Kluwer's
International Series

Operations Research • Management Science

ADVANCING THE
STATE-OF-THE-ART

HANDBOOK OF TRANSPORTATION SCIENCE

Second Edition

Volume Contributors:

Richard Arnott
Cynthia Barnhart
Martin Beckmann
Moshe Ben-Akiva
Chandra R. Bhat
Michel Bierlaire
Lawrence Bodin
Arnab Bose
Michael J. Cassidy
Amy M. Cohn
Teodor Gabriel Crainic
Mark S. Daskin
Leonard Evans
Michael Florian
Randolph W. Hall
Donald Hearn
Petros Ioannou
Ellis L. Johnson
Diego Klabjan
Frank S. Koppelman
Marvin Kraus
Vittorio Maniezzo
Aristide Mingozzi
George L. Nemhauser
Peter Nijkamp
Susan H. Owen
Markos Papageorgiou
Tönu Puu
Piet Rietveld
Kalyan Talluri
Pamela Vance
Garrett van Ryzin

edited by
Randolph W. Hall

HANDBOOK OF TRANSPORTATION SCIENCE

Second Edition

INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE

Frederick S. Hillier, Series Editor Stanford University

Weyant, J. / *ENERGY AND ENVIRONMENTAL POLICY MODELING*
Shanthikumar, J.G. & Sumita, U. / *APPLIED PROBABILITY AND STOCHASTIC PROCESSES*
Liu, B. & Esogbue, A.O. / *DECISION CRITERIA AND OPTIMAL INVENTORY PROCESSES*
Gal, T., Stewart, T.J., Hanne, T. / *MULTICRITERIA DECISION MAKING: Advances in MCDM Models, Algorithms, Theory, and Applications*
Fox, B.L. / *STRATEGIES FOR QUASI-MONTE CARLO*
Hall, R.W. / *HANDBOOK OF TRANSPORTATION SCIENCE*
Grassman, W.K. / *COMPUTATIONAL PROBABILITY*
Pomerol, J-C. & Barba-Romero, S. / *MULTICRITERION DECISION IN MANAGEMENT*
Axsäter, S. / *INVENTORY CONTROL*
Wolkowicz, H., Saigal, R., & Vandenberghe, L. / *HANDBOOK OF SEMI-DEFINITE PROGRAMMING: Theory, Algorithms, and Applications*
Hobbs, B.F. & Meier, P. / *ENERGY DECISIONS AND THE ENVIRONMENT: A Guide to the Use of Multicriteria Methods*
Dar-El, E. / *HUMAN LEARNING: From Learning Curves to Learning Organizations*
Armstrong, J.S. / *PRINCIPLES OF FORECASTING: A Handbook for Researchers and Practitioners*
Balsamo, S., Personé, V., & Onvural, R./ *ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING*
Bouyssou, D. et al. / *EVALUATION AND DECISION MODELS: A Critical Perspective*
Hanne, T. / *INTELLIGENT STRATEGIES FOR META MULTIPLE CRITERIA DECISION MAKING*
Saaty, T. & Vargas, L. / *MODELS, METHODS, CONCEPTS and APPLICATIONS OF THE ANALYTIC HIERARCHY PROCESS*
Chatterjee, K. & Samuelson, W. / *GAME THEORY AND BUSINESS APPLICATIONS*
Hobbs, B. et al. / *THE NEXT GENERATION OF ELECTRIC POWER UNIT COMMITMENT MODELS*
Vanderbei, R.J. / *LINEAR PROGRAMMING: Foundations and Extensions, 2nd Ed.*
Kimms, A. / *MATHEMATICAL PROGRAMMING AND FINANCIAL OBJECTIVES FOR SCHEDULING PROJECTS*
Baptiste, P., Le Pape, C. & Nuijten, W. / *CONSTRAINT-BASED SCHEDULING*
Feinberg, E. & Shwartz, A. / *HANDBOOK OF MARKOV DECISION PROCESSES: Methods and Applications*
Ramík, J. & Vlach, M. / *GENERALIZED CONCAVITY IN FUZZY OPTIMIZATION AND DECISION ANALYSIS*
Song, J. & Yao, D. / *SUPPLY CHAIN STRUCTURES: Coordination, Information and Optimization*
Kozan, E. & Ohuchi, A. / *OPERATIONS RESEARCH/ MANAGEMENT SCIENCE AT WORK*
Bouyssou et al. / *AIDING DECISIONS WITH MULTIPLE CRITERIA: Essays in Honor of Bernard Roy*
Cox, Louis Anthony, Jr. / *RISK ANALYSIS: Foundations, Models and Methods*
Dror, M., L'Ecuyer, P. & Szidarovszky, F. / *MODELING UNCERTAINTY: An Examination of Stochastic Theory, Methods, and Applications*
Dokuchaev, N. / *DYNAMIC PORTFOLIO STRATEGIES: Quantitative Methods and Empirical Rules for Incomplete Information*
Sarker, R., Mohammadian, M. & Yao, X. / *EVOLUTIONARY OPTIMIZATION*
Demeulemeester, R. & Herroelen, W. / *PROJECT SCHEDULING: A Research Handbook*
Gazis, D.C. / *TRAFFIC THEORY*
Zhu, J. / *QUANTITATIVE MODELS FOR PERFORMANCE EVALUATION AND BENCHMARKING*
Ehrgott, M. & Gandibleux, X. / *MULTIPLE CRITERIA OPTIMIZATION: State of the Art Annotated Bibliographical Surveys*
Bienstock, D. / *Potential Function Methods for Approx. Solving Linear Programming Problems*
Matsatsinis, N.F. & Siskos, Y. / *INTELLIGENT SUPPORT SYSTEMS FOR MARKETING DECISIONS*
Alpern, S. & Gal, S. / *THE THEORY OF SEARCH GAMES AND RENDEZVOUS*

