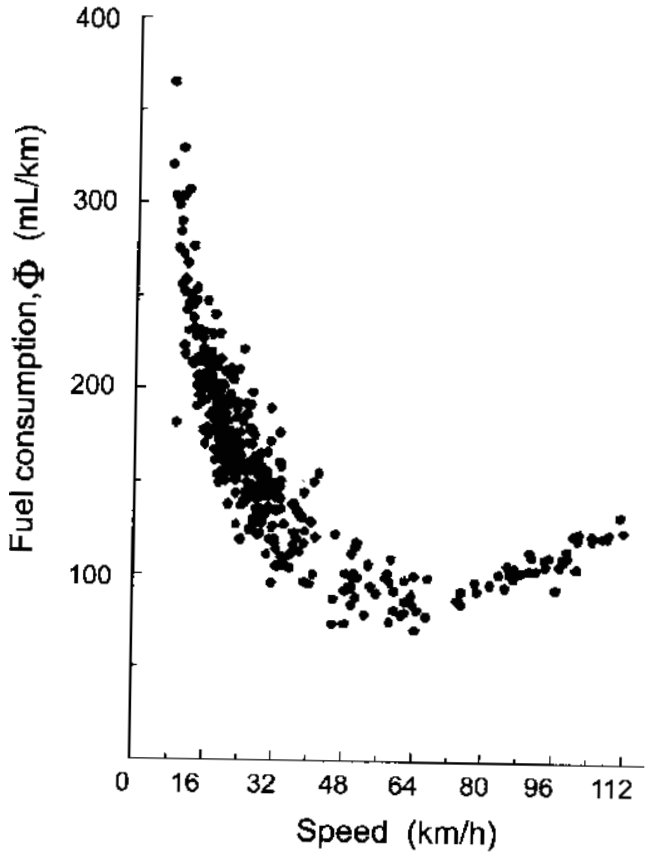


Figure 7.4
Fuel Consumption Data for a Ford Fairmont (6-Cyl.)
Data Points represent both City and Highway Conditions.

to the vehicle mass (Figure 7.5). K_2 (in mL/sec) is a function In an effort to improve the accuracy of the Elemental Model, other researchers have considered additional independent variables. Among them, Akcelik and co-workers (Akcelik 1981; Richardson and Akcelik1983) proposed a model which separately estimates the fuel consumed in each of the three portions of an urban driving cycle, namely, during cruising, idling, and deceleration-acceleration cycle. Hence, the fuel consumed along an urban roadway section is estimated as:

$$F = f_1 X_s + f_2 d_s + f_3 h, \qquad (7.11)$$

where,

F = average fuel consumption per roadway section (mL)
X_s = total section distance (km)
d_s = average stopped delay per vehicle (secs)
h = average number of stops per vehicle
f_1 = fuel consumption rate while cruising (mL/km)
f_2 = fuel consumption rate while idling (mL/sec)
f_3 = excess fuel consumption per vehicle stop (mL)

The model is similar to that used in the TRANSYT-7F simulation package (Wallace 1984).

Herman and Ardekani (1985), through extensive field studies, have shown that delay and number of stops should not be used

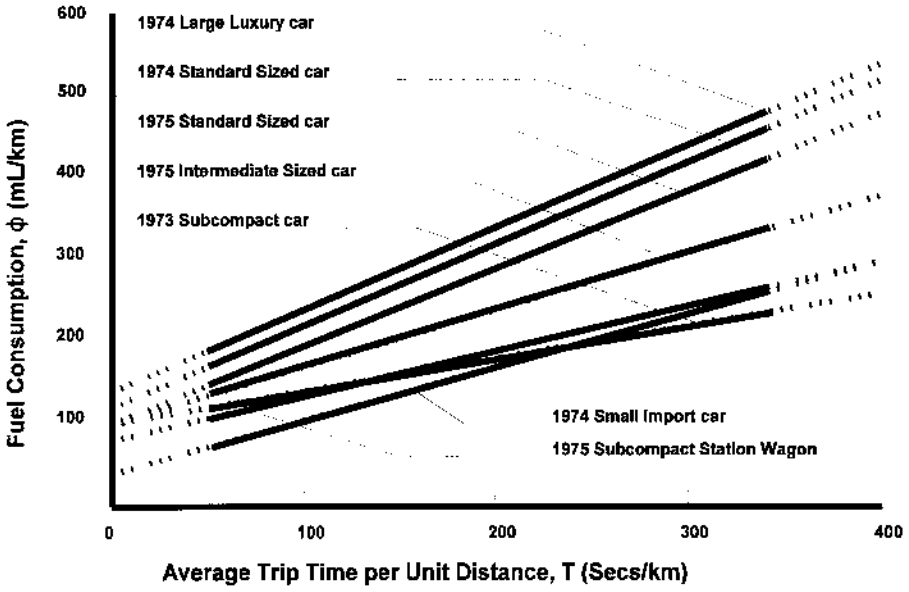


Figure 7.5
Fuel Consumption Versus Trip Time per Unit Distance for a Number of Passenger Car Models.

together in the same model as estimator variables. This is due to the tendency for the number of stops per unit distance to be highly correlated with delay per unit distance under urban traffic conditions. They propose an extension of the elemental model of Equation 7.5 in which a correction is applied to the fuel consumption estimate based on the elemental model depending on whether the number of stops made is more or less than the expected number of stops for a given average speed (Herman and Ardekani 1985).

A yet more elaborate urban fuel consumption model has been set forth by Watson et al. (1980). The model incorporates the changes in the positive kinetic energy during acceleration as a predictor variable, namely,

$$F = K_1 + K_2/V_s + K_3 V_s + K_4 PKE, \qquad (7.12)$$

where,

F = fuel consumed (Lit/km)
V_s = space mean speed (km/hr)

The term *PKE* represents the sum of the positive kinetic energy changes during acceleration in m/sec^2, and is calculated as follows:

$$PKE = \Sigma(V_f^2 - V_i^2)/(12{,}960\, X_s) \qquad (7.13)$$

where,

V_f = final speed (km/hr)
V_i = initial speed (km/hr)
X_s = total section length (km)

A number of other urban fuel consumption models are discussed by other researchers, among which the work by Hooker et al. (1983), Fisk (1989), Pitt et al. (1987), and Biggs and Akcelik (1986) should be mentioned.

7.2.4 Highway Models

Highway driving corresponds to driving conditions under which average speeds are high enough so that the aero-dynamic effects on fuel consumption become significant. This occurs at average speeds over about 55 km/h (Evans et al. 1976). Two highway models based on constant cruising speed are those by Vincent et al. (1980) and Post et al. (1981). The two models are valid at any speed range, so long as a relatively constant cruise speed can be maintained (steady state speeds). The steady-state speed requirement is, of course, more easily achievable under highway driving conditions.

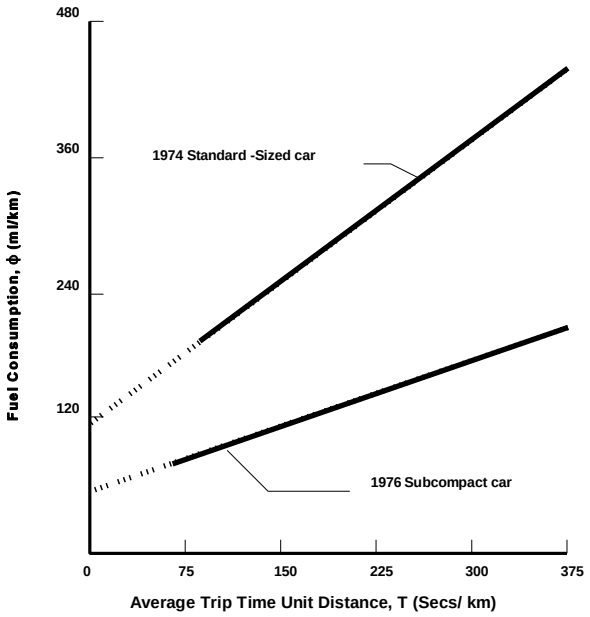


Figure 7.6
Fuel Consumption Data and the Elemental Model Fit for Two Types of Passenger Cars (Evans and Herman 1978).

The model by Vincent, Mitchell, and Robertson is used in the TRANSYT-8 computer package (Vincent et al. 1980) and is in the form:

$$f_c = a + b\,V_c + c\,V_c^2, \tag{7.14}$$

where,

V_c = steady-state cruising speed (km/hr)
f_c = steady-state fuel consumption rate at cruising speed (mL/km), calibration of this model for a mid-size passenger car yields

a = 170 mL/km,
b = -4.55 mL-hr/km^2; and
c = 0.049 mL-hr^2/km^3 (Akcelik 1983).

A second steady-state fuel model formulated by Post et al. (1981) adds a V^2 term to the elemental model of Equation 7.5 to account for the aero-dynamic effects, namely,

$$f_c = b_1 + b_2/V_c + b_3\,V_c^2. \tag{7.15}$$

Calibration of this model for a Melbourne University test car (Ford Cortina Wagon, 6-Cyl, 4.1L, automatic transmission) yields (Akcelik 1983) the following parameter values (Also see Figure 7.7):

b_1 = 15.9 mL/km
b_2 = 2,520 mL/hr
b_3 = 0.00792 mL-hr^2/km^3.

Instantaneous fuel consumption models may also be used to estimate fuel consumption for non-steady-state speed conditions under both urban or highway traffic regimes. These models are used in a number of microscopic traffic simulation packages such as NETSIM (Lieberman et al. 1979) and MULTSIM (Gipps and Wilson 1980) to estimate fuel consumption based on instantaneous speeds and accelerations of individual vehicles.

By examining a comprehensive form of the instantaneous model, Akcelik and Bayley (1983) find the following simpler form of the function to be adequate, namely,

$$f = K_1 + K_2\,V + K_3\,V^3 + /K_4\,aV + K_5\,a^2V/ \tag{7.16}$$

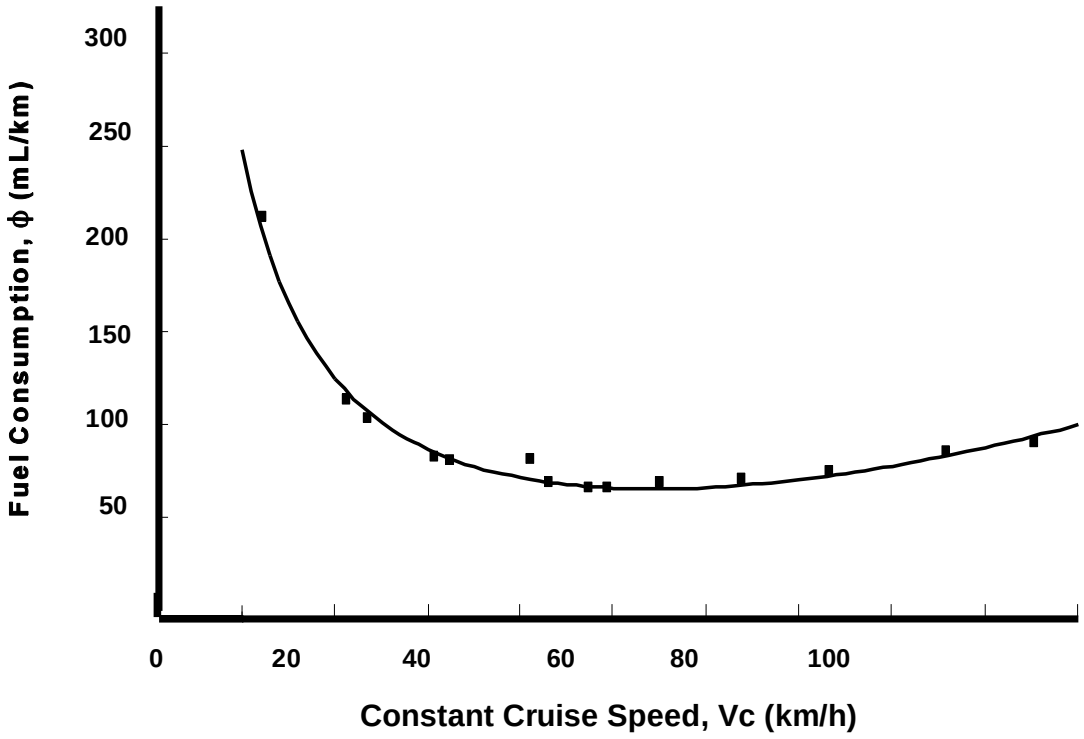


Figure 7.7
Constant-Speed Fuel Consumption per Unit Distance for the Melbourne University Test Car (Akcelik 1983).

where,

f	=	instantaneous fuel consumption (mL/sec)
V	=	instantaneous speed (km/hr)
a	=	instantaneous acceleration rate (a > 0)(km/hr/sec)
K_1	=	parameter representing idle flow rate (mL/sec)
K_2	=	parameter representing fuel consumption to overcome rolling resistance
K_3	=	parameter representing fuel consumption to overcome air resistance
K_4, K_5	=	parameters related to fuel consumption due to positive acceleration

The above model has been used by Kenworthy et al. (1986) to assess the impact of speed limits on fuel consumption.

7.2.5 Discussion

During the past two decades, the U.S. national concerns over dependence on foreign oil and air quality have renewed interest in vehicle fuel efficiency and use of alternative fuels. Automotive engineers have made major advances in vehicular fuel efficiency (Hickman and Waters 1991; Greene and Duleep 1993; Komor et al. 1993; Greene and Liu 1988).

Such major changes do not however invalidate the models presented, since the underlying physical laws of energy consumption remain unchanged. These include the relation between energy consumption rate and the vehicle mass, engine size, speed, and speed noise. What does change is the need to recalibrate the model parameters for the newer mix of vehicles. This argument is equally applicable to alternative fuel vehicles (DeLuchi et al. 1989), with the exception thatthere may also be a need to redefine the variable units, for example, from mL/sec or Lit/km to KwH/sec or KwH/km, respectively.

7.3 Air Quality Models

7.3.1 Introduction

Transportation affects the quality of our daily lives. It influences our economic conditions, safety, accessibility, and capability to reach people and places. Efficient and safe transportation satisfies us all but in contrast, the inefficient and safe use of our transportation system and facilities which result in traffic congestions and polluted air produces personal frustration and great economic loss.

The hazardous air pollutants come from both mobile and stationary sources. Mobile sources include passenger cars, light and heavy trucks, buses, motorcycles, boats, and aircraft. Stationary sources range from oil refineries to dry cleaners and iron and steel plants to gas stations.

This section concentrates on mobile source air pollutants which comprise more than half of the U.S. air quality problems. Transportation and tailpipe control measure programs in addition to highway air quality models are discussed in the section.

Under the 1970 U.S. Clean Air Act, each state must prepare a State Implementation Plan (SIP) describing how it will controll emissions from mobile and stationary sources to meet the requirements of the National Ambient Air Quality Standards (NAAQS) for six pollutants: (1) particulate matter (formerly known as total suspended particulate (TSP) and now as PM10 which emphasizes the smaller particles), (2) sulfur dioxide (SO2), (3) carbon monoxide (CO), (4) nitrogen dioxide (NO2), (5) ozone (O3), and (6) lead (Pb).

The 1990 U.S. Clean Air Act requires tighter pollution standards especially for cars and trucks and would empower the Environmental Protection Agency (EPA) to withhold highway funds from states which fail to meet the standards for major air pollutants (USDOT 1990). The 1970 and 1990 federal emission standards for motor vehicles are shown in Table 7.1.

This section concentrates on mobile source air pollutants which comprise more than half of the U.S. air quality problems. Exhaust from these sources contain carbon monoxide, volatile organic compounds (VOCs), nitrogen oxides, particulates, and lead. The VOCs along with nitrogen oxides are the major elements contributing to the formation of "smog".

Carbon monoxide which is one of the main pollutants is a colorless, and poisonous gas which is produced by the incomplete burning of carbon in fuels. The NAAQS standard for ambient CO specifies upper limits for both one-hour and eight-hour averages that should not be exceeded more than once per year. The level for one-hour standard is 35 parts per million (ppm), and for the eight-hour standard is 9 ppm. Most information and trends focus on the 8-hour average because it is the more restrictive limit (EPA 1990).

7.3.2 Air Quality Impacts of Transportation Control Measures

Some of the measures available for reducing traffic congestion and improving mobility and air quality is documented in a report prepared by Institute of Transportation Engineers (ITE) in 1989. This "toolbox" cites, as a primary cause of traffic congestion, the increasing number of individuals commuting by automobile in metropolitan areas, to and from locations dispersed throughout a wide region, and through areas where adequate highway capacity does not exist. The specific actions that can be taken to improve the situation are categorized under five components as follows:

1) Getting the most out of the existing highway system
   - Intelligent Transportation Systems (ITS)
   - urban freeways (ramp metering, HOV's)
   - arterial and local streets (super streets, parking management)
   - enforcement
2) Building new capacity (new highway, reconstruction)
3) Providing transit service (paratransit service, encouraging transit use)
4) Managing transportation demand
   - strategic approaches to avoiding congestion (road pricing)
   - mitigating existing congestion (ridesharing)
5) Funding and institutional measures
   - funding (fuel taxes, toll roads)
   - institutional measures (transportation management associations)

Table 7.1
Federal Emission Standards

	1970 Standards (grams/km)	1990 Standards (grams/km)
Light Duty Vehicles[1] (0-3, 340 Kgs)		
Carbon Monoxide (CO)	2.11	2.11
Hydrocarbons (HC)	0.25	0.16
Oxides of Nitrogen (NOx)	0.62	0.25[2]
Particulates	0.12	0.05
Light Duty Vehicles[1] (1,700-2,600 Kgs)		
Carbon Monoxide (CO)	2.11	2.73
Hydrocarbons (HC)	0.25	0.20
Oxides of Nitrogen (NOx)	0.62	0.44[2]
Particulates	0.12	0.05

	1970 Standards (grams/km)	1990 Standards (grams/km)
Light Duty Trucks (over 2,600 Kgs GVWR)		
Carbon Monoxide (CO)	6.22	3.11
Hydrocarbons (HC)	0.50	0.24
Oxides of Nitrogen (NOx)	1.06	0.68
Particulates	0.08	0.05

[1] ***Light duty vehicles include light duty trucks.***
[2] ***The new emission standards specified in this table are for useful life of 5 years or 80,000 Kms whichever first occurs.***

7.3.3 Tailpipe Control Measures

It is clear that in order to achieve the air quality standards and to reduce the pollution from the motor vehicle emissions, substantial additional emission reduction measures are essential. According to the U.S. Office of Technology Assessment (OTA), the mobile source emissions are even higher in most sources - responsible for 48 percent of VOCs in non-attainment U.S. cities in 1985, compared to other individual source categories (Walsh 1989).

The following set of emission control measures, if implemented, has the potential to substantially decrease exhaust emissions from major air pollutants:

- limiting gasoline volatility to 62.0 KPa (9.0 psi) RVP;
- adopting "onboard" refueling emissions controls;
- "enhanced" inspection and maintenance programs; requiring "onboard diagnostics" for emission control systems;
- adopting full useful life (160,000 kms) requirements;

- requiring alternative fuel usage;
- utilizing oxygenated fuels/reformulated gasoline measures;
- adopting California standards (Tier I and II).

Some of the measures recommended in California standards include: improved inspection and maintenance such as "centralized" inspection and maintenance programs, heavy duty vehicle smoke enforcement, establishing new diesel fuel quality standards, new methanol-fueled buses, urban bus system electrification, and use of radial tires on light duty vehicles.

In the case of diesel-fueled LDT's (0-3,750 lvw) and light-duty vehicles, before the model year 2004, the applicable standards for NOx shall be 0.62 grams/km for a useful life as defined above.

7.3.4 Highway Air Quality Models

As discussed earlier, federal, state, and local environmental regulations require that the air quality impacts of transportation - related projects be analyzed and be quantified. For this purpose, the Federal Highway Administration (FHWA) has issued guidelines to ensure that air quality effects are considered during planning, design, and construction of highway improvements, so that these plans are consistent with State Implementation Plans (SIPs) for achieving and maintaining air quality standards .

For example, the level of CO associated with a given project is a highway - related air quality impact that requires evaluation. In general, it must be determined whether the ambient standards for CO (35 ppm for 1 hour and 9 ppm for 8 hours, not to be exceeded more than once per year) will be satisfied or exceeded as a result of highway improvements. This requirement calls for estimating CO concentrations on both local and areawide scales.

A number of methods of varying sophistication and complexity are used to estimate air pollutant levels. These techniques include simple line-source-oriented Gaussian models as well as more elaborate numerical models (TRB 1981). The databases used in most models could be divided into the following categories:

- meteorological data
  - wind speed and direction
  - temperature
  - humidity
  - wind and temperature fluctuations
- traffic data
  - traffic volume
  - vehicle speed
  - vehicle length or type
- site types
  - at-grade sites
  - elevated sites
  - cut sites
- period of measurement

Some of these models which estimate the pollutant emissions from highway vehicles are discussed in more detail in the following sections.

7.3.4.1 UMTA Model

The most simple of these air quality models is the one which relates vehicular speeds and emission levels (USDOT 1985). The procedure is not elaborate but is a quick-response technique for comparison purposes. This UMTA (now Federal Transit Administration) model contains vehicular emission factors related to speed of travel for freeways and surfaced arterials.

The model uses a combination of free flow and restrained (peak period) speeds. It assumes that one-third of daily travel would occur in peak hours of flow reflecting restrained (congested) speeds, while two-thirds would reflect free-flow speed characteristics. For a complete table of composite emission factors categorized by autos and trucks for two calendar years of 1987 and 1995 refer to Characteristics of Urban Transportation System, U.S. Department of Transportation, Urban Mass Transportation Administration, October 1985.

7.3.4.2 CALINE-4 Dispersion Model

This line source air quality model has been developed by the California Department of Transportation (FHWA 1984). It is based on the Gaussian diffusion equation and employs a mixing zone concept to characterize pollutant dispersion over the roadway.

The model assesses air quality impacts near transportation facilities given source strength, meteorology, and site geometry. CALINE-4 can predict pollution concentrations for receptors

located within 500 meters of the roadway. It also has special options for modeling air quality near intersections, street canyons, and parking facilities.

CALINE-4 uses a composite vehicle emission factor in grams per vehicle-mile and converts it to a modal emission factor. The Environmental Protection Agency (EPA) has developed a series of computer programs, the latest of which is called Mobile4.1 (EPA 1991), for estimating composite mobile emission factors given average route speed, percent cold and hot-starts, ambient temperature, vehicle mix, and prediction year. These emission factors are based on vehicle distribution weighted by type, age, and operation mode, and were developed from certification and surveillance data, mandated future emission standards, and special emission studies.

Composite emission factors represent the average emission rate over a driving cycle. The cycle might include acceleration, deceleration, cruise, and idle modes of operation. Emission rates specific to each of these modes are called modal emission factors. The speed correction factors used in composite emission factor models, such as MOBILE4, are derived from variable driving cycles representative of typical urban trips. The Federal Test Procedures (FTP) for driving cycle are the basis for most of these data.

Typical input variables for the CALINE-4 model are shown in Table 7.2. In case of an intersection, the following assumptions are made for determining emission factors:

- uniform vehicle arrival rate;
- constant acceleration and deceleration rates; constant time rate of emissions over duration of each mode;
- deceleration time rate of emissions equals 1.5 times the idle rate;
- an "at rest" vehicle spacing of 7 meters; and
- all delayed vehicles come to a full stop.

In addition to composite emission factor at 26 km/hr (EFL), the following variables must be quantified for each intersection link:

- arrival volume in vehicles per hour;
- departure volume in vehicles per hour;
- average number of vehicles per cycle per lane for the dominant; movement
- average number or vehicles delayed per cycle per lane for the dominant movement (NDLA);
- distance from link endpoints to stopline;
- acceleration and deceleration times (ACCT, DCLT);
- idle times at front and end of queue;
- cruise speed (SPD); and
- idle emission rate (EFI).

The following computed variables are determined for each link from the input variables:

- acceleration rate;
- deceleration rate;
- acceleration length;
- deceleration length;
- acceleration-speed product;
- FTP-75 (BAG2) time rate emission factor;
- acceleration emission factor;
- cruise emission factor;
- deceleration emission factor; and
- queue length (LQU=NDLA*VSP), where VSP is the "at rest" vehicle spacing.

The cumulative emission profiles (CEP) for acceleration, deceleration, cruise, and idle modes form the basis for distributing the emissions. These profiles are constructed for each intersection link, and represent the cumulative emissions per cycle per lane for the dominant movement. The CEP is developed by determining the time in mode for each vehicle during an average cycle/lane event multiplied by the modal emission time rate and summed over the number of vehicles. CALINE4 can predict concentrations of relatively inert pollutants such as carbon monoxide (CO), and other pollutants like nitrogen dioxide (NO2) and suspended particles.

7.3.4.3 Mobie Source Emission Factor Model

MOBILE4.1 is the latest version of mobile source emission factor model developed by the U.S. Environmental Protection Agency (EPA). It is a computer program that estimates hydrocarbon (HC), carbon monoxide (CO), and oxides of nitrogen (NOx) emission factors for gasoline-fueled and diesel-fueled highway motor vehicles.

MOBILE4.1 calculates emission factors for eight vehicle types in two regions (low- and high-altitude). Its emission estimates depend on various conditions such as ambient temperature, speed, and mileage accrual rates. MOBILE4.1 will estimate emission factors for any calendar year between 1960 and 2020.

Table 7.2
Standard Input Values for the CALINE4

I. Site Variables	
Temperature	©
Wind Speed	(m/s)
Wind Direction	(deg)
Directional Variability	(deg)
Atmospheric Stability	(F)
Mixing Height	(m)
Surface Roughness	(cm)
Settling Velocity	(m/s)
Deposition Velocity	(m/s)
Ambient Temperature	©
II. Link Variables	
Traffic Volume	(veh/hr)
Emission Factor	(grams/veh-mile)
Height	(m)
Width	(m)
Link Coordinates	(m)
III. Receptor Locations	(m)

The 25 most recent model years are considered to be in operation in each calendar year. It is to be used by the states in the preparation of the highway mobile source portion of the 1990 base year emission inventories required by the Clean Air Act Amendments of 1990.

MOBILE4.1 calculates emission factors for gasoline-fueled light-duty vehicles (LDVs), light-duty trucks (LDTs), heavy-duty vehicles (HDVs), and motorcycles, and for diesel LDVs, LDTs, and HDVs. It also includes provisions for modeling the effects of oxygenated fuels (gasoline-alcohol and gasoline- ether blends) on exhaust CO emissions. Some of the primary input variables and their ranges are discussed below.

Speed correction factors are used by the model to correct exhaust emissions for average speeds other than that of the FTP (32 km/hr). MOBILE4.1 uses three speed correction models: low speeds (4-32 km/hr), moderate speeds (32-77 km/hr), and high speeds (77-105 km/hr). The pattern of emissions as a function of vehicle speed is similar for all pollutants, technologies, and model year groups. Emissions are greatest at the minimum speed of 4 km/hr, decline relatively rapidly as speeds increase from 4 to 32 km/hr, decline more slowly as speeds increase from 32 to 77 km/hr, and then increase with increasing speed to the maximum speed of 105 km/hr.

The vehicle miles traveled (VMT) mix is used to specify the fraction of total highway VMT that is accumulated by each of the eight regulated vehicle types. The VMT mix is used only to calculate the composite emission factor for a given scenario on the basis of the eight vehicle class-specific emission factors. Considering the dependence of the calculated VMT mix on the annual mileage accumulation rates and registration distributions by age, EPA expects that states develop their own estimates of VMT by vehicle type for specific highway facility, sub-zones, time of day, and so on.

Many areas of the country have implemented inspection and maintenance (I/M) programs as a means of further reducing mobile source air pollution. MOBILE4.1 has the capability of modeling the impact of an operating I/M program on the calculated emission factors, based on user specification of certain parameters describing the program to be modeled. Some of the parameters include:

- program start year and stringency level;
- first and last model years of vehicles subject to program;
- program type (centralized or decentralized);
- frequency of inspection (annual or biennial); and
- test types.

MOBILE4.1 (EPA 1991) has the ability to model uncontrolled levels of refueling emissions as well as the impacts of the implementation of either or both of the major types of vehicle recovery systems. These include the "Stage II" (at the pump) control of vehicle refueling emissions or the "onboard" (on the vehicle) vapor recovery systems (VRS).

The minimum and maximum daily temperatures are used in MOBILE4.1 in the calculation of the diurnal portion of evaporative HC emissions, and in estimating the temperature of dispensed fuel for use in the calculation of refueling emissions. The minimum temperature must be between -18 C to 38 C (0 F and 100 F), and the maximum temperature must be between -12 C to 49 C (10 F and 120 F) inclusive.

The value used for calendar year in MOBILE4.1 defines the year for which emission factors are to be calculated. The model has the ability to model emission factors for the year 1960 through 2020 inclusive. The base year (1990) inventories are based on a typical day in the pollutant season, most commonly Summer for ozone and Winter for CO. The base year HC inventories are based on interpolation of the calendar year 1990 and 1991 MOBILE4.1 emission factors.

One important determinant of emissions performance is the mode of operation. The EPA's emission factors are based on testing over the FTP cycle, which is divided into three segments or operating modes: cold start, stabilized, and hot start. Emissions generally are highest when a vehicle is in the cold-start mode: the vehicle, engine, and emission control equipment are all at ambient temperature and thus not performing at optimum levels. Emissions are generally somewhat lower in hot start mode, when the vehicle is not yet completely warmed up but has not been sitting idle for sufficient time to have cooled completely to ambient temperature. Finally, emissions generally are lowest when the vehicle is operating in stabilized mode, and has been in continuous operation long enough for all systems to have attained relatively stable, fully "warmed-up" operating temperatures.

The EPA has determined through its running loss emission test programs that the level of running loss emissions depends on several variables: the average speed of the travel, the ambient temperature, the volatility (RVP) of the fuel, and the length of the trip. "Trip length" as used in MOBILE4.1 refers to the duration of the trip (how long the vehicle has been traveling), not on the distance traveled in the trip (how far the vehicle has been driven). Test data show that for any given set of conditions (average speed, ambient temperature, and fuel volatility), running loss emissions are zero to negligible at first, but increase significantly as the duration of the trip is extended and the fuel tank, fuel lines, and engine become heated.

7.3.4.4 MICRO2

MICRO2 is an air quality model which computes the air pollution emissions near an intersection. The concentration of the pollutants in the air around the intersection is not computed. In order to determine the pollution concentration, a dispersion model which takes weather conditions such as wind, speed, and direction into account should be used (Richards 1983).

MICRO2 bases its emissions on typical values of the FTP performed for Denver, Colorado in the early 1980s. They are:

- FTP (1) HC 6.2 grams/veh/km
- FTP (2) CO 62.2 grams/veh/km

- FTP (3) NOx 1.2 grams/veh/km

For lower than Denver altitudes or years beyond 1980's, emission rates may be lower and should change from these initial values.

The emission formulas as a function of acceleration and speeds are as follow:

$$HC\ Emission\ (gram/sec) = 0.018 + 5.668*10^{-3}\ (A*S) + 2.165*10^{-4}\ (A*S^2) \quad (7.17a)$$

$$CO\ Emission\ (gram/sec) = 0.182 - 8.587*10^{-2}\ (A*S) + 1.279*10^{-2}\ (A*S^2) \quad (7.17b)$$

$$NOx\ Emission\ (gram/sec) = 3.86*10^{-3} + 8.767*10^{-3}(A*S)\ (for\ A*S>0) \quad (7.17c)$$

$$NOx\ Emission\ (gram/sec) = 1.43*10^{-3} - 1.830*10^{-4}(A*S)\ (for\ A*S<0) \quad (7.17d)$$

where,

A = acceleration (meters/sec^2) and
S = speed (meters/sec).

7.3.4.5 The TRRL Model

This model has been developed by the British Transport and Road Research Laboratory (TRRL) and it predicts air pollution from road traffic (Hickman and Waterfield 1984). The estimations of air pollution are in the form of hourly average concentrations of carbon monoxide at selected locations around a network of roads. The input data required are the configuration of the road network, the location of the receptor, traffic volumes and speeds, wind speed, and wind direction.

The concentration of carbon monoxide may be used as to approximate the likely levels of other pollutants using the following relations:

$$HC\ (ppm) = 1.8\ CO\ (ppm) * R + 4.0 \quad (7.18a)$$

$$NOx\ (ppm) = CO\ (ppm) * R + 0.1 \quad (7.18b)$$

where R is the ratio of pollutant emission rate to that of carbon monoxide for a given mean vehicle speed,

Mean Speed (km/hr)	NOx (ppm)	HC (ppm)
20	0.035	0.205
30	0.050	0.240
40	0.070	0.260
50	0.085	0.280
60	0.105	0.290
70	0.120	0.305

A quick but less accurate estimate of the annual maximum 8 hour CO concentration can also be obtained from the average peak hour CO concentration estimate, as follows:

$$C_8 = 1.85\ C_1 + 1.19 \quad (7.19)$$

where C_8 is the annual maximum 8 hour concentration, and C_1 is the average peak hour concentration.

A graphical screening test is introduced by which any properties likely to experience an air pollution problem are identified. The procedure first reduces the network to a system of long roads and roundabouts (if any). Then from a graph, the concentration of carbon monoxide for standard traffic conditions for locations at any distance from each network element may be determined. Factors are then applied to adjust for the traffic conditions at the site and the sum of the contributions from each element gives an estimate of the likely average peak hour concentration. An example of the graphical screening test results is shown in Table 7.3.

7.3.5 Other Mobile Source Air Quality Models

There are many other mobile source models which estimate the pollutant emission rates and concentrations near highway and arterial streets. Most of these models relate vehicle speeds and other variables such as vehicle year model, ambient temperature, and traffic conditions to emission rates. A common example of this type of relation could be found in Technical Advisory #T6640.10 of EPA report, "Mobile Source Emission Factor Tables for MOBILE3." Other popular models include HIWAY2 and CAL3QHC. HIWAY2 model has been developed by U.S. EPA to estimate hourly concentrations of non-reactive pollutants, like CO, downwind of roadways. It is usually used

Table 7.3
Graphical Screening Test Results for Existing Network

Distance from Centerline (m)	13	45	53	103
CO for 1000 vehs/hr at 100 km/h (ppm)	1.08	0.49	0.40	0.11
Traffic Flow (vehs/hr)	2,600	1,400	800	400
Speed (km/hr)	40	20	20	50
Speed Correction Factor	2.07	3.59	3.59	1.73
CO for Actual Traffic Conditions (ppm)	5.81	2.46	1.15	0.08
Total 1-Hour CO = 9.50(ppm) Equivalent Annual Maximum 8-Hour = 18.76(ppm)				

for analyzing at- grade highways and arterials in uniform wind conditions at level terrain as well as at depressed sections (cuts) of roadways. The model cannot be used if large obstructions such as buildings or large trees hinder the flow of air. The simple terrain requirement makes this model less accurate for urban conditions than CALINE4 type of models.

The CAL3QHC model has the ability to account for the emissions generated by vehicles traveling near roadway intersections. Because idling emissions account for a substantial portion of the total emissions at an intersection, this capability represents a significant improvement in the prediction of pollutant concentrations over previous models. This EPA model is especially designed to handle near-saturated and/or over-capacity traffic conditions and complex intersections where major roadways interact through ramps and elevated highways. The model combines the CALINE3 line source dispersion model with an algorithm that internally estimates the length of the queues formed by idling vehicles at signalized intersections. The inputs to the model includeinformation and data commonly required by transportation models such as roadway geometries, receptor locations, vehicular emissions, and meteorological conditions. Emission factors used in the model should be obtained from mobile source emission factor models such as MOBILE4.

References

Akcelik, R. (1981). *Fuel Efficiency and Other Objectives in Traffic System Management.* Traffic Engineering & Control, Vol. 22, pp. 54-65.

Akcelik, R. And C. Bayley (1983). *Some Results on Fuel Consumption Models. Appeared in Progress in Fuel Consumption Modelling for Urban Traffic Management.* Research Report ARR No. 124, Austrailian Road Research Board, pp. 51-56.

Andreassen, D. (1991). *Population and Registered Vehicle Data Vs. Road Deaths.* Accident Analysis & Prev., Vol. 23, No. 5, pp. 343-351.

Belmont, D. M. (1953). *Effect of Average Speed And Volume On Motor-vehicle Accidents On Two-lane Tangents.* Proc. Highway Research Board, 32, pp. 383-395.

Biggs, D. C. and R. Akcelik (1986). *An Energy-Related Model for Instantaneous Fuel Consumption.* Traffic Engineering & Control, Vol.27, No.6, pp. 320-325.

Bonneson, J. A., and P. T. McCoy (1993). *Estimation of Safety At Two-way Stop-controlled Intersections On Rural Highways*. Transportation Research Record 1401, pp. 83-89, Washington DC.

Brundell-Freij, K. and L. Ekman (1991). *Flow and Safety*. Presented at the 70th Annual Meeting of the Transportation Research Board, Washington, DC.

Brüde, U. and J. Larsson (1993). *Models For Predicting Accidents At Junctions Where Pedestrians And Cyclists Are Involved. How Well Do They Fit?* Accident Analysis & Prev. Vol. 25, No. 6, pp. 499-509.

Ceder, A. and M. Livneh (1978). *Further Evaluation Of The Relationships Between Road Accidents and Average Daily Traffic.* Accident Analysis & Prev., Vol 10, pp. 95-109.

Ceder, A. (1982). *Relationship Between Road Accidents And Hourly Traffic Flow-II*. Accident Analysis & Prev., Vol 14, No. 1, pp. 35-44.

Ceder, A. and M. Livneh (1982). *Relationship Between Road Accidents And Hourly Traffic Flow-I*. Accident Analysis & Prev., Vol. 14, No. 1.

Chang, M. F., L. Evans, R. Herman, and P. Wasielewski (1976). *Gasoline Consumption in Urban Traffic*. Transportation Research Record 599, Transportation Research Board, pp. 25-30.

Chang, M. F. and R. Herman (1981). *Trip Time Versus Stop Time and Fuel Consumption Characteristics in Cities.* Transportation Science, Vol. 15, No. 3, pp. 183-209.

Cleveland, D. E., L. P. Kostyniuk, and K.-L. Ting (1985). *Design and Safety On Moderate Volume Two-lane Roads*. Transportation Research Record, 1026, pp. 51-61.

DeLuchi, M., Q. Wang, and D. Sperling (1989). *Electric Vehicles: Performance, Life-Cycle Costs, Emissions, and Recharging Requirements*. Transportation Research, Vol. 23A, No. 3.

Evans, L., R. Herman, and T. Lam (1976). *Multivariate Analysis of Traffic Factors Related to Fuel Consumption in Urban Driving*. Transportation Science, Vol. 10, No. 2, pp. 205-215.

Evans, L. and R. Herman (1978). *Automobile Fuel Economy on Fixed Urban Driving Schedules*. Transportation Science, Vol. 12, No. 2, pp. 137-152.

Everall, P. F. (1968). *The Effect of Road and Traffic Conditions on Fuel Consumption*. RRL Report LR 226, Road Research Laboratory, Crowthorne, England.

Federal Highway Administration (1984). *CALINE4 - A Dispersion Model for Predicting Air Pollutant Concentrations Near Roadways*. Report FHWA/CA/TL - 84/15.

Fisk, C. S. (1989). *The Australian Road Research Board Instantaneous Model of Fuel Consumption*. Transportation Research, Vol. 23B, No. 5, pp. 373-385.

Gipps, P. G. and B. G. Wilson (1980). *MULTISIM: A Computer Package for Simulating Multi-Lane Traffic Flows*. Proceedings of the 4th Biennial Conference of Simulation Society of Australia.

Greene, D. L. and J. T. Liu (1988). *Automotive Fuel Economy Improvements and Consumers' Surplus*. Transportation Research, Vol. 22A, No. 3.

Greene, D. L. and K. G. Duleep (1993). *Costs and Benefits of Automotive Fuel Economy Improvement: A Partial Analysis*. Transportation Research, Vol. 27A, No.3.

Hauer, E. (1971). *Accident Overtaking and Speed Control*. Accident Analysis and Prevention, Vol. 3, No. 1, pp. 1-13.

Hauer, E. (1992). *Empirical Bayes Approach To The Estimation Of "Unsafety"; The Multivariate Regression Method*. Accident Analysis & Prev., Vol. 24, No. 5, pp. 457-477.

Hauer, E., D. Terry, and M. S. Griffith (1994). *The Effect Of Resurfacing On The Safety Of Rural Roads In New York State.* Paper 940541 presented at the 73rd Annual Meeting of the Transportation Research Board, Washington, DC.

Hall, J. W. And O. J. Pendleton (1990). *Rural Accident Rate Variation with Traffic Volume.* Transportation Research Record 1281, TRB, NRC, Washington, DC, pp. 62-70.

Herman, R. and S. Ardekani (1985). *The Influence of Stops on Vehicle Fuel Consumption in Urban Traffic*. Transportation Science, Vol. 19, No. 1, pp. 1-12.

Hickman, A. J. and V. H. Waterfield (1984). *A User's Guide to the Computer Programs for Predicting Air Pollution from Road Traffic*. TRRL Supplementary Report 806.

Hickman, A. J. and M. H. L. Waters (1991). *Improving Automobile Fuel Economy*. Traffic Engineering & Control, Vol. 23, No. 11.

Hooker, J. N., A. B. Rose, and G. F. Roberts (1983). *Optimal Control of Automobiles for Fuel Economy*. Transportation Science, Vol. 17, No. 2, pp. 146-167.

Institute of Transportation Engineers (1989). *A Toolbox for Alleviating Traffic Congestion*. Publication No. IR.054A.

Jovanis, P. P. and H-L. Chang (1987). *Modelling The Relationship Of Accidents To Miles Travelled*. Transportation Research Record 1068, Washington, DC, pp. 42-51.

Kenworthy, J. R., H. Rainford, P. W. G. Newman, and T. J. Lyons (1986). *Fuel Consumption, Time Saving and Freeway Speed Limits*. Traffic Engineering & Control, Vol. 27, No. 9, pp. 455-459.

Kihlberg, J. K. and K. J. Tharp (1968). *Accident Rates As Related To Design Elements Of Rural Highways.* National Cooperative Highway Research Program Report, 47, Highway Research Board.

Komor, P., S. F. Baldwin, and J. Dunkereley (1993). *Technologies for Improving Transportation Energy Efficiency in the Developing World*. Transportation Research, Vol 27A, No. 5.

Lam, T. N. (July 1985). *Estimating Fuel Consumption From Engine Size*. Journal of Transportation Engineering, Vol. 111, No. 4, pp. 339-357.

Leutzbach, W., W. Siegener, and R. Wiedemann (1970). *Über den Zusammenhang Zwischen Verkehrsunfällen Und Verkehrsbelastung Auf Eienem deutschen Autobahnabschnitt* (About the Relationship between Traffic Accidents and Traffic Flow on a Section of a German Freeway). Accident Analysis & Prev. Vol. 2, pp. 93-102.

Lieberman, E., R. D. Worrall, D. Wicks, and J. Woo (1979). *NETSIM Model (5 Vols.)*. Federal Highway Administration, Report No. FHWA-RD-77-41 to 77-45, Washington, DC.

Lundy, R. A. (1965). *Effect of Traffic Volumes And Number Of Lanes On Freeway Accident Rates.* Highway Research Record 99, HRB, National Research Council, Washington, DC, pp. 138-147.

Mahalel, D. (1986). *A Note on Accident Risk.* Transportation Research Record, 1068, National Research Council, Washington, DC, pp. 85-89.

McDonald, J. W. (1953). *Relationship between Number of Accidents and Traffic Volume at Divided-Highway Intersections.* Bulletin No. 74, pp. 7-17.

Miaou, S-P. and H. Lum (1993). *Modeling Vehicle Accidents And Highway Geometric Design Relationships*. Accident Analysis & Prev. Vol. 25, No. 6, pp. 689-709.

Ng, J. C. N. and E. Hauer (1989). *Accidents on Rural Two-Lane Roads: Differences Between Seven States*. Transportation Research Record, 1238, pp. 1-9.

Ontario Ministry of Transportation and Communications (1982). *A Technical Background Document For Automotive Fuel Economy*. The Transportation Energy Management Program (TEMP) Report No. DRS-82-01, Ontario, Canada.

Pelensky, E., W. R. Blunden, and R. D. Munro (1968). *Operating Costs of Cars in Urban Areas*. Proceedings, Fourth Conference of the Australian Road Research Board, Vol. 4, part 1, pp. 475-504.

Persaud B. and Dzbik, L. (1993). *Accident Prediction Models For Freeways.* Transportation Research Record 1401, TRB, NRC, Washington, DC, pp. 55-60.

Pfundt, K. (1968). *Comparative Studies Of Accidents On Rural Roads.* Strassenbau und Strassenverkehrstechnik, Vol. 82, Federal Ministry of Transport, Bonn.

Pfundt, K. (1969). *Three Difficulties In The Comparison Of Accident Rates*. Accident Analysis & Prevention, Vol. 1, pp. 253-259.

Pitt, D. R., T. J. Lyons, P. W. G. Newman, and J. R. Kenworthy (1987). *Fuel Consumption Models: An Evaluation Based on a Study of Perth's Traffic Patterns*. Traffic Engineering & Control, Vol. 27, No. 2, pp. 60-68.

Post, K., J. Tomlin, D. Pitt, N. Carruthers, A. Maunder, J. H. Kent, and R. W. Bilger (1981). *Fuel Economy and Emissions Research Annual Report*. Charles Kolling Research Laboratory Technical Note: ER36, University of Sydney.

Quaye, K., L. Leden, and E. Hauer (1993). *Pedestrian Accidents And Left-turning Traffic At Signalized Intersections.* AAA foundation for Traffic Safety, Washington, DC.

Richards, G. G. (1983). *MICRO2 - An Air Quality Intersection Model*. Colorado Department of Highways.

Richardson, A. J. and R. Akcelik (1983). *Fuel Consumption and Data Needs for the Design and Evaluation of Urban Traffic Systems.* Research Report ARR No. 124, Austrailian Road Research Board, pp. 51-56.

Roosmark, P.O. (1966). *Trafikolyckor i trevagskorsningar.* Statens Vaginstitut, Preliminary Report 22, Stockholm.

Satterthwaite, S. P. (1981). *A Survey into Relationships between Traffic Accidents and Traffic Volumes.* Transport and Road Research Laboratory Supplementary Report, SR 692, pp. 41.

Slatterly, G. T. and D. E. Cleveland (1969). *Traffic Volume. Traffic Control And Roadway Elements.* Chapter 2, Automotive Safety Foundation, pp. 8.

Tanner, J. C. (1953). *Accidents At Rural Three-way Junctions.* J. Instn. of Highway Engrs. II(11), pp. 56-67.

Thorson, O. (1967). *Traffic Accidents and Road Layout.* Copenhagen Technical University of Denmark.

Transportation Research Board (1981). *Methodology for Evaluating Highway Air Pollution Dispersion Models*. NCHRP Report No. 245, National Research Council.

U.S. DOT (1990). *A Statement of National Transportation Policy, Strategies for Action, Moving America - New Directions, New Opportunities.*

Urban Mass Transportation Administration (1985). *Characteristics of Urban Transportation System*. U.S. Department of Transportation.

U. S. EPA (1990). *National Air Quality and Emissions Trends Report*. Office of Air Quality.

U. S. EPA (1991). *MOBILE4.1 User's Guide - Mobile Source Emission Factor Model*. Report EPA-AA-TEB-91-01.

Vincent, R. A., A. I. Mitchell, and D. I. Robertson (1980). *User Guide to TRANSYT Version 8*. Transport and Road Research Lab Report No. LR888.

Wallace, C. E., K. G. Courage, D. P. Reaves, G. W. Schoene, and G. W. Euler (1984). *TRANSY-7F User's Manual*. U.S. Dept. of Transportation, Office of Traffic Operations.

Walsh, M. P. (1989). *Statement Before the U.S. Senate Environment and Public Works Committee.*

Watson, H. C., E. E. Milkins, and G. A. Marshal (1980). *A Simplified Method for Quantifying Fuel Consumption of Vehicles in Urban Traffic*. SAE-Australia, Vol. 40, No. 1, pp. 6-13.

Webb, G. M. (1955). *The Relationship Between Accidents and Traffic Volumes at Signalized Intersections*. Institute of Traffic Engineers. Proceedings.

Welbourne, E. R. (1979). *Exposure Effects on the Rate and Severity of Highway Collisions at Different Travel Speeds.* Technical Memorandum TMVS 7903, Vehicle Systems, Transport Canada.

Zegeer, C. V., J. Hummer, D. Reinfurt, L. Herf, and W. Hunter (1986). *Safety Effects of Cross-Section Design for Two-Lane Roads.* Volume I, Final Report, FHWA-RD-87/008, Federal Highway Administration, Transportation Research Board.

UNSIGNALIZED INTERSECTION THEORY

BY ROD J. TROUTBECK[13]
WERNER BRILON[14]

[13] Professor, Head of the School, Civil Engineering, Queensland University of Technology, 2 George Street, Brisbane 4000 Australia.

[14] Professor, Institute for Transportation, Faculty of Civil Engineering, Ruhr University, D 44780 Bochum, Germany.

Chapter 8 - Frequently used Symbols

b_i = proportion of volume of movement *i* of the total volume on the shared lane
C_w = coefficient of variation of service times
D = total delay of minor street vehicles
D_q = average delay of vehicles in the queue at higher positions than the first
$E(h)$ = mean headway
$E(t_c)$ = the mean of the critical gap, t_c
$f(t)$ = density function for the distribution of gaps in the major stream
$g(t)$ = number of minor stream vehicles which can enter into a major stream gap of size, t
L = logarithm
m = number of movements on the shared lane
n = number of vehicles
η_c = increment, which tends to 0, when $Var(t_c)$ approaches 0
η_f = increment, which tends to 0, when $Var(t_f)$ approaches 0
q = flow in veh/sec
q_s = capacity of the shared lane in veh/h
$q_{m,i}$ = capacity of movement *i,* if it operates on a separate lane in veh/h
q_m = the entry capacity
q_m = maximum traffic volume departing from the stop line in the minor stream in veh/sec
q_p = major stream volume in veh/sec
t = time
t_c = critical gap time
t_f = follow-up times
t_m = the shift in the curve
$Var(t_c)$ = variance of critical gaps
$Var(t_f)$ = variance of follow-up-times
$Var\ (W)$ = variance of service times
W = average service time. It is the average time a minor street vehicle spends in the first position of the queue near the intersection
W_1 = service time for vehicles entering the empty system, i.e no vehicle is queuing on the vehicle's arrival
W_2 = service time for vehicles joining the queue when other vehicles are already queuing

8.
UNSIGNALIZED INTERSECTION THEORY

8.1 Introduction

Unsignalized intersections are the most common intersection type. Although their capacities may be lower than other intersection types, they do play an important part in the control of traffic in a network. A poorly operating unsignalized intersection may affect a signalized network or the operation of an Intelligent Transportation System.

The theory of the operation of unsignalized intersections is fundamental to many elements of the theory used for other intersections. For instance, queuing theory in traffic engineering used to analyze unsignalized intersections is also used to analyze other intersection types.

8.1.1 The Attributes of a Gap Acceptance Analysis Procedure

Unsignalized intersections give no positive indication or control to the driver. He or she is not told when to leave the intersection. The driver alone must decide when it is safe to enter the intersection. The driver looks for a safe opportunity or "gap" in the traffic to enter the intersection. This technique has been described as gap acceptance. Gaps are measured in time and are equal to headways. At unsignalized intersections a driver must also respect the priority of other drivers. There may be other vehicles that will have priority over the driver trying to enter the traffic stream and the driver must yield to these drivers.

All analysis procedures have relied on gap acceptance theory to some extent or they have understood that the theory is the basis for the operation even if they have not used the theory explicitly.

Although gap acceptance is generally well understood, it is useful to consider the gap acceptance process as one that has two basic elements.