HANDBOOK OF TRANSPORTATION SCIENCE

Second Edition

edited by

Randolph W. Hall
University of Southern California

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-48058-1
Print ISBN: 1-4020-7246-5

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

CONTENTS

Human Elements in Transportation

Flows and Congestion

Spatial Models

Routing and Network Models

Economic Models

PREFACE TO THE SECOND EDITION

The Second Edition of the *Handbook of Transportation Science* is a compendium of the fundamental concepts, methods and principles underlying transportation. It has been expanded from the first edition through the addition of four chapters. Chapter 15 extends the networks section of the book by addressing supply chains, distribution networks and logistics. While the emphasis is on freight transportation, the principles for network design extend to other applications, such as public transportation. Chapters 16 through 18 fall in a new section on transportation economics. Chapter 16 addresses revenue management, a relatively recent topic in transportation, that has had substantial impact on the airline industry in particular. Chapter 17 presents spatial interaction models, which provides a mechanism for analyzing patterns of development. Lastly, Chapter 18 provides the principles of transportation economics, with emphasis on pricing and public policy. In addition to the new chapters, several of the original chapters have been updated and revised. We hope that the Second Edition continues to inspire research into the science of transportation.

This page intentionally left blank

ACKNOWLEDGMENTS

David Boyce, Carlos Daganzo, Michael Florian, and Nigel Wilson provided valuable guidance in suggesting chapter topics and authors. Cenk Caliskan and Georgia Lum provided a tremendous help in formatting chapters. The work was supported, in part, by a grant from the National Science Foundation (DMI-9732878), and support from the United States Department of Transportation for the METRANS University Transportation Center.

This page intentionally left blank

1 TRANSPORTATION SCIENCE

Randolph W. Hall

> "The whole value of science consists in the power which it confers upon us of applying to one object the knowledge acquired from like objects"

Stanley Jevons wrote these words more than a century ago in *The Principles of Science* (Jevons, 1958, p. 1). Yet even today, *The Principles of Science* is a guidepost for defining what science is and how it is conducted.

Though "Transportation Science" did not exist as a discipline in the time of Jevons, his insights provide a motivation for *The Handbook of Transportation Science.* The premise for our book is that transportation can be defined as a scientific discipline that transcends transportation technology and methods. Whether by car, truck, airplane -- or by a mode of transportation that has not yet been conceived -- transportation obeys fundamental properties. The science of transportation defines these properties, and demonstrates how our knowledge of one mode of transportation can be used to explain the behavior of another.

Like any of the natural sciences, transportation science as a discipline arose out of human curiosity, and the desire for explanations for how the world around us behaves. In the words of famed physicist Max Planck,

> "The beginning of every act of knowing, and therefore the starting point of every science, must be in our own personal experience … They form the first and most real hook on which we fasten the thought-chain of science." (Planck, 1932, p. 66)

And so is the case for transportation science. When one looks back to the earliest publications on the subject from the 1950s and early 1960s, we first see a desire to understand the dynamics of roadway traffic. Then and now, there is hardly a person in the profession who does not view a trip on the highway as a scientific experiment, seeking to understand why traffic flows as it does, how bottlenecks appear and disappear, and what causes the myriad of driving behaviors. Many of the early pioneers were, in fact, trained in natural sciences, such as physics, and cleverly combined knowledge of natural phenomena, such as thermodynamics and fluid mechanics, with their observations on traffic flow.

Transportation scientists are motivated by the desire to explain spatial interactions that result in movement of people or objects from place to place. Its heritage includes research in the fields of geography, economics and location theory, dating over several centuries. Its methodologies draw from physics, operations research, probability and control theory. It is fundamentally a quantitative discipline, relying on mathematical models and optimization algorithms to explain the phenomena of transportation.

Publications in transportation science appear in many places, but they are most concentrated in the journals *Transportation Research B* and *Transportation Science,* and also in the proceedings of the International Symposium on Transportation and Traffic Flow Theory. Transportation scientists perform both empirical and theoretical work (many do both), and use real transportation systems as their laboratories. They interact frequently with practitioners, with scientific findings resulting from examination of real problems.

Fundamentally, transportation science recognizes that all modes of transportation have the same essential elements: vehicles, guideways, and terminals, operating under some control policy. Vehicles comprise mobile resources that accompany persons or shipments (P/S) as they travel from place to place. They provide the motive power to propel P/Ss on their trip, and provide the carrying space to ensure a safe and/or comfortable journey. Guideways are stationary resources that define feasible paths of travel and provide the physical infrastructure to support vehicles and P/Ss. They add to safety by restricting movements to defined paths, and provide an efficient surface for movement. Terminals are stationary resources that reside at discrete location. They offer the capability to sort vehicles, persons and objects among incoming and outgoing transportation routes. Lastly, control represents the rules, regulations and algorithms that determine movements and trajectories within transportation systems.

Many years ago, transportation occurred by human, animal and natural (e.g., wind, currents, gravity) power, in simple vehicles (or none at all), on guideways that required little in the way of construction. Terminals, if they could be called that, were the market towns, caravansaries or trading posts, and control was executed through the minds of individual travelers. By contrast, today most movement depends on propulsion by motors or engines, built guideways and terminals, and, to some degree, computer control. So in many respects, one might say that transportation modes of the late 20th century have little in common with their ancestors.

Nevertheless, similarities abound. For any given mode of transportation, vehicles, guideways, terminals and control are configured to perform several basic functions. All modes of transportation provide the capability to propel, brake and steer. Most (even animal and human) provide mechanisms to store energy for propulsion, to sort persons and objects at terminals, to couple shipments together into efficient loads, and to contain these shipments as they travel from place to place. The way that a mode of transportation accomplishes these functions may be unique, but the basic tasks are the same (Hall, 1995).

As mentioned, this book is concerned with the properties and characteristics that transcend individual modes of transportation, and collectively define a science of transportation. The chapters and structure of this book are intended to elucidate these properties on a subject-by-subject basis, and not by mode. We begin with the human element of transportation. On a day-to-day basis, individuals are presented with a plethora of transportation choices, some of which are determined by ingrained habits and circumstances; others of which result from deliberation. The route followed, the time of travel and, to some degree, the choice of destination and mode, are all daily decisions, and constitute short-term traveler behavior (Chapter 2). These decisions are imbedded within the broader context of how we plan and organize our activities, the subject of Chapter 3. And the way we operate our vehicles is the main determinant of Transportation Safety, as covered in Chapter 4.