- First is the extent drivers find the gaps or opportunities of a particular size useful when attempting to enter the intersection.
- Second is the manner in which gaps of a particular size are made available to the driver. Consequently, the proportion of gaps of a particular size that are offered to the entering driver and the pattern of the inter-arrival times are important.

This chapter describes both of these aspects when there are two streams. The theory is then extended to intersections with more than two streams.

8.1.2 Interaction of Streams at Unsignalized Intersections

A third requirement at unsignalized intersections is that the interaction between streams be recognized and respected. At all unsignalized intersections there is a hierarchy of streams. Some streams have absolute priority, while others have to yield to higher order streams. In some cases, streams have to yield to some streams which in turn have to yield to others. It is useful to consider the streams as having different levels of priority or ranking. For instance:

Rank 1 stream - has absolute priority and does not need to yield right of way to another stream,

Rank 2 stream - has to yield to a rank 1 stream,

Rank 3 stream - has to yield to a rank 2 stream and in turn to a rank 1 stream, and

Rank 4 stream - has to yield to a rank 3 stream and in turn to a rank 2 stream and to a rank 1 stream.

8.1.3 Chapter Outline

Sections 8.2 discusses gap acceptance theory and this leads to Section 8.3 which discusses some of the common headway distributions used in the theory of unsignalized intersections.

Most unsignalized intersections have more than two interacting streams. Roundabouts and some merges are the only examples of two interacting streams. Nevertheless, an understanding of the operation of two streams provides a basis to extend the knowledge to intersections with more than two streams. Section

8.4 discusses the performance of intersections with two interacting streams.

Section 8.5 to 8.8 discuss the operation of more complex intersections. Section 8.9 covers other theoretical treatments of unsignalized intersections. In many cases, empirical approaches have been used. For instance the relationships for AWSC (All Way Stop Controlled) intersections are empirical. The time between successive departures of vehicles on the subject roadway are related to the traffic conditions on the other roadway elements.

The theory described in this chapter is influenced by the human factors and characteristics as described in Chapter 3, and in particular, Sections 3.13 and 3.15. The reader will also note similarities between the material in this chapter and Chapter 9 dealing with signalized intersections. Finally, unsignalized intersections can quickly become very complicated and often the subject of simulation programs. The comments in Chapter 10 are particularly relevant here.

8.2 Gap Acceptance Theory

8.2.1 Usefulness of Gaps

The gap acceptance theory commonly used in the analysis of unsignalized intersections is based on the concept of defining the extent drivers will be able to utilize a gap of particular size or duration. For instance, will drivers be able to leave the stop line at a minor road if the time between successive vehicles from the left is 10 seconds; and, perhaps how many drivers will be able to depart in this 10 second interval ?

The minimum gap that all drivers in the minor stream are assumed to accept at all similar locations is the critical gap. According to the driver behavior model usually assumed, no driver will enter the intersection unless the gap between vehicles in a higher priority stream (with a lower rank number) is at least equal to the critical gap, t_c. For example, if the critical gap was 4 seconds, a driver would require a 4 second gap between Rank 1 stream vehicles before departing. He or she will require the same 4 seconds at all other times he or she approaches the same intersection and so will all other drivers at that intersection.

Within gap acceptance theory, it is further assumed that a number of drivers will be able to enter the intersection from a minor road in very long gaps. Usually, the minor stream vehicles (those yielding right of way) enter in the long gaps at headways often referred to as the "follow-up time", t_f.

Note that other researchers have used a different concept for the critical gap and the follow-up time. McDonald and Armitage (1978) and Siegloch (1973) independently described a concept where a lost time is subtracted from each major stream gap and the remaining time is considered 'useable.' This 'useable' time divided by the saturation flow gives an estimate of the absorption capacity of the minor stream. As shown below, the effect of this different concept is negligible.

In the theory used in most guides for unsignalized intersections around the world, it is assumed that drivers are both consistent and homogeneous. A consistent driver is expected to behave the same way every time at all similar situations. He or she is not expected to reject a gap and then subsequently accept a smaller gap. For a homogeneous population, all drivers are expected to behave in exactly the same way. It is, of course, unreasonable to expect drivers to be consistent and homogeneous.

The assumptions of drivers being both consistent and homogeneous for either approach are clearly not realistic. Catchpole and Plank (1986), Plank and Catchpole (1984), Troutbeck (1988), and Wegmann (1991) have indicated that if drivers were heterogeneous, then the entry capacity would be decreased. However, if drivers are inconsistent then the capacity would be increased. If drivers are assumed to be both consistent and homogeneous, rather than more realistically inconsistent and heterogeneous, then the difference in the predictions is only a few percent. That is, the overall effect of assuming that drivers are consistent and homogeneous is minimal and, for simplicity, consistent and homogeneous driver behavior is assumed.

It has been found that the gap acceptance parameters t_c and t_f may be affected by the speed of the major stream traffic (Harders 1976 and Troutbeck 1988). It also expected that drivers are influenced by the difficulty of the maneuver. The more difficult

a maneuver is, the longer are the critical gap and follow-up time parameters. There has also been a suggestion that drivers require a different critical gap when crossing different streams within the one maneuver. For instance a turn movement across a number of different streams may require a driver having a different critical gap or time period between vehicles in each stream (Fisk 1989). This is seen as a unnecessary complication given the other variables to be considered.

8.2.2 Estimation of the Critical Gap Parameters

The two critical gap parameters that need to be estimated are the critical gap t_c and the follow-up time t_f. The techniques used to estimate these parameters fit into essentially two different groups. The first group of techniques are based on a regression analysis of the number of drivers that accept a gap against the gap size. The other group of techniques estimates the distribution of follow-up times and the critical gap distribution independently. Each group is discussed below.

Regression techniques.
If there is a continuous queue on the minor street, then the technique proposed by Siegloch (1973) produces acceptable results because the output matches the assumptions used in a critical gap analysis. For this technique, the queue must have at least one vehicle in it over the observation period. The process is then:

- Record the size of each gap, *t*, and the number of vehicles, *n*, that enter during this gap;
- For each of the gaps that were accepted by only *n* drivers, calculate the average gap size, *E(t)* (See Figure 8.1);
- Use linear regression on the average gap size values (as the dependent variable) against the number of vehicles that enter during this average gap size, *n*; and

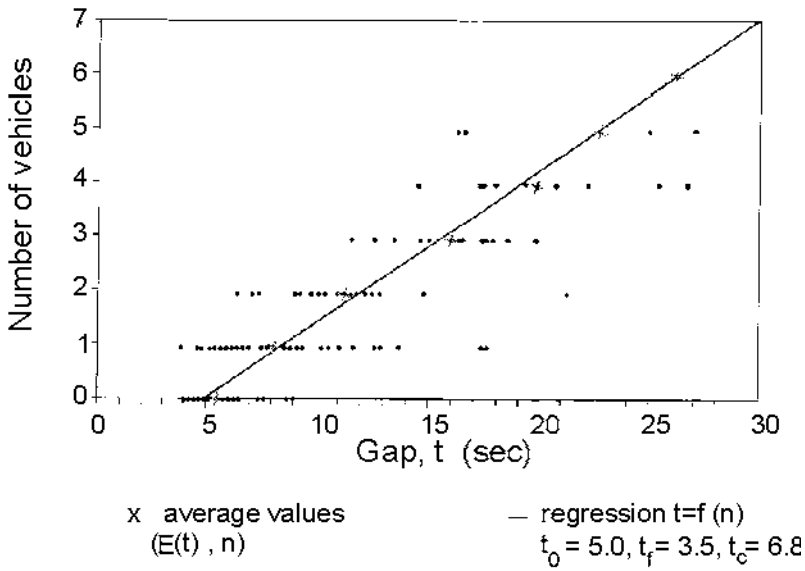


Figure 8.1
Data Used to Evaluate Critical Gaps and Move-Up Times
(Brilon and Grossmann 1991).

- Given the slope is t_f and the intercept of the gap size axis is t_o, then the critical gap t_c is given by

$$t_c = t_o + t_f/2 \qquad (8.1)$$

The regression line is very similar to the stepped line as shown in Figure 8.2. The stepped line reflects the assumptions made by Tanner (1962), Harders (1976), Troutbeck (1986), and others. The sloped line reflects the assumptions made by Siegloch (1973), and McDonald and Armitage (1978).

Independent assessment of the critical gap and follow-up time
If the minor stream does not continuously queue, then the regression approach cannot be used. A probabilistic approach must be used instead.

The follow-up time is the mean headway between queued vehicles which move through the intersection during the longer gaps in the major stream. Consider the example of two major stream vehicles passing by an unsignalized intersection at times 2.0 and 42.0 seconds. If there is a queue of say 20 vehicles wishing to make a right turn from the side street, and if 17 of these minor street vehicles depart at 3.99, 6.22, 8.29, 11.13, 13.14, and so on, then the headways between the minor street vehicles are 6.22-3.99, 8.29-6.22, 11.13-8.29 and so on. The average headway between this group of minor stream vehicles is 2.33 sec. This process is repeated for a number of larger major stream gaps and an overall average headway between the queued minor stream vehicles is estimated. This average headway is the follow-up time, t_f . If a minor stream vehicle was not in a queue then the preceding headway would not be included. This quantity is similar to the saturation headway at signalized intersections.

The estimation of the critical gap is more difficult. There have been numerous techniques proposed (Miller 1972; Ramsey and Routledge 1973; Troutbeck 1975; Hewitt 1983; Hewitt 1985). The difficulty with the estimation of the critical gap is that it cannot be directly measured. All that is known is that a driver's individual critical gap is greater than the largest gap rejected and shorter than the accepted gap for that driver. If the accepted gap was shorter than the largest rejected gap then the driver is considered to be inattentive. This data is changed to a value just below the accepted gap. Miller (1972) gives an alternative method of handling this inconsistent data which uses the data as recorded. The difference in outcomes is generally marginal.

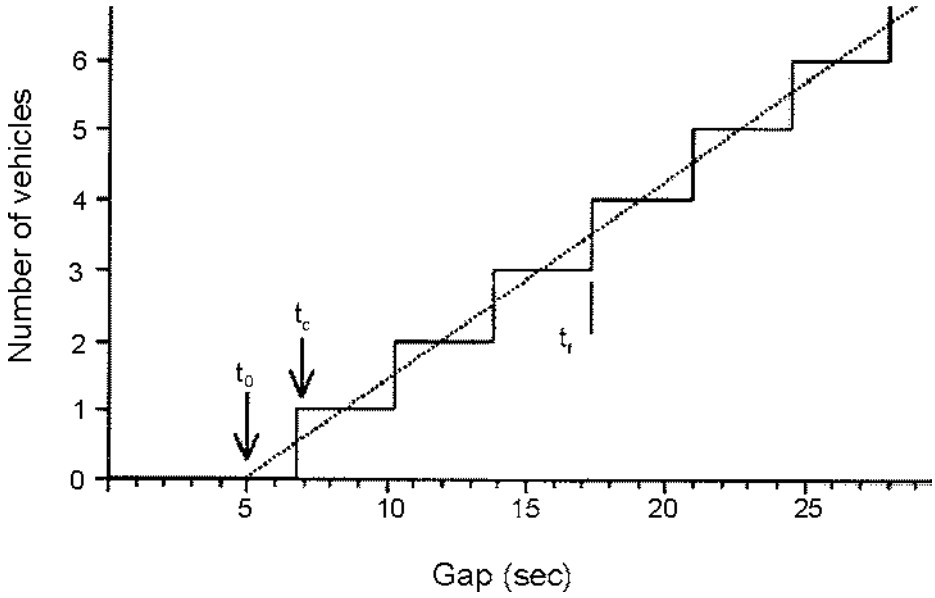


Figure 8.2
Regression Line Types.

Miller (1972), and later Troutbeck (1975) in a more limited study, used a simulation technique to evaluate a total of ten different methods to estimate the critical gap distribution of drivers. In this study the critical gaps for 100 drivers were defined from a known distribution. The arrival times of priority traffic were simulated and the appropriate actions of the "simulated" drivers were noted. This process was repeated for 100 different sets of priority road headways, but with the same set of 100 drivers. The information recorded included the size of any rejected gaps and the size of the accepted gap and would be similar to the information able to be collected by an engineer at the road side. The gap information was then analyzed using each of the ten different methods to give an estimate of the average of the mean of the drivers' critical gaps, the variance of the mean of the drivers' critical gaps, mean of the standard deviation of the drivers' critical gaps and the variance of the standard deviation of the drivers' critical gaps. These statistics enabled the possible bias in predicting the mean and standard deviation of the critical gaps to be estimated. Techniques which gave large variances of the estimates of the mean and the standard deviation of the critical gaps were considered to be less reliable and these techniques were identified. This procedure found that one of the better methods is the Maximum Likelihood Method and the simple Ashworth (1968) correction to the prohibit analysis being a strong alternative. Both methods are documented here. The Probit or Logit techniques are also acceptable, particularly for estimating the probability that a gap will be accepted (Abou-Henaidy et al. 1994), but more care needs to be taken to properly account for flows. Kyte et al (1996) has extended the analysis and has found that the Maximum Likelihood Method and the Hewitt (1983) models gave the best performance for a wide range of minor stream and major stream flows.

The maximum likelihood method of estimating the critical gap requires that the user assumes a probabilistic distribution of the critical gap values for the population of drivers. A log-normal is a convenient distribution. It is skewed to the right and does not have non-negative values. Using the notation:

- a_i = the logarithm of the gap accepted by the *i*th driver,
- a_i = ∞ if no gap was accepted,
- r_i = the logarithm of the largest gap rejected by the ith driver,
- r_i = 0 if no gap was rejected,
- μ and σ^2 are the mean and variance of the logarithm of the individual drivers critical gaps (assuming a log-normal distribution), and
- $f()$ and $F()$ are the probability density function and the cumulative distribution function respectively for the normal distribution.

The probability that an individual driver's critical gap will be between r_i and a_i is $F(a_i) - F(r_i)$. Summing over all drivers, the likelihood of a sample of n drivers having accepted and largest rejected gaps of (a_i, r_i) is

$$\prod_{i=1}^{n} [F(a_i) - F(r_i)] \tag{8.2}$$

The logarithm, L, of this likelihood is then

$$L = \sum_{i=1}^{n} \ln[F(a_i) - F(r_i)] \tag{8.3}$$

The maximum likelihood estimators, μ and σ^2, that maximize L, are given by the solution to the following equations.

$$\frac{\partial L}{\partial \mu} = 0 \tag{8.4}$$

and

$$\frac{\partial L}{\partial \sigma^2} = 0 \tag{8.5}$$

Using a little algebra,

$$\frac{\partial F(x)}{\partial \mu} = -f(x) \tag{8.6}$$

$$\frac{\partial F(x)}{\partial \sigma^2} = -\frac{x-\mu}{2\sigma^2} f(x) \tag{8.7}$$

This then leads to the following two equations which must be solved iteratively. It is recommended that the equation

$$\sum_{i=1}^{n} \frac{f(r_i) - f(a_i)}{F(a_i) - F(r_i)} = 0 \qquad (8.8)$$

should be used to estimate μ given a value of σ^2. An initial value of σ^2 is the variance of all the a_i and r_i values. Using this estimate of μ from Equation 8.8, a better estimate of σ^2 can be obtained from the equation,

$$\sum_{i=1}^{n} \frac{(r_i - \hat{\mu}) f(r_i) - (a_i - \hat{\mu}) f(a_i)}{F(a_i) - F(r_i)} = 0 \qquad (8.9)$$

where $\hat{\mu}$ is an estimate of μ.

A better estimate of the μ can then be obtained from the Equation 8.8 and the process continued until successive estimates of μ and σ^2 do not change appreciably.

The mean, $E(t_c)$, and the variance, $Var(t_c)$, of the critical gap distribution is a function of the log normal distribution parameters, viz:

$$E(t_c) = e^{\mu + 0.5\sigma^2} \qquad (8.10)$$

and

$$Var(t_c) = E(t_c)^2 \ (e^{\sigma^2} - 1) \qquad (8.11)$$

The critical gap used in the gap acceptance calculations is then equal to $E(t_c)$. The value should be less than the mean of the accepted gaps.

This technique is a complicated one, but it does produce acceptable results. It uses the maximum amount of information, without biasing the result, by including the effects of a large number of rejected gaps. It also accounts for the effects due to the major stream headway distribution. If traffic flows were light, then many drivers would accept longer gaps without rejecting gaps. On the other hand, if the flow were heavy, all minor stream drivers would accept shorter gaps. The distribution of accepted gaps is then dependent on the major stream flow. The maximum likelihood technique can account for these different conditions. Unfortunately, if all drivers accept the first gap offered without rejecting any gaps, then Equations 8.8 and 8.9 give trivial results. The user should then look at alternative methods or preferably collect more data.

Another very useful technique for estimating the critical gap is Ashworth's (1968) procedure. This requires that the user identify the characteristics of the probability distribution that relates the proportion of gaps of a particular size that were accepted to the gap size. This is usually done using a Probit analysis applied to the recorded proportions of accepted gaps. A plot of the proportions against the gap size on probability paper would also be acceptable. Again a log normal distribution may be used and this would require the proportions to be plotted against the natural logarithm of the gap size. If the mean and variance of this distribution are $E(t_a)$ and $Var_a(t)$, then Ashworth's technique gives the critical gap as

$$E(t_c) = E(t_a) - q_p \ Var(t_a) \qquad (8.12)$$

where q_p is the major stream flow in units of veh/sec. If the log normal function is used, then $E(t_a)$ and $Var(t_a)$ are values given by the generic Equations 8.10 and 8.11. This is a very practical solution and one which can be used to give acceptable results in the office or the field.

8.2.3 Distribution of Gap Sizes

The distribution of gaps between the vehicles in the different streams has a major effect on the performance of the unsignalized intersection. However, it is important only to look at the distribution of the larger gaps; those that are likely to be accepted. As the shorter gaps are expected to be rejected, there is little point in modeling these gaps in great detail.

A common model uses a random vehicle arrival pattern, that is, the inter-arrival times follow an exponential distribution. This distribution will predict a large number of headways less than 1 sec. This is known to be unrealistic, but it is used because these small gaps will all be rejected.

This exponential distribution is known to be deficient at high flows and a displaced exponential distribution is often recommended. This model assumes that vehicle headways are at least t_m sec.

Better models use a dichotomized distribution. These models assume that there is a proportion of vehicles that are free of interactions and travel at headways greater than t_m. These vehicles are termed "free" and the proportion of free vehicles is α. There is a probability function for the headways of free vehicles. The remaining vehicles travel in platoons and again there is a headway distribution for these bunched vehicles. One such dichotomized headway model is Cowan's (1975) M3 model which assumes that a proportion, α, of all vehicles are free and have an displaced exponential headway distribution and the 1-α bunched vehicles have the same headway of only t_m.

In this chapter, the word "queues" is used to refer to a line of stopped vehicles. On the other hand, a platoon is a group of traveling vehicles which are separated by a short headway of t_m. When describing the length of a platoon, it is usual to include a platoon leader which will have a longer headway in front of him or her. A platoon of length one is a single vehicle travelling without any vehicles close-by. It is often useful to distinguish between free vehicles (or platoon leaders) and those vehicles in the platoon but behind the leader. This latter group are called bunched vehicles. The benefits of a number of different headway models will be discussed later.

8.3 Headway Distributions Used in Gap Acceptance Calculations

8.3.1 Exponential Headways

The most common distribution is the negative exponential distribution which is sometimes referred to as simply the "exponential distribution". This distribution is based on the assumption that vehicles arrive at random without any dependence on the time the previous vehicle arrived. The distribution can be derived from assuming that the probability of a vehicle arriving in a small time interval $(t, t+\delta t)$ is a constant. It can also be derived from the Poisson distribution which gives the probability of n vehicles arriving in time t, that is:

$$P(n) = (qt)^n \frac{e^{-qt}}{n!} \qquad (8.13)$$

where q is the flow in veh/sec. For n = 0 this equation gives the probability that no vehicle arrives in time t. The headway, h, must be then greater than t and the probability, from Equation 8.13 is

$$P(h>t) = e^{-qt} \qquad (8.14)$$

The cumulative probability function of headways is then

$$P(h \le t) = 1 - e^{-qt} \qquad (8.15)$$

The probability distribution function is then

$$f(t) = \frac{d[P(h \le t)]}{dt} = q\ e^{-qt} \qquad (8.16)$$

This is the equation for the negative exponential distribution. The parameter q can be estimated from the flow or the reciprocal of the average headway. As an example, if there were 228 headways observed in half an hour, then the flow is 228/1800 i.e. q = 0.127 veh/sec. The proportion of headways expected to be greater than 5 seconds is then

$$\begin{aligned} P(h>5) &= e^{-q\,t} \\ &= e^{-5*0.127} \\ &= 0.531 \end{aligned}$$

The expected number of headways greater than 5 seconds observed in half an hour is then 0.531•228 or 116.

If the flow was 1440 veh/h or 0.4 veh/sec then the number of headways less than 0.1 seconds is then $q * [P(h>0.1)] * 3600$ or 56 per hour. This over-estimation of the number of very short headways is considered to be unrealistic and the displaced exponential distribution is often used instead of the negative exponential distribution.

8.3.2 Displaced Exponential Distribution

The shifted or displaced exponential distribution assumes that there is a minimum headway between vehicles, t_m. This time can be considered to be the space around a vehicle that no other

vehicle can intrude divided by the traffic speed. If the flow is q veh/h then in one hour q vehicles will pass and there are $t_m \bullet q$ seconds lost while these vehicles pass. The remaining time must then be distributed randomly after each vehicle and the average random component is $(1-t_m \bullet q)/q$ seconds. The cumulative probability distribution of headways is then:

$$F(h) = 1 - e^{-\lambda(h-t_m)} \qquad (8.17)$$

where,

$$\lambda = \frac{q}{1-t_m q} \qquad (8.18)$$

There, the terms, λ and t_m need to be evaluated. These can be estimated from the mean and the variance of the distribution. The mean headway, *E(h)*, is given by:

$$E(h) = 1/q = t_m + \frac{1}{\lambda} \qquad (8.19)$$

The variance of headways is $1/\lambda^2$. These two relationships can then be used to estimate λ and t_m.

This distribution is conceptually better than the negative exponential distribution but it does not account for the platooning that can occur in a stream with higher flows. A dichotomized headway distribution provides a better fit.

8.3.3 Dichotomized Headway Distributions

In most traffic streams there are two types of vehicles, the first are bunched vehicles; these are closely following preceding vehicles. The second group are free vehicles that are travelling without interacting with the vehicles ahead. There have been a number of dichotomized headway distributions developed over time. For instance, Schuhl (1955) proposed a distribution

$$p(h \le t) = 1 - \alpha e^{-t/\bar{h}_f} + (1-\alpha) e^{-(t-t_m)/(\bar{h}_b - t_m)} \qquad (8.20)$$

where there are α vehicles that are free (not in platoons);

there are $(1-\alpha)$ bunched vehicles;

$\bar{h}_f$ is the average headway for free vehicles;

$\bar{h}_b$ is the average headway for bunched or constrained vehicles;

t_m is the shift in the curve.

Other composite headway models have been proposed by Buckley (1962; 1968). However, a better headway model for gap acceptance is the M3 model proposed by Cowan (1975). This model does not attempt to model the headways between the bunched vehicles as these are usually not accepted but rather models the larger gaps. This headway model has a cumulative probability distribution:

$$p(h \le t) = 1 - \alpha e^{-\lambda(t-t_m)} \qquad for\ t > t_m \quad (8.21)$$

and

$$p(h \le t) = 0 \qquad \text{otherwise.}$$

Where λ is a decay constant given by the equation

$$\lambda = \frac{\alpha q}{(1-t_m q)} \qquad (8.22)$$

Cowan's headway model is rather general. To obtain the displaced exponential distribution set α to 1.0. For the negative exponential distribution, set α to 1.0 and t_m to 0. Cowan's model can also give the headway distribution used by Tanner (1962) by setting α to $1-t_m q$, however the distribution of the number of vehicles in platoons is not the same. This is documented below.

Brilon (1988) indicated that the proportion of free vehicles could be estimated using the equation,

$$\alpha = e^{-Aq_p} \qquad (8.23)$$

where A values ranged from 6 to 9. Sullivan and Troutbeck (1993) found that this equation gave a good fit to data, from more than 600 of hours of data giving in excess of 400,000 vehicle headways, on arterial roads in Australia. They also found that the A values were different for different lanes and for different lane widths. These values are listed in Table 8.1.

Table 8.1
"A" Values for Equation 8.23 from Sullivan and Troutbeck (1993).

	Median Lane	All other lanes
Lane width < 3.0 meters	7.5	6.5
3.0 ≤ Lane width ≤ 3.5 meters	7.5	5.25
Lane width> 3.5 meters	7.5	3.7

Typical values of the proportion of free vehicles are given in Figure 8.3.

The hyper-Erlang distribution is also a dichotomized headway distribution that provides an excellent fit to headway data. It is useful in simulation programs but has not been used in traffic theory when predicting capacity or delays. The hyper-Erlang distribution given by Dawson (1969) is:

$$p(h \le t) = 1 - \alpha e^{-(t - t_{mf}/\bar{h}_f - t_{mf})} + (1-\alpha) e^{-(t - t_{mb}/\bar{h}_b - t_{mb})} \sum_{x=0}^{k-1} \frac{k\left[\frac{t - t_{mb}}{\bar{h}_b - t_{mb}}\right]^x}{x!} \quad (8.24)$$

8.3.4 Fitting the Different Headway Models to Data

If the mean headway is 21.5 seconds and standard deviation is 19.55 seconds, then the flow is 1/21.5 or 0.0465 veh/seconds (167 veh/hour). A negative exponential curve that would fit this data is then,

$$p(h \le t) = 1 - e^{-0.0465t}$$

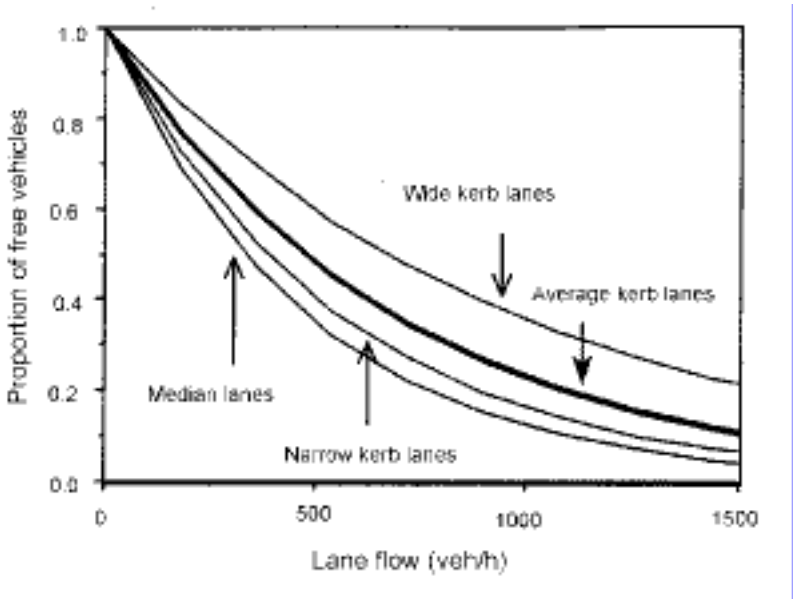


Figure 8.3
Typical Values for the Proportion of Free Vehicles.

To estimate the parameters for the displaced exponential distribution, the difference between the mean and the standard deviation is the displacement, that is t_m is equal to 21.49 – 19.55 or 1.94 seconds. The constant λ used in Equation 8.21 is the reciprocal of the standard deviation. In this case, λ is equal to 1/19.55 or 0.0512 veh/sec. The appropriate equation is then:

$$p(h \le t) = 1 - e^{-0.0512(t-1.94)}$$

The data and these equations are shown in Figure 8.4 which indicates the form of these distributions. The reader should not make any conclusions about the suitability of a distribution from this figure but should rather test the appropriateness of the model to the data collected.

In many cases there are a substantial number of very short headways and a dichotomized headway distribution performs better. As only the larger gaps are likely to be accepted by drivers, there is no point in modeling the shorter gaps in great detail. An example of Cowan's M3 model and headway data from an arterial road is shown in Figure 8.5. Figure 8.6 gives the same data and the hyper-Erlang distribution.

$$P(n) = (1-\alpha)^{n-1}\alpha \qquad (8.25)$$

Under these conditions the mean platoon size is

$$\bar{n} = \frac{1}{\alpha} \qquad (8.26)$$

and the variance by

$$Var(n) = \frac{1-\alpha}{\alpha^2} \qquad (8.27)$$

Another distribution of platoons used in the analysis of unsignalized intersections is the Borel-Tanner distribution. This platooning distribution comes from Tanner's (1962) assumptions where the major stream gaps are the outcome of a queuing process with random arrivals and a minimum inter-departure time of t_m. Although the distribution of these 'revised' major stream gaps is given by Equation 8.21 with α equal to $1-t_m q$,

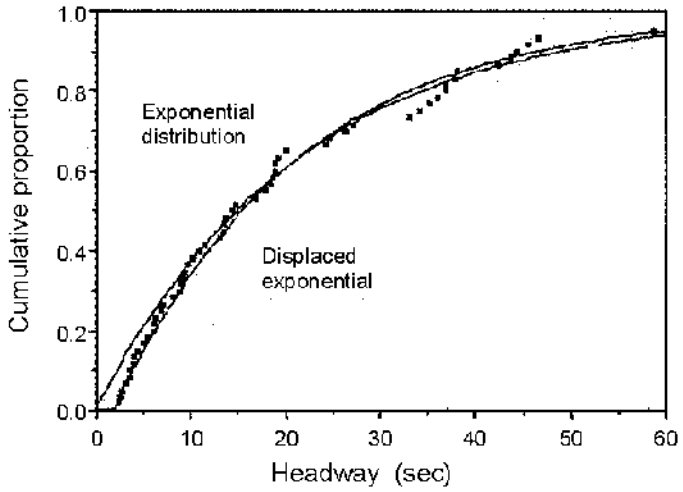


Figure 8.4
Exponential and Displaced Exponential Curves
(Low flows example).

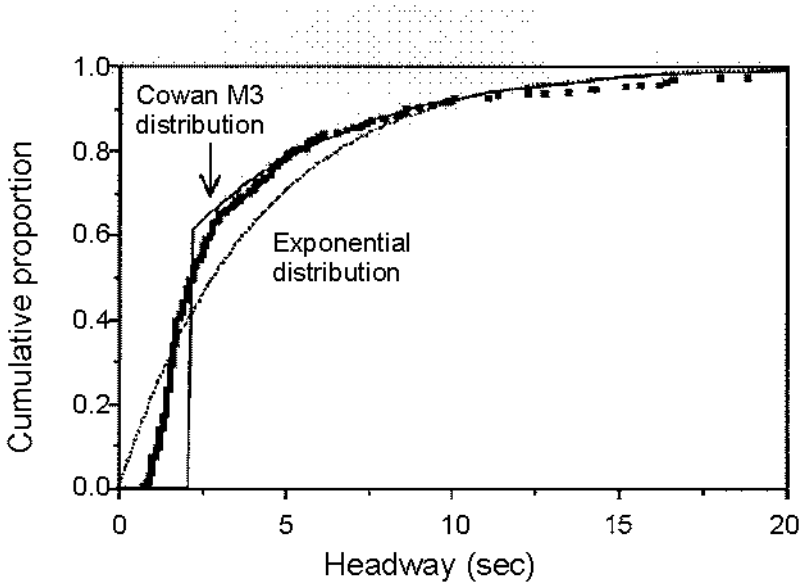


Figure 8.5
Arterial Road Data and a Cowan (1975) Dichotomized Headway Distribution (Higher flows example).

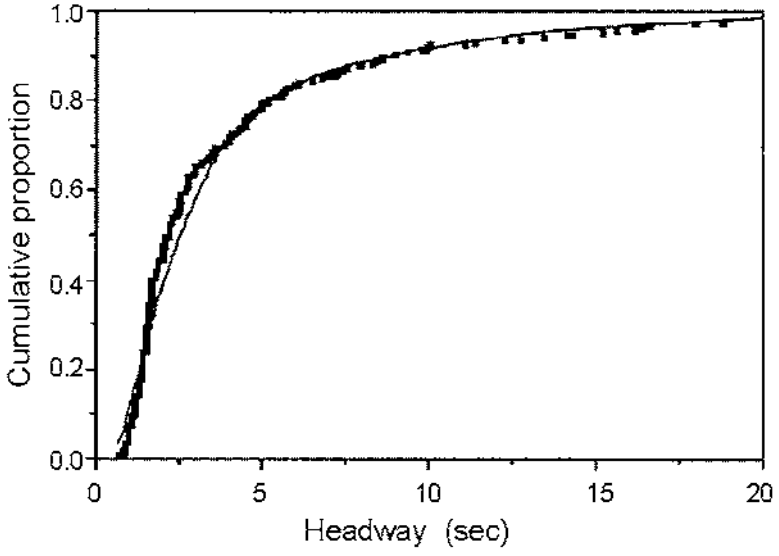


Figure 8.6
Arterial Road Data and a Hyper-Erlang Dichotomized Headway Distribution (Higher Flow Example).

the distribution of the platoon length is Borel-Tanner (Borel 1942; Tanner 1953; 1961; and Haight and Breuer 1960). Again, q is the flow in veh/sec. The Borel-Tanner distribution of platoons gives the probability of a platoon of size n as

$$P(n) = \frac{e^{-nt_m q}(nt_m q)^{n-1}}{n!} \qquad (8.28)$$

where n is an integer.

Haight and Breuer (1960) found the mean platoon size to be $1 / (1 - t_m q)$ or $1/\alpha$ and the variance to be $t_m q / (1 - t_m q)^3$ or $(1- \alpha) /\alpha^3$. For the same mean platoon size, the Borel-Tanner distribution has a larger variance and predicts a greater number of longer platoons than does the geometric distribution. Differences in the platoon size distribution does not affect an estimate of capacity but it does affect the average delay per vehicle as shown in Figure 8.13.

8.4 Interaction of Two Streams

For an easy understanding of traffic operations at an unsignalized intersection it is useful to concentrate on the simplest case first (Figure 8.7).

All methods of traffic analysis for unsignalized intersections are derived from a simple queuing model in which the crossing of two traffic streams is considered. A priority traffic stream (major stream) of the volume q_p (veh/h) and a non-priority traffic stream (minor stream) of the volume q_n (veh/h) are involved in this queuing model. Vehicles from the major stream can cross the conflict area without any delay. Vehicles from the minor stream are only allowed to enter the conflict area, if the next vehicle from the major stream is still t_c seconds away (t_c is the critical gap), otherwise they have to wait. Moreover, vehicles from the minor stream can only enter the intersection t_f seconds after the departure of the previous vehicle (t_f is the follow-up time).

8.4.1 Capacity

The mathematical derivation of the capacity q_m for the minor stream is as follows. Let $g(t)$ be the number of minor stream vehicles which can enter into a major stream gap of duration t. The expected number of these t-gaps per hour is $3600 q_p f(t)$ where,

- $f(t)$ = statistical density function of the gaps in the major stream and
- q_p = volume of the major stream.

Therefore, the amount of capacity which is provided by t-gaps per hour is $3600\ q_p f(t)\ g(t)$.

To get the total capacity, expressed in veh/second, we have to integrate over the whole range of major stream gaps:

$$q_m = q_p \int_0^\infty f(t) \cdot g(t) dt \qquad (8.29)$$

where,

- q_m = maximum traffic volume departing from the stop line in the minor stream in veh/sec,
- q_p = major stream volume in veh/sec,
- $f(t)$ = density function for the distribution of gaps in the major stream, and
- $g(t)$ = number of minor stream vehicles which can enter into a major stream gap of size, t.

Based on the gap acceptance model, the capacity of the simple 2-stream situation (Figure 8.7) can be evaluated by elementary probability theory methods if we assume:

(a) constant t_c and t_f values,

(b) exponential distribution for priority stream headways (cf. Equation 8.15), and

(c) constant traffic volumes for each traffic stream.

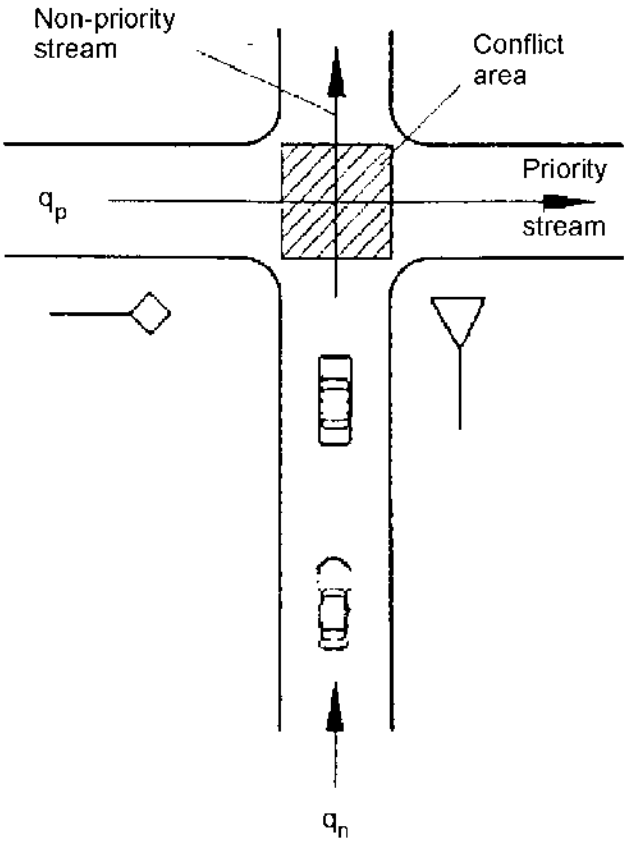


Figure 8.7
Illustration of the Basic Queuing System.

Within assumption (a), we have to distinguish between two different formulations for the term *g(t)*. These are the reason for two different families of capacity equations. The first family assumes a stepwise constant function for *g(t)* (Figure 8.2):

$$g(t) = \sum_{n=0}^{\infty} n \cdot p_n(t) \qquad (8.30)$$

where,

$p_n(t)$= probability that n minor stream vehicles enter a gap in the major stream of duration t,

$$p_n(t) = \begin{cases} 1 & for \quad t_c+(n-1)\cdot t_f \le t < t_c+n\cdot t_f \\ 0 & elsewhere \end{cases}$$

The second family of capacity equations assumes a continuous linear function for *g(t)* . This is an approach which has first been used by Siegloch (1973) and later also by McDonald and Armitage (1978).

$$g(t) = \begin{cases} 0 \; for \; t < t_0 \\ \dfrac{t-t_0}{t_f} \; for \; t \ge t_0 \end{cases} \qquad (8.31)$$

where,

$$t_0 = t_c - \frac{t_f}{2}$$

Once again it has to be emphasized that both in Equations 8.30 and 8.31, t_c and t_f are assumed to be constant values for all drivers.

Both approaches for *g(t)* produce useful capacity formulae where the resulting differences are rather small and can normally be ignored for practical applications (cf. Figure 8.8).

If we combine Equations 8.29 and 8.30, we get the capacity equation used by Drew (1968), Major and Buckley (1962), and by Harders (1968), which these authors however, derived in a different manner:

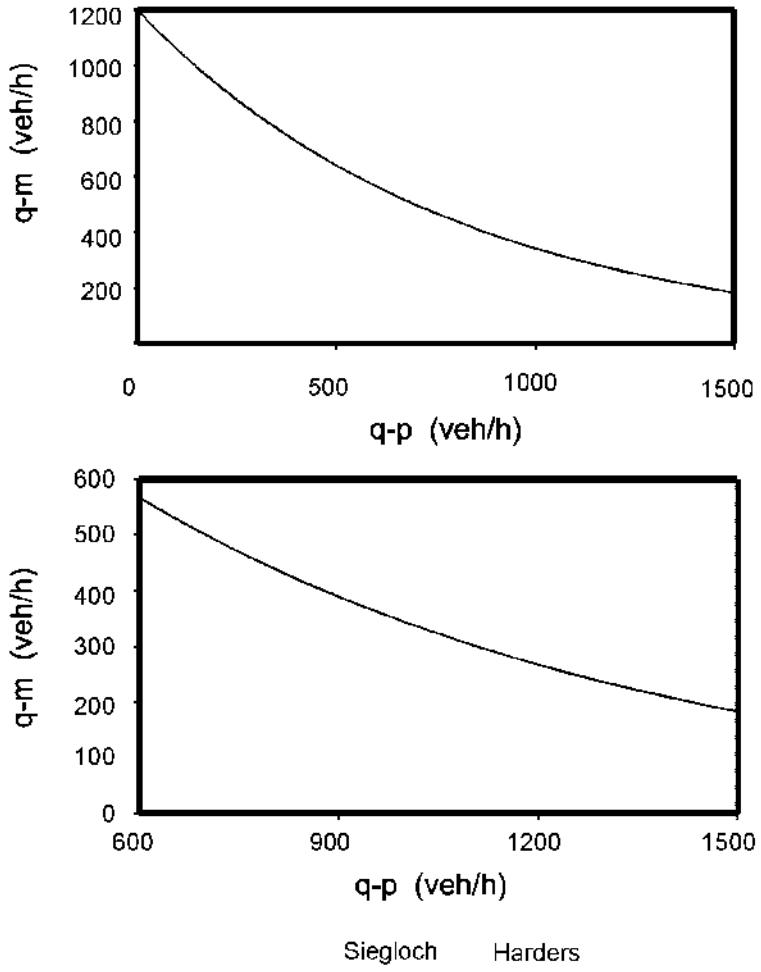


Figure 8.8
Comparison Relation Between Capacity (q-m) and Priority Street Volume (q-p).

$$q_m = q_p \frac{e^{-q_p \cdot t_c}}{1 - e^{-q_p \cdot t_f}} \qquad (8.32)$$

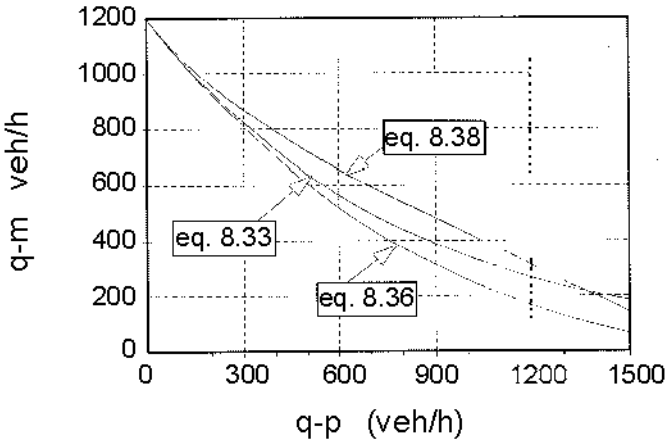
If we combine Equations 8.29 and 8.31 we get Siegloch's (1973) formula,

$$q_m = \frac{1}{t_f} e^{-q_p \cdot t_c} \qquad (8.33)$$

These formulae result in a relation of capacity versus conflicting flow illustrated by the curves shown in Figure 8.8.

The idealized assumptions, mentioned above as (a), (b), (c), however, are not realistic. Therefore, different attempts to drop one or the other assumption have been made. Siegloch (1973) studied different types of gap distributions for the priority stream (cf. Figure 8.9) based on analytical methods. Similar studies have also been performed by Catchpole and Plank (1986) and Troutbeck (1986). Grossmann (1991) investigated these effects by simulations. These studies showed

- If the constant t_c and t_f values are replaced by realistic distributions (cf. Grossmann 1988) we get a decrease in capacity.
- Drivers may be inconsistent; i.e. one driver can have different critical gaps at different times; A driver might reject a gap that he may otherwise find acceptable. This effect results in an increase of capacity.
- If the exponential distribution of major stream gaps is replaced by more realistic headway distributions, we get an increase in capacity of about the same order of magnitude as the effect of using a distribution for t_c and t_f values (Grossmann 1991 and Troutbeck 1986).
- Many unsignalized intersections have complicated driver behavior patterns, and there is often little to be gained from using a distribution for the variables t_c and t_f or complicated headway distributions. Moreover, Grossmann could show by simulation techniques that these effects compensate each other so that the simple capacity equations, 8.32 and 8.33, also give quite realistic results in practice.



Note: Comparison of capacities for different types of headway distributions in the main street traffic flow for t_c = 6 seconds and t_f = 3 seconds. For this example, t_m has been set to 2 seconds.

Figure 8.9
Comparison of Capacities for Different Types of Headway Distributions in the Main Street Traffic Flow.

More general solutions have been obtained by replacing the exponential headway distribution used in assumption (b) with a more realistic one e. g. a dichotomized distribution (cf. Section 8.3.3). This more general equation is:

$$q_m = \frac{\alpha q_p e^{-\lambda(t_c - t_m)}}{1 - e^{-\lambda t_f}} \qquad (8.34)$$

where

$$\lambda = \frac{\alpha q_f}{(1 - t_m q_f)} \qquad (8.35)$$

This equation is illustrated in Figure 8.10. This is also similar to equations reported by Tanner (1967), Gipps (1982), Troutbeck (1986), Cowan (1987), and others. If α is set to 1 and t_m to 0, then Harders' equation is obtained. If α is set to $1 - q_p * t_m$, then this equation reduces to Tanner's (1962) equation:

$$q_m = (1 - q_p \cdot t_m) \cdot \frac{q_p \cdot e^{-q_p \cdot (t_c - t_m)}}{1 - e^{-q_p \cdot t_f}} \qquad (8.36)$$

If the linear relationship for *g(t)* according to Equation 8.37 is used, then the associated capacity equation is

$$q_m = \frac{\alpha \cdot q_p \cdot e^{-\lambda(t_0 - t_m)}}{\lambda \cdot t_f} \qquad (8.37)$$

or

$$q_m = \frac{(1 - q_p t_m) \cdot e^{-\lambda(t_0 - t_m)}}{t_f} \qquad (8.38)$$

This was proposed by Jacobs (1979) .

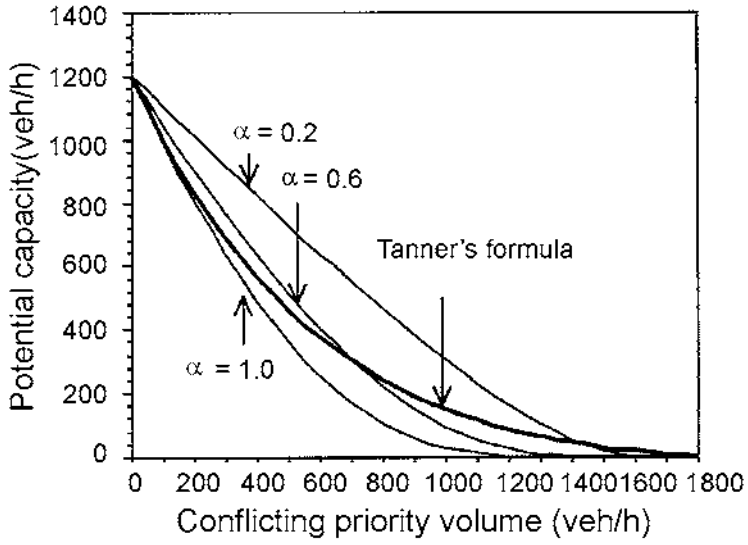


Figure 8.10
The Effect of Changing α in Equation 8.31 and Tanner's Equation 8.36.

Tanner (1962) analyzed the capacity and delay at an intersection where both the major and minor stream vehicles arrived at random; that is, their headways had a negative exponential distribution. He then assumed that the major stream vehicles were restrained such that they passed through the intersection at intervals not less than t_m sec after the preceding major stream vehicle. This allowed vehicles to have a finite length into which other vehicles could not intrude. Tanner did not apply the same constraint to headways in the minor stream. He assumed the same gap acceptability assumptions that are outlined above. Tanner considered the major stream as imposing 'blocks' and 'anti-blocks' on the minor stream. A block contains one or more consecutive gaps less than t_c sec; the block starts at the first vehicle with a gap of more than t_c sec in-front of it and ends t_c sec after the last consecutive gap less than t_c sec. Tanner's equation for the entry capacity is a particular case of a more general equation.

An analytical solution for a realistic replacement of assumptions (a) and (b) within the same set of formulae is given by Plank and Catchpole (1984):

$$q_m = \beta \cdot \frac{q_p \cdot e^{-q_p \cdot t_c}}{1 - e^{-q_p \cdot t_f}} \qquad (8.39)$$

where

$$\beta = 1 + \frac{1}{2} \cdot q_p^2 \left[Var(t_c) + \frac{Var(t_f)}{(e^{q_p \cdot t_f} - 1)} \right] + \eta_c + \eta_f \qquad (8.40)$$

$Var(t_c)$ = variance of critical gaps,
$Var(t_f)$ = variance of follow-up-times,
η_c = increment, which tends to 0, when $Var(t_c)$ approaches 0, and
η_f = increment, which tends to 0, when $Var(t_f)$ approaches 0.

Wegmann (1991) developed a universal capacity formula which could be used for each type of distribution for the critical gap, for the follow-up time and for each type of the major stream headway distribution.

$$q_m = \frac{1 + E\left[z(G - t_c)/\tau\right]}{E(C) \cdot E(1/\tau)} \qquad (8.41)$$

where,

$E(C)$ = mean length of a "major road cycle" C,
C = G + B,
G = gap,
B = block,
τ = probability ($G > t_c$), and
$z(t)$ = expected number of departures within the time interval of duration t.

Since these types of solutions are complicated many researchers have tried to find realistic capacity estimations by simulation studies. This applies especially for the German method (FGSV 1991) and the Polish method.

8.4.2 Quality of Traffic Operations

In general, the performance of traffic operations at an intersection can be represented by these variables (measures of effectiveness, MOE):

(a) average delay,
(b) average queue lengths,
(c) distribution of delays,
(d) distribution of queue lengths (i.e number of vehicles queuing on the minor road),
(e) number of stopped vehicles and number of accelerations from stop to normal velocity, and
(f) probability of the empty system (p_o).

Distributions can be represented by:

- standard deviations,
- percentiles, and
- the whole distribution.

To evaluate these measures, two tools can be used to solve the problems of gap acceptance:

- queuing theory and
- simulation.

Each of these MOEs are a function of q_p and q_n; the proportion of "free" vehicles and the distribution of platoon size length in both the minor and major streams. Solutions from queuing theory in the first step concentrate on average delays.

A general form of the equation for the average delay per vehicle is

$$D = D_{\min}\left(1+\frac{\gamma+\varepsilon x}{1-x}\right) \qquad (8.42)$$

where γ and ε are constants
x is the degree of saturation = q_n/q_m

and D_{min} has been termed Adams' delay after Adams (1936). Adams' delay is the average delay to minor stream vehicles when minor stream flow is very low. It is also the minimum average delay experienced by minor stream vehicles.

Troutbeck (1990) gives equations for γ, ε and D_{min} based on the formulations by Cowan (1987). If stream 2 vehicles are assumed to arrive at random, then γ is equal to 0. On the other hand, if there is platooning in the minor stream, then γ is greater than 0.

For random stream 2 arrivals, ε is given by

$$\varepsilon = \frac{e^{q_p t_f} - q_p t_f - 1 + q_p(e^{q_p t_f}-1)D_{\min}}{q_p(e^{q_p t_f}-1)D_{\min}} \qquad (8.43)$$

Note that ε is approximately equal to 1.0. D_{min} depends on the platooning characteristics in stream 1. If the platoon size distribution is geometric, then

$$D_{\min} = \frac{e^{\lambda(t_c-t_m)}}{\alpha q_p} - t_c - \frac{1}{\lambda} + \frac{\lambda t_m^2 - 2t_m + 2t_m\alpha}{2(t_m\lambda+\alpha)} \qquad (8.44)$$

(Troutbeck 1986).

Tanner's (1962) model has a different equation for Adams' delay, because the platoon size distribution in stream 1 has a Borel-Tanner distribution. This equation is

$$D_{\min} = \frac{e^{q_p(t_c-t_m)}}{(1-t_m q_p)q_p} - t_c - \frac{1}{q_p} + \frac{q_p t_m^2(2t_m q_p - 1)}{2(1-t_m q_p)^2} \qquad (8.45)$$

Another solution for average delay has been given by Harders (1968). It is not based on a completely sophisticated queuing theory. However, as a first approximation, the following equation for the average delay to non-priority vehicles is quite useful.

$$D = \frac{1-e^{-(q_p t_c + q_n t_f)}}{q_m/3600 - q_n} + t_f \qquad (8.46)$$

with q_m calculated using Equation. 8.34 or similar.

M/G/1 Queuing System - A more sophisticated queuing theory model can be developed by the assumption that the simple two-streams system (Figure 8.7) can be represented by a M/G/1 queue. The service counter is the first queuing position on the minor street. The input into the system is formed by the vehicles approaching from the minor street which are assumed to arrive at random, i.e. exponentially distributed arrival headways (i.e. "M"). The time spent in the first position of the queue is the service time. This service time is controlled by the priority stream, with an unknown service time distribution. The "G" is for a general service time. Finally, the "1" in M/G/1 stands for one service channel, i.e. one lane in the minor street.

For the M/G/1 queuing system, in general, the Pollaczek-Khintchine formula is valid for the average delay of customers in the queue

$$D_q = \frac{xW(1+C_w^2)}{2(1-x)} \qquad (8.47)$$

where

W = average service time. It is the average time a minor street vehicle spends in the first position of the queue near the intersection

C_w = coefficient of variation of service times

$$C_w = \frac{\sqrt{Var(W)}}{W}$$

$Var\ (W)$ = variance of service times

The total average delay of minor street vehicles is then

$$D = D_q + W.$$

In general, the average service time for a single-channel queuing system is: l/capacity. If we derive capacity from Equations 8.32 and following and if we include the service time W in the total delay, we get

$$D = \frac{1}{q_m}\left(1 + \frac{x}{1-x}C\right) \qquad (8.48)$$

where

$$C = \frac{1 + C_w^2}{2}$$

Up to this point, the derivations are of general validity. The real problem now is to evaluate C. Only the extremes can be defined which are:

- Regular service: Each vehicle spends the same time in the first position. This gives $Var(W) = 0$, $C_w^2 = 0$, and $C = 0.5$

This is the solution for the M/D/l queue.

- Random service: The times vehicles spend in the first position are exponentially distributed. This gives $Var(W) = E(W)$, $C_w^2 = 1$, and $C = 1.0$

This gives the solution for the M/M/1 queue.

Unfortunately, neither of these simple solutions applies exactly to the unsignalized intersection problem. However, as an approximation, some authors recommend the application of Equation 8.48 with $C = 1$.

Equation 8.42 can be further transformed to

$$D = D_{\min}\ (1+\gamma)\left(1 + \frac{\gamma+\varepsilon}{1+\gamma} * \frac{x}{1-x}\right) \qquad (8.49)$$

where ϵ and γ are documented in Troutbeck (1990).

This is similar to the Pollaczek-Khintchine formula (Equation 8.48). The randomness constant C is given by $(\gamma+\epsilon)/(1+\gamma)$ and the term $1/D_{min}*(1+\gamma)$ can be considered to be an equivalent 'capacity' or 'service rate.' Both terms are a function of the critical gap parameters t_c and t_f and the headway distributions. However, C, γ, and ϵ values are not available for all conditions.

For the M/G/1 system as a general property, the probability p_o of the empty queue is given by

$$p_o = 1 - x \qquad (8.50)$$

This formula is of sufficient reality for practical use at unsignalized intersections.

M/G2/1 queuing system - Different authors found that the service time distribution in the queuing system is better described by two types of service times, each of which has a specific distribution:

W_1 = service time for vehicles entering the empty system, i.e no vehicle is queuing on the vehicle's arrival

W_2 = service time for vehicles joining the queue when other vehicles are already queuing.

Again, in both cases, the service time is the time the vehicle spends waiting in the first position near the stop line. The first ideas for this solution have been introduced by Kremser (1962; 1964) and in a comparable way by Tanner (1962), as well as by Yeo and Weesakul (1964).

The average time which a customer spends in the queue of such a system is given by Yeo's (1962) formula:

$$D_q = \frac{q_n}{2} * \left(\frac{E(W_1^2) - E(W_2^2)}{v} + \frac{E(W_2^2)}{y}\right) \qquad (8.51)$$

where,

D_q = average delay of vehicles in the queue at higher positions than the first,
$E(W_1)$ = expectation of W_1,
$E(W_1^2)$ = expectation of $(W_1 * W_1)$
$E(W_2)$ = expectation of W_2,
$E(W_2^2)$ = expectation of $(W_2 * W_2)$,
v = $y + z$,
y = $1 - q_n E(W_2)$, and
z = $q_n E(W_1)$.

The probability p_o of the empty queue is

$$p_o = y/v \tag{8.52}$$

The application of this formula shows that the differences against Equation 8.50 are quite small (< 0.03). Refer to Figure 8.11.

If we also include the service time (= time of minor street vehicles spent in the first position) in the total delay, we get

$$D = \frac{E(W_1^2)}{v} + \frac{q_n}{2} * \left(\frac{y * E(W_1^2) + z * E(W_2^2)}{v * y} \right) \tag{8.53}$$

(Brilon 1988):Formulae for the expectations of W_1 and W_2 respectively have been developed by Kremser (1962):

$$E(W_1) = \frac{1}{q_p}(e^{q_p t} \tag{8.54}$$

$$E(W_2) = \frac{e^{q_p t_c}}{q_p}(1 - e^{-q_p t_f})$$

$$E(W_1^2) = \frac{2}{q_p}(e^{q_p t_c} - 1 - q_p t_c)(\frac{e^{q_p t_c}}{q_p} + t_f - t_c) + t_f^2 - t_c^2$$

$$E(W_2^2) = \frac{2 \cdot e^{q_p t_c}}{q_p^2}(e^{q_p t_c} - q_p t_c)(1 - e^{-q_p t_f}) - q_p t_f e^{-}$$

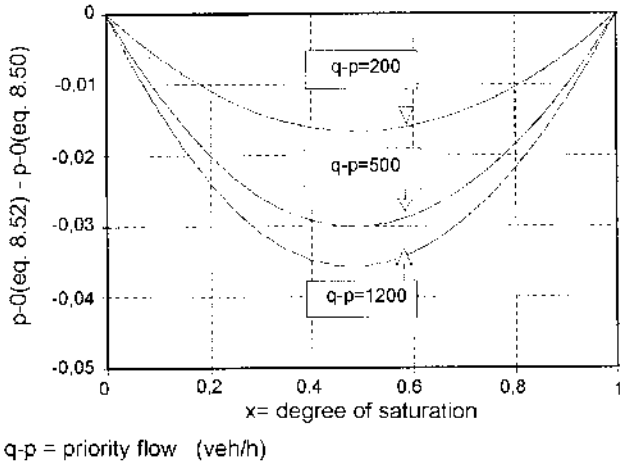


Figure 8.11
Probability of an Empty Queue: Comparison of Equations 8.50 and 8.52.

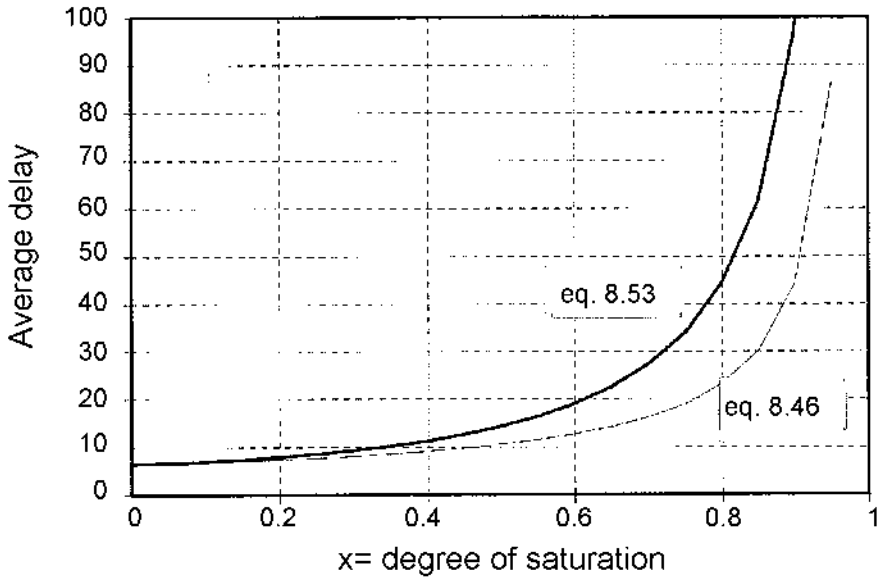
Kremser (1964), however, showed that the validity of these equations is restricted to the special case of $t_c = t_f$, which is rather unrealistic for two-way-stop-control unsignalized intersections. Daganzo (1977) gave an improved solution for $E(W_2)$ and $E(W_2^2)$ which again was extended by Poeschl (1983). These new formulae were able to overcome Kremer's (1964) restrictions. It can, however be shown that Kremer's first approach (Equation 8.56) also gives quite reliable approximate results for t_c and t_f values which apply to realistic unsignalized intersections. The following comments can also be made about the newer equations.

- The formulae are so complicated that they are far from being suitable for practice. The only imaginable application is the use in computer programs.

- Moreover, these formulae are only valid under assumptions (a), (b), and (c) in Section 8.4.1 of the paper. That means that for practical purposes, the equations can only be regarded as approximations and only apply for undersaturated conditions and steady state conditions.

Figure 8.12 gives a graphical comparison for some of the delay formulae mentioned.

Differences in the platoon size distribution affects the average delay per vehicle as shown in Figure 8.13. Here, the critical gap was 4 seconds, the follow-up time was 2 seconds, and the priority stream flow was 1000 veh/h. To emphasize the point, the average delay for a displaced exponential priority stream is 4120 seconds, when the minor stream flow was 400 veh/h. This is much greater than the values for the Tanner and exponential headway examples which were around 11.5 seconds for the same major stream flow. The average delay is also dependent on the average platoon size as shown in Figure 8.14. The differences in delays are dramatically different when the platoon size is changed.



Note: For this example; q_p = 600 veh/h, t_c = 6 sec , and t_f = 3 sec.

Figure 8.12
Comparison of Some Delay Formulae.

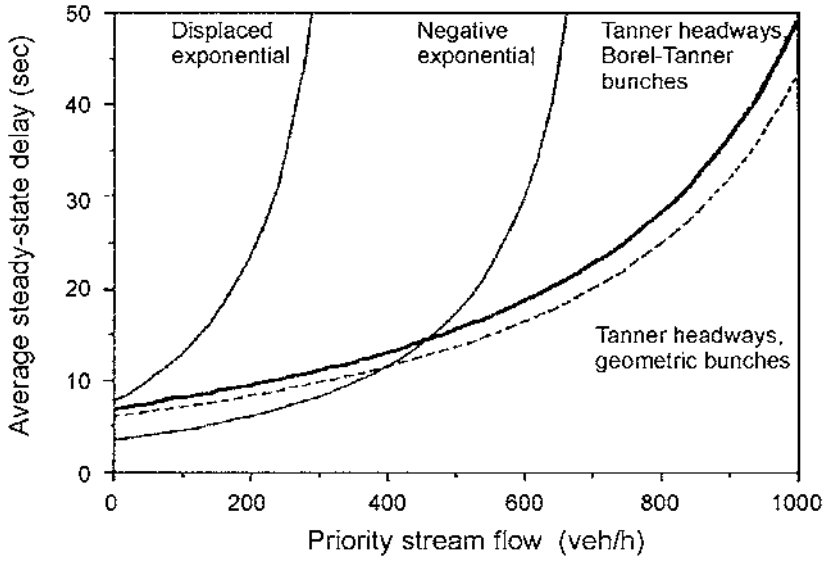


Figure 8.13
Average Steady State Delay per Vehicle
Calculated Using Different Headway Distributions.

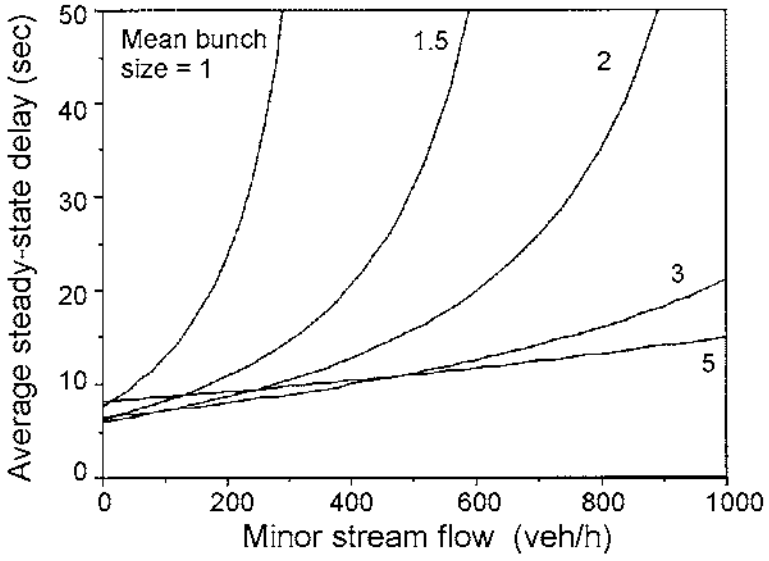


Figure 8.14
Average Steady State Delay per Vehicle by Geometric
Platoon Size Distribution and Different Mean Platoon Sizes.

8.4.3 Queue Length

In each of these queuing theory approaches, the average queue length (L) can be calculated by Little's rule (Little 1961):

$$L = q_n D \tag{8.55}$$

Given that the proportion of time that a queue exists is equal to the degree of saturation, the average queue length when there is a queue is:

$$L_q = q_n D/x = q_m D \tag{8.56}$$

The distribution of queue length then is often assumed to be geometric.

However, a more reliable derivation of the queue length distribution was given by Heidemann (1991). The following version contains a correction of the printing mistakes in the original paper (there: Equations 8.30 and 8.31).

$$p(0) = h_1 \cdot h_3 \cdot (q_p + q_n) \tag{8.57}$$

$$p(1) = p(0) \cdot h_3 \cdot q_n \left[e^{q_n t_f} - (t_c - t_f) \cdot h_2 \right] - q_n \cdot h_1 \cdot h_3$$

$$p(n) = p(n-1) \cdot h_3 \cdot q_n \left[e^{q_n t_f} - (t_c - t_f) \cdot h_2 \right] - h_3 \cdot \sum_{m=0}^{n-2} p(m) \cdot \left[h_2 \frac{(t_c - t_f \cdot q_n)^{n-m}}{(n-m)!} + \frac{(-q_n t_f)^{n-m} \cdot e^{q_n t_f}}{t_f \cdot (n-m-1)!} \right.$$

$p(n)$ = probability that n vehicles are queuing on the minor street

$$h_1 = e^{-q_p t_c} + (e^{-q_p t_f} - 1) \frac{q_n}{q_p}$$

$$h_2 = q_p e^{-q_p t_c - q_n (t_c - t_f)}$$

$$\frac{1}{h_3} = h_2 \cdot q_n \cdot e^{-q_p t_f}$$

These expressions are based on assumptions (a), (b), and (c) in Section 8.4.1. This solution is too complicated for practical use. Moreover, specific percentiles of the queue length is the desired output rather than probabilities. This however, can not be calculated from these equations directly. Therefore, Wu (1994) developed another set of formulae which approximate the above mentioned exact equations very closely:

$$p(0) = 1 - x^{a}$$
$$p(n) = p(0) \cdot x^{a(b(n-1)+1)} \tag{8.58}$$

$p(n)$ = probability that n vehicles are queuing on the minor street

where,

$$x = \frac{q_n}{q_m}$$

x = degree of saturation (q_m according to Equation 8.33).

$$a = \frac{1}{1 + 0.45 \cdot \frac{t_c - t_f}{t_f} \cdot q_p}$$

$$b = \frac{1.51}{1 + 0.68 \cdot \frac{t_c}{t_f} \cdot q_p}$$

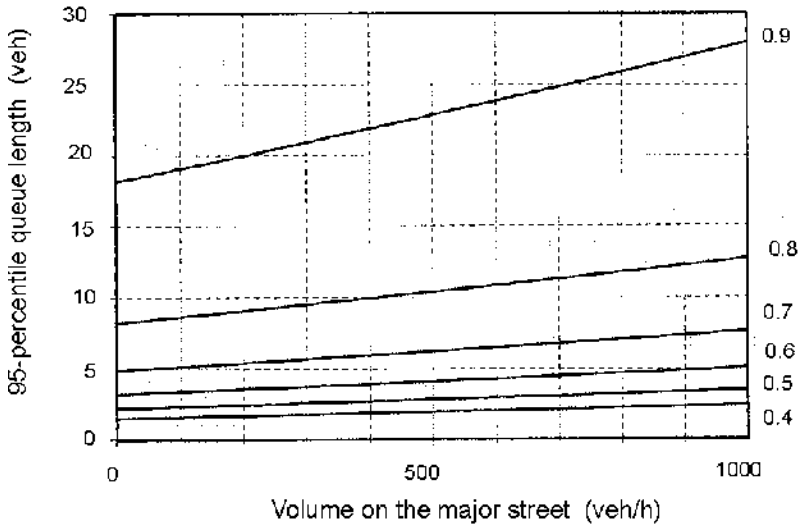
For the rather realistic approximation $t_c \approx 2\, t_f$, we get :

$$a = \frac{1}{1 + 0.45 \cdot q_p} \qquad b = \frac{1.51}{1 + 1.36 \cdot q_p}$$

From Equation 8.58 we get the cumulative distribution function

$$F(n) = p(L \le n) = 1 - x^{a(b \cdot n + 1)} \tag{8.59}$$

For a given percentile, S, (e.g. $S = F(n) = 0.95$) this equation can be solved for n to calculate the queue length which is only exceeded during (1-S)*100 percent of the time (Figure 8.15). For practical purposes, queue length can be calculated with sufficient precision using the approximation of the M/M/1 queuing system and, hence, Wu's equation. The 95-percentile-queue length based on Equation 8.59 is given in Figure 8.15.



The parameter of the curves (indicated on the right side) is the degree of saturation (x).

Figure 8.15
95-Percentile Queue Length Based on Equation 8.59 (Wu 1994).

8.4.4 Stop Rate

The proportion of drivers that are stopped at an unsignalized intersection with two streams was established by Troutbeck (1993). The minor stream vehicles were assumed to arrive at random whereas the major stream headways were assumed to have a Cowan (1975) M3 distribution. Changes of speed are assumed to be instantaneous and the predicted number of stopped vehicles will include those drivers who could have adjusted their speed and avoided stopping for very short periods.

The proportion stopped, $P(x,0)$, is dependent upon the degree of saturation, x, the headways between the bunched major stream vehicles, t_m, the critical gap, t_c. and the major stream flow, q. The appropriate equation is:

$$P(x,0) = 1-(1-x)(1-t_m q_p)e^{-\lambda(t_c-t_m)} \qquad (8.60)$$

where λ is given by $\alpha q_p/(1-t_m q_p)$. The proportion of drivers stopped for more than a short period of t, where t is less than the follow-up time t_f, increases from some minimum value, $P(0,t)$, to 1 as the degree of saturation increases from 0 to 1.

The proportion of drivers stopped for more than a short period t, $P(x,t)$, is given by the empirical equation:

$$P(x,t) = P(0,t)+A\{1-P(0,t)\}x+(1-A)\{1-P(0,t)\}x^2 +(1-A)(1-B)(1-x)x \qquad (8.61)$$

where

$$B = 1-(1-\frac{t}{t_f})(1-t_m q_p)e^{-\lambda(t_a-t_m)}$$

$$A = 1-a_0 e^{-\lambda(t_a-t_m)}$$

and

$$P(0,t) = P(0,0)-q_p t\alpha e^{-\lambda(t_a-t_m)} \qquad (8.62)$$

or

$$P(0,t) = 1 - (1 - t_m q_p + q_p t\alpha) e^{-\lambda(t_a - t_m)}$$

If the major stream is random then a_0 is equal to 1.25 and for bunched major stream traffic, it is 1.15. The vehicles that are stopped for a short period may be able to adjust their speed and these vehicles have been considered to have a "partial stop." Troutbeck (1993) also developed estimates of the number of times vehicles need to accelerate and move up within the queue.

8.4.5 Time Dependent Solution

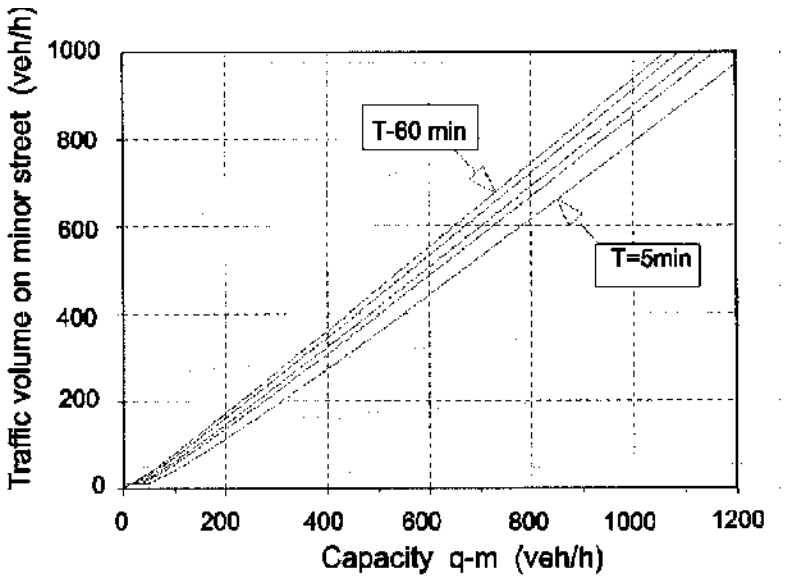
Each of the solutions given by the conventional queuing theory above is a steady state solution. These are the solutions that can be expected for non-time-dependent traffic volumes after an infinitely long time, and they are only applicable when the degree of saturation x is less than 1. In practical terms, this means, the results of steady state queuing theory are only useful approximations if T is considerably greater than the expression on the right side of the following equation.

$$T > \frac{1}{\left(\sqrt{q_m} - \sqrt{q_n}\right)^2} \qquad (8.63)$$

with T = time of observation over which the average delay should be estimated in seconds,

after Morse (1962).

This inequality can only be applied if q_m and q_n are nearly constant during time interval T. The threshold given by Equation 8.63 is illustrated by Figure 8.16. The curves are given for time intervals T of 5, 10, 15, 30, and 60 minutes. Steady state conditions can be assumed if q_n is below the curve for the corresponding T-value. If this condition (Equation 8.63) is not fulfilled, time-dependent solutions should be used. Mathematical solutions for the time dependent problem have been developed by Newell (1982) and now need to be made



Note: The curves are given for time intervals T of 5, 10, 15, 30, and 60 minutes. Steady state conditions can be assumed if q_n is below the curve for the corresponding T-value.

Figure 8.16
Approximate Threshold of the Length of Time Intervals For the Distinction Between Steady-State Conditions and Time Dependent Situations.

more accessible to practicing engineers. There is, however, a heuristic approximate solution for the case of the peak hour effect given by Kimber and Hollis (1979) which are based on the ideas of Whiting, who never published his work.

During the peak period itself, traffic volumes are greater than those before and after that period. They may even exceed capacity. For this situation, the average delay during the peak period can be estimated as:

$$
\begin{aligned}
D &= D_1 + E + \frac{1}{q_m} \\
D_1 &= \frac{1}{2}\left(\sqrt{F^2+G} - F\right) \\
F &= \frac{1}{q_{mo}-q_{no}}\left[\frac{T}{2}(q_m-q_n)y + C(y+\frac{h}{q_m})\right] + E \\
G &= \frac{2Ty}{q_{mo}-q_{no}}\left[C\frac{q_n}{q_m} - (q_m-q_n)E\right] \\
E &= \frac{Cq_{no}}{q_{mo}(q_{mo}-q_{no})} \\
h &= q_m - q_{mo} + q_{no} \\
y &= 1 - \frac{h}{q_n}
\end{aligned}
\qquad (8.64)
$$

q_m = capacity of the intersection entry during the peak period of duration T,
q_{mo} = capacity of the intersection entry before and after the peak period,
q_n = minor street volume during the peak period of duration T, and
q_{no} = minor street volume before and after the peak period

(each of these terms in veh/sec; delay in sec).

C is again similar to the factor C mentioned for the M/G/1 system, where

C = 1 for unsignalized intersections and
C = 0.5 for signalized intersections (Kimber and Hollis 1979).

This delay formula has proven to be quite useful to estimate delays and it has a quite reliable background particularly for temporarily oversaturated conditions.

A simpler equation can be obtained by using the same co-ordinate transfer method. This is a more approximate method. The steady state solution is fine for sites with a low degree of saturation and the deterministic solution is satisfactory for sites with a very high degree of saturation say, greater than three or four. The co-ordinate transfer method is one technique to provide estimates between these two extremes. the reader should also refer to Section 9.4.

The steady state solution for the average delay to the entering vehicle is given by Equation 8.42. The deterministic equation for delay, D_d, on the other hand is

$$
D_d = D_{\min} + \frac{2L_0 + (x_d - 1)q_m T}{2q_m} \qquad x > 1 \qquad (8.65)
$$

and $D_d = 0$ otherwise,

where L_0 is the initial queue,
T is time the system is operating in seconds, and
q_m is the entry capacity.

These equations are diagrammatically illustrated in Figure 8.17. For a given average delay the co-ordinate transformation method gives a new degree of saturation, x_t, which is related to the steady state degree of saturation, x_s, and the deterministic degree of saturation, x_d, such that

$$
x_d - x_t = 1 - x_s = a \qquad (8.66)
$$

Rearranging Equations 8.42 and 8.65 gives two equations for x_s and x_d as a function of the delays D_d and D_s. These two equations are:

$$
x_s = \frac{D_s - D_{\min} - \gamma D_{\min}}{D_s - D_{\min} + \varepsilon D_{\min}} \qquad (8.67)
$$

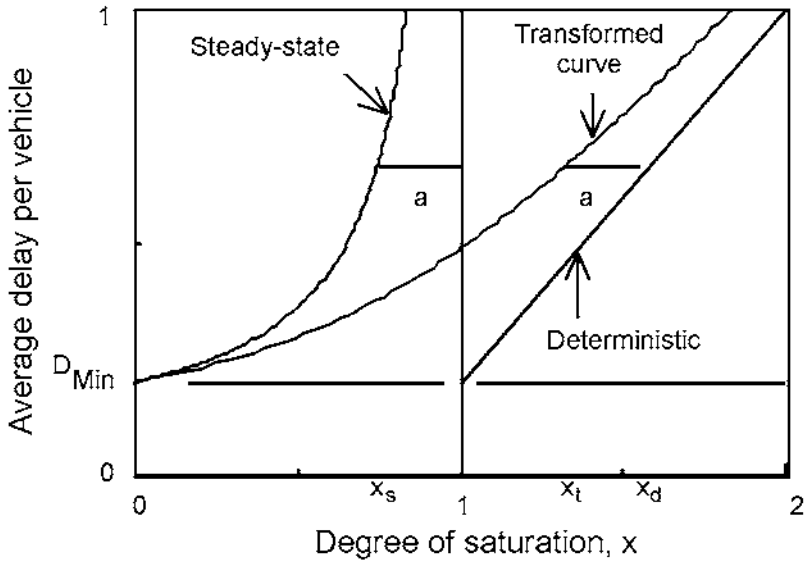


Figure 8.17
The Co-ordinate Transform Technique.

and

$$x_d = \frac{2(D_d - D_{\min}) - 2L_0/q_m}{T} + 1 \qquad (8.68)$$

Using Equation 8.66, x_t is given by:

$$x_t = \frac{2(D_d - D_{\min}) - 2L_0/q_m}{T} - \frac{D_s - D_{\min} - \gamma D_{\min}}{D_s - D_{\min} + \varepsilon D_{\min}} \qquad (8.69)$$

Rearranging Equation 8.69 and setting $D = D_s = D_d$, $x = x_e$ gives:

$$D_t = \frac{1}{2}\left\{\sqrt{A^2 + B} - A\right\} \qquad (8.70)$$

where

$$A = \frac{T(1-x)}{2} - \frac{L_0}{q_m} - D_{\min}(2 - \varepsilon) \qquad (8.71)$$

and

$$B = 4D_{\min}\left\{\frac{T(1-x)(1+\gamma)}{2} + \frac{Tx(\varepsilon+\gamma)}{2} - (1-\varepsilon)\left[\frac{L_0}{q_m} + D_{\min}\right]\right\} \qquad (8.72)$$

Equation 8.66 ensures that the transformed equation will asymptote to the deterministic equation and gives a family of relationships for different degrees of saturation and period of operation from this technique (Figure 8.18).

A simpler equation was developed by Akçelik in Akcelik and Troutbeck (1991). The approach here is to rearrange Equation 8.42 to give:

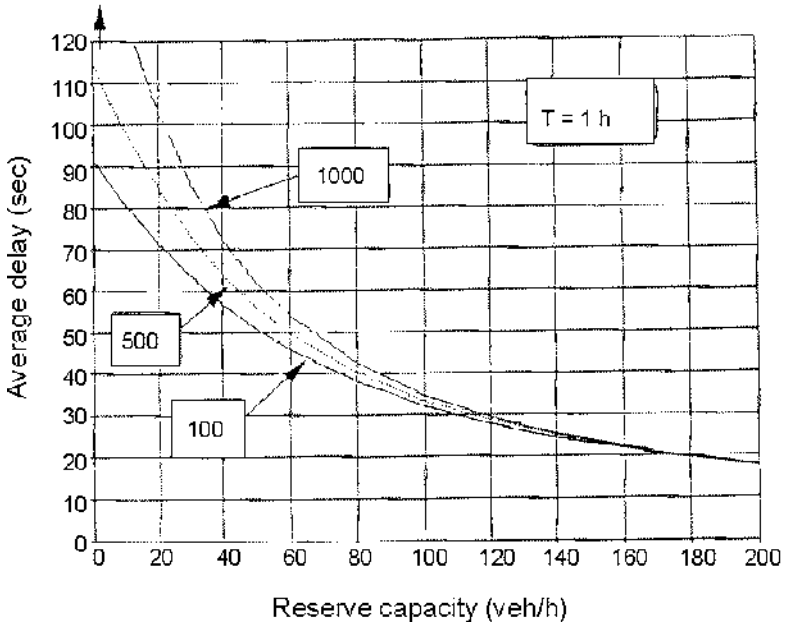


Figure 8.18
A Family of Curves Produced from the Co-Ordinate Transform Technique.

$$a = 1 - x_s = \frac{D_{\min}(\gamma + \varepsilon x_s)}{D_s - D_{\min}}$$

and this is approximately equal to:

$$a \approx \frac{D_{\min}(\gamma + \varepsilon x_t)}{D_s - D_{\min}} \qquad (8.73)$$

If this is used in Equation 8.66 and then rearranged then the resulting equation of the non-steady state delay is:

$$D - D_{\min} = \frac{1}{2}\left[\frac{L_0}{q_m} + \frac{(x-1)T}{4} + \sqrt{\left[\frac{L_0}{2q_m} + \frac{(x-1)T}{4}\right]^2 + \frac{TD_{\min}(\varepsilon x + \gamma)}{2}}\right] \qquad (8.74)$$

A similar equation for M/M/1 queuing system can be obtained if ε is set to 1, γ is set to zero, and D_{min} is set to $1/q_m$; the result is:

$$D = \frac{1}{q_m} + \frac{T}{4}\left\{(x-1) + \sqrt{(x-1)^2 + \frac{8x}{q_m T}}\right\} \qquad (8.75)$$

The average delay predicted by Equation 8.74 is dependent on the initial queue length, the time of operation, the degree of saturation, and the steady state equation coefficients. This equation can then be used to estimate the average delay under oversaturated conditions and for different initial queues. The use of these and other equations are discussed below.

8.4.6 Reserve Capacity

Independent of the model used to estimate average delays, the reserve capacity (R) plays an important role

$$R = q_{emax} - q_n \qquad (8.76)$$

In the 1985 edition of the HCM but not the 1994 HCM, it is used as the measure of effectiveness. This is based on the fact that average delays are closely linked to reserve capacity. This close relationship is shown in Figure 8.19. In Figure 8.19, the average delay, D, is shown in relation to reserve capacity, R. The delay calculations are based on Equation 8.64 with a peak hour interval of duration T= 1 hour. The parameters (100, 500, and 1000 veh/hour) indicate the traffic volume, q_p, on the major street. Based on this relationship, a good approximation of the average delay can also be expressed by reserve capacities. What we also see is that - as a practical guide - a reserve capacity

$R > 100$ pcu/h generally ensures an average delay below 35 seconds.

Brilon (1995) has used a coordinate transform technique for the "Reserve Capacity" formulation for average delay with oversaturated conditions. His set of equations can be given by

$$D = -B + \sqrt{B^2 + b} \qquad (8.77)$$

where

$$B = \frac{1}{2}\left(bR - \frac{L_0}{q_m} \right) \qquad (8.78)$$

$$b = \left\{ \frac{1}{q_m - R_f} \left[\frac{L_0 - R_f T}{2} \left(1 - \frac{R_f}{R_0} \right) - \frac{L_0}{q_m} \right] \right\} \frac{1}{|R_f|} \qquad (8.79)$$

$$R_f = \frac{100 \cdot 3600}{T} \qquad (8.80)$$

$$L_0 = \frac{q_{n0}}{R_0} = \frac{q_{m0} - R_0}{R_0} \qquad (8.81)$$

T	=	duration of the peak period
q_m	=	capacity during the peak period
q_n	=	minor street flow during the peak period
R	=	reserve capacity during the peak period $= q_{emax} - q_n$
L_0	=	average queue length in the period before and after the peak period
q_{n0}	=	minor street flow in the period before and after the peak period
q_{m0}	=	capacity in the period before and after the peak period
R_0	=	reserve capacity in the period before and after the peak period

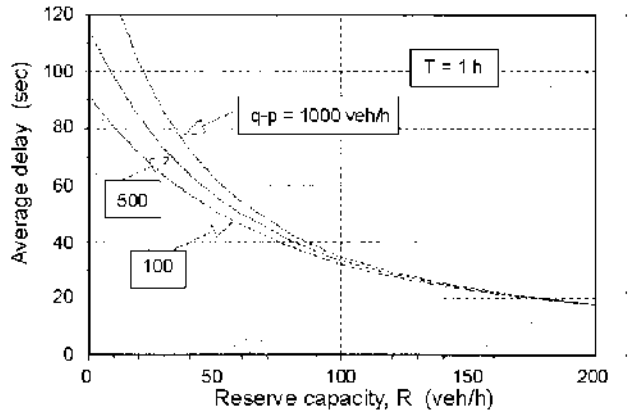


Figure 8.19
Average Delay, D, in Relation to Reserve Capacity R.

All variables in these equations should be used in units of seconds (sec), number of vehicles(veh), and veh/sec. Any capacity formula to estimate q_m and q_{m0} from Section 8.4.1 can be used.

The numerical results of these equations as well as their degree of sophistication are comparable with those of Equation 8.75.

8.4.7 Stochastic Simulation

As mentioned in the previous chapters, analytical methods are not capable of providing a practical solution, given the complexity and the assumptions required to be made to analyze unsignalized intersections in a completely realistic manner. The modern tool of stochastic simulation, however, is able to overcome all the problems very easily. The degree of reality of the model can be increased to any desired level. It is only restricted by the efforts one is willing to undertake and by the available (and tolerable) computer time. Therefore, stochastic simulation models for unsignalized intersections were developed very early (Steierwald 1961a and b; Boehm, 1968). More recent solutions were developed in the U. K. (Salter 1982), Germany (Zhang 1988; Grossmann 1988; Grossmann 1991), Canada (Chan and Teply 1991) and Poland (Tracz 1991).

Speaking about stochastic simulation, we have to distinguish two levels of complexity:

1) ***Point Process Models*** - Here cars are treated like points, i.e. the length is neglected. As well, there is only limited use of deceleration and acceleration. Cars are regarded as if they were "stored" at the stop line. From here they depart according to the gap acceptance mechanism. The effect of limited acceleration and deceleration can, of course, be taken into account using average vehicle performance values (Grossmann 1988). The advantage of this type of simulation model is the rather shorter computer time needed to run the model for realistic applications. One such model is KNOSIMO (Grossmann 1988, 1991). It is capable of being operated by the traffic engineer on his personal computer during the process of intersection design. A recent study (Kyte et al , 1996) pointed out that KNOSIMO provided the most realistic representation of traffic flow at unsignalized intersections among a group of other models.

   KNOSIMO in its present concept is much related to German conditions. One of the specialities is the restriction to single-lane traffic flow for each direction of the main street. Chan and Teply (1991) found some easy modifications to adjust the model to Canadian conditions as well. Moreover, the source code of the model could easily be adjusted to traffic conditions and driver behavior in other countries.

2) ***Car Tracing Models*** - These models give a detailed account of the space which cars occupy on a road together with the car-following process but are time consuming to run. An example of this type of model is described by Zhang (1988).

Both types of models are useful for research purposes. The models can be used to develop relationships which can then be represented by regression lines or other empirical evaluation techniques.

8.5 Interaction of Two or More Streams in the Priority Road

The models discussed above have involved only two streams; one being the priority stream and the second being a minor stream. The minor stream is at a lower rank than the priority stream. In some cases there may be a number of lanes that must be given way to by a minor stream driver. The capacity and the delay incurred at these intersections have been looked at by a number of researchers. A brief summary is given here.

If the headways in the major streams have a negative exponential distribution then the capacity is calculated from the equation for a single lane with the opposing flow being equal to the sum of the lane flows. This results in the following equation for capacity in veh/h:

$$q_{emax} = \frac{3600 q e^{-q t_a}}{1 - e^{-q t_f}} \qquad (8.82)$$

where q is the total opposing flow.

Tanner (1967) developed an equation for the capacity of an intersection where there were *n* major streams. The traffic in each lane has a dichotomized headway distribution in which there is a proportion of vehicles in bunches and the remaining vehicles free of interaction. All bunched vehicles are assumed to have a headway of t_m and the free vehicles have a headway equal to the t_m plus a negative exponentially distributed (or random) time. This is the same as Cowan's (1975) M3 model. Using the assumption that headways in each lane are independent, Tanner reviewed the distribution of the random time periods and estimated the entry capacity in veh/h as:

$$q_{emax} = \frac{3600\left[\lambda(1-t_{m1}q_1)(1-t_{m2}q_2)\cdots(1-t_{m1}q_1)e^{-\lambda(t_a-t_m)}\right]}{1-e^{-\lambda t_f}} \qquad (8.83)$$

where $$\lambda = \lambda_1 + \lambda_2 + \ldots\ldots + \lambda_n \qquad (8.84)$$

$$\lambda_i = \alpha_i q_i / (1-t_m q_i) \qquad (8.85)$$

q_i is the flow in the major stream *i* in veh/sec.
α_i is the proportion of free vehicles in the major stream *i*.

This equation by Tanner is more complicated than an earlier equation (Tanner 1962) based on an implied assumption that the proportion of free vehicles, α_i, is a function of the lane flow. That is

$$\alpha_i = (1-t_m q_i)$$

and then λ_i reduces to q_i. Fisk (1989) extended this earlier work of Tanner (1962) by assuming that drivers had a different critical gap when crossing different streams. While this would seem to be an added complication it could be necessary if drivers are crossing some major streams from the left before merging with another stream from the right when making a left turn.

Her equation for capacity is:

$$q_{emax} = \frac{3600q\left[\prod_{i=1}^{n}(1-t_{mi}q_i)e^{-q_i t_{ai}}\right]e^{qt_m}}{1-e^{-qt_f}} \qquad (8.86)$$

where $$q = q_1 + q_2 + \ldots\ldots + q_n$$

8.5.1 The Benefit of Using a Multi-Lane Stream Model

Troutbeck (1986) calculated the capacity of a minor stream to cross two major streams which both have a Cowan (1975) dichotomized headway distribution. The distribution of opposing headways is:

$$F(t) = \frac{2q_1q_2t}{(q_1+q_2)} \qquad for \quad t < t_m \qquad (8.87)$$

and

$$F(t) = 1-\alpha' e^{-\lambda'(t-t_m)} \qquad for \quad t > t_m \qquad (8.88)$$

where

$$\alpha' = \frac{\alpha_1 q_1(1-q_2t_m)+\alpha_2 q_2(1-q_1t_m)}{(q_1+q_2)} \qquad (8.88a)$$

or after a little algebra,

$$\alpha' q = \lambda' \prod_{i=1}^{n}(1-q_i t_m) \qquad (8.88b)$$

and

$$\lambda' = \lambda_1 + \lambda_2 \qquad (8.89)$$

As an example, if there were two identical streams then the distribution of headways between vehicles in the two streams is given by Equations 8.87 and 8.88. This is also shown in figures from Troutbeck (1991) and reported here as Figure 8.20.

Gap acceptance procedures only require that the longer headways or gaps be accurately represented. The shorter gaps need only be noted.

Consequently the headway distribution from two lanes can be represented by a single Cowan M3 model with the following properties:

$$F(t) = 1-\alpha * e^{-\lambda'(t-t_m^*)} \qquad t > t_m^* \qquad (8.90)$$

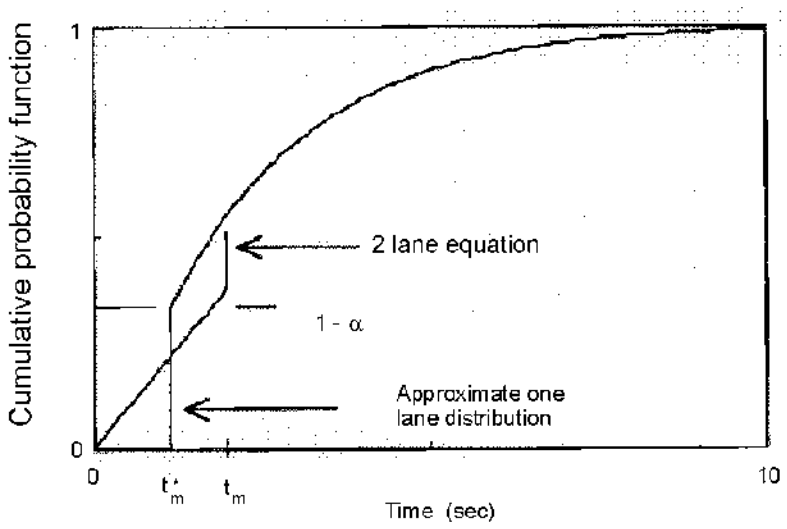


Figure 8.20
Modified 'Single Lane' Distribution of Headways (Troutbeck 1991).

and otherwise *F(t)* is zero. This modified distribution is also illustrated in Figure 8.20. Values of α* and t_m^* must be chosen to ensure the correct proportions and the correct mean headway are obtained. This will ensure that the number of headways greater than *t*, 1–*F(t)*, is identical from either the one lane or the two lane equations when *t* is greater than t_m^*.

Troutbeck (1991) gives the following equations for calculating α* and t_m^* which will allow the capacity to be calculated using a modified single lane model which are identical to the estimate from a multi-lane model.

The equations

$$(1-t_m^* q_1 - t_m^* q_2)e^{\lambda' t_m^*} = (1-t_m q_1)(1-t_m^* q_2)e^{\lambda' t_m} \quad (8.91)$$

and

$$\alpha * e^{\lambda' t_m^*} = \alpha' e^{\lambda' t_m} \quad (8.92)$$

are best solved iteratively for t_m with $t_{m,i}$ being the *i*th estimate. The appropriate equation is

$$t_m^*, i+1 = \frac{1-(1-t_m q_1)(1-t_m q_2)e^{\lambda'(t_m - t_{m,i}^*)}}{q_1 + q_2} \quad (8.93)$$

α* is then found from Equation 8.93.

Troutbeck (1991) also indicates that the error in calculating Adams' delay when using the modified single lane model instead of the two lane model is small. Adams' delay is the delay to the minor stream vehicles when the minor stream flow is close to zero. This is shown in Figure 8.21. Since the modified distribution gives satisfactory estimates of Adams' delay, it will also give satisfactory estimates of delay.

In summary, there is no practical reason to increase the complexity of the calculations by using multi-lane models and a single lane dichotomized headway model can be used to represent the distribution of headways in either one or two lanes.

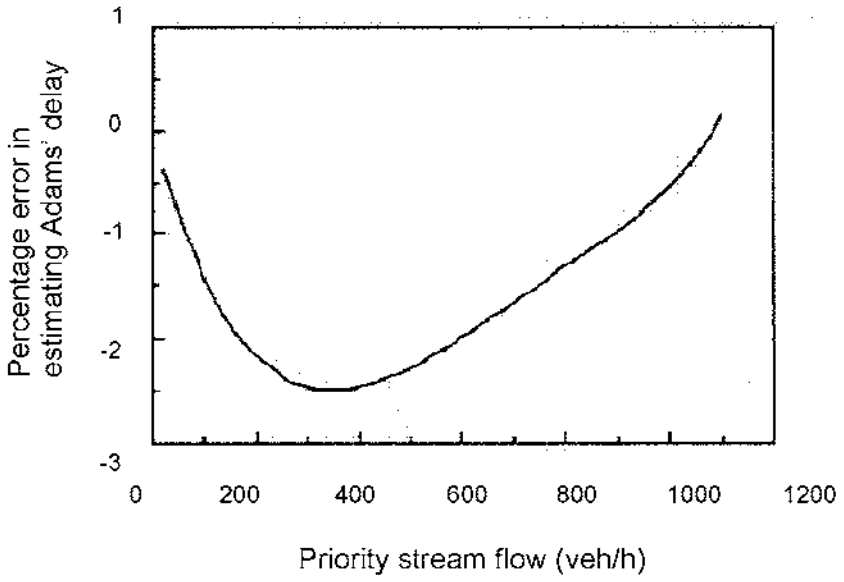


Figure 8.21
Percentage Error in Estimating Adams' Delay Against the Major Stream Flow for a Modified Single Lane Model (Troutbeck 1991).

8.6 Interaction of More than Two Streams of Different Ranking

8.6.1 Hierarchy of Traffic Streams at a Two Way Stop Controlled Intersection

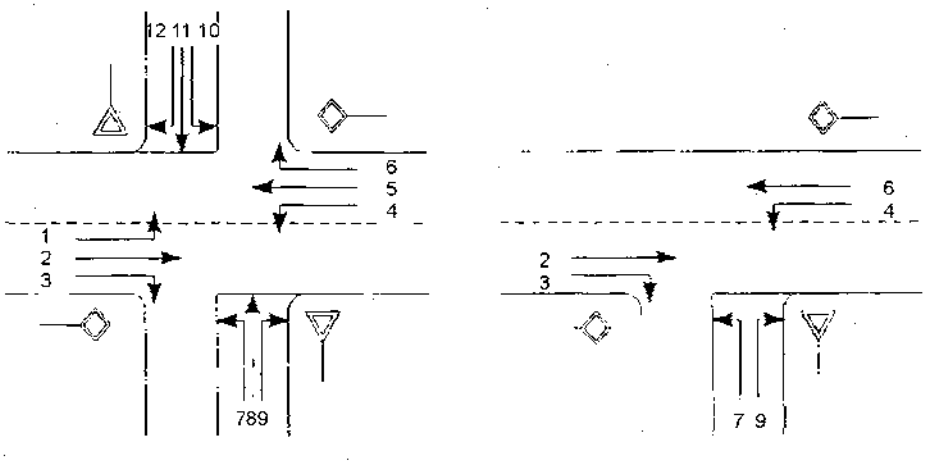
At all unsignalized intersections except roundabouts, there is a hierarchy of streams. Some streams have absolute priority (Rank 1), while others have to yield to higher order streams. In some cases, streams have to yield to some streams which in turn have to yield to others. It is useful to consider the streams as having different levels of priority or ranking. These different levels of priority are established by traffic rules. For instance,

Rank 1 stream	has absolute priority and does not need to yield right of way to another stream,
Rank 2 stream	has to yield to a Rank 1 stream,
Rank 3 stream	has to yield to a Rank 2 stream and in turn to a Rank 1 stream, and
Rank 4 stream	has to yield to a Rank 3 stream and in turn to Rank 2 and Rank 1 streams (left turners from the minor street at a cross-intersection).

This is illustrated in Figure 8.22 produced for traffic on the right side. The figure illustrates that the left turners on the major road have to yield to the through traffic on the major road. The left turning traffic from the minor road has to yield to all other streams but is also affected by the queuing traffic in the Rank 2 stream.

8.6.2 Capacity for Streams of Rank 3 and Rank 4

No rigorous analytical solution is known for the derivation of the capacity of Rank-3-movements like the left-turner from the minor street at a T-junction (movement 7 in Figure 8.22, right side). Here, the gap acceptance theory uses the impedance factors p_0 as an approximation. p_0 for each movement is the probability that no vehicle is queuing at the entry. This is given with sufficient accuracy by Equation 8.50 or better with the two service time Equation 8.52. Only during the part $p_{0,rank\text{-}2}$ of the total time, vehicles of Rank 3 can enter the intersection due to highway code regulations.



Note: The numbers beside the arrows indicate the enumeration of streams given by the Highway Capacity Manual (1994, Chapter 10).

Figure 8.22
Traffic Streams And Their Level Of Ranking.

Therefore, for Rank-3-movements, the basic value q_m for the potential capacity must be reduced to $p_0 \cdot q_m$ to get the real potential capacity q_e:

$$q_{e,rank\text{-}3} = p_{0,rank\text{-}2} \cdot q_{m,rank\text{-}3} \tag{8.94}$$

For a T-junction, this means

$$q_{e,7} = p_{0,4} \cdot q_{m,7}$$

For a cross-junction, this means

$$q_{e,8} = p_x \cdot q_{m,8} \tag{8.95}$$

$$q_{e,11} = p_x \cdot q_{m,11} \tag{8.96}$$

with

$$p_x = p_{0,1} \cdot p_{0,4}$$

Here the index numbers refer to the index of the movements according to Figure 8.22. Now the values of $p_{0,8}$ and $p_{0,11}$ can be calculated according to Equation 8.50.

For Rank-4-movements (left turners at a cross-intersection), the dependency between the p_0 values in Rank-2 and Rank-3-movements must be empirical and can not be calculated from analytical relations. They have been evaluated by numerous simulations by Grossmann (1991; cf. Brilon and Grossmann 1991). Figure 8.23 shows the statistical dependence between queues in streams of Ranks 2 and 3.

In order to calculate the maximum capacity for the Rank-4-movements (numbers 7 and 10), the auxiliary factors, $p_{z,8}$ and $p_{z,11}$, should be calculated first:

$$p_{z,i} = 0.65 p_{y,i} - \frac{p_{y,i}}{p_{v,i}+3} + 0.6\sqrt{p_{y,i}} \tag{8.97}$$

diminished to calculate the actual capacities, q_e. Brilon (1988, cf. Figures 8.7 and 8.8) has discussed arguments which support this double introduction.

The reasons for this are as follows:

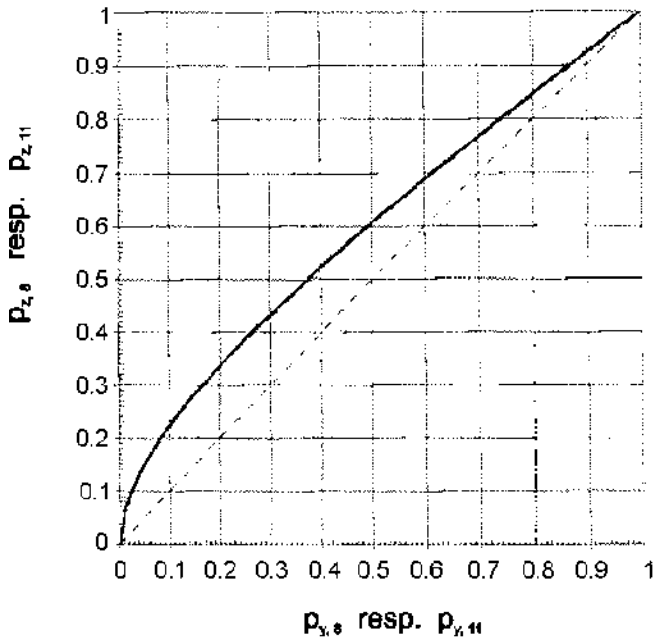


Figure 8.23
Reduction Factor to Account for the Statistical Dependence Between Streams of Ranks 2 and 3.