Another important property of most transportation networks is that travel time depends on traffic flows, as well as system design and system control. Chapters 5 (queueing) and Chapter 6 (traffic flow theory) show how congestion originates on transportation networks, and how vehicles, travelers and shipments interact as they travel across the network. Chapter 5 addresses congestion and delay in a broad context, spanning all types of transportation, whereas Chapter 6 focuses on movement along links of a network where vehicles interact with each other.

At a more microscopic level, vehicle flows and trajectories depend on how their speed and direction are controlled. Until fairly recently, this was a human task, but increasingly vehicle control is automated, though electronic sensing and computer processors. Chapter 7 covers automated techniques for controlling trajectories, whereas Chapter 8 covers control at the more macroscopic scale of regulating flows. Macroscopic control is typically executed by conveying messages to vehicle operators (e.g., visible signals at intersections, network entrances and along lanes of travel.). In the future, it is not hard to imagine a coalescence of vehicle and traffic control within a single automated system.

The next two topics are at the historical core of transportation science: continuous-space models (Chapter 9) and transportation location (Chapter 10). The continuous-space approach has been used extensively as an explanatory tool for optimal network design, both with respect to physically constructed networks (roadways, railroads, etc.) and operational networks (vehicle routes). It draws from spatial economic theory and continuum models in physics. Transportation location also addresses system design, largely from the perspective of placing discrete facilities, such as terminals and points of production. It is the first of five chapters that include transportation optimization.

One of the ways that traveler behavior is revealed is in the flow of traffic along links in the transportation network. And one of the most studied, and most challenging, areas of research in transportation science is network assignment, or the estimation and prediction of these flows (Chapter 11). Network assignment uses optimization methods to predict the consequences of traveler behavior.

The next four chapters -- 12, 13, 14 and 15 -- describe different aspects of routing and networks, represented by the assignment of persons/shipments to vehicles and terminals, and the sequencing of stops along routes. The emphasis of Chapter 12 is local routing, represented by vehicle tours that can be accomplished within the span of a single day. The emphasis of Chapter 13 is routing freight over longhaul networks, represented by tours that travel from city to city, and last more than one day. Chapter 14 is concerned with routing the crews that operate vehicles on longhaul networks, with focus on the personnel constraints that dictate feasible tours. Chapter 15 addresses the design of transportation networks and supply chains, including the use of vehicles and terminals for shipment consolidation.

The final section of the book – Chapters 16, 17 and 18 – address transportation economics. Chapter 16 focuses on the recent topic of revenue management, or how transportation companies can use pricing to maximize their returns on investment. In Chapter 17, research is presented on spatial interaction, which provides a framework for predicting patterns of development, in light of transportation services and infrastructure. The final chapter covers transportation economics in general, with emphasis on pricing, markets, and public policy.

We wrote this book with the intention of documenting the core knowledge of transportation science. As would be the case for any other science, this book cannot provide ultimate conclusions. But it can record the methods and issues that define the discipline of transportation science as it exists at the end of the 20th century. In the words of the noted philosopher Karl Popper (1959, p. 281):

> "Science never pursues the illusory aim of making its answers final, or even probable. Its advance is, rather, towards the infinite yet attainable aim of ever discovering new, deeper and more general problems, and of subjecting its ever tentative answers to ever renewed and ever more rigorous tests."

We hope that the *Handbook of Transportation Science* provides this inspiration for the transportation scientists of the future.

References

Hall, R.W. (1995). The architecture of transportation systems, *Transportation Research C,* **3***,* 129-142.
Jevons, W.S. (1958). *The Principles of Science,* Dover Publishing, New York.
Planck, M. (1932). *Where is Science Going,* W.W. Norton and Company, New York.
Popper, K.R. (1959). *The Logic of Scientific Discovery,* Basic Books, New York.