- During times of queuing in Rank-2 streams (e.g. left turners from the major street), the Rank-3 vehicles (e.g. left turners from the minor street at a T-junction) cannot enter the intersection due to traffic regulations and the highway code. Since the portion of time provided for Rank-3 vehicles is p_0, the basic capacity calculated from Section 8.4.1 for Rank-3 streams has to be diminished by the factor p_0 for the corresponding Rank-2 streams (Equations 8.95 to 8.99).
- Even if no Rank-2 vehicle is queuing, these vehicles influence Rank-3 operations, since a Rank-2 vehicle approaching the intersection within a time of less than t_c prevents a Rank-3 vehicle from entering the intersection.

Grossmann (1991) has proven that among the possibilities considered, the described approach is the easiest and quite realistic in the range of traffic volumes which occur in practical applications.

8.7 Shared Lane Formula

8.7.1 Shared Lanes on the Minor Street

If more than one minor street movement is operating on the same lane, the so-called "shared lane equation" can be applied. It calculates the total capacity q_s of the shared lane, if the capacities of the corresponding movements are known. (Derivation in Harders, 1968 for example.)

$$\frac{1}{q_s} = \sum_{i=1}^{m} \frac{b_i}{q_{m,i}} \qquad (8.100)$$

q_s = capacity of the shared lane in veh/h,
$q_{m,i}$ = capacity of movement *i,* if it operates on a separate lane in veh/h,

b_i = proportion of volume of movement i of the total volume on the shared lane,
m = number of movements on the shared lane.

The formula is also used by the HCM (1994, Equation 10-9).

This equation is of general validity regardless of the formula for the estimation of q_m and regardless of the rank of priority of the three traffic movements. The formula can also be used if the overall capacity of one traffic stream should be calculated, if this stream is formed by several partial streams with different capacities, e.g. by passenger cars and trucks with different critical gaps. Kyte at al (1996) found that this procedure for accounting for a hierarchy of streams, provided most realistic results.

8.7.2 Shared Lanes on the Major Street

In the case of a single lane on the major street shared by right-turning and through movements (movements no. 2 and 3 or 5 and 6 in Figure 8.22), one can refer to Table 8.2.

If left turns from the major street (movements no. 1 and 4 in Figure 8.22) have no separate turning lanes, vehicles in the priority l movements no. 2 and 3, and no. 5 and 6 respectively in Figure 8.21 may also be obstructed by queuing vehicles in those streams. The factors $p_{0,1}$* and $p_{0,4}$* indicate the probability that there will be no queue in the respective shared lane. They might serve for a rough estimate of the disturbance that can be expected and can be approximated as follows (Harders 1968):

$$p_{o,i}^{*} = 1 - \frac{1 - p_{0,i}}{1 - q_j t_{Bj} - q_k t_{Bk}} \qquad (8.101)$$

where: $i = 1, j = 2$ and $k = 3$ (cf. Figure 8.22)

or

$i = 4, j = 5$ and $k = 6$ (cf. Figure 8.22)

q_j = volume of stream j in veh/sec,
q_k = volume of stream k in veh/sec, and
t_{Bj} and t_{Bk} = follow-up time required by a vehicle in stream j or k (s).
(1.7 sec < t_B < 2.5 sec, e.g. t_B = 2 sec)

In order to account for the influence of the queues in the major street approach lanes on the minor street streams no. 7, 8, 10, and 11, the values $p_{0,1}$ and $p_{0,4}$, according to Equation 8.47 have to be replaced by the values $p_{0,1}$* and $p_{0,4}$* according to Equation 8.101. This replacement is defined in Equations 8.95 to 8.97.

8.8 Two-Stage Gap Acceptance and Priority

At many unsignalized intersections there is a space in the center of the major street available where several minor street vehicles can be stored between the traffic flows of the two directions of the major street, especially in the case of multi-lane major traffic (Figure 8.24). This storage space within the intersection enables the minor street driver to cross the major streams from each direction at different behavior times. This behavior can contribute to an increased capacity. This situation is called two-stage priority. The additional capacity being provided by these wider intersections can not be evaluated by conventional capacity calculation models.

Brilon et al. (1996) have developed an analytical theory for the estimation of capacities under two-stage priority conditions. It is based on an earlier approach by Harders (1968). In addition to the analytical theory, simulations have been performed and were the basis of an adjustment factor α. The resulting set of equations for the capacity of a two-stage priority situation are:

$$c_T = \frac{\alpha}{y^{k+1} - 1} \left\{ y(y^k - 1) \cdot [c(q_5) - q_1] + (y - 1) \cdot c(q_1 + q_2 + q_5) \right\}$$

for $y \neq 1$

$$c_{T(y=1)} = \frac{\alpha}{k+1} \left\{ k[c(q_5) - q_1] + (c(q_1 + q_2 + q_5) \right\} \qquad (8.104)$$

for $y = 1$

c_T = total capacity of the intersection for minor through traffic (movement 8)

Table 8.2
Evaluation of Conflicting Traffic Volume q_p
Note: The indices refer to the traffic streams denoted in Figure 8.22.

Subject Movement	No.	Conflicting Traffic Volume q_p
Left Turn from Major Road	1	$q_5 + q_6$ [3]
	7	$q_2 + q_3$ [3]
Right Turn from Minor Road	6	q_2 [2] $+ 0.5\, q_3$ [1]
	12	q_5 [2] $+ 0.5\, q_6$ [1]
Through Movement from Minor Road	5	$q_2 + 0.5\, q_3$ [1] $+ q_5 + q_6$ [3] $+ q_1 + q_4$
	11	$q_2 + q_3$ [3] $+ q_5 + 0.5\, q_6$ [1] $+ q_1 + q_4$
Left Turn from Minor Road	4	$q_2 + 0.5\, q_3$ [1] $+ q_5 + q_1 + q_4 + q_{12}$ [4)5)6)] $+ q_{11}$ [5]
	10	$q_5 + 0.5\, q_6$ [1] $+ q_2 + q_1 + q_4 + q_6$ [4)5)6)] $+ q_8$ [5]

Notes

1) ***If there is a right-turn lane, q_3 or q_6 should not be considered.***
2) ***If there is more than one lane on the major road, q_2 and q_5 are considered as traffic volumes on the right lane.***
3) ***If right-turning traffic from the major road is separated by a triangular island and has to comply with a YIELD- or STOP-Sign, q_6 and q_3 need not be considered.***
4) ***If right-turning traffic from the minor road is separated by a triangular island and has to comply with a YIELD- or STOP-sign, q_9 and q_{12} need not be considered.***
5) ***If movements 11 and 12 are controlled by a STOP-sign, q_{11} and q_{12} should be halved in this equation. Similarly, if movements 8 and 9 are controlled by a STOP-sign, q_8 and q_9 should be halved.***
6) ***It can also be justified to omit q_9 and q_{12} or to halve their values if the minor approach area is wide.***

where

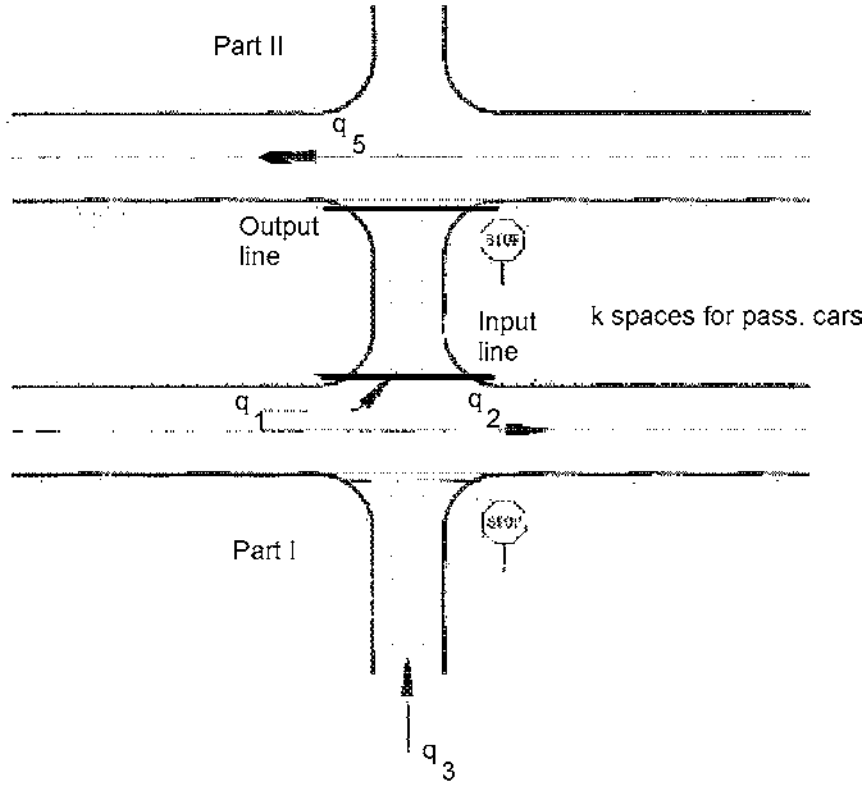
$$y = \frac{c(q_1 + q_2) - c(q_1 + q_2 + q_5)}{c(q_5) - q_1 - c(q_1 + q_2 + q_5)}$$

q_1 = volume of priority street left turning traffic at part I
q_2 = volume of major street through traffic coming from the left at part I
q_5 = volume of the sum of all major street flows coming from the right at part II.

$$\alpha = 1 \qquad \text{for } k=0$$

$$a = 1 - 0.32\exp(-1.3 \cdot \sqrt{k}) \qquad \text{for } k > 0 \qquad (8.105)$$

Of course, here the volumes of all priority movements at part II have to be included. These are: major right (6, except if this movement is guided along a triangular island separated from the through traffic) , major through (5), major left (4); numbers of movements according to Figure 8.22.



Note: The theory is independent of the number of lanes in the major street.

Figure 8.24
Minor Street Through Traffic (Movement 8) Crossing the Major Street in Two Phases.

$c(q_1+q_2)$	=	capacity at part I
$c(q_5)$	=	capacity at part II
$c(q_1+q_2+q_5)$	=	capacity at a cross intersection for minor through traffic with a major street traffic volume of $q_1+q_2+q_5$

(All of these capacity terms are to be calculated by any useful capacity formula, e.g. the Siegloch-formula, Equation 8.33)

The same set of formulas applies in analogy for movement 7. If both movements 7 and 8 are operated on one lane then the total capacity of this lane has to be evaluated from c_{T7} and c_{T8} using the shared lane formula (Equation 8.95). Brilon et al. (1996) provide also a set of graphs for an easier application of this theory.

8.9 All-Way Stop Controlled Intersections

8.9.1 Richardson's Model

Richardson (1987) developed a model for all-way stop controlled intersections (AWSC) based on M/G/1 queuing theory. He assumed that a driver approaching will either have a service time equal to the follow-up headway for vehicles in this approach if there are no conflicting vehicles on the cross roads (to the left and right). The average service time is the time between successive approach stream vehicles being able to depart. If there were conflicting vehicles then the conflicting

vehicles at the head of their queues will depart before the approach stream being analysed. Consequently, Richardson assumed that if there were conflicting vehicles then the average service time is the sum of the clearance time, T_c, for conflicting vehicles and for the approach stream.

For simplicity, Richardson considered two streams; northbound and westbound. Looking at the northbound drivers, the probability that there will be a conflicting vehicle on the cross road is given by queuing theory as ρ_w. The average service time for northbound drivers is then

$$s_n = t_m (1-\rho_w) + T_c \rho_w \quad (8.106)$$

A similar equation for the average service time for westbound drivers is

$$s_w = t_m (1-\rho_e) + T_c \rho_e \quad (8.107)$$

where,

ρ_i is the utilization ratio and is $q_i s_i$,
q_i is the flow from approach i,
s_i is the service time for approach i
t_m is the minimum headway, and
T_c is the total clearance time.

These equations can be manipulated to give a solution for s_n as

$$s_n = \frac{q_w t_m T_c + t_m - q_w t_m^2}{1 - q_w q_n (T_c^2 - 2 t_m T_c + t_m^2)} \quad (8.108)$$

If there are four approaches then very similar equations are obtained for the average service time involving the probability there are no cars on either conflicting stream. For instance,

$$s_n = t_m (1-\rho_{ew}) + T_c \rho_{ew} \quad (8.109)$$

$$s_s = t_m (1-\rho_{ew}) + T_c \rho_{ew} \quad (8.110)$$

$$s_e = t_m (1-\rho_{ns}) + T_c \rho_{ns} \quad (8.111)$$

$$s_w = t_m (1-\rho_{ns}) + T_c \rho_{ns} \quad (8.112)$$

The probability of no conflicting vehicles being $1-\rho_{ns}$ given by

$$1-\rho_{ns} = (1-\rho_s)(1-\rho_n) \quad (8.113)$$

hence,

$$\rho_{ns} = 1 - (1-q_n s_n)(1-q_s s_s) \quad (8.114)$$

and

$$\rho_{ew} = 1 - (1-q_e s_e)(1-q_w s_w) \quad (8.115)$$

Given the flows, q_n, q_s, q_e, and q_w and using an estimate of service times, p_{ns} and p_{ew} can be estimated using Equations 8.114 and 8.115. The iterative process is continued with Equations 8.109 to 8.112 providing a better estimate of the service times, s_n, s_s, s_e, and s_w.

Richardson used Herbert's (1963) results in which t_m was found to be 4 sec and T_c was a function of the number of cross flow lanes to be crossed. The equation was

$$t_c = 3.6 * 0.1 \; number\ of\ lanes$$

and T_c is the sum of the t_c values for the conflicting and the approach streams.

The steady-state average delay was calculated using the Pollaczek-Khintchine formula with Little's equation as:

$$W_s = \frac{2\rho - \rho^2 + q^2 Var(s)}{2(1-\rho)q} \quad (8.116)$$

or

$$W_s = \frac{\rho}{q}\left[1 + \frac{\rho\left(1 + \frac{q^2}{\rho^2} Var(s)\right)}{2(1-\rho)}\right]$$

This equation requires an estimate of the variance of the service times. Here Richardson has assumed that drivers either had a service time of h_m or T_c. For the northbound traffic, there were $(1-\rho_{ew})$ proportion of drivers with a service time of exactly t_m and ρ_{ew} drivers with a service time of exactly T_c. The variance is then

$$Var(s)_n = t_m^2(1-\rho_{ew}) + T_c^2\rho - s_n^2 \qquad (8.117)$$

and

$$\rho = \frac{s_n - t_m}{T_c - t_m} \qquad (8.118)$$

This then gives

$$Var(s)_n = t_m^2 \frac{T_c - s_n}{T_c - t_m} + T_c^2 \frac{s_n - t_m}{T_c - t_m} - s_n^2 \qquad (8.119)$$

for the northbound traffic. Similar equations can be obtained for the other approaches. An example of this technique applied to a four way stop with single lane approaches is given in Figure 8.25. Here the southbound traffic has been set to 300 veh/h. The east-west traffic varies but with equal flows in both directions. In accordance with the comments above, t_m was 4 sec and T_c was 2*t_c or 7.6 sec.

Richardson's approach is satisfactory for heavy flows where most drivers have to queue before departing. His approach has been extended by Horowitz (1993), who extended the number of maneuver types and then consequently the number of service time values. Horowitz has also related his model to Kyte's (1989) approach and found that his modified Richardson model compared well with Kyte's empirical data.

Figure 8.25 from Richardson's research, gives the performance as the traffic in one set approaches (north-south or east-west)increases. Typically, as traffic flow in one direction increases so does the traffic in the other directions. This will usually result in the level of delays increasing at a more rapid rate than the depicted in this figure.

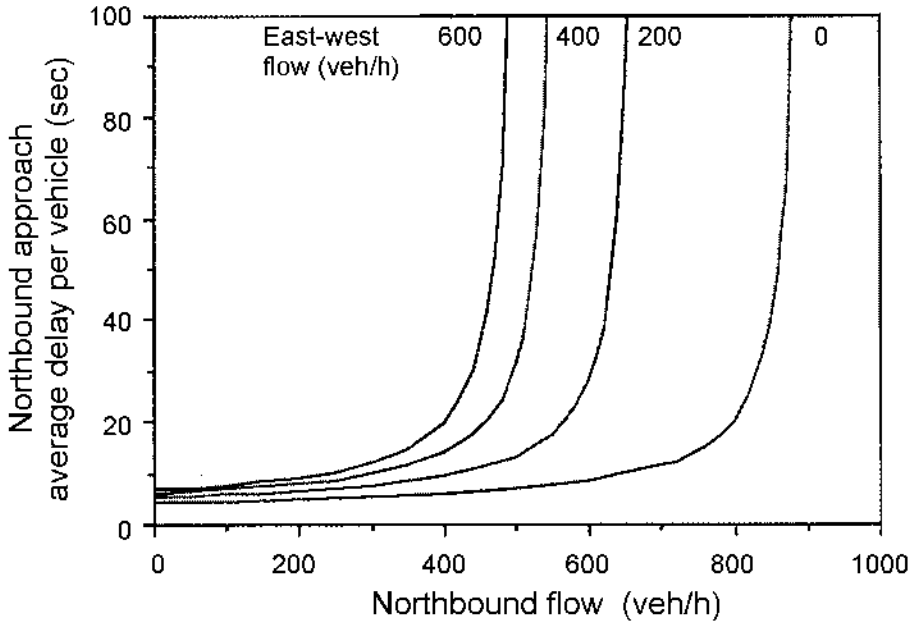


Figure 8.25
Average Delay For Vehicles on the Northbound Approach.

8.10 Empirical Methods

Empirical models often use regression techniques to quantify an element of the performance of the intersection. These models, by their very nature, will provide good predictions. However, at times they are not able to provide a cause and effect relationships.

Kimber and Coombe (1980), of the United Kingdom, have evaluated the capacity of the simple 2-stream problem using empirical methods. The fundamental idea of this solution is as follows: Again, we look at the simple intersection (Figure 8.7) with one priority traffic stream and one non-priority traffic stream during times of a steady queue (i.e. at least one vehicle is queuing on the minor street). During these times, the volume of traffic departing from the stop line is the capacity. This capacity should depend on the priority traffic volume q_p during the same time period. To derive this relationship, observations of traffic operations of the intersection have to be made during periods of oversaturation of the intersection. The total time of observation then is divided into periods of constant duration, e.g. 1 minute. During these 1-minute intervals, the number of vehicles in both the priority flow and the entering minor street traffic are counted. Normally, these data points are scattered over a wide range and are represented by a linear regression line. On average, half of the variation of data points results from the use of one-minute counting intervals. In practice, evaluation intervals of more than 1-minute (e.g. 5-minutes) cannot be used, since this normally leads to only few observations.

As a result, the method would produce linear relations for q_m:

$$q_m = b - c \cdot q_p \qquad (8.120)$$

Instead of a linear function, also other types of regression could be used as well, e.g.

$$q_m = A \cdot e^{-Bx} \quad . \qquad (8.121)$$

Here, the regression parameters A and B could be evaluated out of the data points by adequate regression techniques. This type of equation is of the same form as Siegloch's capacity formula (Equation 8.33). This analogy shows that A=3600/t_f.

In addition to the influence of priority stream traffic volumes on the minor street capacity, the influence of geometric layout of the intersection can be investigated. To do this, the constant values b and c or A and B can be related to road widths or visibility or even other characteristic values of the intersection layout by another set of linear regression analysis (see e.g. Kimber and Coombe 1980).
The advantages of the empirical regression technique compared to gap acceptance theory are:

- there is no need to establish a theoretical model.
- reported empirical capacities are used.
- influence of geometrical design can be taken into account.
- effects of priority reversal and forced priority are taken into account automatically.
- there is no need to describe driver behavior in detail.

The disadvantages are:

- transferability to other countries or other times (driver behavior might change over time) is quite limited: For application under different conditions, a very big sample size must always be evaluated.
- no real understanding of traffic operations at the intersection is achieved by the user.
- the equations for four-legged intersections with 12 movements are too complicated.
- the derivations are based on driver behavior under oversaturated conditions.
- each situation to be described with the capacity formulae must be observed in reality. On one hand, this requires a large effort for data collection. On the other hand, many of the desired situations are found infrequently, since congested intersections have been often already signalized.

8.10.1 Kyte's Method

Kyte (1989) and Kyte et al. (1991) proposed another method for the direct estimation of unsignalized intersection capacity for both AWSC and TWSC intersections. The idea is based on the fact that the capacity of a single-channel queuing system is the inverse of the average service time. The service time, t_W, at the unsignalized intersection is the time which a vehicle spends in the first position of the queue. Therefore, only the average of these times (t_W) has to be evaluated by observations to get the capacity.

Under oversaturated conditions with a steady queue on the minor street approach, each individual value of this time in the first

position can easily be observed as the time between two consecutive vehicles crossing the stop line. In this case, however, the observations and analyses are equivalent to the empirical regression technique .

Assuming undersaturated conditions, however, the time each of the minor street vehicles spends in the first position could be measured as well. Again, the inverse of the average of these times is the capacity. Examples of measured results are given by Kyte et al. (1991).

From a theoretical point of view, this method is correct. The problems relate to the measurement techniques (e.g. by video taping). Here it is quite difficult to define the beginning and the end of the time spent in the first position in a consistent way. If this problem is solved, this method provides an easy procedure for estimating the capacity for a movement from the minor street even if this traffic stream is not operating at capacity.

Following a study of AWSC intersections, Kyte et al. (1996) developed empirical equations for the departure headways from an approach for different levels of conflict.

$$h_i = h_{b\text{-}i} + h_{LT\text{-}adj} P_{LT} + h_{RT\text{-}adj} P_{RT} + h_{HV\text{-}adj} P_{HV} \qquad (8.122)$$

where:

h_i	is the adjusted saturation headway for the degree of conflict case *i*;
$h_{b\text{-}i}$	is the base saturation headway for case *i*;
$h_{LT\text{-}adj}$ and $h_{RT\text{-}adj}$	are the headway adjustment factors for left and right turners respectively;
P_{LT} and P_{RT}	are the proportion of left and right turners;
$h_{HV\text{-}adj}$	is the adjustment factor for heavy vehicles; and
P_{HV}	is the proportion of heavy vehicles.

The average departure headway, $\overline{d}$, is first assumed to be four seconds and the degree of saturation, x , is the product of the flow rate, V and $\overline{d}$. A second iterative value of $\overline{d}$ is given by the equation:

$$\overline{d} = \sum_{i=1}^{5} P(C_i) h_i$$

where $P(C_i)$ is the probability that conflict Ç occurs. These values also depend on estimates of $\overline{d}$ and the h_i values. The service time is given by the departure headway minus the move-up time.

Kyte et al. (1996) recognizes that capacity can be evaluated from two points of view. First, the capacity can be estimated assuming all other flows remain the same. This is the approach that is typically used in Section 8.4.1. Alternatively capacity can be estimated assuming the ratio of flow rates for different movements for all approaches remain constant. All flows are incrementally increased until one approach has a degree of saturation equal to one.

The further evaluation of these measurement results corresponds to the methods of the empirical regression techniques. Again, regression techniques can be employed to relate the capacity estimates to the traffic volumes in those movements with a higher rank of priority.

8.11 Conclusions

This chapter describes the theory of unsignalized intersections which probably have the most complicated intersection control mechanism. The approaches used to evaluate unsignalized intersections fall into three classes.

(a) Gap acceptance theory which assumes a mechanism for drivers departure. This is generally achieved with the notion of a critical gap and a follow on time. This attributes of the conflicting stream and the non priority stream are also required. This approach has been successfully used to predict capacity (Kyte et al. 1996) and has been extended to predict delays in the simpler conditions.

(b) Queuing theory in which the service time attributes are described. This is a more abstract method of describing driver departure patterns. The advantages of using queuing theory is that measures of delay (and queue lengths) can be more easily defined for some of the more complicated cases.

(c) Simulation programs. These are now becoming more popular. However, as a word of caution, the output from these models is only as good as the algorithms used in the model, and some simpler models can give excellent results. Other times, there is a temptation to look at the output from models without relating the results to the existing theory. This chapter describes most of the theories for unsignalised intersections and should assist simulation modelers to indicate useful extension to theory.

Research in these three approaches will undoubtably continue. New theoretical work highlights parameters or issues that should be considered further. At times, there will be a number of counter balancing effects which will only be identified through theory.

The issues that are likely to be debated in the future include the extent that one stream affects another as discussed in Section 8.6; the similarities between signalized and unsignalized intersections; performance of oversaturated intersection and variance associated with the performance measures.

References

Abou-Henaidy, M., S. Teply, and J. D. Hunt (1994). *Gap Acceptance Investigations in Canada*. Proceedings of the Second International Symposium on Highway Capacity. Australian Road Research Board- Transportation Research Board, Vol 1. pp. 1-20.

Adams (1936). *Road Traffic Considered as a Random Series.* Journal of the Institute of Civil Engineers, London. Vol. 4, pp. 121-130.

Akcelik, R. and R. J. Troutbeck (1991). *Implementation of the Australian Roundabout Analysis Method in SIDRA*. *In:* U. Brannolte (ed.) *Highway Capacity and Level of Service, Proceedings of the International Symposium on Highway Capacity*, Karlsruhe, A. A. Balkema, Rotterdam, pp.17-34.

Ashworth, R. (1968). *A Note on the Selection of Gap Acceptance Criteria for Traffic Simulation Studies*. Transportation Research, 2, pp. 171- 175, 1968.

Boehm, H. (1968). *Die Anwendung der Monte-Carlo-Methode in der Strassenverkehrstechnik. Teil I: Untersuchungen an ungesteuerten Knoten (Application of the Monte- Carlo-Method in Traffic Engineering. Part I: Investigations at Unsignalized Intersections).* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 73. Bonn 1968.

Brilon, W. (1988). *Recent Developments in Calculation Methods for Unsignalized Intersections in West Germany*. In: Intersections Without Traffic Signals (Ed.: W. Brilon). Springer publications, Berlin 1988.

Brilon, W. (Ed.) (1991). *Intersections Without Traffic Signals II.* Springer Publications, Berlin.

Brilon, W. (1995). *Delays at Oversaturated Unsignalized Intersections based on Reserve Capacities*. Paper presented at the TRB Annual Meeting, Washington D.C.

Brilon, W. and M. Grossmann (1991). *The New German Guideline for Capacity of Unsignalized Intersections.* In: Intersections without Traffic Signals II (Ed.: W. Brilon). Springer Publications, Berlin.

Brilon, W, N. Wu, and K. Lemke (1996). *Capacity at Unsignalized Two-Stage Priority Intersections*. Paper 961280. Presented at the 75th TRB Annual Meeting, Washington D.C.

Borel, G. (1942). *Sur L'emploi du Théorème De Bernoulli Pour Faciliter Le Calcul D'un Infinité De Co-Efficients. Application Au Problème De L'attente À Un Guichet.* C. R. Acad. Sci. Paris Vol. 214, pp. 452-456.

Buckley, D. J. (1962). *Road Traffic Headway Distribution*. Proceedings, 1st ARRB Conf., Vol. 1(1), pp. 153-186.

Buckley, D. J. (1968). *A Semi-Poisson Model of Traffic Flow*. Transportation Science, Vol. 2(2), pp. 107-132.

Catchpole, E. A. and A. W. Plank (1986). *The Capacity of a Priority Intersection*. Transportation Research Board, 20B (6), pp. 441-456

Chan, P. L. and S. Teply, S. (1991). *Simulation of Multilane Stop-Controlled T-Intersections by Knosimo in Canada*. In: Intersections without Traffic Signals II (Ed.: W. Brilon). Springer Publications, Berlin.

Cowan, R. J. (1975). *Useful Headway Models*. Transportation Research, 9(6), pp. 371-375.

Cowan, R. J. (1987). *An Extension of Tanner's Results on Uncontrolled Intersections.* Queuing Systems, Vol 1., pp. 249-263.

Daganzo, C. F. (1977). *Traffic Delay at Unsignalized Intersections: Clarification of Some Issues*. Transportation Science, Vol. 11.

Dawson, R. F. (1969). *The Hyperlang Probability Distribution - A Generalized Traffic Headway Model*. Proceedings of the FourthISTTT in Karsruhe, Strassenbau und Strassenverkehehrstechnik, No 89 1969, pp 30-36. Drew, D. R. (1968). *Traffic Flow Theory and Control.* McGraw-Hill Book Company, New York.

FGSV (1991). *Merkblatt zur Berechnung der Leistungsfähigkeit von Knotenpunkten ohne Lichtsignalanlagen (Guideline for the Estimation of Capacity at Unsignalized Intersections).* German Association for Roads and Traffic (FGSV), Cologne.

Fisk, C. S. (1989). *Priority Intersection Capacity - A Generalization of Tanner's Formula.* Transportation Research, Vol. 23 B, pp. 281-286.

Gipps, P. G. (1982). *Some Modified Gap Acceptance Formulae*. Proceedings of the 11th ARRB Conference, Vol. 11, part 4, p.71.

Grossmann, M. (1988). *Knosimo - a Practicable Simulation Model for Unsignalized Intersections*. In: Intersections without Traffic Signals (Ed.: W. Brilon). Springer Publications, Berlin.

Grossmann, M. (1991). *Methoden zur Berechnung und Beurteilung von Leistungsfähigkeit und Verkehrsqualitat an Knotenpunkten ohne Lichtsignalanlagen (Methods for Calculation and Judgement of Capacity and Traffic Quality at Intersections Without Traffic Signals).* Ruhr-University Bochum, Germany, Chair of Traffic Engineering, Vol. 9,.

Haight, F. A. and M. A. Breuer, (1960). *The Borel-Tanner Distribution.* Biometrika, Vol. 47 (1 and 2), pp. 143-150.

Harders, J. (1968). *Die Leistungsfähigkeit Nicht Signalgeregelter Städtischer Verkehrsknoten (The Capacity of Unsignalized Urban Intersections).* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 76.

Harders, J. (1976). *Grenz- und Folgezeitlücken als Grundlage für die Leistungsfähigkeit von Landstrassen (Critical Gaps and Move-Up Times as the Basis of Capacity Calculations for Rural Roads).* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 216.

HCM (1985). *Highway Capacity Manual.* Transportation Research Board, Special Report 209, Washington, DC.

Herbert, J. (1963). *A Study of Four Way Stop Intersection Capacities*. Highway Research Record, 27, Highway Research Board, Washington, DC.

Hewitt, R. H. (1983). *Measuring Critical Gap*. Transportation Science, 17(1), pp. 87-109.

Hewitt, R. H. (1985). *A Comparison Between Some Methods of Measuring Critical Gap.* Traffic Engineering & Control, 26(1), pp. 13-21.

Heidemann, D. (1991). *Queue Lengths and Waiting-Time Distributions at Priority Intersections.* Transportation Research B, Vol 25B, (4) pp. 163-174.

Horowitz, A. J. (1993). *Revised Queueing Model of Delay at All-Way Stop-Controlled Intersections.* Transportation Research Record, 1398, pp. 49-53.

Jacobs, F. (1979). *Leistungsfähigkeitsberechnung von Knotenpunkten ohne Lichtsignalanlagen (Capacity Calculations for Unsignalised Intersections).* Vorlesungsmanuskript, Stuttgart.

Kimber, R. M. and E. M. Hollis (1979). *Traffic Queues and Delays at Road Junctions*. TRRL Laboratory Report, LR909.

Kimber, R. M. and R. D. Coombe (1980). *The Traffic Capacity of Major/Minor Priority Junctions.* TRRL Report SR 582.

Kremser, H. (1962). *Ein Zusammengesetztes Wartezeitproblem Bei Poissonschen Verkehrsströmen (A Complex Problem of Delay with Poisson-Distributed Traffic Flows)*. Osterreichisches Ingenieur-Archiv, 16.

Kremser, H. (1964). *Wartezeiten Und Warteschlangen Bei Einfadelung Eines Poissonprozesses in Einen Anderen Solchen Prozess (Delays and Queues with One Poisson Process Merging into Another One)*. Österreichisches Ingenieur-Archiv, 18.

Kyte, M. (1989). *Estimating Capacity and Delay at an All-Way Stop-Controlled Intersection*. Research Report, TRANSNOW, Moscow/Idaho.

Kyte, M., J. Zegeer, and K. Lall (1991). *Empirical Models for Estimating Capacity and Delay at Stop-Controlled Intersections in the United States.* In: Intersections Without Traffic Signals II (Ed.: W. Brilon), Springer Publications, Berlin.

Kyte et al. (1996). *Capacity and Level of Service at Unsignalized Intersections*. Final Report for National Cooperative Highway Research Program Project 3-46.

Little, J. (1961). *A Proof of the Queueing Formula L = W.* Operations Research 9, pp. 383-387.

McDonald, M. and D. J. Armitage (1978). *The Capacity of Roundabouts*. Traffic Engineering & Control. Vol. 19(10), pp. 447-450.

Miller, A. J. (1972). *Nine Estimators of Gap Acceptance Parameters.* In: Traffic Flow and Transportation (Ed. Newell). Proceedings International Symposium on the Theory of Traffic Flow and Transportation, American Elsevier Publishing Co.

Morse, P. M. (1962). *Queues, Inventories and Maintenance*. John Wiley.

Newell, G. F. (1971). *Applications of Queueing Theory.* Chapman and Hall Ltd., London.

Newell, G. F. (1982). *Applications of Queueing Theory.* 2nd Ed. Chapman and Hall Ltd., London.

Plank, A. W. and E. A. Catchpole (1984). *A General Capacity Formula for an Uncontrolled Intersection.* Traffic Engineering Control. 25(6), pp. 327-329.

Poeschl, F. J. (1983). *Die Nicht Signalgesteuerte Nebenstrassenzufahrt Als Verallgemeinertes M/G/ 1-Warteschlangensystem (The Unsignalized Minor Street Entry as a Generalized M/G/1 Queueing System)*. Zeitschrift für Operations Research, Vol. 27 B.

Ramsey, J. B. H. and I. W. Routledge (1973). *A New Approach to the Analysis of Gap Acceptance Times.* Traffic Engineering Control, 15(7), pp. 353-357.

Richardson, A. J. (1987). *A Delay Model for Multiway Stop-Sign Intersections*. Transportation Research Record, 1112, pp. 107-112.

Salter, R. J. (1982). *Simulation of Delays and Queue Lengths at Over-Saturated Priority Highway Junctions.* Australian Road Research, Vol. 12. No. 4.

Schuhl, A. (1955). *The Probability Theory Applied to the Distribution of Vehicles on Two-Lane Highways.* Poisson and Traffic. The Eno Foundation for Highway Traffic Control. Sangatuck, CT.

Siegloch, W. (1973). *Die Leistungsermittlung an Knotenpunkten Ohne Lichtsignalsteuerung (Capacity Calculations for Unsignalized Intersections).* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 154.

Steierwald, G. (1961). *Die Anwendung Der Monte-Carlo-Methode in Der Strassenverkehrsforschung (The Application of the Monte-Carlo Method in Traffic Engineering Research).* Habilitation thesis, RWTH Aachen.

Steierwald, G. (1961). *Die Leistunsfähigkeit von Knotenpurlkten des Strassenverkehrs.* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 11, Bonn.

Tanner, J. C. (1953). *A Problem of Interference Between Two Queues.* Biometrika, 40, pp. 58-69.

Tanner, J. C. (1961) *A Derivation of the Borel Distribution.* Biometrica, Vol. 48, p. 222.

Tanner, J. C. (1962). *A Theoretical Analysis of Delays At An Uncontrolled Intersection*. Biometrica 49(1 and 2), pp. 163-70.

Tanner, J. C. (1967). *The Capacity of an Uncontrolled Intersection*. Biometrica, 54(3 and 4), pp. 657-658.

Tracz, M. (1991). *Polish Guidelines for Capacity Analysis of Priority Intersections.* In: Intersections Without Traffic Signals II (Ed.: W. Brilon), Springer Publications, Berlin.

Troutbeck, R. J. (1975). *A Review of the Ramsey-Routledge Method for Gap Acceptance Times*. Traffic Engineering & Control, 16(9), pp. 373-375.

Troutbeck. R. J. (1986). *A verage Delay at an Unsignalized Intersection with Two Major Streams Each Having a Dichotomized Headway Distribution.* Transportation Science, 20(4), pp. 272-286.

Troutbeck, R. J. (1988). *Current and Future Australian Practices for the Design of Unsignalized Intersections.* In: Intersections without Traffic Signals (Ed.: W. Brilon), Springer Publications, Berlin.

Troutbeck, R. J. (1990). *Roundabout Capacity and Associated Delay*. In: Transportation and Traffic Theory. Proceedings of the Eleventh International Symposium on Transportation and Traffic Theory, in Yokohama, Japan (Ed: M Koshi), Elsevier.

Troutbeck, R. J. (1991). *Unsignalized Intersection and Roundabouts in Australia: Recent Developments*. In: Intersections without Traffic Signals II (Ed.: W. Brilon), Springer Publications, Berlin.

Sullivan, D. and R. J. Troutbeck (1993). *Relationship Between the Proportion of Free Vehicles and Flow Rate on Arterial Roads.* Physical Infrastructure Centre Report, 92-21, Queensland University of Technology, Brisbane.

Wegmann, H. (1991). *A General Capacity Formula for Unsignalized Intersections.* In: Intersections without Traffic Signals II (Ed.: W. Brilon), Springer Publications, Berlin 1991.

Wu, N. (1994). *An Approximation for the Distribution of Queue Lengths at Signalized Intersections.* Proceedings of the Second International Symposium on Highway Capacity. Australian Road Research Board - Transportation Research Board, Vol 2., pp. 717-736.

Yeo, G. F. (1962). *Single-Server Queues with Modified Service Mechanisms*. Journal Australia Mathematics Society, Vol 2, pp. 499-502.

Yeo, G. F. and B. Weesakul (1964). *Delays to Road Traffic at an Intersection.* Journal of Applied Probability, pp. 297-310.

Zhang, X. (1988). *The Influence of Partial Constraint on Delay at Priority Junctions.* In: Intersections without Traffic Signals (Ed.: W. Brilon), Springer Publications, Berlin.

Additional References (Not cited in the Chapter 8)

Anveden, J. (1988). *Swedish Research on Unsignalized Intersections.* In: Intersections Without Traffic Signals (Ed.: W. Brilon). Springer Publications, Berlin.

Ashworth, R. (1969). *The Capacity of Priority-Type Intersections with a Non-Uniform Distribution of Critical Acceptance Gaps.* Transportation Research, Vol. 3.

Ashworth, R. (1970). *The Analysis and Interpretation of Gap Acceptance Data. Transportation Research*. No. 4, pp. 270-280.

Baas, K. G. (1987). *The Potential Capacity of Unsignalized Intersections.* ITE-Journal, pp. 43-56.

Bang, K. L., A. Hansson, and B. Peterson (1978). *Swedish Capacity Manual*. Transportation Research Record 667, pp 1-28.

Brilon, W. (1981). *Anmerkungen Zur Leistungsermittlung von Knotenpunkten Ohne Lichtsignalanlagen. (Some Notes on Capacity Calculations for Unsignalized Intersections).* Strassenverkehrstechnik, Heft 1, pp. 20-25.

Brilon, W. (Ed.) (1988). *Intersections Without Traffic Signals.* Springer Publications, Berlin.

Brilon, W. and M. Grossmann (1989). *Entwicklung Eines Simulationsmodells Für Knotenpunkte Ohne Lichtsignalanlagen (Development of a Simulation Model for Intersections Without Traffic Signals)*. Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 554, Bonn.

Brilon, W. and M. Grossmann (1991). *Aktualisiertes Berechnungsverfahren Für Knotenpunkte Ohne Lichtsignalanlagen (Actualized Calculation Procedure for Intersections Without Traffic Signals).* Schriftenreihe Strassenbau und Strassenverkehrstechnik, Vol. 596, Bonn.

Brilon, W., M. Grossmann, and B. Stuwe (1991). *Towards a New German Guideline for Capacity of Unsignalized Intersections.* Transportation Research Record, No. 1320, Washington, DC.

Catling, I. (1977). *A Time Dependent Approach to Junction Delays*. Traffic Engineering & Control, Vol. 18(11). Nov. 77, pp. 520-523, 536.

Chodur, J. and S. Gaca (1988). *Simulation Studies of the Effects of Some Geometrical and Traffic Factors on the Capacity of Priority Intersections.* In: Intersections Without Traffic Signals (Ed.: W. Brilon). Springer Publications, Berlin.

Cowan, R. J. (1984). *Adams' Formula Revised.* Traffic Engineering & Control, Vol. 25(5), pp 272- 274.

Fisk C. S. and H. H. Tan (1989). *Delay Analysis for Priority Intersections.* Transportation Research, Vol. 23 B, pp 452-469.

Fricker, J. D., M. Gutierrez, and D. Moffett (1991). *Gap Acceptance and Wait Time at Unsignalized Intersections.* In: Intersections without Traffic Signals II (Ed.: W. Brilon). Springer Publications, Berlin.

Hansson, A. (1978). *Swedish Capacity Manual.* Statens Vägverk (National Road Administration) Intern Rapport NR 24.

Hansson, A. (1987). *Berakningsmetoder För Olika Effektmått I Korningar, Del Iii: Korsningar Utan Trafiksignaler. (Procedures for Estimating Performance Measures of Intersections. Part III: Intersections Without Signal Control.)* Statens Vägverk (National Road Administration).

Hansson, A. and T. Bergh (1988). *A New Swedish Capacity Manual/CAPCAL 2.* Proceedings 14th Australian Road Research Board Conference, Vol. 14(2), pp. 38-47.

Hawkes, A. G. (1965). *Queueing for Gaps in Traffic.* Biometricia 52 (1 and 2). pp. 79-85.

Hondermarcq, H. (1968). *Essais De Priorité Á Gancle. 9th Intentional Study Week on Road Traffic Flow and Safety.* Munich.

Japan Society of Traffic Engineers (1988). *The Planning and Design of At-Grade Intersections*.

Jessen, G. D. (1968). *Ein Richtlinienvorschlag Für Die Behandlung Der Leistungsfähigkeit von Knotenpunkten Ohne Signalregelung (A Guideline Suggested for Capacity Calculations for Unsignalized Intersections)*. Strassenverkehrstechnik, No. 7/8.

Jirava, P. and P. Karlicky (1988). *Research on Unsignalized Intersections with Impact on the Czechoslovak Design Standard.* In: Intersections without Traffic Signals (Ed.: W. Brilon). Springer Publications, Berlin.

Kimber, R. M., M. Marlow, and E. W. Hollis (1977). *Flow/Delay Relationships for Major/Minor Priority Junctions.* Traffic Engineering & Control, 18(11), pp. 516-519.

Kimber, R. M., I. Summersgill, and I. J. Burrow (1986). *Delay Processes at Unsignalized Junctions: The Interrelation Between Queueing and Geometric and Queueing Delay.* Transportation Research Board, 20B(6), pp. 457-476.

Lassarre, S., P. Lejeune, and J. C. Decret (1991). *Gap Acceptance and Risk Analysis at Unsignalized Intersections.* In: Intersections Without Traffic Signals II, Springer-Verlag, Berlin.

Middelham, F. (1986). *Manual for the Use of FLEXSYT-I With FLEXCOL-76*. Ministerie van Verkeer en Waterstaat, The Hague.

Semmens, M. C. (1985). *PICADY 2: an Enhanced Program to Model Capacities, Queues and Delays at Major/Minor Priority Junctions.* TRRL Report RR36.

Siegloch, W. (1974). *Ein Richtlinienvorschlag Zur Leistungsermittlung an Knotenpunkten Ohne Lichtsignalsteuerung (Capacity Calculations for Unsignalized Intersections)*. Strassenverkehrstechnik, Vol. 1.

Tracz, M. (1988). *Research of Traffic Performance of Major/Minor Priority Intersections.* In: Intersections Without Traffic Signals (Ed.: W. Brilon), Springer Publications, Berlin.

Tracz, M., J. Chodur, and St. Gondek (1990). *The Use of Reserve Capacity and Delay as the Complementary Measures of Junction Effectiveness.* Proceedings of the International Symposium 'Transportation and Traffic Theory", Yokohama.

Tracz, M. and J. Chodur (1991). *Comparative Analysis of Major/Minor Priority Intersection Capacity Methods.* In: Highway Capacity and Level of Service, A. A. Balkema, Rotterdam.

TRB (1991). *Interim Material on Unsignalized Intersection Capacity.* Transportation Research Circular No. 373.

Troutbeck, R. J. (1989). *Evaluating the Performance of a Roundabout*. Australian Road Research Board Special Report, 45.

Troutbeck, R. J. (1990). *Traffic Interactions at Roundabouts.* Proceedings of the 15th Australian Road Research Board Conference, 15(5), pp. 17-42.

Troutbeck, R. J. (1992). *Estimating the Critical Acceptance Gap from Traffic Movements.* Research Report, 92-5.

Turner, D. J., R. Singh, and Y. H. Cheong (1984). *The Development of Empirical Equations for Capacity Analysis at Priority Junctions in Singapore*. Singapore Transport, Vol. 3, No. 4, pp. 17-21.

Vasarhely, B. (1976). *Stochastic Simulation of the Traffic of an Uncontrolled Road Intersection.* Transportation Research, Vol. 10.

TRAFFIC FLOW AT SIGNALIZED INTERSECTIONS

BY NAGUI ROUPHAIL[15]
ANDRZEJ TARKO[16]
JING LI[17]

[15] Professor, Civil Engineering Department, North Carolina State University, Box 7908, Raleigh, NC 276-7908

[16] Assistant Professor, Purdue University, West LaFayette, IN 47907

[17] Principal, TransSmart Technologies, Inc., Madison, WI 53705

Chapter 9 - Frequently used Symbols

$$I = \frac{variance\ of\ the\ number\ of\ arrivals\ per\ cycle}{mean\ number\ of\ arrivals\ per\ cycle}$$

I_i = cumulative lost time for phase i (sec)
L = total lost time in cycle (sec)
q = $A(t)$ = cumulative number of arrivals from beginning of cycle starts until t,
B = index of dispersion for the departure process,

$$B = \frac{variance\ of\ number\ of\ departures\ during\ cycle}{mean\ number\ of\ departures\ during\ cycle}$$

c = cycle length (sec)
C = capacity rate (veh/sec, or veh/cycle, or veh/h)
d = average delay (sec)
d_1 = average uniform delay (sec)
d_2 = average overflow delay (sec)
$D(t)$ = number of departures after the cycle starts until time t (veh)
e_g = green extension time beyond the time to clear a queue (sec)
g = effective green time (sec)
G = displayed green time (sec)
h = time headway (sec)
i = index of dispersion for the arrival process
q = arrival flow rate (veh/sec)
Q_0 = expected overflow queue length (veh)
$Q(t)$ = queue length at time t (veh)
r = effective red time (sec)
R = displayed red time (sec)
S = departure (saturation) flow rate from queue during effective green (veh/sec)
t = time
T = duration of analysis period in time dependent delay models
U = actuated controller unit extension time (sec)
$Var(.)$ = variance of (.)
W_i = total waiting time of all vehicles during some period of time i
x = degree of saturation, $x = (q/S) / (g/c)$, or $x = q/C$
y = flow ratio, $y = q/S$
Y = yellow (or clearance) time (sec)
Δ = minimum headway

9.
TRAFFIC FLOW AT SIGNALIZED INTERSECTIONS

9.1 Introduction

The theory of traffic signals focuses on the estimation of delays and queue lengths that result from the adoption of a signal control strategy at individual intersections, as well as on a sequence of intersections. Traffic delays and queues are principal performance measures that enter into the determination of intersection level of service (LOS), in the evaluation of the adequacy of lane lengths, and in the estimation of fuel consumption and emissions. The following material emphasizes the theory of *descriptive models of traffic flow,* as opposed to prescriptive (i.e. signal timing) models. The rationale for concentrating on descriptive models is that a better understanding of the interaction between demand (i.e. arrival pattern) and supply (i.e. signal indications and types) at traffic signals is a prerequisite to the formulation of optimal signal control strategies. Performance estimation is based on assumptions regarding the characterization of the traffic arrival and service processes. In general, currently used delay models at intersections are described in terms of a deterministic and stochastic component to reflect both the fluid and random properties of traffic flow.

The deterministic component of traffic is founded on the fluid theory of traffic in which demand and service are treated as continuous variables described by flow rates which vary over the time and space domain. A complete treatment of the fluid theory application to traffic signals has been presented in Chapter 5 of the monograph.

The stochastic component of delays is founded on steady-state queuing theory which defines the traffic arrival and service time distributions. Appropriate queuing models are then used to express the resulting distribution of the performance measures. The theory of unsignalized intersections, discussed in Chapter 8 of this monograph, is representative of a purely stochastic approach to determining traffic performance.

Models which incorporate both deterministic (often called uniform) and stochastic (random or overflow) components of traffic performance are very appealing in the area of traffic signals since they can be applied to a wide range of traffic intensities, as well as to various types of signal control. They are approximations of the more theoretically rigorous models, in which delay terms that are numerically inconsequential to the final result have been dropped. Because of their simplicity, they have received greater attention since the pioneering work by Webster (1958) and have been incorporated in many intersection control and analysis tools throughout the world.

This chapter traces the evolution of delay and queue length models for traffic signals. Chronologically speaking, early modeling efforts in this area focused on the adaptation of steady-state queuing theory to estimate the random component of delays and queues at intersections. This approach was valid so long as the average flow rate did not exceed the average capacity rate. In this case, stochastic equilibrium is achieved and expectations of queues and delays are finite and therefore can be estimated by the theory. Depending on the assumptions regarding the distribution of traffic arrivals and departures, a plethora of steady-state queuing models were developed in the literature. These are described in Section 9.3 of this chapter.

As traffic flow rate approaches or exceeds the capacity rate, at least for a finite period of time, the steady-state models assumptions are violated since a state of stochastic equilibrium cannot be achieved. In response to the need for improved estimation of traffic performance in both under and oversaturated conditions, and the lack of a theoretically rigorous approach to the problem, other methods were pursued. A prime example is the time-dependent approach originally conceived by Whiting (unpublished) and further developed by Kimber and Hollis (1979). The time-dependent approach has been adopted in many capacity guides in the U.S., Europe and Australia. Because it is currently in wide use, it is discussed in some detail in Section 9.4 of this chapter.

Another limitation of the steady-state queuing approach is the assumption of certain types of arrival processes (e.g Binomial, Poisson, Compound Poisson) at the signal. While valid in the case of an isolated signal, this assumption does not reflect the impact of adjacent signals and control which may alter the pattern and number of arrivals at a downstream signal. Therefore performance in a system of signals will differ considerably from that at an isolated signal. For example, signal coordination will tend to reduce delays and stops since the arrival process will be different in the red and green portions of the phase. The benefits of coordination are somewhat subdued due to the dispersion of platoons between signals. Further, critical signals in a system could have a metering effect on traffic which proceeds

downstream. This metering reflects the finite capacity of the critical intersection which tends to truncate the arrival distribution at the next signal. Obviously, this phenomenon has profound implications on signal performance as well, particularly if the critical signal is oversaturated. The impact of upstream signals is treated in Section 9.5 of this chapter.