This page intentionally left blank

2 DISCRETE CHOICE MODELS WITH APPLICATIONS TO DEPARTURE TIME AND ROUTE CHOICE

Moshe Ben-Akiva and Michel Bierlaire

2.1 Introduction

The analysis of travel behavior is typically disaggregate, meaning that the models represent the choice behavior of individual travelers. Discrete choice analysis is the methodology used to analyze and predict travel decisions. Therefore, we begin this chapter with a review of the theoretical and practical aspects of discrete choice models. After a brief discussion of general assumptions, we introduce the random utility model, which is the most common theoretical basis of discrete choice models. We then present the alternative discrete choice model forms such as Logit, Nested Logit, Generalized Extreme Value and Probit, as well as more recent developments such as Hybrid Logit and the Latent Class choice model. Finally, we elaborate on the applications of these models to two specific short-term travel decisions: route choice and departure time choice.

2.2 Discrete Choice Models

We provide here a brief overview of the general framework of discrete choice models. We refer the reader to Ben-Akiva and Lerman (1985) for detailed discussion.

General Modeling Assumptions

The framework for a discrete choice model can be presented by a set of general assumptions. We distinguish among assumptions regarding the:

1. decision-maker -- defining the decision-making entity and its characteristics;
2. alternatives -- determining the options available to the decision-maker;

3. attributes -- measuring the benefits and costs of an alternative to the decision-maker; and
4. decision rule -- describing the process used by the decision-maker to choose an alternative.

Decision-maker Discrete choice models are also referred to as disaggregate models, meaning that the decision-maker is assumed to be an individual. The "individual" decision-making entity depends on the particular application. For instance, we may consider that a group of persons (a household or an organization, for example) is the decision-maker. In doing so, we may ignore all internal interactions within the group, and consider only the decisions of the group as a whole. We refer to "decision-maker" and "individual" interchangeably throughout this chapter. To explain the heterogeneity of preferences among decision-makers, a disaggregate model must include their characteristics such as the socio-economic variables of age, gender, education and income.

Alternatives Analyzing individual decision making requires not only knowledge of what has been chosen, but also of what has not been chosen. Therefore, assumptions must be made about available options, or alternatives, that an individual considers during a choice process. The set of considered alternatives is called the choice set.

A discrete choice set contains a finite number of alternatives that can be explicitly listed. The choice of a travel mode is a typical example of a choice from a discrete choice set. The identification of the list of alternatives is a complex process usually referred to as *choice set generation*. The most widely used method for choice set generation uses deterministic criteria of alternative availability. For example, the possession of a driver's license determines the availability of the auto drive option.

The universal choice set contains all potential alternatives in the application's context. The choice set is the subset of the universal choice set considered by, or available to, a particular individual. Alternatives in the universal choice set that are not available to the individual are therefore excluded from the choice set.

In addition to availability, the decision-maker's awareness of the alternative could also affect the choice set. The behavioral aspects of awareness introduce uncertainty in modeling the choice set generation process and motivate the use of probabilistic choice set generation models that predict the probability of each feasible choice set within the universal set. A discrete choice model with a probabilistic choice set generation model is described later in this chapter as a special case of the latent class choice model.

Attributes Each alternative in the choice set is characterized by a set of attributes. Note that some attributes may be generic to all alternatives, and some may be alternative-specific.

An attribute is not necessarily a directly measurable quantity. It can be any function of available data. For example, instead of considering travel time as an attribute of a transportation mode, the logarithm of the travel time may be used, or the effect of out-of-pocket cost may be represented by the ratio between the out-of-pocket cost and the income of the individual. Alternative definitions of attributes as functions of available data must usually be tested to identify the most appropriate.

Decision Rule The decision rule is the process used by the decision-maker to evaluate the alternatives in the choice set and determine a choice. Most models used for travel behavior applications are based on *utility theory,* which assumes that the decision-maker's preference for an alternative is captured by a value, called utility, and the decision-maker selects the alternative in the choice set with the highest utility.