With the proliferation of traffic-responsive signal control technology, a treatise on signal theory would not be complete without reference to their impact on signal performance. The manner in which these controls affect performance is quite diverse and therefore difficult to model in a generalized fashion. In this chapter, basic methodological approaches and concepts are introduced and discussed in Section 9.6. A complete survey of adaptive signal theory is beyond the scope of this document.

9.2 Basic Concepts of Delay Models at Isolated Signals

As stated earlier, delay models contain both deterministic and stochastic components of traffic performance. The deterministic component is estimated according to the following assumptions: a) a zero initial queue at the start of the green phase, b) a uniform arrival pattern at the arrival flow rate (q) throughout the cycle c) a uniform departure pattern at the saturation flow rate (S) while a queue is present, and at the arrival rate when the queue vanishes, and d) arrivals do not exceed the signal capacity, defined as the product of the approach saturation flow rate (S) and its effective green to cycle ratio (g/c). The effective green time is that portion of green where flows are sustained at the saturation flow rate level. It is typically calculated at the displayed green time minus an initial start-up lost time (2-3 seconds) plus an end gain during the clearance interval (2-4 seconds depending on the length of the clearance phase).

A simple diagram describing the delay process in shown in Figure 9.1. The queue profile resulting from this application is shown in Figure 9.2. The area under the queue profile diagram represents the total (deterministic) cyclic delay. Several

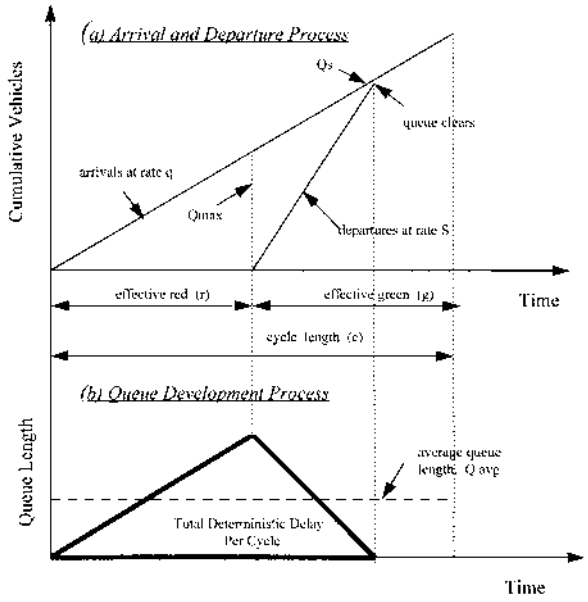


Figure 9.1
Deterministic Component of Delay Models.

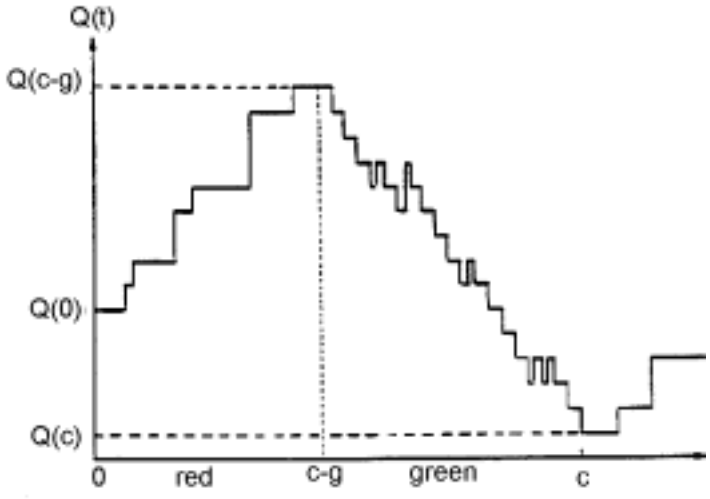


Figure 9.2
Queuing Process During One Signal Cycle
(Adapted from McNeil 1968).

performance measures can be derive including the average delay per vehicle (total delay divided by total cyclic arrivals) the number of vehicle stopped (Q_s), the maximum number of vehicles in the queue (Q_{max}) , and the average queue length (Q_{avg}). Performance models of this type are applicable to low flow to capacity ratios (up to about 0.50), since the assumption of zero initial and end queues is not violated in most cases.

As traffic intensity increases, however, there is a increased likelihood of "cycle failures". That is, some cycles will begin to experience an overflow queue of vehicles that could not discharge from a previous cycle. This phenomenon occurs at random, depending on which cycle happens to experience higher-than-capacity flow rates. The presence of an initial queue (Q_o) causes an additional delay which must be considered in the estimation of traffic performance. Delay models based on queue theory (e.g. M/D/n/FIFO) have been applied to account for this effect.

Interestingly, at extremely congested conditions, the stochastic queuing effect are minimal in comparison with the size of oversaturation queues. Therefore, a fluid theory approach may be appropriate to use for highly oversaturated intersections. This leaves a gap in delay models that are applicable to the range of traffic flows that are numerically close to the signal capacity. Considering that most real-world signals are timed to operate within that domain, the value of time-dependent models are of particular relevance for this range of conditions.

In the case of vehicle actuated control, neither the cycle length nor green times are known in advance. Rather, the length of the green is determined partly by controller-coded parameters such as minimum and maximum green times, and partly by the pattern of traffic arrivals. In the simplest case of a basic actuated controller, the green time is extended beyond its minimum so long as a) the time headway between vehicle arrivals does not exceed the controller's unit extension (U), and b) the maximum green has not been reached. Actuated control models are discussed further in Section 9.6.

9.3 Steady-State Delay Models

9.3.1 Exact Expressions

This category of models attempts to characterize traffic delays based on statistical distributions of the arrival and departure processes. Because of the purely theoretical foundation of the models, they require very strong assumptions to be considered valid. The following section describes how delays are estimated for this class of models, including the necessary data requirements.

The expected delay at fixed-time signals was first derived by Beckman (1956) with the assumption of the binomial arrival process and deterministic service:

$$d = \frac{c-g}{c(1-q/S)}\left[\frac{Q_o}{q}+\frac{c-g+1}{2}\right] \qquad (9.1)$$

where,

c = signal cycle,
g = effective green signal time,
q = traffic arrival flow rate,
S = departure flow rate from queue during green,
Q_o = expected overflow queue from previous cycles.

The expected overflow queue used in the formula and the restrictive assumption of the binomial arrival process reduce the practical usefulness of Equation 9.1. Little (1961) analyzed the expected delay at or near traffic signals to a turning vehicle crossing a Poisson traffic stream. The analysis, however, did not include the effect of turners on delay to other vehicles. Darroch (1964a) studied a single stream of vehicles arriving at a fixed-time signal. The arrival process is the generalized Poisson process with the Index of Dispersion:

$$I = \frac{var(A)}{q\,h} \qquad (9.10)$$

where,

$var(.)$ = variance of (.)
q = arrival flow rate,
h = interval length,
A = number of arrivals during interval $h = qh$.

The departure process is described by a flexible service mechanism and may include the effect of an opposing stream by defining an additional queue length distribution caused by this factor. Although this approach leads to expressions for the expected queue length and expected delay, the resulting models are complex and they include elements requiring further modeling such as the overflow queue or the additional queue component mentioned earlier. From this perspective, the formula is not of practical importance. McNeil (1968) derived a formula for the expected signal delay with the assumption of a general arrival process, and constant departure time. Following his work, we express the total vehicle delay during one signal cycle as a sum of two components

$$W = W_1 + W_2, \qquad (9.3)$$

where

W_1 = total delay experienced in the red phase and
W_2 = total delay experienced in the green phase.

$$W_1 = \int_0^{(c-g)} [Q(0) + A(t)]\,dt \qquad (9.4)$$

and

$$W_2 = \int_{(c-g)}^{c} Q(t)dt \qquad (9.5)$$

where,

$Q(t)$ = vehicle queue at time t,
$A(t)$ = cumulative arrivals at t,

Taking expectations in Equation 9.4 it is found that:

$$E(W_1) = (c-g)\,Q_o + \frac{1}{2}q\,(c-g)^2. \qquad (9.6)$$

Let us define a random variable Z_2 as the total vehicle delay experienced during green when the signal cycle is infinite. The

variable Z_2 is considered as the total waiting time in a busy period for a queuing process $Q(t)$ with compound Poisson arrivals of intensity q, constant service time $1/S$ and an initial system state $Q(t=t_0)$. McNeil showed that provided $q/S<1$:

$$E(Z_2) = \frac{(1 + Iq/S - q/S)\, E[Q(t_0)]}{2S(1 - q/S)^2} + \frac{E[Q^2(t_0)]}{2S(1 - q/S)}. \qquad (9.7)$$

Now W_2 can be expressed using the variable Z_2:

$$E(W_2) = E[Z_2 \mid Q(t=c-g)] - E[Z_2 \mid Q(t=c)] \qquad (9.8)$$

and

$$E[W_2] = \frac{(1 + Ig/S + q/S)\{E[Q(c-g) - Q(c)]\}}{2S(1 - q/S)^2} + \frac{E[Q^2(c-g)] - E[Q^2(c)]}{2S(1 - q/S)}. \qquad (9.9)$$

The queue is in statistical equilibrium, only if the degree of saturation x is below 1,

$$x = \frac{q/S}{g/c} < 1. \qquad (9.10)$$

For the above condition, the average number of arrivals per cycle can discharge in a single green period. In this case $E[\,Q(0)\,] = E[\,Q(c)\,]$ and $E[\,Q^2(0)\,] = E[\,Q^2(c)\,]$. Also $Q(c-g) = Q(0) + A(c)$, so that:

$$E[Q(c-g) - Q(c)] = E[A(c-g)] = q(c-g) \qquad (9.11)$$

and

$$E[Q^2(c-g) - Q^2(c)] = 2E[A(c-g)]E[Q(0)] + E[A^2(c-g)] = 2q(c-g)Q_o + q^2(c-g)^2 + q(c-g)I \qquad (9.12)$$

Equations 9.9, 9.11, and 9.12 yield:

$$E(W_2) = \frac{1}{2S(1 - q/S)^2}[(1 + Iq/S - q/S)g(c-g) + (1 - q/S)(2q(c-g)Q_o + q^2(c-g)^2 + q(c-g)I)] \qquad (9.13)$$

and using Equations 9.3, 9.4 and 9.13, the following is obtained:

$$E(W) = \frac{(c-g)}{c(1 - q/S)}\left[\frac{Q_0}{q} + \left(\frac{c-g}{2}\right)qc + \frac{1}{S}\left(1 + \frac{I}{1-q/S}\right)\right] \qquad (9.14)$$

The average vehicle delay d is obtained by dividing $E(W)$ by the average number of vehicles in the cycle (qc):

$$d = \frac{c-g}{2c(1 - q/S)}[(c-g) + \frac{2}{q}Q_o + \frac{1}{S}(1 + \frac{I}{1-q/S})] \qquad (9.15)$$

which is in essence the formula obtained by Darroch when the departure process is deterministic. For a binomial arrival process $I=1-q/S$, and Equation 9.15 becomes identical to that obtained by Beckmann (1956) for binomial arrivals. McNeil and Weiss (in Gazis 1974) considered the case of the compound Poisson arrival process and general departure process obtaining the following model:

$$d = \frac{(c-g)}{2c(1-q/S)}\left\{(c-g) + \frac{2}{q}\left[1 + \frac{(1-q/S)(1-B^2)}{2S}\right]Q_o + \frac{1}{S}(1 + \frac{I + B^2 q/S}{1-q/S})\right\} \qquad (9.16)$$

An examination of the above equation indicates that in the case of no overflow ($Q_o = 0$), and no randomness in the traffic process ($I=0$), the resultant delay becomes the uniform delay component. This component can be derived from a simple input-output model of uniform arrivals throughout the cycle and departures as described in Section 9.2. The more general case in Equation 9.16 requires knowledge of the size of the average overflow queue (or queue at the beginning of green), a major limitation on the practical usefulness of the derived formulae, since these are usually unknown.

A substantial research effort followed to obtain a closed-form analytical estimate of the overflow queue. For example, Haight (1959) specified the conditional probability of the overflow queue at the end of the cycle when the queue at the beginning of the cycle is known, assuming a homogeneous Poisson arrival process at fixed traffic signals. The obtained results were then modified to the case of semi-actuated signals. Shortly thereafter, Newell (1960) utilized a bulk service queuing model with an underlying binomial arrival process and constant departure time, using generating function technique. Explicit expressions for overflow queues were given for special cases of the signal split.

Other related work can be found in Darroch (1964a) who used a more general arrival distribution but did not produce a closed form expression of queue length, and Kleinecke (1964), whose work included a set of exact but complicated series expansion for Q_o, for the case of constant service time and Poisson arrival process.

9.3.2 Approximate Expressions

The difficulty in obtaining exact expressions for delay which are reasonably simple and can cover a variety of real world conditions, gave impetus to a broad effort for signal delay estimation using approximate models and bounds. The first, widely used approximate delay formula was developed by Webster (1961, reprint of 1958 work with minor amendments) from a combination of theoretical and numerical simulation approaches:

$$d = \frac{c(1-g/c)^2}{2[1-(g/c)x]} + \frac{x^2}{2q(1-x)} - 0.65\left(\frac{c}{q^2}\right)^{\frac{1}{3}} x^{2+5(g/c)} \qquad (9.17)$$

where,

d = average delay per vehicle (sec),
c = cycle length (sec),
g = effective green time (sec),
x = degree of saturation (flow to capacity ratio),
q = arrival rate (veh/sec).

The first term in Equation 9.17 represents delay when traffic can be considered arriving at a uniform rate, while the second term makes some allowance for the random nature of the arrivals. This is known as the "random delay", assuming a Poisson arrival process and departures at constant rate which corresponds to the signal capacity. The latter assumption does not reflect actual signal performance, since vehicles are served only during the effective green, obviously at a higher rate than the capacity rate. The third term, calibrated based on simulation experiments, is a corrective term to the estimate, typically in the range of 10 percent of the first two terms in Equation 9.17.

Delays were also estimated indirectly, through the estimation of Q_o, the average overflow queue. Miller (1963) for example obtained a approximate formulae for Q_o that are applicable to any arrival and departure distributions. He started with the general equality true for any general arrival and departure processes:

$$Q(c) = Q(0) + A - C + \Delta C \qquad (9.18)$$

where,

$Q(c)$ = vehicle queue at the end of cycle,
$Q(0)$ = vehicle queue at the beginning of cycle,
A = number of arrivals during cycle,
C = maximum possible number of departures during green,
ΔC = reserve capacity in cycle equal to $(C\text{-}Q(0)\text{-}A)$ if $Q(0)+A < C$, zero otherwise.

Taking expectation of both sides of Equation 9.18, Miller obtained:

$$E(\Delta C) = E(C-A), \qquad (9.19)$$

since in equilibrium $Q(0) = Q(c)$.

Now Equation 9.18 can be rewritten as:

$$Q(c) - [\Delta C - E(\Delta C)] = Q(0) - [C - A - E(C-A)] \qquad (9.20)$$

Squaring both sides, taking expectations, the following is obtained:

$$E[Q(c)]^2 + 2E[Q(c)]E(\Delta C] + Var(\Delta C) = E[Q(0)]^2 + Var(C-A) \qquad (9.21)$$

For equilibrium conditions, Equation 9.21 can be rearranged as follows:

$$Q_o = \frac{Var(C-A) - Var(\Delta C)}{2E(C-A)} \qquad (9.22)$$

where,

C = maximum possible number of departures in one cycle,
A = number of arrivals in one cycle,
ΔC = reserve capacity in one cycle.

The component $Var(\Delta C)$ is positive and approaches 0 when $E(C)$ approaches $E(A)$. Thus an upper bound on the expected overflow queue is obtained by deleting that term. Thus:

$$Q_o \leq \frac{Var(C-A)}{2E(C-A)} \qquad (9.23)$$

For example, using Darroch's arrival process (i.e. $E(A)=qc$, $Var(A)=Iqc$) and constant departure time during green ($E(C)=Sg$, $Var(C)=0$) the upper bound is shown to be:

$$Q_o \leq \frac{Ix}{2(1-x)} \qquad (9.24)$$

where $x=(qc)/(Sg)$.

Miller also considered an approximation of the excluded term $Var(\Delta C)$. He postulates that:

$$I \approx \frac{Var(\Delta C)}{E(C-A)} \qquad (9.25)$$

and thus, an approximation of the overflow queue is

$$Q_o \approx \frac{(2x-1)I}{2(1-x)} \quad , \quad x \geq 0.50 \qquad (9.26)$$

which can now be substituted in Equation 9.15. Further approximations of Equation 9.15 were aimed at simplifying it for practical purposes by neglecting the third and fourth terms which are typically of much lower order of magnitude than the first two terms. This approach is exemplified by Miller (1968a) who proposed the approximate formula:

$$d = \frac{(1-g/c)}{2(1-q/s)}\left[c(1-g/c) + \frac{2Q_0}{q}\right] \qquad (9.27)$$

which can be obtained by deleting the second and third terms in McNeil's formula 9.15. Miller also gave an expression for the overflow queue formula under Poisson arrivals and fixed service time during the green:

$$Q_o = \frac{\exp\left[-1.33\sqrt{Sg(1-x)/x}\right]}{2(1-x)}. \qquad (9.28)$$

Equations 9.15, 9.16, 9.17, 9.27, and 9.28 are limited to specific arrival and departure processes. Newell (1965) aimed at developing delay formulae for general arrival and departure distributions. First, he concluded from a heuristic graphical argument that for most reasonable arrival and departure processes, the total delay per cycle differs from that calculated with the assumption of uniform arrivals and fixed service times (Clayton, 1941), by a negligible amount if the traffic intensity is sufficiently small. Then, by assuming a queue discipline LIFO (Last In First Out) which does not effect the average delay estimate, he concluded that the expected delay when the traffic is sufficiently heavy can be approximated:

$$d = \frac{c(1-g/c)^2}{2(1-q/S)} + \frac{Q_o}{q}. \qquad (9.29)$$

This formula gives identical results to formula (Equation 9.15) if one neglects components of $1/S$ order in (Equation 9.15) and when $1-q/S=1-g/c$. The last condition, however, is never met if equilibrium conditions apply. To estimate the overflow queue, Newell (1965) defines F_Q as the cumulative distribution of the overflow queue length, F_{A-D} as the cumulative distribution of the overflow in the cycle, where the indices A and D represent cumulative arrivals and departures, respectively. He showed that under equilibrium conditions:

$$F_Q(x) = \int_0^{\infty} F_Q(z) dF_{A-D}(x-z) \qquad (9.30)$$

The integral in Equation 9.30 can be solved only under the restrictive assumption that the overflow in a cycle is normally distributed. The resultant Newell formula is as follows:

$$Q_o = \frac{qc(1-x)}{\Pi} \int_0^{\Pi/2} \frac{\tan^2\theta}{-1+\exp[Sg(1-x)^2/(2\cos^2\theta)]} d\theta. \qquad (9.31)$$

A more convenient expression has been proposed by Newell in the form:

$$Q_o = \frac{IH(\mu)x}{2(1-x)}. \qquad (9.32)$$

where,

$$\mu = \frac{Sg-qc}{(ISg)^{1/2}}. \qquad (9.33)$$

The function $H(\mu)$ has been provided in a graphical form.

Moreover, Newell compared the results given by expressions (Equation 9.29) and (Equation 9.31) with Webster's formula and added additional correction terms to improve the results for medium traffic intensity conditions. Newell's final formula is:

$$d = \frac{c(1-g/c)^2}{2(1-q/S)} + \frac{Q_o}{q} + \frac{(1-g/c)I}{2S(1-q/S)^2}. \qquad (9.34)$$

Table 9.1
Maximum Relative Discrepancy between the Approximate Expressions and Ohno's Algorithm (Ohno 1978).

Approximate Expressions (Equation #, Q_0 computed according to Equation #)	Range of y = 0.0 ~ 0.5				Range of *g/c* = 0.4 ~ 1.0	
	s = 0.5 v/s		s = 1.5 v/s		s = 0.5 v/s	s = 1.5 v/s
	c = 90 s	c = 30 s	c = 90 s	c = 30 s	c = 30 s	c = 90 s
	g = 46 s	g = 16 s	g = 45.33 s	g = 15.33 s	q = 0.2 s	q = 0.6 s
Modified Miller's expression (9.15, 9.28)	0.22	2.60	-0.53	0.22	2.24	0.26
Modified Newell's expression (9.15, 9.31)	0.82	2.53	0.25	0.82	2.83	0.25-
McNeil's expression (9.15, Miller 1969)	0.49	1.79	0.12	0.49	1.51	0.08
Webster's full expression (9.17)	-8.04	-21.47	3.49	-7.75	119.24	1381.10
Newell's expression (9.34, 9.31)	-4.16	10.89	-1.45	-4.16	-15.37	-27.27

More recently, Cronje (1983b) proposed an analytical approximation of the function $H(\mu)$:

$$H(\mu) = \exp[-\mu-(\mu^2/2)] \qquad (9.35)$$

where,

$$\mu=(1-x)(Sg)^{1/2}. \qquad (9.36)$$

He also proposed that the correction (third) component in Equation 9.34 could be neglected.

Earlier evaluations of delay models by Allsop (1972) and Hutchinson (1972) were based on the Webster model form. Later on, Ohno (1978) carried out a comparison of the existing delay formulae for a Poisson arrival process and constant departure time during green. He developed a computational procedure to provide the basis for evaluating the selected models, namely McNeil's expression, Equation 9.15 (with overflow queue calculated with the method described by Miller 1969), McNeil's formula with overflow queue according to Miller (Equation 9.28) (modified Miller's expression), McNeil's formula with overflow queue according to Newell (Equation 9.31) (modified Newell's expression), Webster expression (Equation 9.17) and the original Newell expression (Equation 9.34). Comparative results are depicted in Table 9.1 and Figures 9.3 and 9.4. Newell's expression appear to be more accurate than Webster, a conclusion shared by Hutchinson (1972) in his evaluation of three simplified models (Newell, Miller, and Webster). Figure 9.3 represents the percentage relative errors of the approximate delay models measured against Ohno's algorithm (Ohno 1978) for a range of flow ratios. The modified Miller's and Newell's expressions give almost exact average delay values, but they are not superior to the original McNeil formula. Figure 9.4 shows the same type of errors, categorized by the g/c ratio. Further efforts to improve on their estimates will not give any appreciable reduction in the errors. The modified Miller expression was recommended by Ohno because of its simpler form compared to McNeil's and Newell's.

9.4 Time-Dependent Delay Models

The stochastic equilibrium assumed in steady-state models requires an infinite time period of stable traffic conditions (arrival, service and control processes) to be achieved. At low flow to capacity ratios equilibrium is reached in a reasonable period of time, thus the equilibrium models are an acceptable approximation of the real-world process. When traffic flow approaches signal capacity, the time to reach statistical equilibrium usually exceeds the period over which demand is sustained. Further, in many cases the traffic flow exceeds capacity, a situation where steady-state models break down. Finally, traffic flows during the peak hours are seldom stationary, thus violating an important assumption of steady-state models. There has been many attempts at circumventing the limiting assumption of steady-state conditions. The first and simplest way is to deal with arrival and departure rates as a function of time in a deterministic fashion. Another view is to model traffic at signals, assuming stationary arrival and departure processes but not necessarily under stochastic equilibrium conditions, in order to estimate the average delay and queues over the modeled period of time. The latter approach approximates the time-dependent arrival profile by some mathematical function (step-function, parabolic, or triangular functions) and calculates the corresponding delay. In May and Keller (1967) delay and queues are calculated for an unsignalized bottleneck. Their work is nevertheless representative of the deterministic modeling approach and can be easily modified for signalized intersections. The general assumption in their research is that the random queue fluctuations can be neglected in delay calculations. The model defines a cumulative number of arrivals $A(t)$:

$$A(t)=\int_0^t q(\tau)d\tau \qquad (9.37)$$

and departures $D(t)$ under continuous presence of vehicle queue over the period $[0,t]$:

$$D(t) = \int_0^t S(\tau)d\tau \qquad (9.38)$$

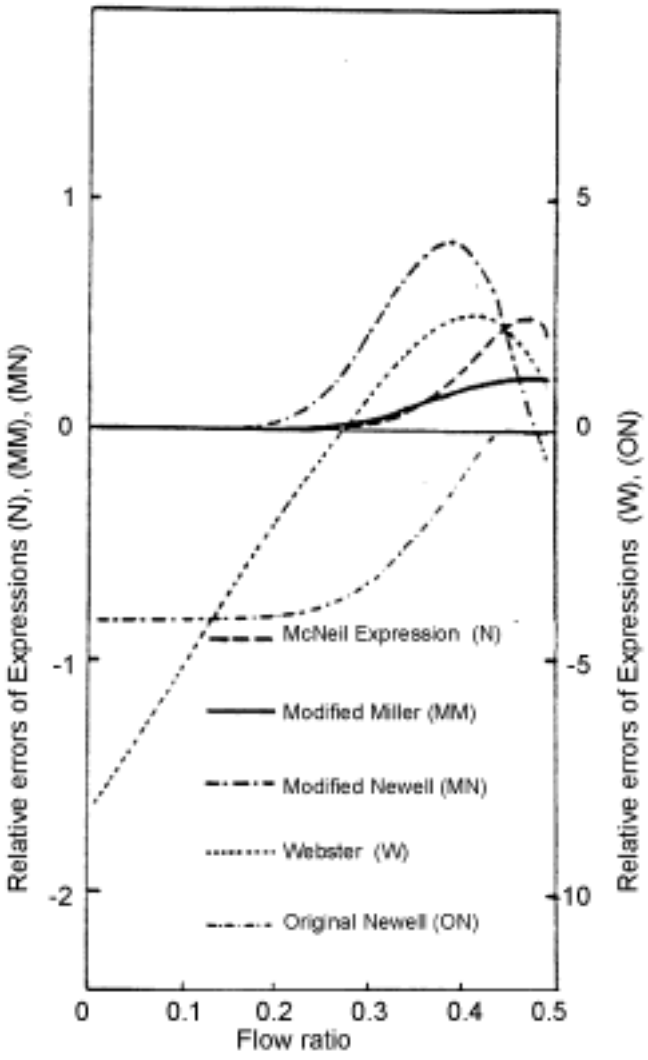


Figure 9.3
Percentage Relative Errors for Approximate Delay Models by Flow Ratios (Ohno 1978).

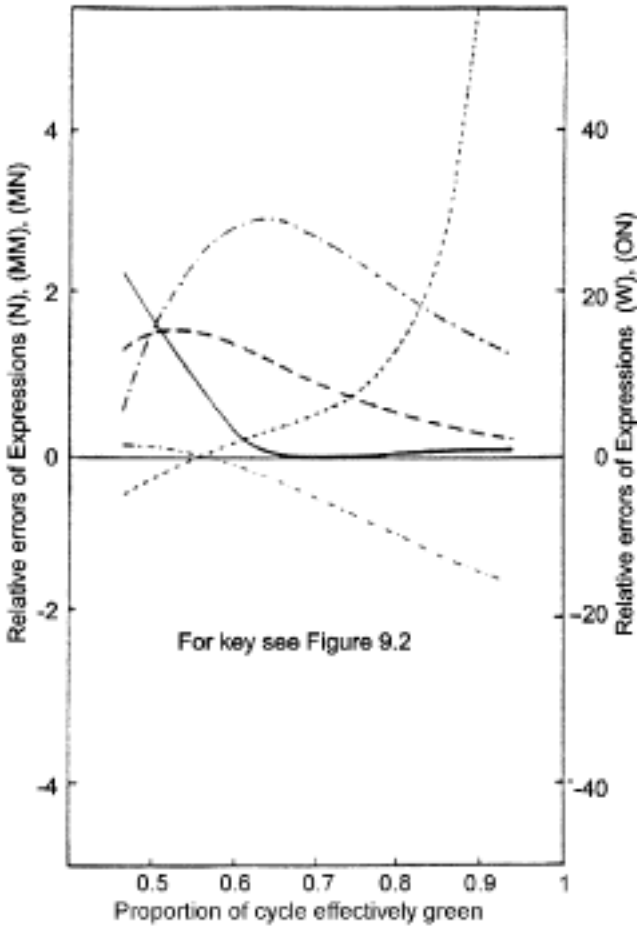


Figure 9.4
Relative Errors for Approximate Delay Models by Green to Cycle Ratios (Ohno 1978).

The current number of vehicles in the system (queue) is

$$Q(t) = Q(0) + A(t) - D(t) \qquad (9.39)$$

and the average delay of vehicles queuing during the time period $[0,T]$ is

$$d = \frac{1}{A(T)} \int_0^T Q(t)dt \qquad (9.40)$$

The above models have been applied by May and Keller to a trapezoidal-shaped arrival profile and constant departure rate. One can readily apply the above models to a signal with known signal states over the analysis period by substituting $C(\tau)$ for $S(\tau)$ in Equation 9.38:

$C(\tau)$ = 0 if signal is red,
= $S(\tau)$ if signal is green and $Q(\tau) > 0$,
= $q(\tau)$ if signal is green and $Q(\tau) = 0$.

Deterministic models of a single term like Equation 9.39 yield acceptable accuracy only when $x<<1$ or $x>>1$. Otherwise, they tend to underestimate queues and delays since the extra queues causes by random fluctuations in q and C are neglected.

According to Catling (1977), the now popular coordinate transformation technique was first proposed by Whiting, who did not publish it. The technique when applied to a steady-state curve derived from standard queuing theory, produces a time-dependent formula for delays. Delay estimates from the new models when flow approaches capacity are far more realistic than those obtained from the steady-state model. The following observations led to the development of this technique.

- At low degree of saturation ($x<<1$) delay is almost equal to that occurring when the traffic intensity is uniform (constant over time).
- At high degrees of saturation ($x>>1$) delay can be satisfactory described by the following deterministic model with a reasonable degree of accuracy:

$$d = d_1 + \frac{T}{2}(x-1) \qquad (9.41)$$

where d_1 is the delay experienced at very low traffic intensity, (uniform delay) T = analysis period over which flows are sustained.

- steady-state delay models are asymptotic to the y-axis (i.e generate infinite delays) at unit traffic intensity (x=1). The coordinate transformation method shifts the original steady-state curve to become asymptotic to the deterministic oversaturation delay line--i.e.-- the second term in Equation 9.41--see Figure 9.5. The horizontal distance between the proposed delay curve and its asymptote is the same as that between the steady-state curve and the vertical line x=1.

There are two restrictions regarding the application of the formula: (1) no initial queue exists at the beginning of the interval $[0,T]$, (2) traffic intensity is constant over the interval $[0,T]$. The time-dependent model behaves reasonably within the period $[0,T]$ as indicated from simulation experiments. Thus, this technique is very useful in practice. Its principal drawback, in addition to the above stated restrictions (1) and (2) is the lack of a theoretical foundation. Catling overcame the latter difficulties by approximating the actual traffic intensity profile with a step-function. Using an example of the time-dependent version of the Pollaczek-Khintchine equation (Taha 1982), he illustrated the calculation of average queue and delay for each time interval starting from an initial, non-zero queue.

Kimber and Hollis (1979) presented a computational algorithm to calculate the expected queue length for a system with random arrivals, general service times and single channel service (M/G/1). The initial queue can be defined through its distribution. To speed up computation, the average initial queue is used unless it is substantially different from the queue at equilibrium. In this case, the full computational algorithm should be applied. The non-stationary arrival process is approximated with a step-function. The total delay in a time period is calculated by integrating the queue size over time. The coordinate transformation method is described next in some detail.

Suppose, at time T=0 there are $Q(0)$ waiting vehicles in queue and that the degree of saturation changes rapidly to x. In a deterministic model the vehicle queue changes as follows:

$$Q(T) = Q(0) + (x-1)CT. \qquad (9.42)$$

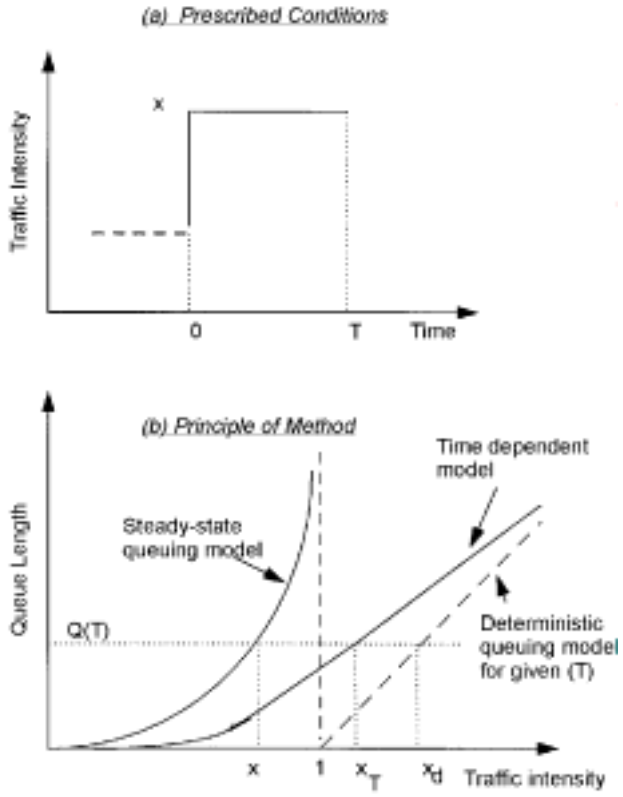


Figure 9.5
The Coordinate Transformation Method.

The steady-state expected queue length from the modified Pollaczek-Khintczine formula is:

$$Q = x + \frac{Bx^2}{1-x} \qquad (9.43)$$

where B is a constant depending on the arrival and departure processes and is expressed by the following equation.

$$B = 0.5\left(1 + \frac{\sigma^2}{\mu^2}\right) \qquad (9.44)$$

where σ^2 and μ are the variance and mean of the service time distribution, respectively.

The following derivation considers the case of exponential service times, for which $\sigma^2 = \mu^2$, $B = 1$. Let x_d be the degree of saturation in the deterministic model (Equation 9.42), x refers to the steady-state conditions in model (Equation 9.44), while x_T refers to the time-dependent model such that $Q(x,T)=Q(x_T,T)$. To meet the postulate of equal distances between the curves and the appropriate asymptotes, the following is true from Figure 9.5:

$$1 - x = x_d - x_T \qquad (9.45)$$

and hence

$$x = x_T - (x_d - 1) \qquad (9.46)$$

and from Equation 9.42:

$$x_d = \frac{Q(T) - Q(0)}{CT} + 1, \qquad (9.47)$$

the transformation is equivalent to setting:

$$x = x_T - \frac{Q(T) - Q(0)}{CT}. \qquad (9.48)$$

From Figure 9.5, it is evident that the queue length at time T, $Q(T)$ is the same at x, x_T, and x_d. By substituting for $Q(T)$ in Equation 9.44, and rewriting Equation 9.48 gives:

$$\frac{Q(T)}{1+Q(T)} = x_T - \frac{Q(T)-Q(0)}{CT} \qquad (9.49)$$

By eliminating the index T in x_T and solving the second degree polynomial in Equation 9.49 for $Q(T)$, it can be shown that:

$$Q(T) = \frac{1}{2}[(a^2+b)^{1/2} - a] \qquad (9.50)$$

where

$$a = (1-x)CT + 1 - Q(0) \qquad (9.51)$$

and

$$b = 4[Q(0) + xCT]. \qquad (9.52)$$

If the more general steady state Equation 9.43 is used, the result for Equation 9.51 and 9.52 is:

$$a = \frac{(1-x)(CT)^2 + [1-Q(0)]CT - 2(1-B)[Q(0)+xCT]}{CT+(1-B)} \qquad (9.53)$$

and

$$b = \frac{4[Q(0) + xCT][CT - (1-B)(Q(0) + xCT)]}{CT + (1-B)}. \qquad (9.54)$$

The equation for the average delay for vehicles arriving during the period of analysis is also derived starting from the average delay per arriving vehicle d_d over the period $[0,T]$,

$$d_d = \frac{[Q(0)+1] + \frac{1}{2}(x-1)CT}{C} \qquad (9.55)$$

and the steady-state delay d_s,

$$d_s = \frac{1}{C}\left(1 + \frac{Bx}{1-x}\right). \qquad (9.56)$$

The transformed time dependent equation is

$$d = \frac{1}{2}[(a^2+b)^{1/2} - a] \qquad (9.57)$$

with the corresponding parameters:

$$a = \frac{T}{2}(1-x) - \frac{1}{C}[Q(0) - B + 2] \qquad (9.58)$$

and

$$b = \frac{4}{C}\left[\frac{T}{2}(1-x) + \frac{1}{2}xTB - \frac{Q(0)+1}{C}(1-B)\right]. \qquad (9.59)$$

The derivation of the coordinate transformation technique has been presented. The steady-state formula (Equation 9.43) does not appear to adequately reflect traffic signal performance, since a) the first term (queue for uniform traffic) needs further elaboration and b) the constant B must be calibrated for cases that do not exactly fit the assumptions of the theoretical queuing models.

Akçelik (1980) utilized the coordinate transformation technique to obtain a time-dependent formula which is intended to be more applicable to signalized intersection performance than Kimber-Hollis's. In order to facilitate the derivation of a time-dependent function for the average overflow queue Q_o, Akçelik used the following expression for undersaturated signals as a simple approximation to Miller's second formula for steady-state queue length (Equation 9.28):

$$Q_o = \begin{cases} \dfrac{1.5(x-x_o)}{1-x} & when\ x>x_o, \\ 0 & otherwise \end{cases} \qquad (9.60)$$

where

$$x_o = 0.67 + \frac{Sg}{600} \qquad (9.61)$$

Akçelik's time-dependent function for the average overflow queue is

$$Q_o = \begin{cases} \dfrac{CT}{4}\left[(x-1)+\sqrt{(x-1)^2+\dfrac{12(x-x_o)}{CT}}\right] & when\ x>x_o, \\ 0 & otherwise. \end{cases} \qquad (9.62)$$

The formula for the average uniform delay during the interval $[0,T]$ for vehicles which arrive in that interval is

$$d = \left\{ \begin{array}{ll} \dfrac{c(1-g/c)^2}{2(1-q/S)} & when\ x<1 \\ (c-g)/2 & when\ x\geq 1 \end{array} \right\} + \frac{Q_o}{C}. \qquad (9.63)$$

Generalizations of Equations 9.60 and 9.61 were discussed by Akçelik (1988) and Akçelik and Rouphail (1994). It should be noted that the average overflow queue, Q_0 is an approximation of the McNeil (Equation 9.15) and Miller (Equation 9.28) formulae applied to the time-dependent conditions, and differs from Newell's approximations Equation 9.29 and Equation 9.34 of the steady-state conditions. According to Akçelik (1980), this approximation is relevant to high degrees of saturation x and its effect is negligible for most practical purposes.

Following certain aspects of earlier works by Haight (1963), Cronje (1983a), and Miller (1968a); Olszewski (1990) used non-homogeneous Markov chain techniques to calculate the stochastic queue distribution using the arrival distribution *P(t,A)* and capacity distribution *P(C)*. Probabilities of transition from a queue of i to j vehicles during one cycle are expressed by the following equation:

$$P_{i,j}(t) = \sum_{C=0}^{\infty} P_{i,j}(t,C)P(C) \qquad (9.64)$$

and

$$P_{i,0}(t,C) = \begin{cases} \displaystyle\sum_{k=0}^{C-i} P(t,A=k) & when\ i\leq C, \\ 0 & otherwise \end{cases} \qquad (9.65)$$

and

$$P_{i,j}(t,C) = \begin{cases} P(t,A=j-i+C) & when\ j\geq i-C, \\ 0 & otherwise. \end{cases} \qquad (9.66)$$

The probabilities of queue states transitions at time t form the transition matrix *P(t)*. The system state at time t is defined with the overflow queue distribution in the form of a row vector $P_Q(t)$. The initial system state variable distribution at time $t=0$ is assumed to be known: $P_Q(0)=[P_1(0), P_2(0),...P_m(0)]$, where $P_i(0)$ is the probability of queue of length i at time zero. The vector of state probabilities in any cycle t can now be found by matrix multiplication:

$$P_Q(t) = P_Q(t-1)\,P(t). \qquad (9.67)$$

Equation 9.67, when applied sequentially, allows for the calculation of queue probability evolution from any initial state.

In their recent work, Brilon and Wu (1990) used a similar computational technique to Olszewski's (1990a) in order to

evaluate existing time-dependent formulae by Catling (1977), Kimber-Hollis (1979), and Akçelik (1980). A comparison of the models results is given in Figures 9.6 and 9.7 for a parabolic arrival rate profile in the analysis period T_o. They found that the Catling method gives the best approximation of the average delay. The underestimation of delays observed in the Akçelik's model is interpreted as a consequence of the authors' using an average arrival rate over the analyzed time period instead of the step function, as in the Catling's method. When the peak flow rate derived from a step function approximation of the parabolic profile is used in Akçelik's formula, the results were virtually indistinguishable from Brilon and Wu's (Akçelik and Rouphail 1993).

Using numeric results obtained from the Markov Chains approach, Brilon and Wu developed analytical approximate (and rather complicated) delay formulae of a form similar to Akçelik's which incorporate the impact of the arrival profile shape (e.g. the peaking intensity) on delay. In this examination of delay models in the time dependent mode, delay is defined according to the path trace method of measurement (Rouphail and Akçelik 1992a). This method keeps track of the departure time of each vehicle, even if this time occurs beyond the analysis period *T*. The path trace method will tend to generate delays that are typically longer than the queue sampling method, in which stopped vehicles are sampled every 15-20 seconds for the duration of the analysis period. In oversaturated conditions, the measurement of delay may yield vastly different results as vehicles may discharge 15 or 30 minutes beyond the analysis period. Thus it is important to maintain consistency between delay measurements and estimation methods. For a detailed discussion of the delay measurement methods and their impact on oversaturation delay estimation, the reader is referred to Rouphail and Akçelik (1992a).

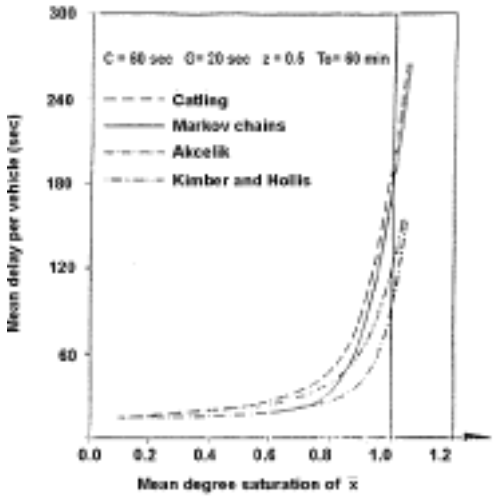


Figure 9.6
Comparison of Delay Models Evaluated by Brilon and Wu (1990) with Moderate Peaking (z=0.50).

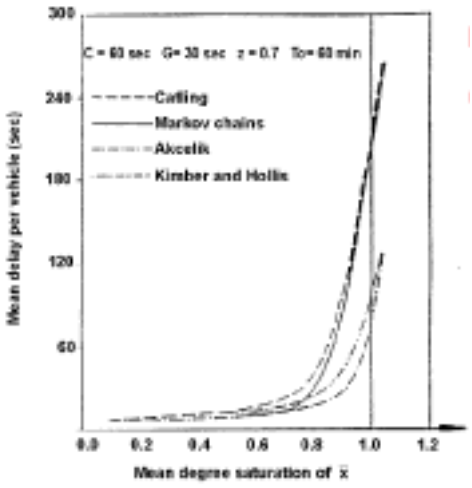


Figure 9.7
Comparison of Delay Models Evaluated by Brilon and Wu (1990) with High Peaking (z=0.70).

9.5 Effect of Upstream Signals

The arrival process observed at a point located downstream of some traffic signal is expected to differ from that observed upstream of the same signal. Two principal observations are made: a) vehicles pass the signal in "bunches" that are separated by a time equivalent to the red signal (platooning effect), and b) the number of vehicles passing the signal during one cycle does not exceed some maximum value corresponding to the signal throughput (filtering effect).

9.5.1 Platooning Effect On Signal Performance

The effect of vehicle bunching weakens as the platoon moves downstream, since vehicles in it travel at various speeds, spreading over the downstream road section. This phenomenon, known as platoon diffusion or dispersion, was modeled by Pacey (1956). He derived the travel time distribution $f(\tau)$ along a road section assuming normally distributed speeds and unrestricted overtaking:

$$f(\tau) = \frac{D}{\tau^2 \sigma \sqrt{2\pi}} \exp\left[- \frac{(\frac{D}{\tau} - \frac{D}{\bar{\tau}})^2}{2\sigma^2} \right] \qquad (9.68)$$

where,

D = distance from the signal to the point where arrivals are observed,
τ = individual vehicle travel time along distance D,
$\bar{\tau}$ = mean travel time, and
σ = standard deviation of speed.

The travel time distribution is then used to transform a traffic flow profile along the road section of distance D:

$$q_2(t_2)\, dt_2 = \int_{t_1} q_1(t_1)\, f(t_2 - t_1)\, dt_1\, dt_2 \qquad (9.69)$$

where,

$q_2(t_2)dt_2$ = total number of vehicles passing some point downstream of the signal in the interval $(t, t+dt)$,
$q_1(t_1)dt_1$ = total number of vehicles passing the signal in the interval $(t, t+dt)$, and
$f(t_2-t_1)$ = probability density of travel time $(t_2 - t_1)$ according to Equation 9.68.

The discrete version of the diffusion model in Equation 9.69 is

$$q_2(j) = \sum_i q_1(i)\, g(j-i) \qquad (9.70)$$

where i and j are discrete intervals of the arrival histograms.

Platoon diffusion effects were observed by Hillier and Rothery (1967) at several consecutive points located downstream of signals (Figure 9.8). They analyzed vehicle delays at pretimed signals using the observed traffic profiles and drew the following conclusions:

- the deterministic delay (first term in approximate delay formulae) strongly depends on the time lag between the start of the upstream and downstream green signals (offset effect);
- the minimum delay, observed at the optimal offset, increases substantially as the distance between signals increases; and
- the signal offset does not appear to influence the overflow delay component.

The TRANSYT model (Robertson 1969) is a well-known example of a platoon diffusion model used in the estimation of deterministic delays in a signalized network. It incorporates the Robertson's diffusion model, similar to the discrete version of the Pacey's model in Equation 9.70, but derived with the assumption of the binomial distribution of vehicle travel time:

$$q_2(j) = \frac{1}{1+a\tau}\, q_1(j) + (1 - \frac{1}{1+a\tau})\, q_2(j-1) \qquad (9.71)$$

where τ is the average travel time and a is a parameter which must be calibrated from field observations. The Robertson model of dispersion gives results which are satisfactory for the

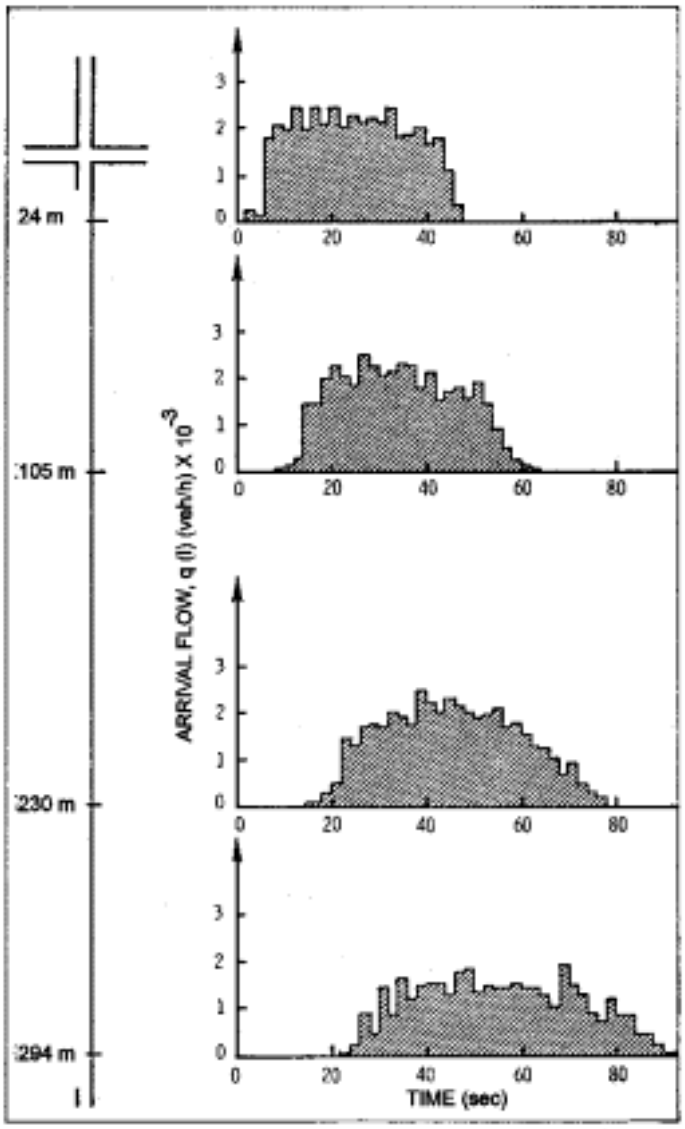


Figure 9.8
Observations of Platoon Diffusion
by Hillier and Rothery (1967).

purpose of signal optimization and traffic performance analysis in signalized networks. The main advantage of this model over the former one is much lower computational demand which is a critical issue in the traffic control optimization for a large size network.

In the TRANSYT model, a flow histogram of traffic served (departure profile) at the stopline of the upstream signal is first constructed, then transformed between two signals using model (Equation 9.71) in order to obtain the arrival patterns at the stopline of the downstream signal. Deterministic delays at the downstream signal are computed using the transformed arrival and output histograms.

To incorporate the upstream signal effect on vehicle delays, the Highway Capacity Manual (TRB 1985) uses a progression factor (*PF*) applied to the delay computed assuming an isolated signal. A PF is selected out of the several values based on a platoon ratio f_p. The platoon ratio is estimated from field measurement and by applying the following formula:

$$f_p = \frac{PVG}{g/c} \tag{9.72}$$

where,

PVG = percentage of vehicles arriving during the effective green,
g = effective green time,
c = cycle length.

Courage et al. (1988) compared progression factor values obtained from Highway Capacity Manual (HCM) with those estimated based on the results given by the TRANSYT model. They indicated general agreement between the methods, although the HCM method is less precise (Figure 9.9). To avoid field measurements for selecting a progression factor, they suggested to compute the platoon ratio f_p from the ratios of bandwidths measured in the time-space diagram. They showed that the proposed method gives values of the progression factor comparable to the original method.

Rouphail (1989) developed a set of analytical models for direct estimation of the progression factor based on a time-space diagram and traffic flow rates. His method can be considered a simplified version of TRANSYT, where the arrival histogram consists of two uniform rates with in-platoon and out-of-platoon traffic intensities. In his method, platoon dispersion is also based on a simplified TRANSYT-like model. The model is thus sensitive to both the size and flow rate of platoons. More recently, empirical work by Fambro et al. (1991) and theoretical analyses by Olszewski (1990b) have independently confirmed the fact that signal progression does not influence overflow queues and delays. This finding is also reflected in the most recent update of the Signalized Intersections chapter of the Highway Capacity Manual (1994). More recently, Akçelik (1995a) applied the HCM progression factor concept to queue length, queue clearance time, and proportion queued at signals.

The remainder of this section briefly summarizes recent work pertaining to the filtering effect of upstream signals, and the resultant overflow delays and queues that can be anticipated at downstream traffic signals.

9.5.2 Filtering Effect on Signal Performance

The most general steady-state delay models have been derived by Darroch (1964a), Newell (1965), and McNeil (1968) for the binomial and compound Poisson arrival processes. Since these efforts did not deal directly with upstream signals effect, the question arises whether they are appropriate for estimating overflow delays in such conditions. Van As (1991) addressed this problem using the Markov chain technique to model delays and arrivals at two closely spaced signals. He concluded that the Miller's model (Equation 9.27) improves random delay estimation in comparison to the Webster model (Equation 9.17). Further, he developed an approximate formula to transform the dispersion index of arrivals, I, at some traffic signal into the dispersion index of departures, B, from that signal:

$$B = I \exp(-1.3 F^{0.627}) \tag{9.73}$$

with the factor F given by

$$F = \frac{Q_o}{\sqrt{I_a q c}} \tag{9.74}$$

This model (Equation 9.73) can be used for closely spaced signals, if one assumes the same value of the ratio I along a road section between signals.

Tarko et al. (1993) investigated the impact of an upstream signal on random delay using cycle-by-cycle macrosimulation. They found that in some cases the ratio I does not properly represent the non-Poisson arrival process, generally resulting in delay overestimation (Figure 9.10).

They proposed to replace the dispersion index I with an adjustment factor f which is a function of the difference between the maximum possible number of arrivals m_c observable during one cycle, and signal capacity Sg:

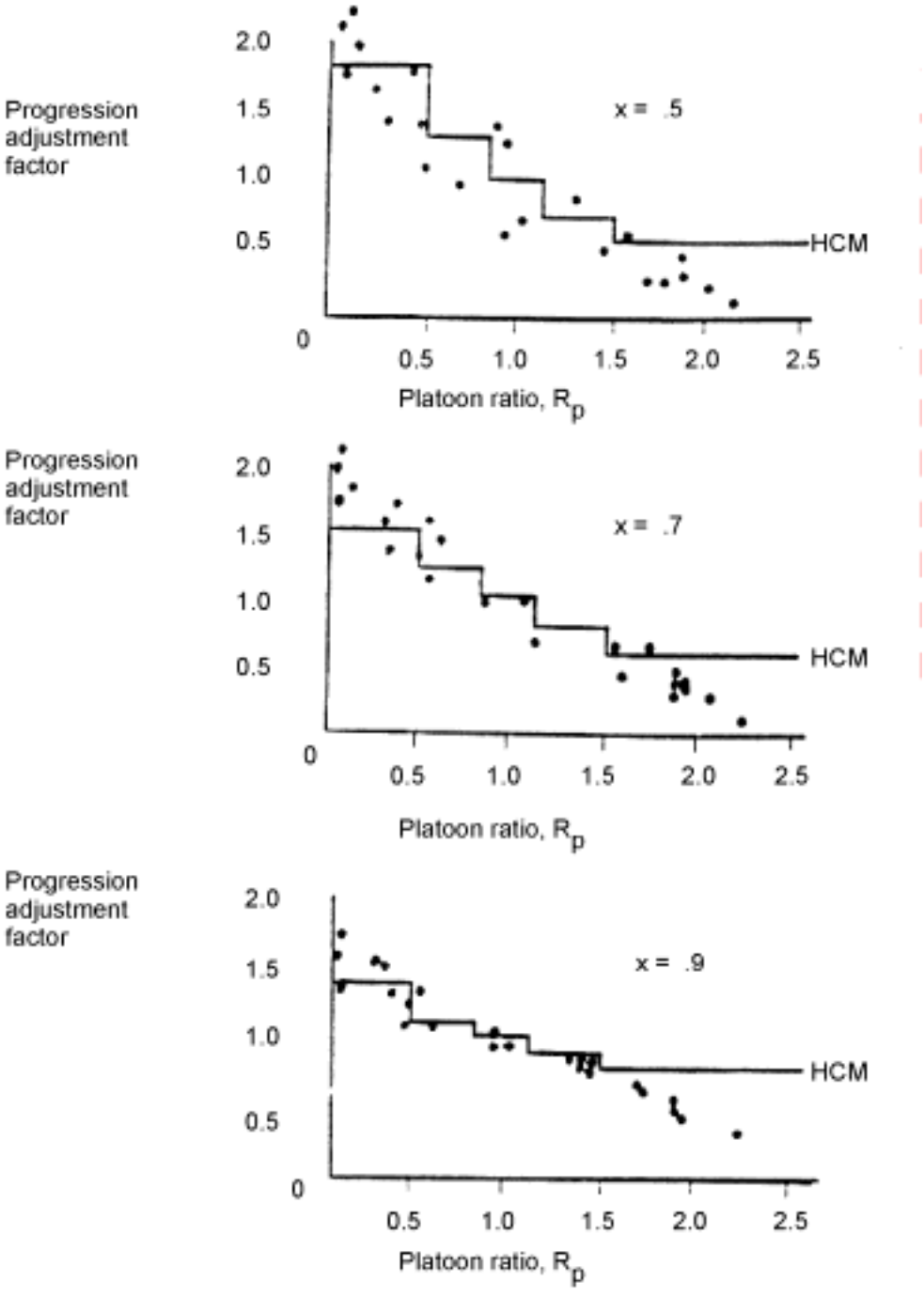


Figure 9.9
HCM Progression Adjustment Factor vs Platoon Ratio
Derived from TRANSYT-7F (Courage et al. 1988).

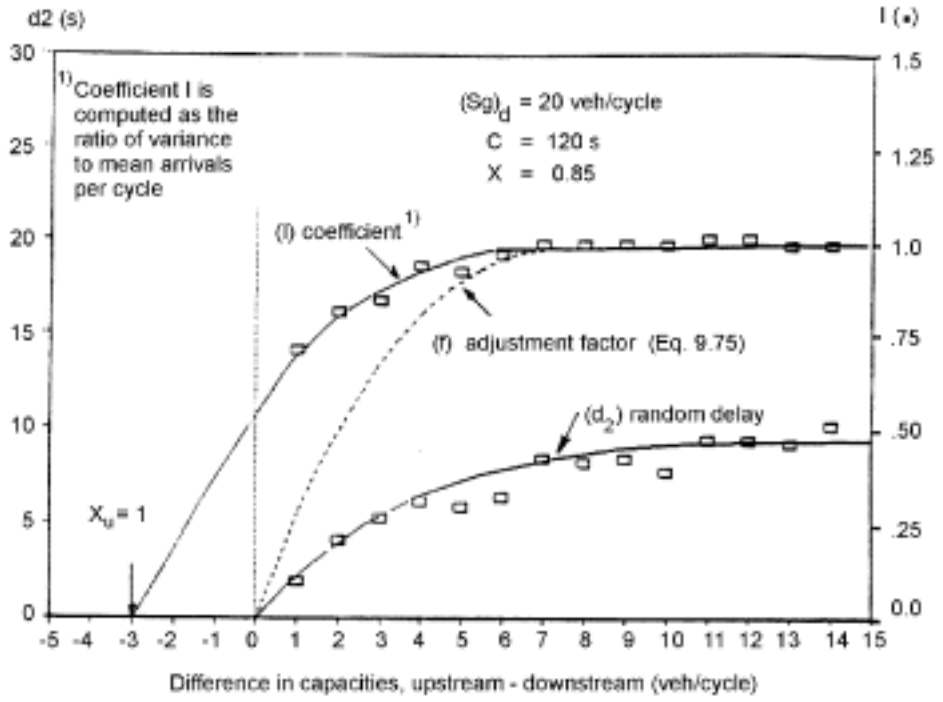


Figure 9.10
Analysis of Random Delay with Respect to the Differential Capacity Factor (f) and Var/Mean Ratio of Arrivals (I)- Steady State Queuing Conditions (Tarko et al. 1993).

$$f = 1 - e^{a(m_c - Sg)} \qquad (9.75)$$

where a is a model parameter, $a < 0$.

A recent paper by Newell (1990) proposes an interesting hypothesis. The author questions the validity of using random delay expressions derived for isolated intersections at internal signals in an arterial system. He goes on to suggest that the sum of random delays at all intersections in an arterial system with no turning movements is equivalent to the random delay at the critical intersection, assuming that it is isolated. Tarko et al. (1993) tested the Newell hypothesis using a computational model which considers a bulk service queuing model and a set of arrival distribution transformations. They concluded that Newell's model estimates provide a close upper bound to the results from their model. The review of traffic delay models at fixed-timed traffic signals indicate that the state of the art has shifted over time from a purely theoretical approach grounded in queuing theory, to heuristic models that have deterministic and stochastic components in a time-dependent domain. This move was motivated by the need to incorporate additional factors such as non-stationarity of traffic demand, oversaturation, traffic platooning and filtering effect of upstream signals. It is anticipated that further work in that direction will continue, with a view towards using the performance-based models for signal design and route planning purposes.

9.6 Theory of Actuated and Adaptive Signals

The material presented in previous sections assumed fixed time signal control, i.e. a fixed signal capacity. The introduction of traffic-responsive control, either in the form of actuated or traffic-adaptive systems requires new delay formulations that are sensitive to this process. In this section, delay models for actuated signal control are presented in some detail, which

incorporate controller settings such as minimum and maximum greens and unit extensions. A brief discussion of the state of the art in adaptive signal control follows, but no models are presented. For additional details on this topic, the reader is encouraged to consult the references listed at the end of the chapter.

9.6.1 Theoretically-Based Expressions

As stated by Newell (1989), the theory on vehicle actuated signals and related work on queues with alternating priorities is very large, however, little of it has direct practical value. For example, "exact" models of queuing theory are too idealized to be very realistic. In fact the issue of performance modeling of vehicle actuated signals is too complex to be described by a comprehensive theory which is simple enough to be useful. Actuated controllers are normally categorized into: fully-actuated, semi-actuated, and volume-density control. To date, the majority of the theoretical work related to vehicle actuated signals is limited to fully and semi-actuated controllers, but not to the more sophisticated volume-density controllers with features such as variable initial and extension intervals. Two types of detectors are used in practice: passage and presence. Passage detectors, also called point or small-area detectors, include a small loop and detect motion or passage when a vehicle crosses the detector zone. Presence detectors, also called area detectors, have a larger loop and detect presence of vehicles in the detector zone. This discussion focuses on traffic actuated intersection analysis with passage detectors only.

Delays at traffic actuated control intersections largely depend on the controller setting parameters, which include the following aspects: unit extension, minimum green, and maximum green. *Unit extension* (also called vehicle interval, vehicle extension, or gap time) is the extension green time for each vehicle as it arrives at the detector. *Minimum green*: summation of the initial interval and one unit extension. The initial interval is designed to clear vehicles between the detector and the stop line. *Maximum green*: the maximum green times allowed to a specific phase, beyond which, even if there are continuous calls for the current phase, green will be switched to the competing approach.

The relationship between delay and controller setting parameters for a simple vehicle actuated type was originally studied by Morris and Pak-Poy (1967). In this type of control, minimum and maximum greens are preset. Within the range of minimum and maximum greens, the phase will be extended for each arriving vehicle, as long as its headway does not exceed the value of unit extension. An intersection with two one-way streets was studied. It was found that, associated with each traffic flow condition, there is an optimal vehicle interval for which the average delay per vehicle is minimized. The value of the optimal vehicle interval decreases and becomes more critical, as the traffic flow increases. It was also found that by using the constraints of minimum and maximum greens, the efficiency and capacity of the signal are decreased. Darroch (1964b) also investigated a method to obtain optimal estimates of the unit extension which minimizes total vehicle delays.

The behavior of vehicle-actuated signals at the intersection of two one-way streets was investigated by Newell (1969). The arrival process was assumed to be stationary with a flow rate just slightly below the saturation rate, i.e. any probability distributions associated with the arrival pattern are time invariant. It is also assumed that the system is undersaturated but that traffic flows are sufficiently heavy, so that the queue lengths are considerably larger than one car. No turning movements were considered. The minimum green is disregarded since the study focused on moderate heavy traffic and the maximum green is assumed to be arbitrarily large. No specific arrival process is assumed, except that it is stationary.

Figure 9.11 shows the evolution of the queue length when the queues are large. Traffic arrives at a rate of q_1, on one approach, and q_2, on the other. r_j, g_j, and Y_j represent the effective red, green, and yellow times in cycle j. Here the signal timings are random variables, which may vary from cycle to cycle. For any specific cycle j, the total delay of all cars W_{ij} is the area of a triangular shaped curve and can be approximated by:

$$\mathrm{E}\{W_{1j}\} = \frac{q_1}{2(1-q_1/S_1)}\left[(E\{r_j\}+Y)^2 + Var(r_j) + \frac{I_1(E\{r_j\}+Y)}{S_1(1-q_1/S_1)} + \frac{V_1}{S_1 q_1}\right] \tag{9.76}$$

$$E\{W_{2j}\} = \frac{q_2}{2(1-q_2/S_2)}\left\{[(E\{g_j\}+Y)^2 + Var(g_j) + \frac{I_2[E(g_j)+Y]}{S_2(1-q_2/S_2)} + \frac{V_2}{S_2 q_2}\right\} \tag{9.77}$$

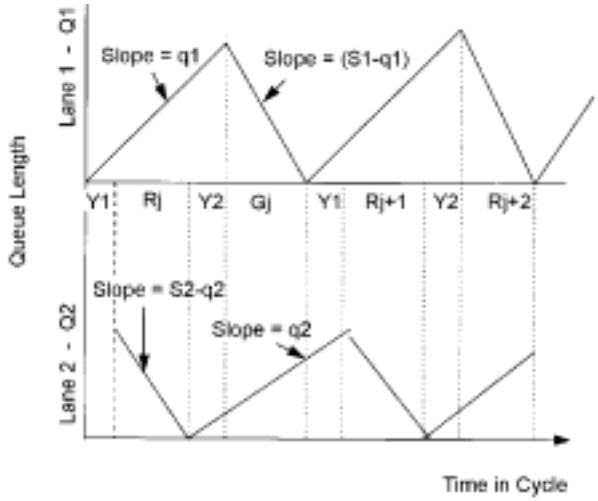


Figure 9.11
Queue Development Over Time Under
Fully-Actuated Intersection Control (Newell 1969).

where

$E\{W_{1j}\}, E\{W_{2j}\}$ = the total wait of all cars during cycle j for approach 1 and 2;
S_1, S_2 = saturation flow rate for approach 1 and 2;
$E\{r_j\}, E\{g_j\}$ = expectation of the effective red and green times;
$\mathrm{Var}(r_j), \mathrm{Var}(g_j)$ = variance of the effective red and green time;
I_1, I_2 = variance to mean ratio of arrivals for approach 1 and 2; and
V_1, V_2 = the constant part of the variance of departures for approach 1 and 2.

Since the arrival process is assumed to be stationary,

$$E\{r_j\} \to E\{r\}, \quad E\{g_j\} \to E\{g\} \tag{9.78}$$

$$Var(r_j) \to Var(r), \quad Var(g_j) \to Var(g) \tag{9.79}$$

$$E\{W_{kj}\} \to E\{W_k\}, \quad k=1,2 \tag{9.80}$$

The first moments of r and g were also derived based on the properties of the Markov process:

$$E\{r\} = \frac{Yq_2/S_2}{1-q_1/S_1-q_2/S_2} \tag{9.81}$$

$$E\{g\} = \frac{Yq_1/S_1}{1-q_1/S_1-q_2/S_2} \tag{9.82}$$

Variances of r and g were also derived, they are not listed here for the sake of brevity. Extensions to the multiple lane case were investigated by Newell and Osuna (1969).

A delay model with vehicle actuated control was derived by Dunne (1967) by assuming that the arrival process follows a binomial distribution. The departure rates were assumed to be constant and the control strategy was to switch the signal when the queue vanishes. A single intersection with two one-lane one-way streets controlled by a two phase signal was considered.

For each of the intervals $(k\tau, k\tau+\tau)$, k=0,1,2... the probability of one arrival in approach i = 1, 2 is denoted by q_i and the probability of no arrival by $p_i=1-q_i$. The time interval, τ, is taken as the time between vehicle departures. Saturation flow rate was assumed to be equal for both approaches. Denote

$W^{(2)}_r$ as the total delay for approach 2 for a cycle having effective red time of length r, then it can be shown that:

$$W^{(2)}_{r+1} = W^{(2)}_r + \mu[\delta_1 + c + \delta_2] \qquad (9.83)$$

where c is the cycle length, δ_1, and δ_2 are increases in delay at the beginning and at the end of the cycle, respectively, when one vehicle arrives in the extra time unit at the beginning of the phase and:

$$\begin{array}{l} \mu = 0 \ \ with\ probability \ \ p_2\ , \\ \ \ 1 \ \ with\ probability \ \ q_2\ . \end{array} \qquad (9.84)$$

Equation 9.83 means that if there is no arrivals in the extra time unit at the beginning of the phase, then $W^{(2)}_{r+1} = W^{(2)}_r$, otherwise $W^{(2)}_{r+1} = W^{(2)}_r + \delta_1 + c + \delta_2$.

Taking the expectation of Equation 9.83 and substituting for $E(\delta_1)$, $E(\delta_2)$:

$$E(W^{(2)}_{r+1}) = E(W^{(2)}_r) + q_2(r+1)/p_2 \qquad (9.85)$$

Solving the above difference equation for the initial condition $W^{(2)}_0 = 0$ gives,

$$E(W^{(2)}_r) = q_2(r^2 + r)/2p_2 \qquad (9.86)$$

Finally, taking the expectation of Equation 9.86 with respect to r gives

$$E(W^{(2)}) = q_2\{var[r]E^2[r] + E[r]\}/(2p_2) \qquad (9.87)$$

Therefore, if the mean and variance of (r) are known, delay can be obtained from the above formula. $E(W^{(i)})$ for approach 2 is obtained by interchanging the subscripts.

Cowan (1978) studied an intersection with two single-lane one-way approaches controlled by a two-phase signal. The control policy is that the green is switched to the other approach at the earliest time, t, such that there is no departures in the interval $[t-\beta_i - 1, t]$. In general $\beta_i \geq 0$. It was assumed that departure headways are 1 time unit, thus the arrival headways are at least 1 time unit. The arrival process on approach j is assumed to follow a bunched exponential distribution. It comprises random-sized bunches separated by inter-bunch headways. All bunched vehicles are assumed to have the same headway of 1 time unit. All inter-bunch headways follow the exponential distribution. Bunch size was assumed to have a general probability distribution with mean, μ_j, and variance, σ_j^2. The cumulative probability distribution of a headway less than t seconds, $F(t)$, is

$$F(t) = \begin{cases} 1 - \varphi e^{-\rho(t-\Delta)} & for\ t \geq \Delta \\ 0 & for\ t < \Delta \end{cases} \qquad (9.88)$$

where,

- Δ = minimum headway in the arrival stream, Δ=1 time unit;
- φ = proportion of free (unbunched) vehicles; and
- ρ = a delay parameter.

Formulae for average signal timings (r and g) and average delays for the cases of $\beta_j = 0$ and $\beta_j > 0$ are derived separately. $\beta_j = 0$ means that the green ends as soon as the queues for the approach clear while $\beta_j > 0$ means that after queues clear there will be a post green time assigned to the approach. By analyzing the property of Markov process, the following formula are derived for the case of $\beta_j = 0$.

$$E(g_1) = \frac{q_1 L}{1 - q_1 - q_2} \qquad (9.89)$$

$$E(g_2) = \frac{q_2 L}{1 - q_1 - q_2} \qquad (9.90)$$

$$E(r_1) = l_2 + \frac{q_2 L}{1 - q_1 - g_2} \qquad (9.91)$$

$$E(r_2) = l_1 + \frac{q_1 L}{1 - q_1 - q_2} \qquad (9.92)$$

where,

$E(g_1)$, $E(g_2)$ = expected effective green for approach 1 and 2;
$E(r_1)$, $E(r_2)$ = expected effective red for approach 1 and 2;
L = lost time in cycle;
l_1, l_2 = lost time for phase 1 and 2; and
q_1, q_2 = the stationary flow rate for approach 1 and 2.

The average delay for approach 1 is:

$$\frac{L(1-q_2)}{2(1-q_1-q_2)} + \frac{q_1^2(1-q_2)^2\rho_2(\sigma_2^2+\mu_2^2)+(1-q_1)^3(1-q_2)\rho_1(\sigma_1^2+\mu_1^2)}{2(1-q_1-q_2)(1-q_1-g_2+2q_1q_2)} \quad (9.93)$$

Akçelik (1994, 1995b) developed an analytical method for estimating average green times and cycle time at a basic vehicle actuated controller that uses a fixed unit extension setting by assuming that the arrival headway follows the bunched exponential distribution proposed by Cowan (1978). In his model, the minimum headway in the arrival stream Δ is not equal to one. The delay parameter, ρ, is taken as $\varphi q_t/\theta$, where q_t is the total arrival flow rate and $\theta=1-\Delta q$. In the model, the free (unbunched) vehicles are defined as those with headways greater than the minimum headway Δ. Further, all bunched vehicles are assumed to have the same headway Δ. Akçelik (1994) proposed two different models to estimate the proportion of free (unbunched) vehicles φ. The total time, g, allocated to a movement can be estimated as where g_{min} is the minimum green time and g_e, the green extension time. This green time, g, is subject to the following constraint

$$g \le g_{\max} \quad \vee \quad g_e < g_{emax} \quad (9.94)$$

where g_{max} and g_{max} are maximum green and extension time settings separately. If it is assumed that the unit extension is set so that a gap change does not occur during the saturated portion of green period, the green time can be estimated by:

$$g = g_s + e_g \quad (9.95)$$

where g_s is the saturated portion of the green period and e_g is the extension time assuming that gap change occurs after the queue clearance period. This green time is subject to the boundaries:

$$g_{\min} \le g \le g_{\max} \quad (9.96)$$

The saturated portion of green period can be estimated from the following formula:

$$g_s = \frac{f_q yr}{1-y} \quad (9.97)$$

where,

f_q = queue length calibration factor to allow for variations in queue clearance time;
S = saturation flow;
r = red time; and
y = q/S, ratio of arrival to saturation flow rate.

The average extension time beyond the saturated portion can be estimated from:

$$e_g = n_g h_g + e_t \quad (9.98)$$

where,

n_g = average number of arrivals before a gap change after queue clearance;
h_g = average headway of arrivals before a gap change after queue clearance; and
e_t = terminating time at gap change (in most case it is equal to the unit extension U).

For the case when $e_t = U$, Equation 9.98 becomes

$$e_g = -\frac{1}{q} + \left(\frac{\Delta}{\varphi} + \frac{1}{q}\right)e^{q(U-\Delta)} \quad (9.99)$$

9.6.2 Approximate Delay Expressions

Courage and Papapanou (1977) refined Webster's (1958) delay model for pretimed control to estimate delay at vehicle-actuated signals. For clarity, Webster's simplified delay formula is restated below.

$$d = 0.9(d_1+d_2) = 0.9\left[\frac{c(1-g/c)^2}{2(1-q/s)} + \frac{x^2}{2q(1-x)}\right] \quad (9.100)$$

Courage and Papapanou used two control strategies: (1) the available green time is distributed in proportion to demand on

the critical approaches; and (2) wasted time is minimized by terminating each green interval as the queue has been properly serviced. They propose the use of the cycle lengths shown in Table 9.2 for delay estimation under pretimed and actuated signal control:

Table 9.2
Cycle Length Used For Delay Estimation for Fixed-Time and Actuated Signals Using Webster's Formula (Courage and Papapanou 1977).

Type of Signal	Cycle Length in 1st Term	Cycle Length in 2nd Term
Pretimed	Optimum	Optimum
Actuated	Average	Maximum

The optimal cycle length, c_0, is Webster's:

$$c_0 = \frac{1.5L + 5}{1 - \Sigma y_{ci}} \qquad (9.101)$$

where L is total cycle lost time and y_{ci} is the volume to saturation flow ratio of critical movement i. The average cycle length, c_a is defined as:

$$c_a = \frac{1.5L}{1 - \Sigma y_{ci}} \qquad (9.102)$$

and the maximum cycle length, c_{max}, is the controller maximum cycle setting. Note that the optimal cycle length under pretimed control will generally be longer than that under actuated control. The model was tested by simulation and satisfactory results obtained for a wide range of operations.

In the U. S. Highway Capacity Manual (1994), the average approach delay per vehicle is estimated for fully-actuated signalized lane groups according to the following:

$$d = d_1 * DF + d_2 \qquad (9.103)$$

$$d_1 = \frac{c(1 - g/c)^2}{2(1 - xg/c)} \qquad (9.104)$$

$$d_2 = 900Tx^2[(x-1) + \sqrt{(x-1)^2 + \frac{mx}{CT}}] \qquad (9.105)$$

where, d, d_1, d_2, g, and c are as defined earlier and

DF = delay factor to account for signal coordination and controller type;
x = q/C, ratio of arrival flow rate to capacity;
m = calibration parameter which depends on the arrival pattern;
C = capacity in veh/hr; and
T = flow period in hours (T=0.25 in 1994 HCM).

The delay factor DF=0.85, reduces the queuing delay to account for the more efficient operation with fully-actuated operation when compared to isolated, pretimed control. In an upcoming revision to the signalized intersection chapter in the HCM, the delay factor will continue to be applied to the uniform delay term only.

As delay estimation requires knowledge of signal timings in the average cycle, the HCM provides a simplified estimation method. The average signal cycle length is computed from:

$$c_a = \frac{Lx_c}{x_c - \Sigma y_{ci}} \qquad (9.106)$$

where x_c = critical q/C ratio under fully-actuated control (x_c=0.95 in HCM). For the critical lane group i, the effective green:

$$g_i = \frac{y_{ci}}{x_c} c_a \qquad (9.107)$$

This signal timing parameter estimation method has been the subject of criticism in the literature. Lin (1989), among others, compared the predicted cycle length from Equation 9.106 with field observations in New York state. In all cases, the observed cycle lengths were higher than predicted, while the observed x_c ratios were lower.

Lin and Mazdeysa(1983) proposed a general delay model of the following form consistent with Webster's approximate delay formula:

$$d=0.9\{\frac{c(1-K_1\frac{g}{c})^2}{2(1-K_1\frac{g}{c}K_2x)}+\frac{3600(K_2x)^2}{2q(1-K_2x)}\} \quad (9.10)$$

where g, c, q, x are as defined earlier and K_1 and K_2 are two coefficients of sensitivity which reflect different sensitivities of traffic actuated and pretimed delay to both g/c and x ratios. In this study, K_1 and K_2 are calibrated from the simulation model for semi-actuated and fully-actuated control separately. More importantly, the above delay model has to be used in conjunction with the method for estimating effective green and cycle length. In earlier work, Lin (1982a, 1982b) described a model to estimate the average green duration for a two phase fully-actuated signal control. The model formulation is based on the following assumption: (1) the detector in use is small area passage detector; (2) right-turn-on-red is either prohibited or its effect can be ignored; and (3) left turns are made only from exclusive left turn lanes. The arrival pattern for each lane was assumed to follow a Poisson distribution. Thus, the headway distribution follows a shifted negative exponential distribution.

Figure 9.12 shows the timing sequence for a two phase fully actuated controller. For phase i, beyond the initial green interval, g_{mini}, green extends for F_i based on the control logic and the settings of the control parameters. F_i can be further divided into two components: (1) e_{ni} — the additional green extended by n vehicles that form moving queues upstream of the detectors after the initial interval G_{mini}; (2) E_{ni} — the additional green extended by n vehicles with headways of no more than one unit extension, U, after G_{mini} or e_{ni}. Note that e_{ni} and E_{ni} are random variables that vary from cycle to cycle. Lin (1982a, 1982b) developed the procedures to estimate e_i and E_i, the expected value of e_{ni} and E_{ni}, as follows. A moving queue upstream of a detector may exist when G_{mini} is timed out in case the flow rate of the critical lane q_c is high. If there are n vehicles arriving in the critical lane during time T_i, then the time required for the nth vehicle to reach the detector after G_{mini} is timed out can be estimated by the following equation:

$$t_n = nw + \sqrt{\frac{2(nL-s_i)}{a}} - G_{mini} \quad (9.109)$$

where w is the average time required for each queuing vehicle to start moving after the green phase starts, L is the average vehicle length, a is the vehicle acceleration rate from a standing position., and s is the detector setback. If $t_n \le 0$, there is no moving queue exists and thus $e_i=0$; otherwise the green will be extended by the moving queue. Let s be the rate at which the queuing vehicle move across the detector. Considering that additional vehicles may join the queue during the time interval t_n, if $t_n>0$ and $s>0$, then:

$$e_{ni}=\frac{st_n}{s-q} \quad (9.110)$$

To account for the probability that no moving queues exist upstream of the detector at the end of the initial interval, the expected value of e_{ni}, e_i is expressed as:

$$e_i = \sum_{n=n_{\min}}^{\infty} \frac{P_j(n/T_i)\,e_{ni}}{1-p_j(n<n_{\min})} \quad (9.111)$$

where n_{min} is the minimum number of vehicles required to form a moving queue.

To estimate E_i, let us suppose that after the initial G_{mini} and additional green e_{ni} have elapsed, there is a sequence of k consecutive headways that are shorter than U followed by a headway longer than U. In this case the green will be extended k times and the resultant green extension time is $kJ+U$ with probability $[F(h \le U)]^k F(h \ge U)$, where J is the average length of each extension and $F(h)$ is the cumulative headway distribution function.

$$J = \frac{\int_{\Delta}^{U_i} tf(t)\,dt}{F(h<U_i)} \quad (9.112)$$

and therefore

$$E_i=\sum_{k=0}^{\infty}(kJ+U)[F(h\le U)]^k F(h>U)=-\frac{1}{q}+[\Delta+\frac{1}{q}]e^{q(U-\Delta)} \quad (9.113)$$

where Δ is the minimum headway in the traffic stream.

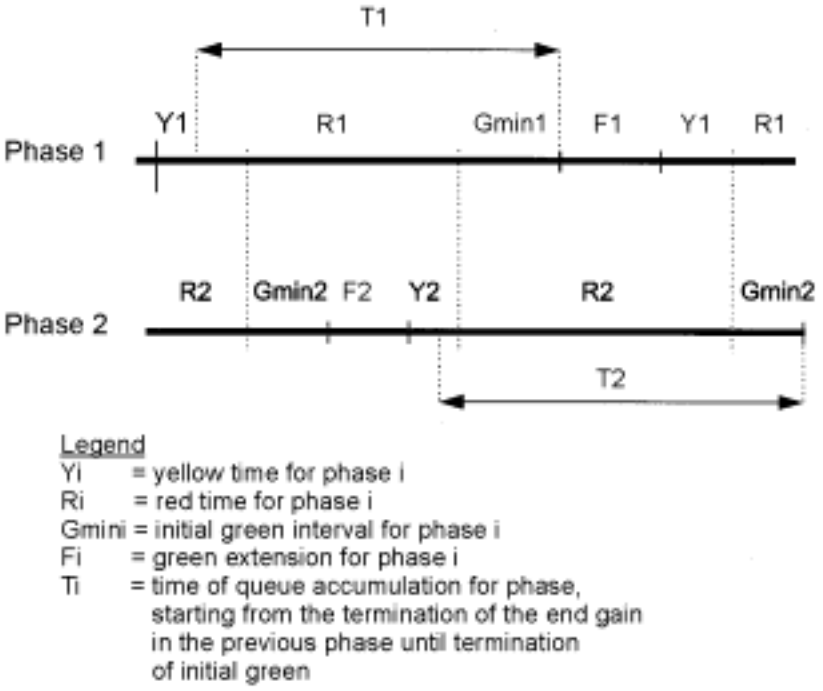


Figure 9.12
Example of a Fully-Actuated Two-Phase Timing Sequence (Lin 1982a).

Referring to Figure 9.12, after the values of T_1 and T_2 are obtained, G_i can be estimated as:

$$G_i = \Sigma_{n=0}^{\infty}(G_{mini}+e_i+E_i)\,P(n/T_i) \qquad (9.114)$$

subject to

$$G_{mini}+e_i+E_i \le (G_{\max})_i \qquad (9.115)$$

where $P(n/T_i)$ is the probability of n arrivals in the critical lane of the ith phase during time interval T_i. Since both T_1 and T_2 are unknown, an iterative procedure was used to determine G_1 and G_2.

Li et al. (1994) proposed an approach for estimating overflow delays for a simple intersection with fully-actuated signal control. The proposed approach uses the delay format in the 1994 HCM (Equations 9.104 and 9.105) with some variations, namely a) the delay factor, DF, is taken out of the formulation of delay model and b) the multiplier x^2 is omitted from the formulation of the overflow delay term to ensure convergence to the deterministic oversaturated delay model. Thus, the overflow delay term is expressed as:

$$d_2 = 900T[(x-1)+\sqrt{(x-1)^2+\frac{8kx}{CT}}] \qquad (9.116)$$

where the parameter (k) is derived from a numerical calibration of the steady-state for of Equation 9.105 as shown below.

$$d_2 = \frac{kx}{C(1-x)} \qquad (9.117)$$

This expression is based on a more general formula by Akçelik (1988) and discussed by Akçelik and Rouphail (1994). The calibration results for the parameter k along with the overall statistical model evaluation criteria (standard error and R^2) are depicted in Table 9.3. The parameter k which corresponds to pretimed control, calibrated by Tarko (1993) is also listed. It is noted that the pretimed steady-state model was also calibrated using the same approach, but with fixed signal settings. The first and most obvious observation is that the pretimed model produced the highest k (delay) value compared to the actuated models. Secondly, the parameter was found to increase with the size of the controller's unit extension (U).

Procedures for estimating the average cycle length and green intervals for semi-actuated signal operations have been developed by Lin (1982b, 1990) and Akçelik (1993b). Recently, Lin (1992) proposed a model for estimating average cycle length and green intervals under semi-actuated signal control operations with exclusive pedestrian-actuated phase. Luh (1991) studied the probability distribution of and delay estimation for semi-actuated signal controllers.

In summary, delay models for vehicle-actuated controllers are derived from assumptions related to the traffic arrival process, and are constrained by the actuated controller parameters. The distribution of vehicle headways directly impact the amount of green time allocated to an actuated phase, while controller parameters bound the green times within specified minimums and maximums. In contrast to fixed-time models, performance models for actuated have the additional requirement of estimating the expected signal phase lengths. Further research is needed to incorporate additional aspects of actuated operations such as phase skipping, gap reduction and variable maximum greens. Further, there is a need to develop generalized models that are applicable to both fixed time and actuated control. Such models would satisfy the requirement that both controls yield identical performance under very light and very heavy traffic demands. Recent work along these lines has been reported by Akçelik and Chung (1994, 1995).

9.6.3 Adaptive Signal Control

Only a very brief discussion of the topic is presented here. Adaptive signal control systems are generally considered superior to actuated control because of their true demand responsiveness. With recent advances in microprocessor technology, the gap-based strategies discussed in the previous section are becoming increasingly outmoded and demonstrably inefficient. In the past decade, control algorithms that rely on explicit intersection/network delay minimization in a time-variant environment, have emerged and been successfully tested. While the algorithms have matured both in Europe and the U.S., evident by the development of the MOVA controller in the U.K. (Vincent et al. 1988), PRODYN in France (Henry et al. 1983), and OPAC in the U.S. (Gartner et al. 1982-1983), theoretical work on traffic performance estimation under adaptive control is somewhat limited. An example of such efforts is the work by Brookes and Bell (1991), who investigated the use of Markov Chains and three heuristic approaches in an attempt to calculate the expected delays and stops for discrete time adaptive signal control. Delays are computed by tracing the queue evolution process over time using a `rolling horizon' approach. The main problem lies in the estimation (or prediction) of the initial queue in the current interval. While the Markov Chain approach yields theoretically correct answers, it is of limited value in practice due to its extensive computational and storage requirements. Heuristics that were investigated include the use of the mean queue length, in the last interval as the starting queue in this interval; the `two-spike' approach, in which the queue length

Table 9.3
Calibration Results of the Steady-State Overflow Delay Parameter (k) (Li et al. 1994).

Control	Pretimed	U=2.5	U=3.5	U=4.0	U=5.0
k (m=8k)	0.427	0.084	0.119	0.125	0.231
s.e.	NA	0.003	0.002	0.002	0.006
R^2	0.903	0.834	0.909	0.993	0.861

distribution has non-zero probabilities at zero and at an integer value closest to the mean; and finally a technique that propagates the first and second moment of the queue length distribution from period to period.

Overall, the latter method was recommended because it not only produces estimates that are sufficiently close to the theoretical estimate, but more importantly it is independent of the traffic arrival distribution.

9.7 Concluding Remarks

In this chapter, a summary and evolution of traffic theory pertaining to the performance of intersections controlled by traffic signals has been presented. The focus of the discussion was on the development of stochastic delay models.

Early models focused on the performance of a single intersection experiencing random arrivals and deterministic service times emulating fixed-time control. The thrust of these models has been to produce point estimates--i.e. expectations of-- delay and queue length that can be used for timing design and quality of service evaluation. The model form typically include a deterministic component to account for the red-time delay and a stochastic component to account for queue delays. The latter term is derived from a queue theory approach.

While theoretically appealing, the steady-state queue theory approach breaks down at high degrees of saturation. The problem lies in the steady-state assumption of sustained arrival flows needed to reach stochastic equilibrium (i.e the probability of observing a queue length of size Q is time-independent) . In reality, flows are seldom sustained for long periods of time and therefore, stochastic equilibrium is not achieved in the field at high degrees of saturation.

A compromise approach, using the coordinate transformation method was presented which overcomes some of these difficulties. While not theoretically rigorous, it provides a means for traffic performance estimation across all degrees of saturation which is also dependent on the time interval in which arrival flows are sustained.

Further extensions of the models were presented to take into account the impact of platooning, which obviously alter the arrival process at the intersection, and of traffic metering which may causes a truncation in the departure distribution from a highly saturated intersection. Next, an overview of delay models which are applicable to intersections operating under vehicle actuated control was presented. They include stochastic models which characterize the randomness in the arrival *and* departure process-- capacity itself is a random variable which can vary from cycle to cycle, and fixed-time equivalent models which treat actuated control as equivalent pretimed models operating at the average cycle and average splits.

Finally, there is a short discussion of concepts related to adaptive signal control schemes such as the MOVA systems in the United Kingdom and OPAC in the U.S. Because these approaches focus primarily on optimal signal control rather than performance modeling, they are somewhat beyond the scope of this document.

There are many areas in traffic signal performance that deserve further attention and require additional research. To begin with, the assumption of uncorrelated arrivals found in most models is not appropriate to describe platooned flow--where arrivals are highly correlated. Secondly, the estimation of the initial overflow queue at a signal is an area that is not well understood and documented. There is also a need to develop queuing/delay models that are constrained by the physical space available for queuing. Michalopoulos (1988) presented such an application using a continuous flow model approach. Finally, models that describe the interaction between downstream queue lengths and upstream departures are needed. Initial efforts in this direction have been documented by Prosser and Dunne (1994) and Rouphail and Akçelik (1992b).

References

Akçelik, R. (1980). *Time-Dependent Expressions for Delay, Stop Rate and Queue Length at Traffic Signals*. Australian Road Research Board, Internal Report, AIR 367-1.

Akçelik, R. (1988). *The Highway Capacity Manual Delay Formula for Signalized Intersections*. ITE Journal 58(3), pp. 23-27.

Akçelik, R. and N. Rouphail (1993). *Estimation of Delays at Traffic Signals for Variable Demand Conditions*. Transportation Research-B, Vol. 27B, No. 2, pp. 109-131.

Akçelik, R. (1994). *Estimation of Green Times and Cycle Time for Vehicle-Actuated Signals*. Transportation Research Board 1457, pp.63-72.

Akçelik, R. and E. Chung (1994). *Traffic Performance Models for Unsignalized Intersections and Fixed-Time Signals*. Proceedings, Second International Symposium on Highway Capacity, Sydney, Australia, ARRB, Volume I, pp. 21-50.

Akçelik, R. and N. Rouphail (1994). *Overflow Queues and Delays with Random and Platoon Arrivals at Signalized Intersections,* Journal of Advanced Transportation, Volume 28(3), pp. 227-251.

Akçelik, R. (1995a). *Extension of the Highway Capacity Manual Progression Factor Method for Platooned Arrivals,* Research Report ARR No. 276, ARRB Transport Research Ltd., Vermont South, Australia.

Akçelik, R. (1995b). *Signal Timing Analysis for Vehicle Actuated Control.* Working Paper WD TE95/007, ARRB Transport Research Ltd., Vermont South, Australia.

Akçelik, R. and E. Chung (1995). *Calibration of Performance Models for Traditional Vehicle-Actuated and Fixed-Time Signals.* Working Paper WD TO 95/103, ARRB Transport Research Ltd., Vermont South, Australia.

Allsop, R. (1972). *Delay at a Fixed Time Traffic Signal-I: Theoretical Analysis*. Transportation Science, 6, pp. 260-285.

Beckmann, M. J., C. B. McGuire, and C. B. Winsten (1956). *Studies in the Economics in Transportation*. New Haven, Yale University Press.

Bell, M. G. H. and D. Brookes (1993). *Discrete Time-Adaptive Traffic Signal Control*. Transportation Research, Vol. 1C, No. 1, pp. 43-55.

Brilon, W. and N. Wu (1990). *Delays At Fixed-time Traffic Signals Under Time Dependent Traffic Conditions*. Traffic Engineering and Control, 31(12), pp. 623-63.

Brookes, D. and M. G. Bell (1991). *Expected Delays and Stop Calculation for Discrete Adaptive Traffic Signal Control*. Proceedings, First International Symposium on Highway Capacity, Karlsruhe, Germany, U. Brannolte, Ed.

Catling, I. (1977). *A Time-Dependent Approach To Junction Delays*. Traffic Engineering and Control, 18(11), pp. 520-523,526.

Clayton, A. (1941). *Road Traffic Calculations*. J. Inst. Civil. Engrs, 16, No.7, pp.247-284, No. 8, pp. 588-594.

Courage, K. G. and P. P. Papapanou (1977). *Estimation of Delay at Traffic-Actuated Signals*. Transportation Research Record, 630, pp. 17-21.

Courage, K. G., C. E. Wallace, and R. Alqasem (1988). *Modeling the Effect of Traffic Signal Progression on Delay*. Transportation Research Record, 1194, TRB, National Research Council, Washington, DC, pp. 139-146.

Cowan, R. (1978). *An Improved Model for Signalized Intersections with Vehicle-Actuated Control*. J. Appl. Prob. 15, pp. 384-396.

Cronje, W. B. (1983a). *Derivation of Equations for Queue Length, Stops, and Delays for Fixed Time Traffic Signals*. Transportation Research Record, 905, pp. 93-95.

Cronje, W. B. (1983b). *Analysis of Existing Formulas for Delay, Overflow, and Stops*. Transportation Research Record, 905, pp. 89-93.

Darroch, J. N. (1964a). *On the Traffic-Light Queue*. Ann. Math. Statist., 35, pp. 380-388.

Darroch, J. N., G. F. Newell, and R. W. J. Morris (1964b). *Queues for a Vehicle-Actuated Traffic Light*. Operational Research, 12, pp. 882-895.

Dunne, M. C. (1967). *Traffic Delay at a Signalized Intersection with Binomial Arrivals*. Transportation Science, 1, pp. 24-31.

Fambro, D. B., E. C. P. Chang, and C. J. Messer (1991). *Effects of the Quality of Traffic Signal Progression on Delay*. National Cooperative Highway Research Program Report 339, Transportation Research Board, National Research Council, Washington, DC.

Gartner, N. (1982-1983). *Demand Responsive Decentralized Urban Traffic Control*. Part I: Single Intersection Policies; Part II: Network Extensions, Office of University Research, U.S.D.O.T.

Gazis, D. C. (1974). *Traffic Science*. A Wiley-Intersection Publication, pp. 148-151.

Gerlough, D. L. and M. J. Huber (1975). *Traffic Flow Theory*. Transportation Research Board Special Report, 165, National Research Council, Washington, DC.

Haight, F. A. (1959). *Overflow At A Traffic Flow*. Biometrika. Vol. 46, Nos. 3 and 4, pp. 420-424.

Haight, F. A. (1963). *Mathematical Theories of Traffic Flow*. Academic Press, New York.

Henry, J., J. Farges, and J. Tuffal (1983). *The PRODYN Real-Time Traffic Algorithm*. 4th IFAC-IFIC--IFORS on Control in Transportation Systems, Baden-Baden.

Hillier, J. A. and R. Rothery (1967). *The Synchronization of Traffic Signals for Minimum Delays*. Transportation Science, 1(2), pp. 81-94.

Hutchinson, T. P. (1972). *Delay at a Fixed Time Traffic Signal-II. Numerical Comparisons at Some Theoretical Expressions*. Transportation Science, 6(3), pp. 286-305.

Kimber, R. M. and E. M. Hollis (1978). *Peak Period Traffic Delay at Road Junctions and Other Bottlenecks*. Traffic Engineering and Control, Vol. 19, No. 10, pp. 442-446.

Kimber, R. and E. Hollis (1979). *Traffic Queues and Delays at Road Junctions*. TRRL Laboratory Report, 909, U.K.

Li, J., N. Rouphail, and R. Akçelik (1994). *Overflow Delay Estimation for Intersections with Fully-Actuated Signal Control*. Presented at the 73rd Annual Meeting of TRB, Washington, DC.

Lighthill, M. H. and G. B. Whitham (1957). *On Kinematic Waves: II. A Theory of Traffic Flow on Long Crowded Roads*. Proceedings of the Royal Society, London Series A229, No. 1178, pp. 317-345.

Little, J. D. C. (1961). *Approximate Expected Delays for Several Maneuvers by Driver in a Poisson Traffic*. Operations Research, 9, pp. 39-52.

Lin, F. B. (1982a). *Estimation of Average Phase Duration of Full-Actuated Signals*. Transportation Research Record, 881, pp. 65-72.

Lin, F. B. (1982b). *Predictive Models of Traffic-Actuated Cycle Splits*. Transportation Research - B, Vol. 16B, No. 5, pp. 361-372.

Lin, F. (1989). *Application of 1985 Highway Capacity Manual for Estimating Delays at Signalized Intersection*. Transportation Research Record, 1225, TRB, National Research Council, Washington DC, pp. 18-23.

Lin, F. (1990). *Estimating Average Cycle Lengths and Green Intervals of Semiactuated Signal Operations for Level-of-Service Analysis*. Transportation Research Record, 1287, TRB, National Research Council, Washington DC, pp. 119-128.

Lin, F. (1992). *Modeling Average Cycle Lengths and Green Intervals of Semi-actuated Signal Operations with Exclusive Pedestrian-actuated Phase*. Transportation Research, Vol. 26B, No. 3, pp. 221-240.

Lin, F. B. and F. Mazdeyasa (1983). *Delay Models of Traffic Actuated Signal Controls*. Transportation Research Record, 905, pp. 33-38.

Luh, J. Z. and Chung Yee Lee (1991). *Stop Probability and Delay Estimations at Low Volumes for Semi-Actuated Traffic Signals*. Transportation Science, 25(1), pp. 65-82.

May, A. D. and H. M. Keller (1967). *A Deterministic Queuing Model*. Transportation Research, 1(2), pp. 117-128.

May, A. D. (1990). *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, New Jersey, pp. 338-375.

McNeil, D. R. (1968). *A Solution to the Fixed-Cycle Traffic Light Problem for Compound Poisson Arrivals*. J. Appl. Prob. 5, pp. 624-635.

Miller, A. J. (1963). *Settings for Fixed-Cycle Traffic Signals*. Operational Research Quarterly, Vol. 14, pp. 373-386.

Miller, A. J. (1968a). *Australian Road Capacity Guide - Provisional Introduction and Signalized Intersections*. Australian Road Research Board Bulletin No.4, (Superseded by ARRB report ARR No. 123, 1981).

Miller, A. J. (1968b). *The Capacity of Signalized Intersections in Australia.* Australian Road Research Board, ARRB Bulletin No.3.

Morris, R. W. T. and P. G. Pak-Boy (1967). *Intersection Control by Vehicle Actuated Signals*. Traffic Engineering and Control, No.10, pp. 288-293.

Newell, G. F. (1960). *Queues for a Fixed-Cycle Traffic Light*. The Annuals of Mathematical Statistics, Vol.31, No.3, pp. 589-597.

Newell, G. F. (1965). A*pproximation Methods for Queues with Application to the Fixed-Cycle Traffic Light*. SIAM Review, Vol.7.

Newell, G. F. (1969). *Properties of Vehicle Actuated Signals: I. One-Way Streets*. Transportation Science, 3, pp. 30-52.

Newell, G. F. (1989). *Theory of Highway Traffic Signals*. UCB-ITS-CN-89-1, Institute of Transportation Studies, University of California.

Newell, G. F. (1990). *Stochastic Delays on Signalized Arterial Highways*. Transportation and Traffic Theory, Elsevier Science Publishing Co.,Inc., M. Koshi, Ed., pp. 589-598.

Newell, G. F. and E. E. Osuna (1969). *Properties of Vehicle-Actuated Signals: II. Two-Way Streets*. Transportation Science, 3, pp. 99-125.

Ohno, K. (1978). *Computational Algorithm for a Fixed Cycle Traffic Signal and New Approximate Expressions for Average Delay*. Transportation Science, 12(1), pp. 29-47.

Olszewski, P. S. (1988). *Efficiency of Arterial Signal Coordination*. Proceedings 14th ARRB Conference, 14(2), pp. 249-257.

Olszewski, P. (1990a). *Modelling of Queue Probability Distribution at Traffic Signals*. Transportation and Traffic Flow Theory, Elsevier Science Publishing Co., Inc., M. Koshi, Ed., pp. 569-588.

Olszewski, P. (1990b). *Traffic Signal Delay Model for Non-Uniform Arrivals*. Transportation Research Record, 1287, pp. 42-53.

Pacey, G. M. (1956). *The Progress of a Bunch of Vehicles Released from a Traffic Signal*. Research Note No. Rn/2665/GMP. Road Research Laboratory, London (mimeo).

Potts, R. B. (1967). *Traffic Delay at a Signalized Intersection with Binomial Arrivals*. pp. 126-128.

Robertson, D. I. (1969). *TRANSYT: A Traffic Network Study Tool*. Road Research Laboratory Report LR 253, Crowthorne.

Rorbech, J. (1968). *Determining The Length of The Approach Lanes Required at Signal-Controlled Intersections on Through Highways*. Transportation Research Record, 1225, TRB, National Research Council, Washington, DC, pp. 18-23.

Rouphail, N. (1989). *Progression Adjustment Factors at Signalized Intersections*. Transportation Research Record, 1225, pp. 8-17.

Rouphail, N. and R. Akçelik (1992a). *Oversaturation Delay Estimates with Consideration of Peaking*. Transportation Research Record, 1365, pp. 71-81.

Rouphail, N. and R. Akçelik(1992b). *A Preliminary Model of Queue Interaction at Signalized Paired Intersections,* Proceedings, 16th ARRB Conference 16(5), Perth, Australia, pp. 325-345.

Stephanopoulos, G., G. Michalopoulos, and G. Stephanopoulos (1979). *Modelling and Analysis of Traffic Queue Dynamics at Signalized Intersections*. Transportation Research, Vol. 13A, pp. 295-307.

Tarko, A., N. Rouphail, and R. Akçelik (1993b). *Overflow Delay at a Signalized Intersection Approach Influenced by an Upstream Signal: An Analytical Investigation*. Transportation Research Record, No. 1398, pp. 82-89.

Transportation Research Board (1985). Highway Capacity Manual, Special Report 209, National Research Council, Washington, DC.

Transportation Research Board (1993). Highway Capacity Manual, Chapter 9 (Signalized Intersections), National Research Council, Washington, DC.

Van As, S. C (1991). *Overflow Delay in Signalized Networks*. Transportation Research - A, Vol. 25A, No. 1, pp. 1-7.

Vincent, R. A. and J. R. Pierce (1988). *MOVA: Traffic Responsive, Self Optimizing Signal Control for Isolated Intersections*. TRRL Lab Research Report, 170.

Webster, F. V. (1958). *Traffic Signal Settings*. Road Research Laboratory Technical Paper No. 39, HMSO, London.

Wirasinghe, S. C. (1978). *Determination of Traffic Delays from Shock-Wave Analysis*. Transportation Research, Vol. 12, pp. 343-348.