This concept, employed by consumer theory of micro-economics, presents strong limitations for practical applications. The underlying assumptions of this approach are often violated in decision-making experiments. The complexity of human behavior suggests that the decision rule should include a probabilistic dimension.

Some models assume that the decision rule is intrinsically probabilistic, and even complete knowledge of the problem would not overcome the uncertainty. Others consider the individuals' decision rules as deterministic, and motivate the uncertainty from the limited capability of the analyst to observe and capture all the dimensions of the choice process, due to its complexity.

Specific families of models can be derived depending on the assumptions about the source of uncertainty. Models with probabilistic decision rules, like the model proposed by Luce (1959), or the "elimination by aspects" approach proposed by Tversky (1972), assume a deterministic utility and a probabilistic decision process. Random utility models, used intensively in econometrics and in travel behavior analysis, are based on deterministic decision rules, where utilities are represented by random variables.

Random Utility Theory

Random utility models assume, as does the economic consumer theory, that the decision-maker has a perfect discrimination capability. However, the analyst is assumed to have incomplete information and, therefore, uncertainty must be taken into account. Manski (1977) identifies four different sources of uncertainty: unobserved alternative attributes; unobserved individual characteristics (also called "unobserved taste variations"); measurement errors; and proxy, or instrumental, variables.

The utility is modeled as a random variable in order to reflect this uncertainty. More specifically, the utility that individual n associates with alternative i in the choice set C_n is given by

$$U_{in} = V_{in} + \varepsilon_{in},$$

where V_{in} is the deterministic (or systematic) part of the utility, and ε_{in} is the random term, capturing the uncertainty. The alternative with the highest utility is chosen. Therefore, the probability that alternative i is chosen by decision-maker n from choice set C_n is

$$P(i|\, C_n) = P[\, U_{in} \geq U_{jn}\ \forall\, j \in C_n] = P[U_{in} = \max_{j \in C_n} U_{jn}].$$

In the following we introduce the assumptions necessary to make a random utility model operational.

Location and scale parameters Considering two arbitrary real numbers α and μ, where $\mu > 0$, we have that

$$P[\, U_{in} \geq U_{jn}\ \ \forall\, j \in C_n] =$$

$$P[\mu U_{in} + \alpha \geq \mu U_{jn} + \alpha\ \ \forall\, j \in C_n] =$$

$$P[\, U_{in} - U_{jn} \geq 0\ \forall\, j \in C_n].$$

The above illustrates the fact that only the signs of the *differences* between utilities are relevant here, and not utilities themselves. The concept of ordinal utility is relative and not absolute. In order to estimate and use a specific model arbitrary values have to be selected for α and μ. The selection of the scale parameter μ is usually based on a convenient normalization of one of the variances of the random terms. The location parameter α is usually set to zero. See also the discussion below of Alternative Specific Constants.

Alternative specific constants The means of the random terms can be assumed to be equal to any convenient value *c* (usually zero, or the Euler constant γ for Logit models). This is not a restrictive assumption. If we denote the mean of the error term of alternative *i* by $m_i = E[\varepsilon_{in}]$, we can define a new random variable $e_{in} = \varepsilon_{in} - m_i + c$ such that $E[e_{in}] = c$. We have

$$P[\, U_{in} \geq U_{jn}\ \forall\, j \in C_n] = P[\, V_{in} + m_i + e_{in} \geq V_{jn} + m_j + e_{jn}\ \forall\, j \in C_n],$$

a model in which the deterministic part of the utilities are $V_{in} + m_i$ and the random terms are e_{in} (with mean *c*). The terms m_i are then included as Alternative Specific Constants (ASC) that capture the means of the random terms. Therefore, we may assume without loss of generality that the error terms of random utility models have a constant mean *c* by including alternative specific constants in the deterministic part of the utility functions.

As only differences between utilities are relevant, only differences between ASCs are relevant as well. It is common practice to define the location parameter α as the negative of one of the ASCs. This is equivalent to constraining that ASC equal zero. From a modeling viewpoint, the choice of the particular alternative whose ASC is constrained is arbitrary. However, Bierlaire, Lotan and Toint (1997) have shown that the speed of convergence of the estimation process may be improved by imposing different constraints.