TRAFFIC SIMULATION

BY EDWARD LIEBERMAN[18]
AJAY K. RATHI

[18]President, KLD Associates, Inc. 300 Broadway, Huntington Station, NY 11746

CHAPTER 10 - Frequently used Symbols

a_f = acceleration response of follower vehicle to some stimulus
v = instantaneous speed of lead vehicle
v_f = instantaneous speed of follower vehicle
d = projected maximum deceleration rate of lead vehicle
d_f = projected maximum deceleration rate of follower vehicle
R_f = reaction time lag of driver in follower vehicle
I_i = *i*th replicate of specified seed, S_0
R_I = *i*th random number
h = headway separating vehicles (sec)
H = mean headway (sec)
h_{min} = minimum headway (sec)
R = random number
X_I = *i*th observation (sample) of an MOE
μ_x = mean of sample
δ^2 = variance
$\hat{\delta}$ = estimate of variance
$t_{n-1,\,1-\alpha/2}$ = upper 1-α/2 critical point of the *t* distribution with *n*-1 degrees of freedom
$\bar{x}_j(m)$ = mean of *m* observation of *j*th batch
$\bar{x}$ = grand sample mean across batches
N_i = number of replications of the *i*th strategy
$var(X)$ = variance of statistic, X

10.
TRAFFIC SIMULATION

10.1 Introduction

Simulation modeling is an increasingly popular and effective tool for analyzing a wide variety of dynamical problems which are not amenable to study by other means. These problems are usually associated with complex processes which can not readily be described in analytical terms. Usually, these processes are characterized by the interaction of many system components or *entities*. Often, the behavior of each entity and the interaction of a limited number of entities, may be well understood and can be reliably represented logically and mathematically with acceptable confidence. However, the complex, simultaneous interactions of many system components cannot, in general, be adequately described in mathematical or logical forms.

Simulation models are designed to "mimic" the behavior of such systems. Properly designed models *integrate* these separate entity behaviors and interactions to produce a detailed, quantitative description of system performance. Specifically, simulation models are mathematical/logical representations (or *abstractions*) of real-world systems, which take the form of software executed on a digital computer in an experimental fashion.

The user of traffic simulation software specifies a "scenario" (e.g., highway network configuration, traffic demand) as model inputs. The simulation model results *describe* system operations in two formats: (1) statistical and (2) graphical. The numerical results provide the analyst with detailed quantitative descriptions of *what* is likely to happen. The graphical and animated representations of the system functions can provide insights so that the trained observer can gain an understanding of *why* the system is behaving this way. However, it is the responsibility of the analyst to properly interpret the wealth of information provided by the model to gain an understanding of cause-and-effect relationships.

Traffic simulation models can satisfy a wide range of requirements:

1. *Evaluation of alternative treatments*
   With simulation, the engineer can control the experimental environment and the range of conditions to be explored. Historically, traffic simulation models were used initially to evaluate signal control strategies, and are currently applied for this purpose as an integral element of the ATMS research and development activity.

2. *Testing new designs*
   Transportation facilities are costly investments. Simulation can be applied to quantify traffic performance responding to different geometric designs before the commitment of resources to construction.

3. *As an element of the design process*
   The classical iterative design paradigm of conceptual design followed by the recursive process of evaluation and design refinement, can benefit from the use of simulation. Here, the simulation model can be used for evaluation; the detailed statistics provided can form the basis for identifying design flaws and limitations. These statistics augmented with animation displays can provide invaluable insights guiding the engineer to improve the design and continue the process.

4. *Embed in other tools*
   In addition to its use as a stand-alone tool, simulation *sub-models* can be integrated within software tools designed to perform other functions. Examples include: (1) the flow model within the TRANSYT-7F signal optimization; (2) the DYNASMART simulation model within a dynamic traffic assignment; (3) the simulation component of the INTEGRATION assignment/control model; (4) the CORSIM model within the Traffic Research Laboratory (TreL) developed for FHWA; and (5) the simulation module of the EVIPAS actuated signal optimization program.

5. *Training personnel*
   Simulation can be used in the context of a real-time laboratory to train operators of Traffic Management Centers. Here, the simulation model, which is integrated with a real-time traffic control computer, acts as a surrogate for the real-world surveillance, communication and traffic environments.

6. *Safety Analysis*
   Simulation models to "recreate" accident scenarios have proven to be indispensable tools in the search to build safer vehicles and roadways. An example is the CRASH program used extensively by NHTSA.

This compilation of applications indicates the variety and scope of traffic simulation models and is by no means exhaustive. Simulation models can also be *supportive* of analytical models such as PASSER, and of computational procedures such as the HCS. While these and many other computerized tools do not include simulation sub-models, users of these tools can enhance their value by applying simulation to evaluate their performance.

This chapter is intended for transportation professionals, researchers, students and technical personnel who either currently use simulation models or who wish to explore their potential. Unlike the other chapters of this monograph, we will not focus exclusively on theoretical developments -- although fundamental simulation building blocks will be discussed. Instead, we will describe the properties, types and classes of traffic simulation models, their strengths and pitfalls, user caveats, and model-building fundamentals. We will emphasize how the user can derive the greatest benefits from simulation through proper interpretation of the results, with emphasis on the need to adequately calibrate the model and to apply rigorous statistical analysis of the results.

10.2 When Should the Use of Simulation Models be Considered?

Since simulation models *describe* a dynamical process in statistical and pictorial formats., they can be used to analyze a wide range of applications wherever...

- Mathematical treatment of a problem is infeasible or inadequate due to its temporal or spatial scale, and/or the complexity of the traffic flow process.
- The assumptions underlying a mathematical formulation (e.g., a linear program) or an heuristic procedure (e.g., those in the Highway Capacity Manual) cast some doubt on the accuracy or applicability of the results.
- The mathematical formulation represents the dynamic traffic/control environment as a simpler quasi steady-state system.
- There is a need to view vehicle animation displays to gain an understanding of *how* the system is behaving in order to explain *why* the resulting statistics were produced.
- Congested conditions persist over a significant time.

It must be emphasized that traffic simulation, by itself, cannot be used *in place of* optimization models, capacity estimation procedures, demand modeling activities and design practices. Simulation can be used to *support* such undertakings, either as embedded submodels or as an auxiliary tool to evaluate and extend the results provided by other procedures. Some representative statistics (called Measures of Effectiveness, MOE) provided by traffic simulation models are listed in Table 10.1.

Such statistics can be presented for each specified highway section (network link) and for each specified time period, to yield a level of detail that is both spatially and temporally disaggregated. Aggregations of these data, by subnetwork and network-wide, and over specified time periods, may also be provided.

10.3 Examples of Traffic Simulation Applications

Given the great diversity of applications that are suitable for the use of traffic simulation models, the following limited number of examples provides only a limited representation of past experience.

10.3.1 Evaluation of Signal Control Strategies

This study (Gartner and Hou, 1992) evaluated and compared the performance of two arterial traffic control strategies,

Table 10.1
Simulation Output Statistics: Measures of Effectiveness

Measure for Each Link and for Entire Network	
Travel: Vehicle-Miles	Bus Travel Time
Travel Time: Vehicle-minutes	Bus Moving Time
Moving Time: Vehicle-minutes	Bus Delay
Delay Time: Vehicle-minutes persons-minutes	Bus Efficiency: Moving Time Total Travel Time
Efficiency: Moving Time Total Travel Time	Bus Speed
Mean Travel Time per Vehicle-Mile	Bus Stops
Mean Delay per Vehicle-Mile	Time bus station capacity exceeded
Mean Travel Time per vehicle	Time bus station is empty
Mean Time in Queue	Fuel consumed
Mean Stopped Time	CO Emissions
Mean Speed	HC Emissions
Vehicle Stops	NOX Emissions
Link Volumes Occupancy	
Mean Link Storage Area Consumed	
Number of Signal Phase Failures	
Average Queue Length	
Maximum Queue Length	
Lane Changes	
Bus Trips	
Bus Person Trips	

MULTIBAND and MAXBAND, employing the TRAF-NETSIM simulation model. The paper describes the statistical analysis procedures, the number of simulation replications executed and the resulting 95 percent confidence intervals, and the results of the analysis.

Figure 10.1 is taken from this paper and illustrates how simulation can provide objective, accurate data sufficient to distinguish between the performance of alternative analytical models, within the framework of a controlled experiment.

10.3.2 Analysis of Equilibrium Dynamic Assignments (Mahmassani and Peeta, 1993)

This large-scale study used the DYNASMART simulation-assignment model to perform both user equilibrium (UE) and system optimal (SO) equilibrium calculations for a specified network, over a range of traffic loading conditions from unsaturated to oversaturated. This is an example of traffic simulation used as a component of a larger model to perform a complex analysis of an ITS initiative.

Figure 10.2 which is taken from the cited paper illustrates how simulation can produce internally consistent results for large scale projects, of sufficient resolution to distinguish between two comparable equilibrium assignment approaches.

10.3.3 Analysis of Corridor Design Alternatives (Korve Engineers 1996)

This analysis employed the WATSim simulation model to evaluate alternative scenarios for increasing capacity and improving traffic flow on a freeway connection, SR242, in

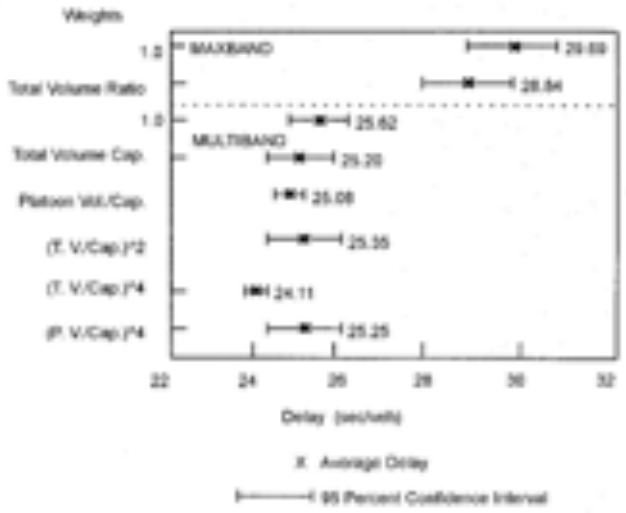


Figure 10.1
Average Delay Comparison, Canal Street, MULTIBAND & MAXBAND (KLD-242).

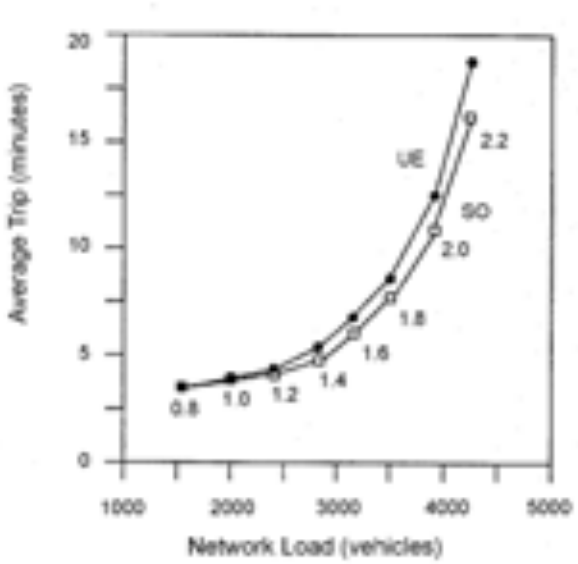


Figure 10.2
Comparison of average trip times (minutes) of SO & UE...(KLD-243).

California and ensuring a balanced design relative to freeway SR4 on the north and I680 to the south. Design alternatives considered for three future periods (years 2000, 2010, 2020) included geometric changes, widening, HOV lanes and ramp metering. Given the scale of this 20-mile corridor and the strong interactions of projected design changes for the three highways, the use of simulation provided a statistical basis for quantifying the operational performance of the corridor sections for each alternative.

This example illustrates the use of simulation as an element of the design process with the capability of analyzing candidate designs of large-scale highway systems in a manner that lies beyond the capabilities of a straight-forward HCM analysis.

10.3.4 Testing New Concepts

The TRAF-NETSIM simulation model was used by Rathi and Lieberman (1989) to determine whether the application of metering control along the periphery of a congested urban area could mitigate the extent and duration of congestion within the area, thereby improving performance and productivity. Here, a new control concept was tested on a real-world test-bed: a section of Manhattan. This example illustrates the value in testing new "high risk" ideas with simulation without exposing the public to possible adverse consequences, and prior to expending resources to implement these concepts.

These examples certainly do not represent the full range of traffic simulation applications. Yet, they demonstrate the application of traffic simulation in the areas of (1) traffic control; (2) transportation planning; (3) design; and (4) research.

10.4 Classification of Simulation Models

Almost all traffic simulation models describe *dynamical* systems -- *time* is always the basic independent variable. *Continuous* simulation models describe how the elements of a system change

state continuously over time in response to continuous stimuli. *Discrete* simulation models represent real-world systems (that are either continuous or discrete) by asserting that their *states* change abruptly at points in time. There are generally two types of discrete models:

- Discrete time
- Discrete event

The first, segments time into a succession of known time intervals. Within each such interval, the simulation model computes the *activities* which change the *states* of selected system elements. This approach is analogous to representing an initial-value differential equation in the form of a finite-difference expression with the independent variable, Δt.

Some systems are characterized by entities that are "idle" much of the time. For example, the state of a traffic signal indication (say, green) remains constant for many seconds until its state changes *instantaneously* to yellow. This abrupt change in state is called an *event*. Since it is possible to accurately describe the operation of the signal by recording its changes in state as a succession of [known or computed] timed events, considerable savings in computer time can be realized by only executing these events rather than computing the state of the signal second-by-second. For systems of limited size or those representing entities whose states change infrequently, discrete event simulations are more appropriate than are discrete time simulation models, and are far more economical in execution time. However, for systems where most entities experience a continuous change in state (*e.g.,* a traffic environment) and where the model objectives require very detailed descriptions, the discrete time model is likely to be the better choice.

Simulation models may also be classified according to the level of detail with which they represent the system to be studied:

- Microscopic (high fidelity)
- Mesoscopic (mixed fidelity)
- Macroscopic (low fidelity)

A *microscopic* model describes both the system entities and their interactions at a high level of detail. For example, a lane-change maneuver at this level could invoke the car-following law for the subject vehicle with respect to its current leader, then with respect to its putative leader *and* its putative follower in the target lane, as well as representing other detailed driver decision processes. The duration of the lane-change maneuver can also be calculated.

A *mesoscopic* model generally represents most entities at a high level of detail but describes their activities and interactions at a much lower level of detail than would a microscopic model. For example, the lane-change maneuver could be represented for individual vehicles as an instantaneous event with the decision based, say, on relative lane densities, rather than detailed vehicle interactions.

A *macroscopic* model describes entities and their activities and interactions at a low level of detail. For example, the traffic stream may be represented in some aggregate manner such as a statistical histogram or by scalar values of flow rate, density and speed. Lane change maneuvers would probably not be represented at all; the model may assert that the traffic stream is properly allocated to lanes or employ an approximation to this end.

High-fidelity microscopic models, and the resulting software, are costly to develop, execute and to maintain, relative to the lower fidelity models. While these detailed models possess the *potential* to be more accurate than their less detailed counterparts, this potential may not always be realized due to the complexity of their logic and the larger number of parameters that need to be calibrated.

Lower-fidelity models are easier and less costly to develop, execute and to maintain. They carry a risk that their representation of the real-world system may be less accurate, less valid or perhaps, inadequate. Use of lower-fidelity simulations is appropriate if:

- The results are not sensitive to microscopic details.
- The scale of the application cannot accommodate the higher execution time of the microscopic model.
- The available model development time and resources are limited.

Within each level of detail, the developer has wide latitude in designing the simulation model. The developer must identify the sensitivity of the model's performance to the underlying features of the real-world process. For example, if the model is to be used to analyze weaving sections, then a detailed treatment of lane-change interactions would be required, implying the need for a micro- or mesoscopic model. On the other hand, if the

model is designed for freeways characterized by limited merging and no weaving, describing the lane-change interactions in great detail is of lesser importance, and a macroscopic model may be the suitable choice.

Another classification addresses the processes represented by the model: (1) Deterministic; and (2) Stochastic. *Deterministic* models have no random variables; all entity interactions are defined by exact relationships (mathematical, statistical or logical). *Stochastic* models have processes which include probability functions. For example, a car-following model can be formulated either as a deterministic or stochastic relationship by defining the driver's reaction time as a constant value or as a random variable, respectively.

Traffic simulation models have taken many forms depending on their anticipated uses. Table 10.2 lists the TRAF family of models developed for the Federal Highway Administration (FHWA), along with other prominent models, and indicates their respective classifications. This listing is necessarily limited. Some traffic simulation models consider a single facility (NETSIM, NETFLO 1 and 2: surface streets; FRESIM, FREFLO: freeways; ROADSIM: two-lane rural roads; (CORSIM) integrates two other simulation models, FRESIM and NETSIM; INTEGRATION, DYNASMART, TRANSIMS are components of larger systems which include demand models and control policies; while CARSIM is a stand-alone simulation of a car-following model. It is seen that traffic simulation models take many forms, each of which satisfies a specific area of application.

Table 10.2
Representative Traffic Simulation Models

Name	Discrete Time	Discrete Event	Micro	Mesoscopic	Macro	Deterministic	Stochastic
NETSIM	X		X				X
NETFLO 1		X		X			X
NETFLO 2	X				X	X	
FREFLO	X				X	X	
ROADSIM	X		X				X
FRESIM	X		X				X
CORSIM	X		X				X
INTEGRATION	X		X				X
DYNASMART	X			X		X	
CARSIM	X		X				X
TRANSIMS	X			X			X

10.5 Building Traffic Simulation Models

The development of a traffic simulation model involves the following activities:

1) Define the Problem and the Model Objectives
   - State the purpose for which the model is being developed.
   - Define the information that the model must provide.

2) Define the System to be Studied
   - Disaggregate the system to identify its major components.
   - Define the major interactions of these components.
   - Identify the information needed as inputs.
   - Bound the domain of the system to be modeled.

3) Develop the Model
   - Identify the level of complexity needed to satisfy the stated objectives.
   - Classify the model and define its inputs and outputs.
   - Define the flow of data within the model.
   - Define the functions and processes of the model components.
   - Determine the calibration requirements and form: scalars, statistical distributions, parametric dependencies.
   - Develop abstractions (*i.e.*, mathematical-logical-statistical algorithms) of each major system component, their activities and interactions.
   - Create a logical structure for integrating these model components to support the flow of data among them.
   - Select the software development paradigm, programming language(s), user interface, presentation formats of model results.
   - Design the software: simulation, structured or object-oriented programming language; database, relational/object oriented.
   - Document the logic and all computational procedures.
   - Develop the software code and debug.

4) Calibrate the Model
   - Collect/acquire data to calibrate the model.
   - Introduce this data into the model.

5) Model Verification
   - Establish that the software executes in accord with the design specification.
   - Perform verification at the model component level.

6) Model Validation
   - Collect, reduce, organize data for purposes of validation.
   - Establish that the model describes the real system at an acceptable level of accuracy over its entire domain of operation; apply rigorous statistical testing methods.

7) Documentation
   - Executive Summary
   - Users Manual
   - Model documentation: algorithms and software

The development of a traffic simulation model is not a "single-pass" process. At each step in the above sequence, the analyst must review the activities completed earlier to determine whether a revision/extension is required before proceeding further. For example, in step 5 the analyst may verify that the software is replicating a model component properly as designed, but that its performance is at a variance with theoretical expectations or with empirical observations. The analyst must then determine whether the calibration is adequate and accurate (step 4); whether the model's logical/mathematical design is correct and complete (step 3); whether all interactions with other model components are properly accounted for and that the specified inputs are adequate in number and accuracy (step 2). This continual feedback is essential; clearly, it would be pointless to proceed with validation (step 6) if it is *known* that the verification activity is incomplete.

Step 3 may be viewed as the most creative activity of the development process. The simulation logic must represent all relevant interactions by suitably exploring the universe of possibilities and representing the likely outcome. These combinations of interactions are called *processes* which

represent specified *functions* and utilize *component models*. A small sample of these is presented below.

10.5.1 Car-Following

One fundamental interaction present in all microscopic traffic simulation models is that between a leader-follower pair of vehicles traveling in the same lane. This interaction takes the form of a stimulus-response mechanism:

$$a_f = F\ (v_l, v_f, s, d_l, d_f, R_f, P_i) \qquad (10.1)$$

where a_f, the acceleration (response) of the follower vehicle, is dependent on a number of (stimulus) factors including:

v_l, v_f	=	Speeds of leader, follower vehicles, respectively.
s	=	Separation distance.
d_l, d_f	=	Projected deceleration rates of the leader, follower vehicles, respectively.
R_f	=	Reaction time of the driver in the following vehicle.
P_i	=	Other parameters specific to the car-following model.
$F(\bullet)$	=	A mathematical and logical formulation relating the response parameter to the stimulus factors.

This behavioral model can be referenced (i.e., executed) to support other behavioral models such as lane-changing, merging, etc.

10.5.2 Random Number Generation

All stochastic models must have the ability to generate random numbers. Generation of random numbers has historically been an area of interest for researchers and practitioners. Before computers were invented, people relied on mechanical devices and their observations to generate random numbers. While numerous methods in terms of computer programs have been devised to generate random numbers, these numbers only "appear" to be random. This is the reason why some call them *pseudo-random* numbers.

The most popular approach for random number generation is the "linear congruential method" which employs a recursive equation to produce a sequence of random integers S as:

$$S_i = (aS_{i-1} + b) \bmod c.$$

where the integers chosen are defined as,
c is the modulus, such that $c > 0$,
a is the multiplier such that $0 < a < c$,
b is the increment such that $0 < b < m$, and
S_o is the starting value or the *Seed* of the random number generator, such that $0 < S_o < c$.

The ith random number denoted by R_i is then generated as

$$R_i = \frac{S_i}{c}.$$

These random number generations are typically used to generate random numbers between 0 and 1. That is, a Uniform (0,1) random number is generated. *Random variates* are usually referred to as the sample generated from a distribution other than the Uniform (0,1). More often than not, these random variates are generated from the Uniform (0,1) random number. A simulation usually needs random variates during its execution. Based on the distribution specified, there are various analytical methods employed by the simulation models to generate the random variates. The reader is referred to Law and Kelton (1991) or Roberts (1983) for a detailed treatment on this topic. As an example, random variates in traffic simulation are used to generate a stream of vehicles.

10.5.3 Vehicle Generation

At the outset of a simulation run, the system is "empty". Vehicles are generated at origin points, usually at the periphery of the analysis network, according to some headway distribution based on specified volumes. For example, the shifted negative exponential distribution will yield the following expression:

$$h = (H - h_{\min})\ [-\ \ln(1-R)] + H - h_{\min}$$

where

h = Headway (sec) separating vehicle emissions
H = Mean headway = 3600/V, where V is the specified volume, vph
h_{min} = Specified minimum headway (e.g., 1.2 sec/veh)
R = Random number in the range (0 to 1.0), obtained from a pseudo-random number generator.

Suppose the specified volume, V (vph), applies for a 15-minute period. If the user elects to *guarantee* that V is explicitly satisfied by the simulation model, it is necessary to generate N values of h using the above formula repeatedly, generating a new random number each time. Here, $N = V/4$ is the *expected* number of vehicles to be emitted in 15 minutes. The model could then calculate the factor, K:

$$K = \frac{15x60}{\sum_{i=1}^{N} h_i} \qquad (10.2)$$

The model would then multiply each of the N values of h_i, by K, so that the resulting sum of (the revised) h_i will be exactly 15 minutes, ensuring that the user's specification of demand volume are satisfied. However, if $K \neq 1.0$, then the resulting distribution of generated vehicles is altered and one element of stochasticity (*i.e.*, the actual number of generated vehicles) is removed. The model developer must either include this treatment (*i.e.*, eq.10.2), exclude it; or offer it as a user option with appropriate documentation.

10.5.4 A Representative Model Component

Consider two *elements* of every traffic environment: (1) a vehicle and (2) its driver. Each element can be defined in terms of its relevant *attributes*:

Vehicle: Length; width; acceleration limits; deceleration limit; maximum speed; type (auto, bus, truck, ...); maximum turn radius, etc.

Driver: Aggressiveness; responsiveness to stimuli; destination (route); other behavioral and decision processes.

Each attribute must be represented by the analyst, some by scalars (e.g., vehicle length); some by a functional relationship (e.g., maximum vehicle acceleration as a function of its current speed); some by a probability distribution (e.g., driver gap acceptance behavior). All must be calibrated.

The driver-vehicle combination forms a model component, or entity. This component is defined in terms of its own elemental attributes and its functionality is defined in terms of the interactions between these elements. For example, the driver's decision to accelerate at a certain rate may be constrained by the vehicle's operational limitations. In addition, this system component interacts with other model entities representing the environment under study, including:

- roadway geometrics
- intersection configurations
- nearby driver-vehicle entities
- control devices
- lane channelization
- conflicting vehicle movements

As an example, the driver-vehicle entity's interaction with a control device depends on the type and current state of the device (*e.g.*, a signal with a red indication), the vehicle's speed, its distance from stop-bar, the driver's aggressiveness, etc. It is the developer's responsibility to design the model components and their interactions in a manner that satisfies the model objectives and is consistent with its fidelity.

10.5.5 Programming Considerations

Programming languages, in the context of this chapter, may be classified as *simulation* and *general-purpose* languages. Simulation languages such as SIMSCRIPT and GPSS/H greatly ease the task of developing simulation software by incorporating many features which compile statistics and perform queuing and other functions common to discrete simulation modeling.

General-purpose languages may be classified as *procedural* (*e.g.*, FORTRAN, PASCAL, C, BASIC), or *object-oriented* (*e.g.*, SMALLTALK, C++, JAVA). Object-oriented languages are gaining prominence since they support the concept of reusable software defining *objects* which communicate with one another to solve a programming task. Unlike procedural languages, where the functions are separated from, and operate upon, the data base, objects *encapsulate* both data describing its state, as well as operations (or "methods") which can change its state and interact with other objects.

While object-oriented languages can produce more reliable software, they require a higher level of programming skill than do procedural languages. The developer should select a language which is hardware independent, is supported by the major operating systems and is expected to have a long life, given the rapid changes in the world of software engineering. Other factors which can influence the language selection process include: (1) the expected life of the simulation model; (2) the skills of the user community; (3) available budget (time and resources) to develop and maintain the software; and (4) a realistic assessment of available software development skills.

10.6 An Illustration of Simulation Model Building

Given the confines of this chapter, we will illustrate the model development process by presenting the highlights of a sample problem, but avoiding exhaustive detail.

1) ***Define the Problem and Model Objectives*** - An existing microscopic stochastic simulation model of freeway traffic does not consider lane-change operations. It has been determined that this model's results are unreliable as a result. The purpose of this project is to introduce additional logic into the model to represent lane-changing operations. This addition should provide improved accuracy in estimating speed and delay; in addition it will compute estimates of lane changes by lane, by vehicle type and by direction (to the left and to the right).

2) ***Define the System*** -

   a) A freeway of up to six lanes -- level tangent
   b) Three vehicle types: passenger car; single-unit truck; tractor-trailer truck
   c) Required inputs: traffic volume (varies with time); distributions of free-speed, of acceptable risk (expressed in terms of deceleration rates if lead vehicle brakes), of motivation to change lanes, all disaggregated by vehicle type
   d) Drivers are randomly assigned an "aggressiveness index" ranging from 1 (very aggressive) to 10 (very cautious) drawn from a uniform distribution to represent the range of human behavior.

   It must be emphasized that model development is an *iterative* process. For example, the need for the indicated input distributions may not have been recognized during this definition phase, but may have emerged later during the logical design. Note also that the problem is bounded -- no grades or horizontal curves are to be considered at this time. See Figure 10.3 for the form of these distributions.

3) ***Develop the Model*** - Since this lane-change model is to be introduced into an existing microscopic stochastic model, using a procedural language, it will be designed to utilize the existing software. The model logic moves each vehicle, each time-step, Δt, starting with the farthest downstream vehicle, then moving the closest upstream vehicle regardless of lane position, etc. At time, t_o, the vehicle *states* are shown in Figure 10.4(a).

   In developing the model, it is essential to identify the independent *functions* that need to be performed and to segregate each function into a separate software module, or routine. Figure 10.5 depicts the *structure* -- not the flow -- of the software. This structure shows which routines are logically connected, with data flowing between them. Some routines reference others more than once, demonstrating the benefits of disaggregating the software into functionally independent modules.

   As indicated in Table 10.3 which presents the algorithm for the Lane Change Executive Routine in both "Structured English" or "pseudo-code" and as a flow chart, traffic

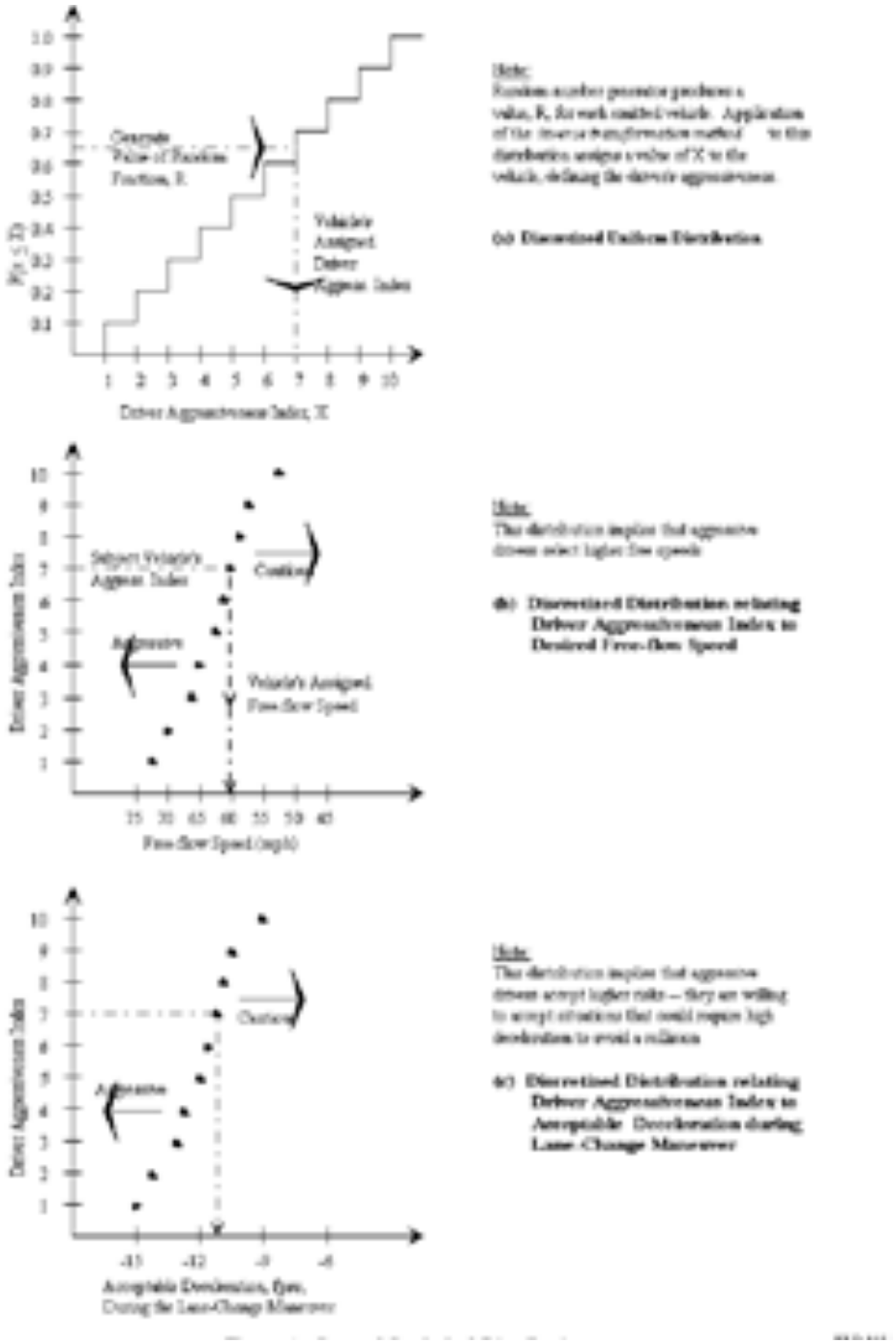
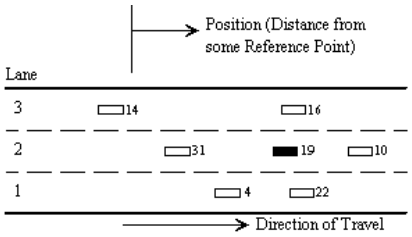
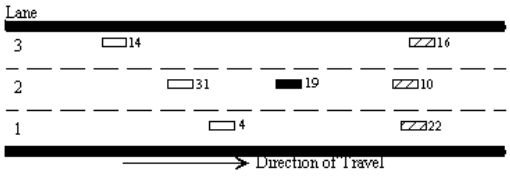


Figure 10.3
Several Statistical Distributions.

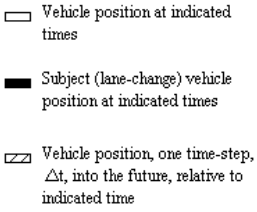


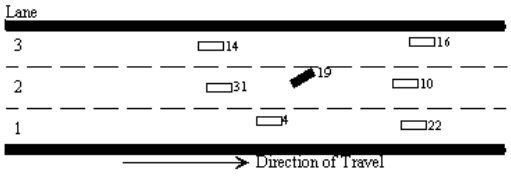
(a) State at time, t_o, before any vehicles have been moved (all vehicle positions are synchronous)

Veh. ID	Longitudinal/Lane Position at time t_o	$t_o+\Delta t$	$t_o+2\Delta t$
10	216/2	181/2	145/2
16	240/3	169/3	119/3
22	243/1	176/1	111/1
19	250/2	216/2+3	174/3
4	276/1	230/1	184/1
31	301/2	255/2	208/2
14	343/3	278/3	215/3

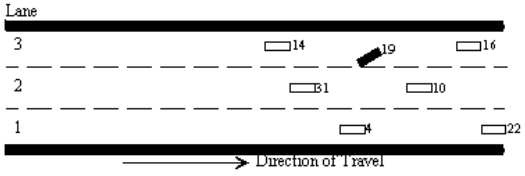


(b) State after vehicles 10, 16 and 22 have been moved (to $t_o + \Delta t$); the others have not been processed as yet (at t_o)





(c) State at time, $t_o + \Delta t$; vehicle positions are synchronous



(d) State at time, $t_o + 2\Delta t$; vehicle positions are synchronous

Figure 10.4
Vehicle Positions during Lane-Change Maneuver

Figure 10.5
Structure Chart of Simulation Modules

Table 10.3
Executive Routine

For each vehicle, I:
CALL routine MOTIV to determine whether this driver is "motivated" to change lanes, now
IF so, THEN
CALL routine CANLN to identify which of neighboring lanes (if either) are
acceptable as potential target lanes
IF the lane to the right is acceptable, THEN
CALL routine CHKLC to determine whether a lane-change is feasible, now.
Set flag if so.
ENDIF
IF the lane to the left is acceptable, THEN
CALL routine CHKLC to determine whether a lane-change is feasible, now.
Set flag, if so.
ENDIF
IF both lane-change flags are set (lane-change is feasible in either direction), THEN
CALL routine SCORE to determine more favorable target lane
ELSE IF one lane-change flag is set, THEN
Identify that lane
ENDIF
IF a [favored] target lane exists, THEN
CALL routine LCHNG to execute the lane-change
Update lane-change statistics
ELSE
CALL routine CRFLW to move vehicle within this lane
Set vehicle's process code (to indicate vehicle has been moved this time-step)
ENDIF
ELSE (no lane-change desired)
CALL routine CRFLW to move vehicle within its current lane
Set vehicle's process code
ENDIF
continue...

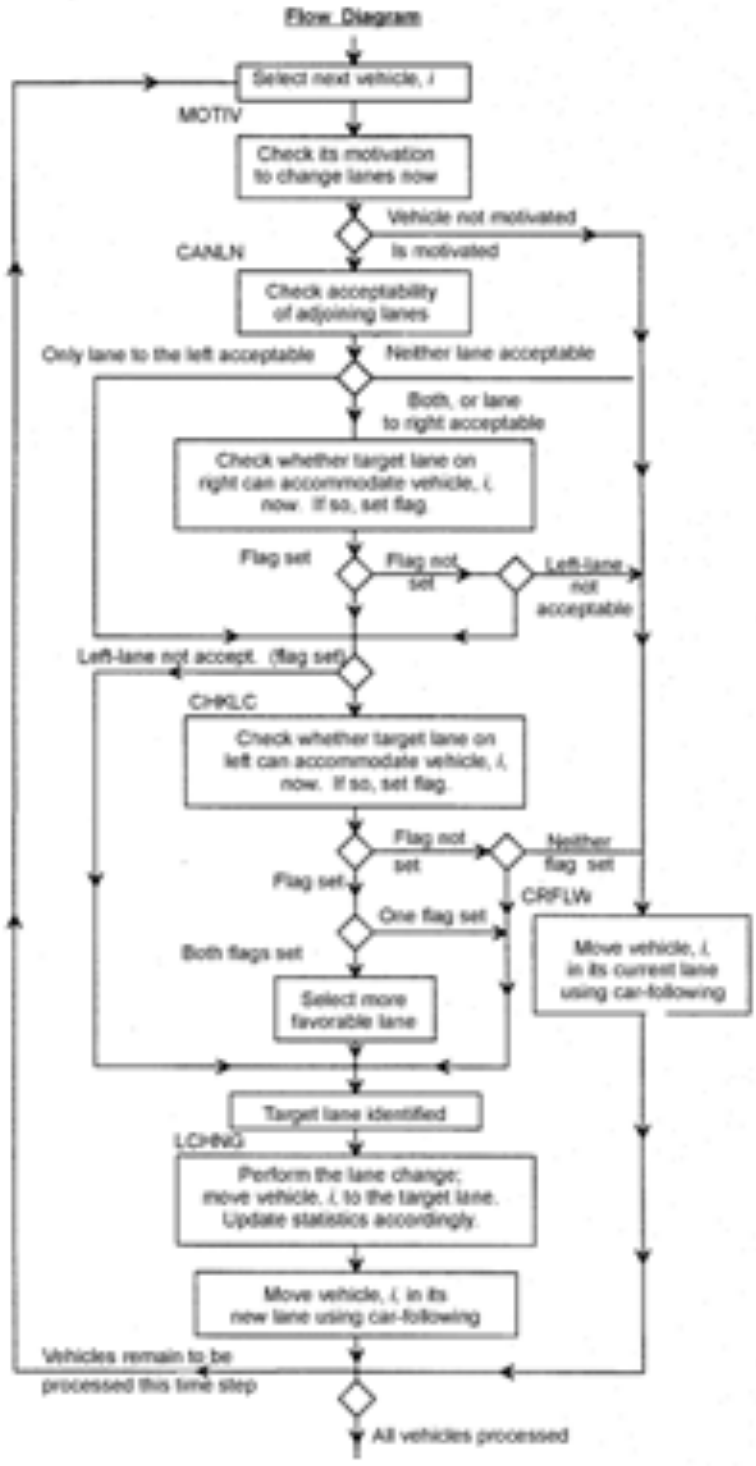


Figure 10.6
Flow Diagram.

simulation models are primarily *logical*, rather than *computational* in context. This property reflects the fact that traffic operations are largely the outcome of driver decisions which themselves are [hopefully!] logical in context. As the vehicles are processed by the model logic, they transition from one *state* to the next. The reader should reference Table 10.3 and Figures 10.3, 10.4, 10.5 and 10.6 to follow the discussion given below for each routine. The "subject vehicle" is shown as number 19 in Figure 10.4.

Executive: Controls the flow of processing, activating (through CALLs) routines to perform the necessary functions. Also updates lane-change statistics.

MOTIV: Determines whether a lane-change is *required* to position the subject vehicle for a downstream maneuver or is *desired* to improve the vehicle's operation (increase its speed).

CANLN: Determines whether either or both adjoining lanes (1 and 3) are suitable for servicing the subject vehicle.

CHKLC: Identifies vehicles 22 and 4 as the leader and follower, respectively, in target Lane 1; and vehicles 16 and 14, respectively, in target Lane 3. The car-following dynamics between the pairs of vehicles, 19 and 22, then 4 and 19 are quantified to assess the prospects for a lane-change to Lane 1. Subsequently, the process is repeated between the pairs of vehicles, 19 and 16; then 14 and 19; for a lane-change to Lane 3. If the gap is inadequate in Lane 1, causing an excessive, and possibly impossible deceleration by either the subject vehicle, 19, or the target follower, 4, to avoid a collision, then Lane 3, would be identified as the only feasible target lane. In any case, CHKLC would identify either Lane 1 or Lane 3, or both, or neither, as acceptable target lanes at this time, depending on safety considerations.

SCORE: If both adjoining target lanes are acceptable, then this *heuristic* algorithm emulates driver reasoning to select the preferable target lane. It is reasonable to expect that the target lane with a higher-speed leader and fewer vehicles -- especially trucks and buses -- would be more attractive. Such *reasonableness* algorithms (which are expressed as "rules" in expert systems), are also common in models which simulate human decisions.

LCHNG: After executing the subject vehicle's lane-change activity, the logic performs some needed bookkeeping:

At time, $t_o + \Delta t$, vehicle 19 acts as the leader for both vehicles 14 and 31, who must "follow" (and are constrained by) its presence.

At time, $t_o + 2\Delta t$, the logic asserts that the lane changer (no. 19) has committed to the lane-change and no longer influences its former follower, vehicle no. 31. Of course, vehicle 14 now follows the lane-change vehicle, 19.

4) ***Calibrate the Model*** - Figures 10.3(b) and (c) are distributions which represent the outcome of a calibration activity. The distribution of free-flow speed is site-specific and can be quantified by direct observation (using paired loop detectors or radar) when traffic conditions are light -- LOS A.

The distribution of acceptable decelerations would be very difficult to quantify by direct observation -- if not infeasible. Therefore, alternative approaches should be considered. For example...

- Gather video data (speeds, distance headway) of lane-change maneuvers. Then apply the car-following model with these data to "back-out" the implied acceptable decelerations. From a sample of adequate size, develop the distribution.

- On a more macro level, gather statistics of lane-changes for a section of highway. Execute the simulation model and adjust this distribution of acceptable deceleration rates until agreement is attained between the lane-changes executed by the model, and those observed in the real world. This is tenuous since it is confounded by the other model features, but may be the best viable approach.

It is seen that calibration -- the process of quantifying model parameters using real-world data -- is often a difficult and costly undertaking. Nevertheless, it is a

necessary undertaking that must be pursued with some creativity and tenacity.

5) ***Model Verification*** - Following *de-bugging*, verification is a structured regimen to provide assurance that the software performs as intended (Note: verification does not address the question, "Are the model components and their interactions *correct*?"). Since simulation models are primarily *logical* constructs, rather than computational ones, the analyst must perform detailed logical path analyses.

Verification is performed at two levels and generally in the sequence given below:

- Each software routine (bottom-up testing)
- Integration of "trees" (top-down testing)...

When completed, the model *developer* should be convinced that the model is performing in accord with expectations over its entire domain of application.

6) ***Model Validation*** - Validation establishes that the model behavior accurately and reliably represents the real-world system being simulated, over the range of conditions anticipated (i.e., the model's "domain"). Model validation involves the following activities:

- Acquiring real-world data which, to the extent possible, extends over the model's domain.
- Reducing and structuring these empirical data so that they are in the same format as the data generated by the model.
- Establishing validation criteria, stating the underlying hypotheses and selecting the statistical tests to be applied.
- Developing the experimental design of the validation study, including a variety of "scenarios" to be examined.
- Performing the validation study:
  - Executing the model using input data and calibration data representing the real-world conditions.
  - Performing the hypothesis testing.
- Identifying the causes for any failure to satisfy the validation tests and repairing the model accordingly.
- Validation should be performed at the component system level as well as for the model as a whole. For example, Figure 10.7 compares the results produced by a car-following model, with field data collected with aerial photographs. Such *face* validation offers strong assurance that the model is valid. This validation activity is *iterative* -- as differences between the model results and the real-world data emerge, the developer must "repair" the model, then revalidate. Considerable skill (and persistence) are needed to successfully validate a traffic simulation model.

7) ***Documentation*** - Traffic simulation models, as is the case for virtually all transportation models, are *data intensive*. This implies that users must invest effort in data acquisition and input preparation to make use of these models. Consequently, it is *essential* that the model be documented for...

- The end user, to provide a "friendly" interface to ease the burden of model application.
- Software maintenance personnel.
- Supervisory personnel who must assess the potential benefits of using the model.

10.7 Applying Traffic Simulation Models

Considerable skill and attention to detail must be exercised by the user in order to derive accurate and reliable results from a simulation-based analysis. The following procedure is recommended:

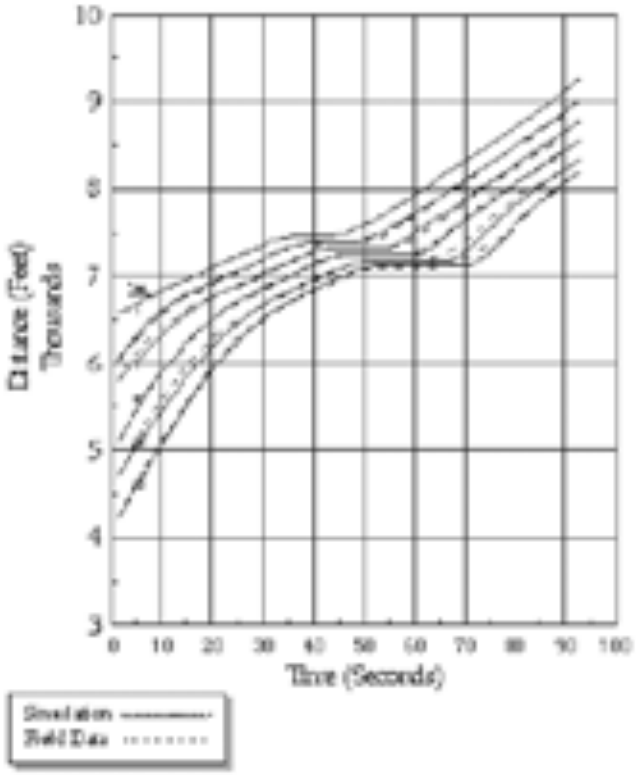


Figure 10.7
Comparison of Trajectories of Vehicles from Simulation Versus Field Data for Platoon 123.

Identify the Problem Domain

What highway facilities are involved?
- Surface streets (grid, arterial, both), freeways, rural roads, toll plaza.

What is the traffic environment?
- Autos, trucks, buses, LRT, HOV...
- Unsaturated, oversaturated conditions.

What is the control environment?
- Signals (fixed time, actuated, computer-controlled), signs
- Route guidance

What is the size of the network and duration of the analysis period?

Define the Purpose of the Study

What information is needed from the simulation model?
- Statistical: MOE sought, level of detail.
- Pictorial: static graphics, animation.

What information is available as input and calibration data?
- Consider expected accuracy and reliability.
- Consider available budget for data acquisition.

Are results needed on a *relative* or *absolute* basis?

What other functions and tools are involved?
- Capacity analysis
- Design
- Demand modeling
- Signal optimization

Is the application real-time or off-line?

Investigate Candidate Traffic Simulation Models

Identify strengths and limitations of each.
- Underlying assumptions

- Computing requirements
- Availability, clarity, completeness of documentation
- Availability, reliability, timeliness of software support

Estimate extent and cost of data collection for calibration and input preparation.

Determine whether model features match problem needs.

Assess level of skill needed to properly apply model.

Determine compatibility with other tools/procedures needed for the analysis.

Assess the Need to Use a Traffic Simulation Model

Is traffic simulation necessary to perform the analysis of the problem?

- Are other tools adequate but less costly?
- Are your skills adequate to properly apply simulation?
- Can the data needed by traffic simulation be acquired?

Is traffic simulation highly beneficial even though not necessary?

- Simulation results can confirm results obtained by other tools.
- Animation displays needed as a presentation medium.

If it is determined that traffic simulation is needed/advisable, continue.

Select Traffic Simulation Model

Relate relevant model attributes to problem needs.

Determine which model satisfies problem needs to the greatest extent. Consider technical, and cost, time, available skills and support, and risks factors.

Data Acquisition

Obtain reliable records of required information.

- Design drawings for geometrics.
- Signal timing plans, actuated controller settings.
- Traffic volume and patterns; traffic composition.
- Transit schedules
- Other, as required.

Confirm the accuracy of these data through field observation.

Undertake field data collection for input and for calibration, as required.

Identify need for accurate operational traffic data: based on model's sensitivity site-specific features; accuracy requirements.

- Select representative locations to acquire these field data.
- Collect data using video or other methods as required: saturation flow rates at intersections; free-flow speeds; acceptable gaps for permitted left-turns, etc.
- Accept model default values or other data from the literature with great care if data collection is infeasible or limited by cost considerations.

Model Calibration

Calibration is the activity of specifying data describing traffic operations and other features that are site-specific. These data may take the form of scalar elemets and of statistical distributions that are referenced by the logic of stochastic simulation models. While traffic simulation models generally provide default value which represent average conditions for these calibration data elements, it is the responsibility of the analyst to quantify these data with field observations to the extent practicable rather than to accept these default values.

Model Execution

The application of a simulation model should be viewed as performing a rigorous statistical experiment. The model must first be executed to *initialize* its database so that the data properly represents the initial state of the traffic environment. This requirement can reliably be realized if the environment is initially at equilibrium.

Thus, to perform an analysis of congested conditions, the analyst should design the experiment so that the initial state of the traffic environment is undersaturated, and then specify the changing

conditions which, over time, censors the congested state which is of interest. Similarly, the final state of the traffic environment should likewise be undersaturated, if feasible.

It is also essential that the analyst properly specify the dynamic (*i.e.*, changing) input conditions which describe the traffic environment. For example, if one-hour of traffic is to be simulated, the analyst should always specify the variation in demand volumes -- and in other variables -- over that hour at an appropriate level of detail rather than specifying average, constant values of volume.

Finally, if animation displays are provided by the model, this option should always be exercised, as discussed below.

Interpretation of Simulation Results

Quite possibly, this activity may be the most critical. It is the analyst who must determine whether the model results constitute a reasonable and valid representation of the traffic environment under study, and who is responsible for any inferences drawn from these results. Given the complex processes taking place in the real-world traffic environment, the analyst must be alert to the possibility that (1) the model's features may be deficient in adequately representing some important process; (2) the input data and/or calibration specified is inaccurate or inadequate; (3) the results provided are of insufficient detail to meet the project objectives; (4) the statistical analysis of the results are flawed (as discussed in the following section); or (5) the model has "bugs" or some of its algorithms are incorrect. Animation displays of the traffic environment (if available) are a most powerful tool for analyzing simulation results. A *careful and thorough* review of this animation can be crucial to the analyst in identifying:

- Cause-and-effect relationships. Specifically the *origins* of congested conditions in the form of growing queues can be observed and related to the factors that caused it.
- Anomolous results (e.g., the creation and growth of queues when conditions are believed to be undersaturated) can be examined and traced to valid, incongested behavior; to errant input specifications; or to model deficiencies.

If the selected traffic simulation lacks an animation feature or if questions remain after viewing the animation, then the following procedures may be applied:

- Execute the model to replicate existing real-world conditions and compare its results with observed behavior. This "face validation", which is recommended regardless of the model selected, can identify model or implementation deficiencies.
- Perform "sensitivity" tests on the study network by varying key variables and observing model responses in a carefully designed succession of model executions.
- Plot these results. A review will probably uncover the perceived anomalies.

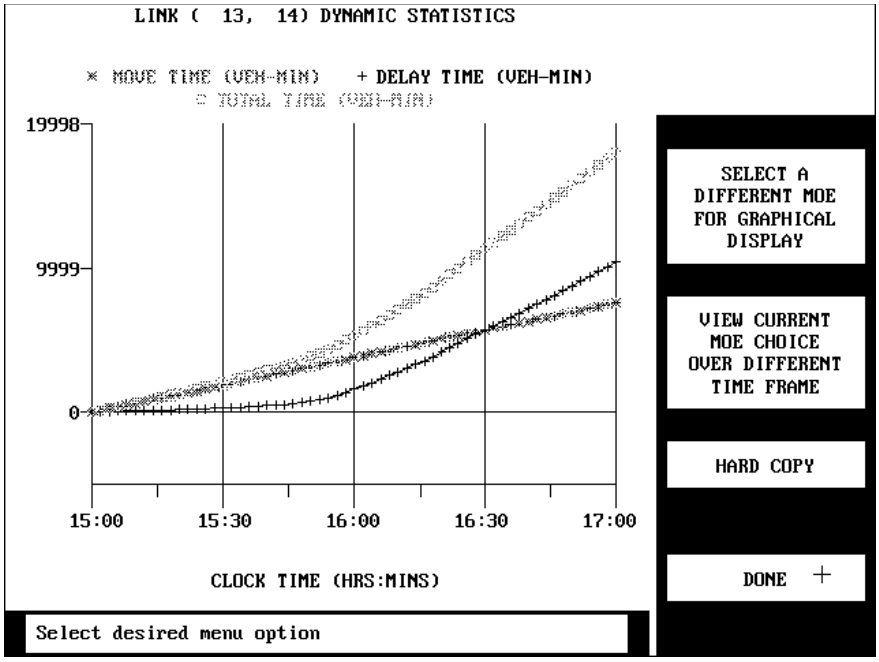
Table 10.1 lists representative data elements provided by traffic simulation models. Figure 10.8 shows typical graph displays while Figure 10.9 displays a "snapshot" of an animation screen.

Note that all the graphical displays can be accessed interactively by the user, thus affording the user an efficient means for extracting the sought *information* and *insights* from the mass of *data* compiled by the simulation model.

Proper output analysis is one of the most important aspects of any simulation study. A variety of techniques are used, particularly for stochastic model output, to arrive at inferences that are supportable by the output. A brief exposition to output analysis of simulation data is paresented next.

10.8 Statistical Analysis of Simulation Data

In most efforts on simulation studies, more often than not a large amount of time and money is spent on model development and coding, but little on analyzing the output data. Simulation practitioners therefore have more confidence in their results than is justified. Unfortunately, many simulation studies begin with a heuristic model, which is then programmed on the computer, and conclude with a single run of the program to yield *an answer*. This is a result of overlooking the fact that a simulation is a sampling experiment implemented on a computer and therefore needs appropriate statistical techniques for its design,



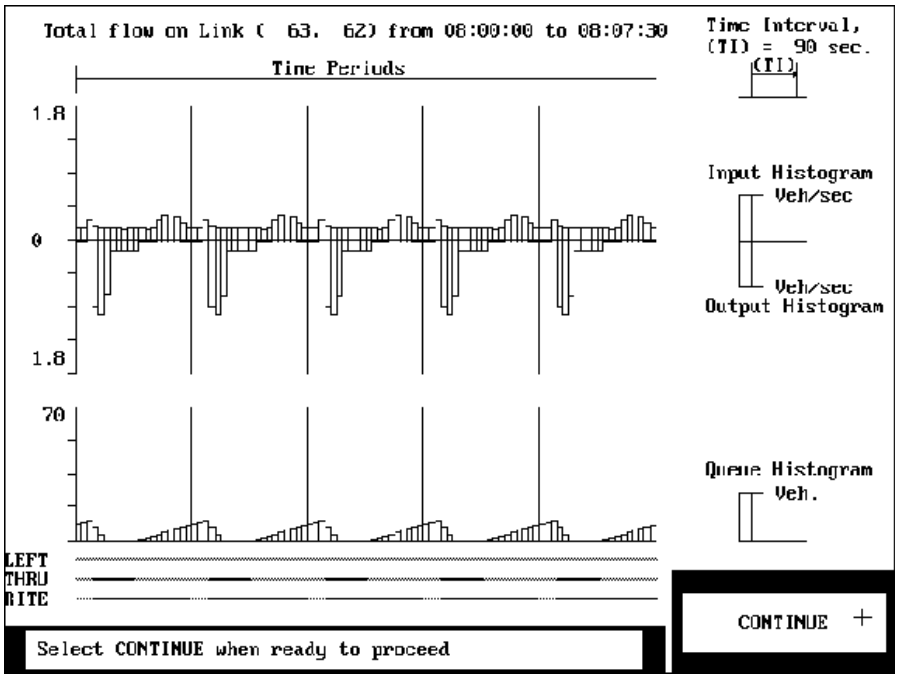


Figure 10.8
Graphical Displays.

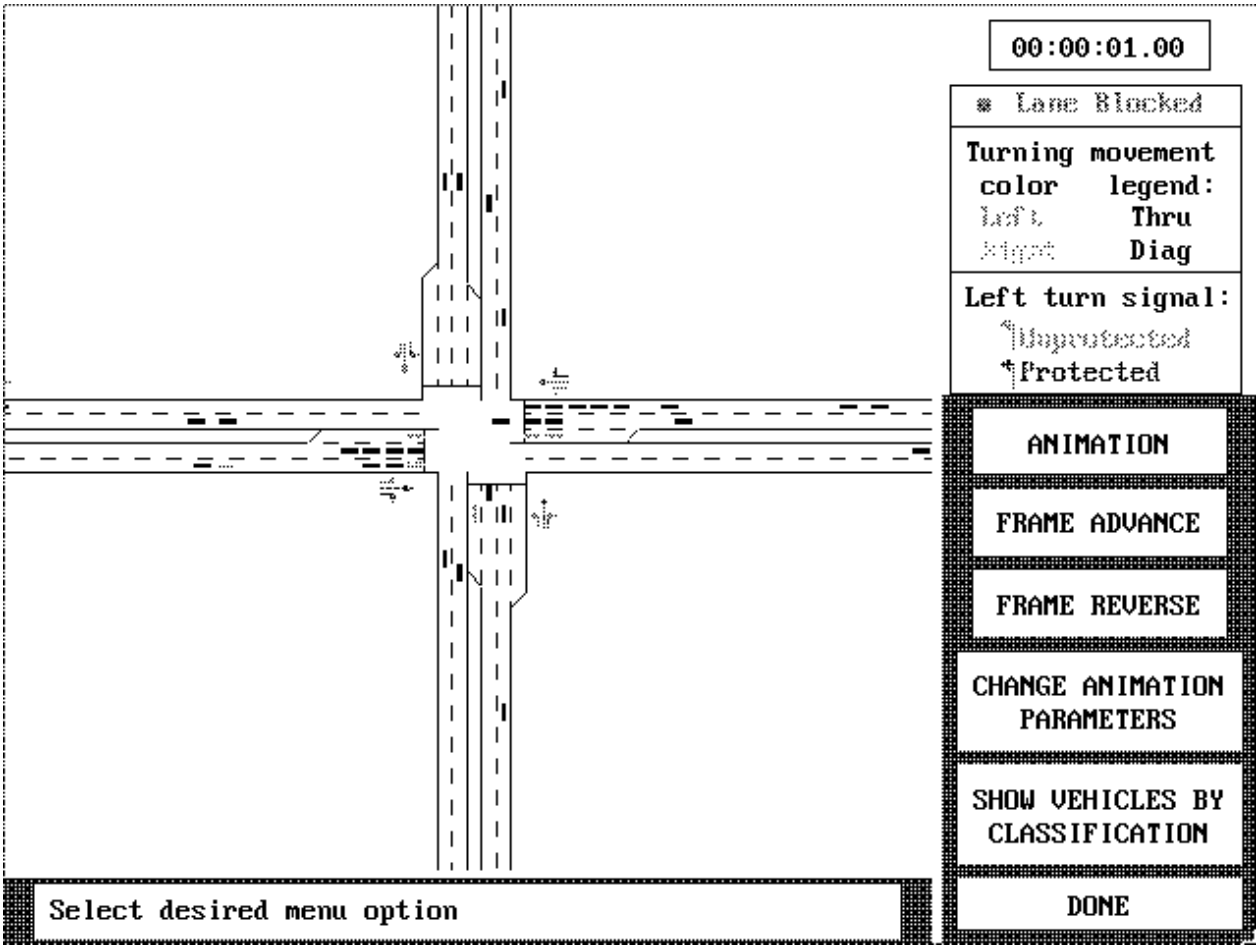


Figure 10.9
Animation Snapshot.

analysis, and implementation. Also, more often than not, output data from simulation experiments are auto correlated and nonstationary. This precludes its analysis using classical statistical techniques which are based on independent and identically distributed (IID) observations.

Typical goals of analyzing output data from simulation experiments are to present point estimates of the measures of effectiveness (MOE) and form confidence intervals around these estimates for one particular system design, or to establish which simulated system is the best among different alternate

configurations with respect to some specified MOE. Point estimates and confidence intervals for the MOEs can be obtained either from one simulation run or a set of replications of the system using independent random number streams across replications.

10.9 Looking to the Future

With the traffic simulation models now mounted on high-performance PCS, and with new graphical user interfaces (GUI) becoming available to ease the burden of input preparation, it is reasonable to expect that usage of these models will continue to increase significantly over the coming years.

In addition, technology-driven advances in computers, combined with the expanding needs of the Intelligent Transportation Systems (ITS) program, suggest that the new applications of traffic simulation can contribute importantly to this program. Specifically:

- Simulation support systems of Advanced Traffic Management Systems (ATMS) in the form of:
  - Off-line emulation to test, refine, evaluate, new real-time control policies.
  - On-line support to evaluate candidate ad-hoc responses to unscheduled events and to advise the operators at the Traffic Management Center (TMC) as to the "best" response. -Real-time component of an advanced control/guidance strategy. That is, the simulation model would be a component of the on-line strategy software.
  - Combining simulation with Artificial Intelligence (AI) software. The simulation can provide the knowledge base in real-time or generate it in advance as an off-line activity.
  - Integrating traffic simulation models with other tools such as: transportation demand models, signal optimization models, GIS, office suites, etc.
  - Providing Internet access.
- Real-time simulators which replicate the performance of TMC operations. These simulators must rely upon simulation models as "drivers" to provide the real-world stimuli to ITS real-time software being tested. Such simulators are invaluable for:
  - Testing new ATMS concepts prior to deployment.
  - Testing interfaces among neighboring TMCs.
  - Training TMC operating personnel.
  - Evaluating different ATMS architectures.
  - Educating practitioners and student.
  - Demonstrating the benefits of ITS programs to state and municipal officials and to the public through animated graphical displays and virtual reality.

References

Gartner, N.H. and D. L. Hou (1992). *Comparative Evaluation of Alternative Traffic Control Strategies,*" Transportation Research Record 1360, Transportation Research Board.

Mahmassani, H.S. and S. Peeta (1993). *Network Performance Under System Optimal and User Equilibrium Dynamic Assignments: Implications for Advanced Traveler Information Systems* Transportation Research Record 1408, Transportation Research Board.

Korve Engineers (1996). *State Route 242 Widening Project - Operations Analysis Report* to Contra Costa Transportation Authority.

Rathi, A.K. and E.B. Lieberman (1989). *Effectiveness of Traffic Restraint for a Congested Urban Network: A Simulation Study.* Transportation Research Record 1232.

KINETIC THEORIES

BY PAUL NELSON*

* Professor, Department of Computer Science, Texas A&M University, College Station, TX 77843-3112

11 Kinetic Theories

Criticisms and accomplishments of the Prigogine-Herman kinetic theory are reviewed. Two of the latter are identified as possible benchmarks, against which to measure proposed novel kinetic theories of vehicular traffic. Various kinetic theories that have been proposed in order to eliminate deficiencies of the Prigogine-Herman theory are assessed in this light. None are found to have yet been shown to meet both of these benchmarks.

Possible objectives and applications for kinetic theories of vehicular traffic are considered. One of these is the traditional application to the development of continuum models, with the resulting microscopically based coefficients. However, modern computing power makes it possible to consider computational solution of kinetic equations *per se,* and therefore direct applications of the kinetic theory (e.g., the kinetic distribution function). It is concluded that the primary applications are likely to be found among situations in which variability between instances is an important consideration (e.g., travel times, or driving cycles).

11.1 Introduction

On page 20 of their well-known monograph on the kinetic theory of vehicular traffic, Prigogine and Herman (1971) summarize possible alternate forms of the relaxation term in their kinetic equation of vehicular traffic. They conclude this discussion by issuing the invitation that "the reader may, if he is so inclined, work out the theory using other forms of the relaxation law." This invitation to explore alternate kinetic models of vehicular traffic has subsequently been accepted by a number of workers, most notably by Paveri-Fontana (1975) and by Phillips (1979, 1977), and more recently by Nelson (1995a), and by Klar and Wegener (1999). The existence of this variety of kinetic models of vehicular traffic raises the issue of how one chooses between them in any particular application; more generally there arises the issue of the types of applications for which any kinetic model has a role. In these lights, the primary objective of this chapter is to address questions related to what might reasonably be expected from a good kinetic theory of vehicular traffic.

The approach presented here is substantially influenced by the work of Nagel (1996), who gave an excellent review of a variety of types of models of vehicular traffic, including continuum ("hydrodynamic"), car-following and particle hopping (cellular automata) models. In particular, he has emphasized that: *i)* any model necessarily represents some compromise in terms of its *fidelity* in describing the reality it is intended to represent; *ii)* different types of models represent engineering judgements as to the relative importance of *resolution,* fidelity and *scale* for the particular application at hand. To some extent, this chapter is intended to address similar issues for *kinetic* models of vehicular traffic. The status of various kinetic models will also be reviewed, in terms of achieving two objectives that seem appropriate to designate as benchmarks, primarily on the basis that the seminal kinetic model of Prigogine and Herman (1971) has been shown to meet those objectives.

It seems appropriate to view kinetic models as occupying a point on the model spectrum that is intermediate between continuum (e.g., hydrodynamic) models and microscopic (e.g., car-following or cellular automata) models.[1] One of the primary applications of kinetic models is to obtain continuum models in a consistent manner from an underlying microscopic model of driver behavior. (See Nelson (1995b) for further thoughts on the role of kinetic models of vehicular traffic as a bridge from microscopic models to macroscopic models.) However, computing power now has advanced to the point that it is practical to consider computational solution of kinetic equations *per se.* This opens the door to the realistic possibility of applying kinetic models directly to the simulation of traffic flow. This is a qualitatively different situation from that prevailing in the 1960's, when kinetic models of vehicular traffic were initially proposed by Prigogine, Herman and co-workers. (See Prigogine and Herman, 1971, and works cited therein.)

The specific further contents of this chapter are as follows. In Section 2 below, the status of the Prigogine-

[1] The word *mesoscopic* has come into recent vogue to describe models that are, in some sense, intermediate between macroscopic and microscopic models.

Herman (1971) kinetic model of vehicular traffic is reviewed. The intent of this review is to provide an evenhanded discussion of both the deficiencies and signal accomplishments of this seminal kinetic theory of vehicular traffic. Two of these accomplishments are suggested as benchmarks that should minimally be met by any proposed novel kinetic model of vehicular traffic. In Section 3 alternate kinetic models that have been proposed in the literature are assessed against these benchmarks, and none are found that yet have been shown to meet both of them. Both of these benchmarks relate to the equilibrium solutions of the Prigogine-Herman kinetic equation, and one of them relates to the recent result of Nelson and Sopasakis (1998) to the effect that under certain circumstances – particularly for sufficiently congested traffic – the Prigogine-Herman model admits a two-parameter family of equilibrium solutions, as opposed to the one-parameter (density) family that would be expected classically.

In Section 4, the role of kinetic equations as a bridge from microscopic to continuum models is considered. Section 5 is devoted to consideration of the potential applications of the solution of kinetic equations *per se.*

11.2 Status of the Prigogine-Herman Kinetic Model

The kinetic model of Prigogine and Herman (1971) is summarized in Subsection 2.1. A number of published criticisms of this model, along with alternative models that have been suggested to overcome some of these criticisms, are reviewed in Subsection 2.2. In Subsection 2.3 two significant accomplishments of the Prigogine-Herman theory are described, and suggested as benchmarks against which novel kinetic theories of vehicular traffic should be measured.

11.2.1 The Prigogine-Herman Model

The kinetic equation of Prigogine and Herman is

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = -\frac{f - f_0}{T} + c(\bar{v} - v)(1 - P)f. \qquad (11.1)$$

Here the various symbols have the following meanings:

a) the zero order moment of *f(x,v,t),*

$$c(x,t) = \int_0^\infty f(x,v,t)\,dv,$$

is vehicular density.

b) the ratio of the first and zero order moments,

$$\bar{v}(x,t) = \int_0^\infty vf(x,v,t)\,dv \Big/ c(x,t)$$

is mean vehicular speed;

c) P is passing probability;

d) T is the relaxation time;

e) f_0 is the corresponding density function for the desired speed of vehicles;

f) f is the density function for the distribution of vehicles in phase space, so that

$$\int_{v_1}^{v_2}\int_{x_1}^{x_2} f(x,v,t)\,dv\,dx$$

is the expected number of vehicles at time t that have position between x_1 and x_2 and speed between v_1 and v_2 ($x_1 \le x_2$ and $v_1 \le v_2$).

The second term on the left-hand side of Eq. (1), the *streaming* term, represents the rate of change of the density function due to motion of the traffic stream, absent any changes of velocity by vehicles. The first term on the right-hand side, which is often called the *relaxation* term, is the contribution to this rate of change that stems from changes of vehicular speed associated with passing or other causes of acceleration. The second term on the right-hand side, the *slowing-down* term, stems from deceleration of vehicles that overtake slower-moving vehicles. The relaxation term is phenomenological in nature, in that it is based on the underlying assumption that increases in vehicular speed cause the actual density to "relax" toward the desired density with some characteristic time T. By contrast, the slowing-down term can be obtained from basic physical arguments, albeit with idealized assumptions such as instantaneous deceleration, treatment of vehicles as point particles (i.e., neglect of the positive length of vehicles), and the validity of what Paveri-Fontana (1975) terms *vehicular chaos.* The validity of both of these particular forms of the rates of change due to changes of speed has been

(1)

questioned, as will be briefly discussed in the following subsection.

A *kinetic equation* generally is an equation that in principle, subject to appropriate initial and boundary conditions, can be solved for the density function *f*, as defined above. Some kinetic equations that are alternatives to that of Prigogine and Herman are discussed in Section 3 following

11.2.2 Criticisms of the Prigogine-Herman Model.

The first published serious critique of the Prigogine-Herman kinetic equation seems to be the work of Munjal and Pahl (1969). These workers raise a number of questions,[2] but the most fundamental of these fall into one of the following two categories:

1. The validity of the slowing-down term (denoted the "interaction term" by these authors) is doubtful in the presence of "queues" (or "platoons") of vehicles. This stems from the fact that the correlation inherent in platoons invalidates the assumption of vehicular chaos (Paveri-Fontana, 1975), which assumption underpins the particular form of the slowing-down term in the Prigogine-Herman kinetic equation.
2. The absence of a derivation of the relaxation term from first principles raises general questions regarding its validity. The validity of the specific expression (in terms of *c*) used by Prigogine et al. for the relaxation time *T* has therefore "not been proven." Further, it is therefore also difficult to "conceive the meaning of the relaxation time" and therefore "define a method for its experimental determination."

In addition to noting the first of these concerns, Paveri-Fontana (1975) argues forcefully that it is fundamentally incorrect to treat the desired speed as a parameter, as in done in the Prigogine-Herman kinetic equation. Rather, he suggests the desired speed must be taken as an additional independent variable, on the same footing as the actual speed, and he provides a modification of the Prigogine-Herman equation that accomplishes precisely that.

Prigogine and Herman (1971, Section 3.4, and 1970) dispute the claim of Munjal and Pahl (1969) to the effect that "the validity of the interaction term (i.e., the Prigogine-Herman slowing-down term) is limited to traffic situations where no vehicles are queuing" (parenthetical clarification added). Current opinion seems inclined to agree with Paveri-Fontana (1975) that on balance Munjal and Pahl have the better of this particular discussion. However, traffic on arterial roads, for which signalized intersections necessarily enforce the formation of platoons, is the only situation that seems thus to be definitely excluded from the domain of the Prigogine-Herman kinetic equation. In particular, it is not *a priori* clear that the same objection is valid for the stop-and-go traffic that seems to characterize congested traffic on freeways. Nelson (1995a) has noted that a *correlation model* is generally needed to obtain a kinetic equation, and vehicular chaos is simply one instance of a correlation model. Other correlation models, which would lead to a kinetic equation other than that of Prigogine and Herman, conceivably could better treat platoons, at least under restricted circumstances. Approaches (e.g., Prigogine and Andrews, 1960; Beylich, 1979), in which multiple-vehicle density functions appear as the unknowns to be determined, also offer the potential ability to treat queues within the spirit of the kinetic theory.

Nelson (1995a) introduced the concepts of a *mechanical model* and a *correlation model* as the fundamental ingredients of any kinetic equation. This work was motivated precisely by the desire to obtain forms of the speeding-up term that are based upon at least the same level of first principles as the classical derivations of the Prigogine-Herman slowing-down term. Klar and Wegener (1999) used this approach to obtain a kinetic equation for traffic flow that accounts for the spatial extent of vehicles. The treatment of vehicles as "points" of zero length is an idealization underlying the Prigogine-Herman kinetic equation that seems not to have been extensively discussed in the earlier literature on traffic flow. Klar and Wegener (1999) show that including the length of vehicles has a significant quantitative effect upon the value of some coefficients in associated continuum models. The observational measurements of the relaxation time by Edie, Herman and Lam (1980) also bear mentioning.

The arguments of Paveri-Fontana (1975) that the desired speed must appear as an independent variable in any kinetic equation, so that the density function depends upon the

[2] Other concerns relate to: *i)* The necessity to include time dependence in the desired speed distribution, owing to the normalization $\int_0^{v_{\max}} f_0(x,v,t)\,dv = c(x,t)$; and *ii)* the interpretation and functional dependence of the passing probability, *P*.

desired speed, as well as position, actual speed and time, seem to be quite convincing. In order to avoid this complexity, some workers (e.g., Nelson, 1995a) choose to focus upon models in which all drivers have the same desired speed. Paveri-Fontana (1975) represents his modification of the Prigogine-Herman equation, to include desired speed as an independent variable, as valid only for dilute traffic. However, as suggested above, it is not completely clear that this restriction is required, unless the dense traffic also includes a significant fraction of the vehicles in platoons.

11.2.3 Accomplishments of the Prigogine-Herman Model

In view of the deficiencies chronicled in the preceding subsection, why would anyone deem the Prigogine-Herman kinetic equation to be of any interest? That question is answered in this subsection, by describing two significant results that stem from the Prigogine-Herman model.

First, Prigogine and Herman (1971, Chap. 4) demonstrated, under the somewhat restrictive assumption that there exist drivers desiring arbitrarily small speeds, that one can obtain traffic stream models (fundamental diagrams, speed/density relations), say $q=Q(c)$, from the equilibrium solutions (i.e., the solutions that are independent of space and time) of their kinetic equation. (Here q is vehicular flow, and c is, as above, vehicular density.) The procedure is precisely analogous to that giving rise to the Maxwellian distribution and the ideal gas law, when applied to the Boltzmann equation of the kinetic theory of gases. Further, at high concentrations the equilibrium solution is *bimodal*; that is, it displays two (local) maxima in speed, in qualitative agreement with the observations of Phillips (1977, 1979). (See the following section for more details of these works.) One of these modes corresponds to a modification of the distribution of desired speeds, and the other (under the assumptions of Prigogine and Herman) to platoon flow in the rather extreme case of stopped traffic (i.e., zero speed). This "multiphase" aspect of congested traffic flow has subsequently been rediscovered by a number of workers. Note that this approach gives rise to a traffic stream model from an underlying microscopic model, via the equilibrium solution of a corresponding kinetic equation. Such a theoretical development contrasts with statements sometimes encountered to the effect that traffic stream models *must* be based upon observational data.

More recently, Nelson and Sopasakis (1998) showed that if one relaxes the assumption of Prigogine and Herman that there exist drivers having arbitrarily small desired speeds, then at sufficiently high densities the equilibrium solution is a two-parameter family. This contrasts with the one-parameter (typically taken as density) family that occurs at low densities, even at all densities under the restrictive assumption of Prigogine and Herman (1971).[3]

The consequence of the equilibrium solutions of Nelson and Sopasakis (1998) for the attendant traffic stream model will now be briefly described. Let

$$F(\zeta;c) := \int_{w_-}^{w_+} \frac{f_0(v)}{c(v-\zeta)}\, dv,$$

where w_- and w_+ are respectively the lower and upper bound on the desired speeds. Then there exists a positive *critical density,* denoted c_{crit}, and defined as the unique root (in c) of the equation $F(0;c) = cT(1-P)$, such that the dependence of mean speed upon density is as follows. Let $\zeta^*=\zeta^*(c)$ be the unique root (in ζ) of $F(\zeta;c) = cT(1-P)$. If $0 \le c \le c_{crit}$, then $\zeta^* \le 0$, and the mean speed is given by

$$\bar{v} = \bar{v}(c) = \frac{1}{cT(1-P)} + \zeta^*.$$

However, if $c > c_{crit}$, then $\zeta^* > 0$, and the mean speed is given by

$$\bar{v} = \bar{v}(c,\zeta) = \frac{1}{cT(1-P)} + \zeta,$$

where now ζ can take on any value such that $0 \le \zeta \le \min\{\zeta^*, w_-\}$. The parameter ζ is the speed of the embedded collective flow, and the preceding equation for the mean speed shows that the overall mean speed increases with increasing speed of the embedded collective flow. Figure 11.1 shows a three-dimensional graphical representation of the resulting "traffic stream model," for a particular hypothetical desired

[3] In some of their work, Prigogine and Herman (1971, Section 4.4, esp. Fig. 4.8 and the related discussion) did permit positive lower bounds for the set of desired speeds, but for reasons that seem unclear at this point their attendant analysis did not identify the full two-parameter range of equilibrium solutions at higher densities.

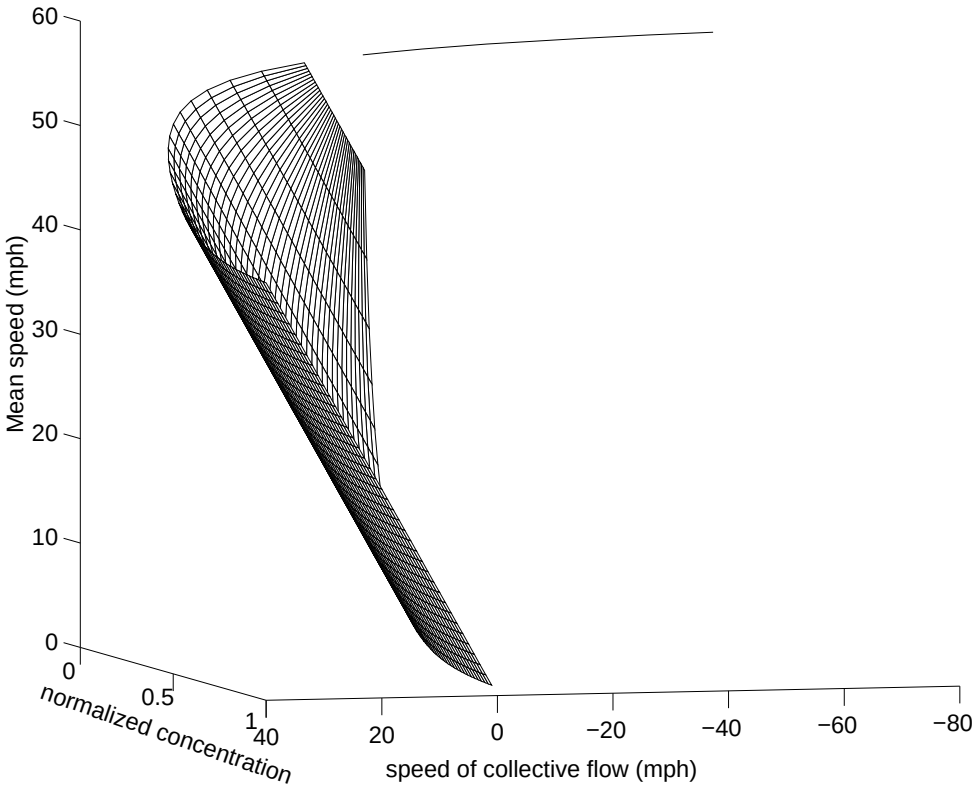


Figure 11-1 Dependence of the mean speed upon density normalized to jam density, $\eta=c/c_P$, for jam density $c_P = 200$ vpm, $P=1-\eta$, $T=\tau\eta/(1-\eta)$, with $\tau=0.003$ hours, and a uniform desired speed distribution from 40 to 80 mph.

speed distribution. See Nelson and Sopasakis (1998) for more details.

The significance of this three-dimensional presentation of a traffic stream model lies in the fact that it is consistent with the well-known tendency (e.g., Drake, Schofer and May, 1967) for traffic flow data to be widely scattered at high densities. The effort to explain this tendency has spawned a number of theories (e.g., Ceder, 1976; Hall, 1987; Disbro and Frame, 1989). The explanation in terms of an embedded collective flow seems possibly preferable to these, in that it derives from the kinetic theory, which is a well-known branch of traffic flow theory, as opposed to requiring some novel *ad hoc* theory.

Thus, the Prigogine-Herman kinetic equation has equilibrium solutions that both reproduce the observed bimodal distribution of speeds at high densities, and provide traffic stream models that reproduce qualitatively the well-known result that at sufficiently high densities mean speeds and flows do not depend exclusively upon vehicular density. One certainly can envision more ambitious objectives for a kinetic theory of vehicular traffic than these two. Some possible such objectives are discussed further in Sections 4 and 5 below. However, given that the seminal Prigogine-Herman kinetic equation of vehicular traffic does accomplish at least these objectives, it seems appropriate to suggest them as minimal benchmarks that should be met by any alternative kinetic equations that might be proposed. In the following section some of the alternative kinetic equations that have been proposed, as described in the preceding subsection, are assessed against these benchmarks.

11.3 Other Kinetic Models

Both benchmarks suggested in the preceding subsection have to do with the equilibrium solutions of the kinetic equation of interest. The equilibrium solutions of the Paveri-Fontana (1975) generalization of the Prigogine-Herman kinetic equation, as described in Subsection 2.2, do not seem to have been definitively ascertained. Indeed, Helbing (1996), who has

extensively applied the Paveri-Fontana kinetic model in his recent works on the kinetic theory of vehicular traffic, states, in regard to these equilibrium solutions, that "unfortunately it seems impossible to find an analytical expression …." He then indicates that "empirical data and microsimulations" suggest these equilibrium solutions are "approximately a Gaussian." Note that Gaussians are *not* bimodal. Thus, the Paveri-Fontana model does not seem to have been shown to satisfy either of the benchmarks suggested in the preceding subsection.

Phillips (1977, 1979) develops yet another kinetic equation that is an alternative to the original Prigogine-Herman kinetic model. However, this development seems predicated on a form of the corresponding equilibrium solution that ignores the considerations that led Prigogine and Herman to the "lower mode" of their bimodal equilibrium solution; cf. Eq. (4) of Phillips, 1979. Phillips compared (sketchily in Phillips, 1979, but exhaustively in Phillips, 1977) the equilibrium solution of his kinetic model against measured speed distributions. With one possibly important exception, the agreement seems reasonable. One therefore expects good agreement between the traffic stream model obtained theoretically from the equilibrium solution and that obtained observationally, although Phillips does not explicitly effect such comparisons. The exception is that a large amount of the data indicates a bimodal equilibrium solution; cf. Figs. 3 and 4 of Phillips, 1979, and numerous figures in Phillips, 1977. Thus, although the bimodal nature of an equilibrium solution is missed by the theoretical analysis, it is supported by the associated observations. In summary, it seems likely that the kinetic equation of Phillips (1979, 1977) meets the first benchmark suggested in the preceding section, and possible that a mathematical reassessment of its equilibrium solutions will reveal that it meets the second of these benchmarks. However, neither of these conclusions has yet been conclusively established.

Nelson (1995a) obtained a specific kinetic equation for purposes of providing a concrete illustration of his proposed general methodology for obtaining speeding-up (and slowing-down) terms based on first principles (i.e., appropriate mechanical and correlation models). In subsequent work (Nelson, Bui and Sopasakis, 1997) it was shown that this kinetic equation provides a theoretical traffic stream model that agrees well with classical traffic stream models, except near jam density. It has further been shown (Bui, Nelson and Sopasakis, 1996) that a simple modification of the underlying correlation model removes the incorrect behavior near jam density. Thus, this kinetic equation has been shown to meet the first of the benchmarks suggested in the preceding section. However, the equilibrium solutions of the kinetic equation of Nelson (1995a) are such that it clearly does *not* meet the second benchmark (i.e., does not predict scattered flow data under congestion). It is possible that the underlying mechanical model could be modified to attain this objective, but that has not been demonstrated.

Klar and Wegener (1999) use numerical techniques to obtain equilibrium solutions of their kinetic equations. They do not explicitly present corresponding traffic stream models. Their numerical equilibrium solutions do not display two modes. It might be difficult to obtain the lower mode, which typically appears as a delta function, by a strictly numerical approach.

Table I summarizes the status of the various kinetic models mentioned here, as regards their ability to meet the two benchmarks delineated in Subsection 2.3.

Table 11-I Status of various kinetic models with respect to the benchmarks of Subsection 11.2.3

Benchmark → / Kinetic Model ↓	Bimodal equilibrium solutions?	Equilibrium with scattered flows at high densities?
Prigogine-Herman (1971)	yes	yes
Paveri-Fontana (1975)	?	?
Phillips (1977, 1979)	no	no
Nelson (1995a)	yes	no
Klar-Wegener (1999)	no?	no

11.4 Continuum Models from Kinetic Equations

Continuum models historically have played an important role in traffic flow theory. They have been obtained either by simply writing them as analogs of some corresponding fluid dynamical system (e.g., Kerner and Konhäuser, 1993), or by rational developments from some presumably more basic microscopic (e.g., car-following) model of traffic flow. In the latter case the continuum equations can be developed either directly from the underlying microscopic model that serves as the starting point, or a kinetic model can play an intermediary role between the microscopic and continuum models. For early examples of the former approach, through the steady-state solutions of car-following models, see numerous references cited in Nelson, 1995b. Nagel (1998) presents a more modern approach, through appropriate formal ("fluid-dynamical") limits of particle-hopping models.

Here the primary interest is, of course, in approaches to continuum models that use a kinetic intermediary to the underlying microscopic model. Such approaches often (e.g., Helbing, 1995) follow the route of first taking the first few (one or two) low-order polynomial moments of the kinetic equation, then achieving closure via *ad hoc* approximations. An alternate approach, via certain formal asymptotic expansions (e.g., Hilbert or Chapman-Enskog expansions) is often used in the kinetic theory of gases (e.g., Grad, 1958). In this approach, the number of polynomial moments of the kinetic equation that are taken tend to be determined by the number of invariants that are defined by the dynamics of the microscopic model of the interaction between the constituent "particles" (vehicles, for traffic flow) of the system. This approach leads to a hierarchy of continuum models (e.g., the Euler/Navier-Stokes/Burnett/super-Burnett equations of fluid dynamics), as opposed to the single continuum equation that tends to result from formal limits of microscopic models. At all levels of this hierarchy the parameters of the resulting continuum model are expressed in terms of those of the underlying microscopic model.

Nelson and Sopasakis (1999) applied the Chapman-Enskog expansion to the Prigogine-Herman (1971) kinetic equation. In the region below the critical density described in Subsection 11.2 the lowest (zero) order expansion was found to be a Lighthill-Whitham (1955) continuum model, with associated traffic stream model corresponding to the one-parameter family of equilibrium solutions. The next highest (first-order) solution was found to be a diffusively corrected Lighthill-Whitham model,

$$\frac{\partial c}{\partial t} + \frac{\partial}{\partial x}\left[Q(c)\right] = \frac{\partial}{\partial x}\left[D(c)\frac{\partial c}{\partial x}\right], \qquad (11.2)$$

where now both the flow function $Q(c)$ and the diffusion coefficient $D(c)$ are known in terms of the density c and the parameters of the Prigogine-Herman kinetic model. This result is perhaps somewhat surprising, as one might reasonably have expected rather a continuum higher-order model of the type suggested by Payne (1971). Figure 11.2 illustrates how an initial discontinuity between an upstream higher-density region and a downstream lower-density region will tend to dissipate according to the diffusively corrected Lighthill-Whitham model, as opposed to the shock wave predicted by ordinary Lighthill-Whitham theory, which is given by Eq. (11.2) with $D \equiv 0$. See Nelson (2000) for more details of the example underlying this figure. Sopasakis (2000) has also developed the zero-order (again Lighthill-Whitham) and first-order Hilbert expansions, and the second-order Chapman-Enskog expansion, for the Prigogine-Herman kinetic model.

Interest in continuum models of traffic flow seems likely to continue, as applications exist within the space of resolution/fidelity/scale requirements for which continuum models are deemed most suitable. Along with this, interest in the use of kinetic models of vehicular traffic as a basis for continuum models seems likely to continue. For example, the venerable Lighthill-Whitham (1955) model is widely viewed as the most basic continuum model of traffic flow. But a suitable traffic stream model is an essential ingredient of the Lighthill-Whitham model. Thus, traffic stream models are an important part of continuum models, as well as being of interest in their own right. Therefore, both of the benchmarks demarcated in the preceding section can be viewed as related to the issue of how well a particular kinetic model performs in terms of providing a particularly low-order continuum model, specifically the Lighthill-Whitham model.

11.5 Direct Solution of Kinetic Equations

Along with the traditional application of kinetic models of vehicular traffic to rational development of continuum models from microscopic models, as described in the preceding section, modern computers permit consideration of the utility of kinetic models in their own right, rather than merely as tools that can be used to construct continuum models. In this respect, there are two substantive issues:

i) How can one solve kinetic equations, to obtain the distribution function (f)?
ii) Given this distribution function, what applications of it can usefully be made?

As regards the first issue, Hoogendoorn and Bovy (to appear) have employed Monte Carlo (i.e., simulation-based) techniques for the computational solution of a kinetic equation of vehicular traffic that builds upon the earlier work of Paveri-Fontana (1975). By contrast, in the kinetic theory of gases there exists a significant body of knowledge (e.g., Neunzert and Struckmeier, 1995, and other works cited therein) relative to the deterministic computational solution of kinetic equations. This knowledge base undoubtedly could be invaluable in attempting to develop similar capabilities for vehicular traffic, but the equations are sufficiently different from those arising in the kinetic theory of gases so that considerable further development is likely to be necessary. This

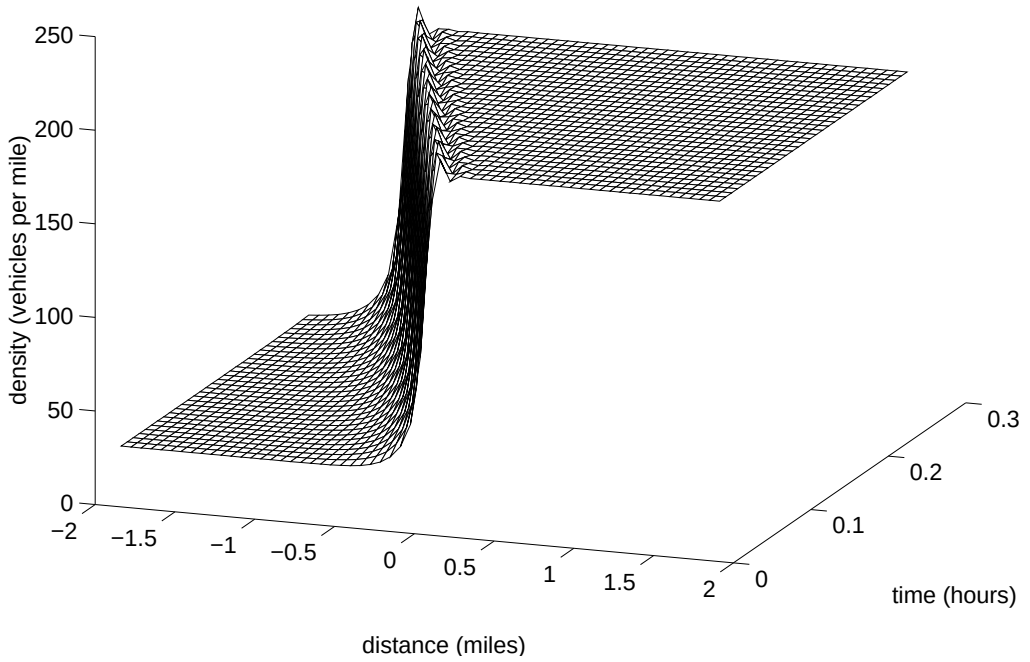


Figure 11-2 Evolution of the flow, according to a diffusively corrected Lighthill-Whitham model, from initial conditions consisting of 190 vpm downstream of *x=0* and 30 vpm upstream. See Nelson (2000) for details of the traffic stream model (flow function) and diffusion coefficient.

development is unlikely to occur in the absence of a relatively clear vision as to the uses that would be made of it. Therefore, the focus here primarily will be on the second of these issues.

Any consideration of applications of the distribution function of a kinetic theory requires a consideration of its interpretation. It is a statistical distribution function. The traditional interpretation of such a distribution function is that it describes the frequency with which certain properties occur among samples drawn from some sample space. In the kinetic theory of vehicular traffic, the samples are vehicles, and the properties of interest are the positions and speeds of the vehicles; however, there are two fundamentally different possible interpretations of the underlying sample space:

1. **Single-instance sampling:** The sample space consists of all vehicles present on a specified road network at a specific designated time.
2. **Ensemble sampling:** The sample space consists of all vehicles present, at a designated time, on one of an *ensemble* of identical road networks.

For example, the Houston freeway network at 5:00 p.m. on Wednesday, July 15, 1998 would be a reasonable sample space for single-instance sampling. On the other hand, the Houston freeway network at 5:00 p.m. on all midweek workdays during 1998 for which dry weather conditions prevailed would be a reasonable approximation of a sample space suitable for ensemble sampling.

The difference between these two interpretations is subtle, but it has profound consequences. Traffic theorists

normally tend to think in terms that are most consistent with single-instance sampling. But any attempt to apply the kinetic theory within that interpretation implies the intention to predict, at some level of approximation, the evolution of traffic for that specific instance, given suitable initial and boundary conditions for the distribution function. It seems somewhat questionable that this is attainable over any significant duration. (The ``rolling horizon'' approach often applied to prediction of traffic flow is a tacit admission of the significance of this issue.) On the other hand, the ensemble sampling interpretation implies only the intent to predict the likelihood with which various outcomes will occur. This intuitively seems much more achievable (cf. p. 10 of Asimov, 1988). Thus the more subtle ensemble sampling interpretation leads to an apparently more achievable objective than does the more obvious single-instance interpretation. For that reason, the ensemble sampling interpretation seems more likely to lead to potential direct applications of the kinetic theory.

The fundamental advantage of kinetic models over continuum models is that the kinetic distribution function provides an estimate of the variability (over various instances of the ensemble, under the ensemble-sampling interpretation) of densities and speed at specific times and locations, whereas continuum models provide estimates of only the mean (presumably over the ensemble) of these quantities. If the quantity (function of position and time) of interest in a particular application is not highly variable between instances within the ensemble, or if that variability is not of interest, then presumably one should choose a continuum model, or perhaps an even more highly aggregated model. On the other hand, if this variability is both of significant magnitude and important to the issue under study, then kinetic models might be a useful alternative to the computationally more expensive possibility of running a sufficiently large number of microscopic simulations so as to capture the nature of the variation between instances.

Some specific instances of quantities for which variability might be of significant interest are travel times and driving cycles. The latter require knowledge about the statistical distribution of accelerations, as well as velocities and densities, but such acceleration information is inherent in the distribution function, along with the mechanical and correlation models that underlie any kinetic model. In fact, kinetic models seem to be the natural connecting link between continuum models, which provide the "cross sectional" view of traffic that most transportation planning is based on, and the "longitudinal" view that underlies the standard driving cycle approach to estimation of fuel emissions (cf. Carson and Austin, 1997).

The crucial question underlying any potential application of kinetic models is whether a kinetic model can be found that has sufficient fidelity and resolution for the particular application, and that can be solved on the necessary scale using available computational resources. The answer to that clearly depends upon the specific details of the particular application, and any such proposed application of a specific kinetic model must be validated against actual observations. However, data of sufficiently high quality to permit such validations are both rare and expensive to obtain. Under these circumstances, it seems appropriate to use microscopic models (e.g., cellular automata) as a framework within which initially to vet proposed kinetic models.

Specifically, it seems worthwhile to employ microscopic models to study the following:

HYPOTHESIS: *The multiparameter family of equilibrium solutions of the Prigogine-Herman kinetic model found by Nelson and Sopasakis (1998), with its attendant traffic stream surface (rather than the classical curve), reflects the fact that actual traffic has a number of spatially homogeneous equilibrium states (with different average speeds) corresponding to the same density.*

If this hypothesis is true, then presumably different initial configurations of a traffic stream have the possibility to approach different states in the long-time limit, even though their densities are the same on the macroscopic scale. Results reported by Nagel (1996, esp. Sec. V) tend to confirm this hypothesis.

References

Asimov, I. (1988). *Prelude to Foundation*. Doubleday, New York.

Beylich, A. E. (1979). *Elements of a Kinetic Theory of Traffic Flow*, Proceedings of the Eleventh International Symposium on Rarefied Gas Dynamics, Commissariat a L'Energie Atomique, Paris, pp. 129-138.

Bui, D. D., P. Nelson and A. Sopasakis (1996). *The Generalized Bimodal Traffic Stream Model and Two-regime Flow Theory*. Transportation and Traffic Theory, Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Pergamon Press (Jean-Baptiste Lesort, Ed.), Oxford, pp. 679-696.

Carson, T. R. and T. C. Austin, *Development of Speed Correction Cycles*. Report prepared for the U. S. Environmental Protection Agency, Contract No. 68-C4-0056,

Work Assignment No. 2-01, Sierra Research, Inc., Sacramento, California, April 30..

Ceder, A. (1976). *A Deterministic Flow Model of the Two-regime Approach.* Transportation Research Record 567*,* TRB, NRC, Washington, D.C., pp. 16-30..

Disbro, J. E. and M. Frame (1989). *Traffic Flow Theory and Chaotic Behavior*. Transportation Research Record 1225*,* TRB, NRC, Washington, D.C., pp. 109-115.

Drake, J. S., J. L. Schofer and A. D. May, Jr. (1967). *A Statistical Analysis of Speed Density Hypothesis*. Highway Research Record 154*,* TRB. NRC, Washington, D.C., pp. 53-87.

Edie, L. C., R. Herman and T. N. Lam (1980*). Observed Multilane Speed Distribution and the Kinetic Theory of Vehicular Traffic*. Transportation Research Volume 9, pp. 225-235,

Grad, H. (1958). *Principles of the Kinetic Theory of Gases*. Handbuch der Physik, Vol. XII, Springer-Verlag, Berlin.

Hall, F. L. (1987). *An Interpretation of Speed-Flow Concentration Relationships Using Catastrophe Theory*. Transportation Research A*,* Volume 21, pp. 191-201.

Helbing, D. (1995). *High-fidelity Macroscopic Traffic Equations*. Physica A, Volume 219, pp. 391-407.

Helbing, D., "Gas-Kinetic Derivation of Navier-Stokes-like Traffic Equations," *Physical Review E*. 53, 2366-2381, 1996.

Hoogendoorn, S.P., and P.H.L. Bovy (to appear). *Non-Local Multiclass Gas-Kinetic Modeling of Multilane Traffic Flow.* Networks and Spatial Theory.

Kerner, B. S. and P. Konhäuser (1993). *Cluster Effect in Initially Homogeneous Traffic Flow*. Physical Review E, Volume 48, pp. R2335-R2338.

Klar, A. and R. Wegener (1999). *A Hierarchy of Models for Multilane Vehicular Traffic* (Part I: Modeling and Part II: Numerical and Stochastic Investigations). SIAM J. Appl. Math, Volume 59, pp. 983-1011.

Lighthill, M. J. and G. B. Whitham (1955). *On Kinematic Waves II: A Theory of Traffic Flow on Long Crowded Roads*. Proceedings of the Royal Society, Series A, Volume 229, pp. 317-345.

Nagel, K. (1996*). Particle Hopping Models and Traffic Flow Theory*. Physical Review E, Volume 53, pp. 4655-4672.

Nagel, K. (1998). *From Particle Hopping Models to Traffic Flow Theory*. Transportation Research Record, Volume 1644, pp. 1-9.

Munjal, P. and J. Pahl (1969). *An Analysis of the Boltzmann-type Statistical Models for Multi-lane Traffic Flow*. Transportation Research, Volume 3, pp. 151-163.

Nelson, P. (1995a). *A Kinetic Model of Vehicular Traffic and its Associated Bimodal Equilibrium Solutions.* Transport Theory and Statistical Physics, Volume 24, pp. 383-409.

Nelson, P. (1995b). *On Deterministic Developments of Traffic Stream Models*. Transportation Research B, Volume 29, pp. 297-302.

Nelson, P. (2000). *Synchronized Flow from Modified Lighthill-Whitham Model*. Physical Review E, Volume 61, pp. R6052-R6055.

Nelson, P., D. D. Bui and A. Sopasakis (1997). *A Novel Traffic Stream Model Deriving from a Bimodal Kinetic Equilibrium*. Preprints of the IFAC/IFIP/IFORS Symposium on Transportation Systems, Technical University of Crete (M. Papageorgiou and A. Pouliezos, Eds.), Volume 2, pp. 799-804.

Nelson, P. and A. Sopasakis (1998). *The Prigogine-Herman Kinetic Model Predicts Widely Scattered Traffic Flow Data at High Concentrations.* Transportation Research B, Volume 32, pp. 589-604.

Nelson, P. and A. Sopasakis (1999). *The Chapman-Enskog Expansion: A Novel Approach to Hierarchical Extension of Lighthill-Whitham Models.* Transportation and Traffic Theory: Proceedings of the 14th International Symposium on Transportation and Traffic Theory, Pergamon (A. Ceder, Ed.), pp. 51-80.

Neunzert, H. and J. Struckmeier (1995). *Particle Methods for the Boltzmann Equation*. Acta Numerica, Volume 4, pp. 417-457.

Paveri-Fontana, S. L. (1975). *On Boltzmann-like Treatments for Traffic Flow: A Critical Review of the Basic Model and an Alternative Proposal for Dilute Traffic Analysis*. Transportation Research, Volume 9, pp. 225-235.

Payne H. J. (1971). *Models of Freeway Traffic and Control.* Simulation Council Procs., Simulations Council, Inc., La Jolla, CA, Vol. 1, pp. 51-61.

Phillips, W. F. (1977). *Kinetic Model for Traffic Flow*. Report DOT/RSPD/DPB/50-77/17, U. S. Department of Transportation.

Phillips, W. F. (1979). *A Kinetic Model for Traffic Flow with Continuum Implications*. Transportation Planning and Technology*,* Volume 5, pp. 131-138.

Prigogine I. and F. C. Andrews (1960). *A Boltzmann-like Approach for Traffic Flow*. Operations Research, Volume 8, pp. 789-797.

Prigogine, I. and R. Hermann (1970). *Comment on "An Analysis of the Boltzmann-type Statistical Models for Multi-lane Traffic Flow.*" Transportation Research, Volume 4, pp. 113-116.

Prigogine, I. and R. Hermann (1971). *Kinetic Theory of Vehicular Traffic,* Elsevier, New York.

Sopasakis, A. (2000). *Developments in the Theory of the Prigogine-Herman Kinetic Equation of Vehicular Traffic.* Ph.D. dissertation (mathematics), Texas A&M University, May.

INDEX

Note: Index has not been updated to reflect revisions to Chapters 2, 5 and new 11.

A

B

C

D

E

F

G

O

P

Q

R

S

T

U

V

W

Y