The deterministic term of the utility The deterministic term V_{in} of each alternative is a function of the attributes of the alternative itself and the characteristics of the decision-maker. That is

$$V_{in} = V(z_{in}, S_n)$$

where z_{in} is the vector of attributes as perceived by individual n for alternative i, and S_n is the vector of characteristics of individual n.

This formulation is simplified using any appropriate vector valued function h that defines a new vector of attributes from both z_{in} and S_n, that is

$$x_{in} = h(z_{in}, S_n).$$

Then we have

$$V_{in} = V(x_{in}).$$

The choice of h is very general, and several forms may be tested to identify the best representation in a specific application. It is usually assumed to be continuous and monotonic in z_{in}. For a linear in the parameters utility specification, h must be a fully determined function (meaning that is does not contain unknown parameters). A linear in the parameters function is denoted as follows

$$V_{in} = \sum_k \beta_k x_{ink} ,$$

or in vector form

$$V_{in} = X_{in}\beta.$$

The deterministic term of the utility is therefore fully specified by the vector of parameters β.

The random part of the utility Among the many potential models that can be derived for the random parts of the utility functions, we describe below the most popular. The models within the Logit family are based on a probability distribution function of the maximum of a series of random variables, introduced by Gumbel (1958). Probit and Probit-like models are based on the Normal distribution motivated by the Central Limit Theorem.

The main advantage of the Probit model is its ability to capture all correlations among alternatives. However, due to the high complexity of its formulation, relatively few applications have been developed. The Logit model has been much more popular, because of its tractability. However, Logit imposes restrictions on the covariance structure that may be unrealistic in some contexts. Other models in the "Logit family" are aimed at relaxing restrictions, while maintaining tractability.

We present first the Generalized Extreme Value Models, a class of random utility model that includes Logit and Nested Logit. Next we present the Probit model and other advanced models including the Generalized Factor Analytical Representation and the Hybrid Logit models (designed to bridge the gap between Logit and Probit models) and the Latent Class Choice model (designed to explicitly include discrete unobserved factors in the model).

The Generalized Extreme Value Models Family

The Generalized Extreme Value (GEV) model has been derived from the random utility model by McFadden (1978). This general model consists of a large family of models. The probability of choosing alternative i within C_n is

$$P(i \mid C_n) = \frac{e^{V_{in}} \dfrac{\partial G}{\partial e^{V_{in}}}\left(e^{V_{1n}},\dots,e^{V_{J_n}}\right)}{\mu G\left(e^{V_{1n}},\dots,e^{V_{J_n}}\right)}.$$

J_n is the number of alternatives in C_n and G is a non-negative differentiable function defined on $\mathrm{IR}_+^{J_n}$ with the following properties:

1. G is homogeneous of degree $\mu > 0$ [1],
2. $\lim\limits_{x_1 \to \infty} G(x_1,\dots,x_i,\dots,x_{J_n}) = \infty, \forall i = 1,\dots,J_n$
3. the k^{th} partial derivative with respect to k distinct x_i is non-negative if k is odd, and non-positive if k is even, that is, for any distinct $i_1,\dots i_k \in \{1,\dots J_n\}$ we have

$$(-1)^k \frac{\partial^k G}{\partial x_{i_1} \dots \partial x_{i_k}}(x) \leq 0 \ \forall x \in \mathrm{IR}_+^{J_n}.$$

As G is homogeneous, Euler's theorem can be invoked to write

$$P(i \mid C_n) = \frac{e^{V_{in}+\ln(G_i(e^{V_{1n}},\dots,e^{V_{J_n n}}))}}{\sum\limits_{j \in C_n} e^{V_{jn}+\ln(G_j(e^{V_{1n}},\dots,e^{V_{J_n n}}))}}$$

where

$$G_i\left(e^{V_{1n}},\dots,e^{V_{J_n n}}\right) = \frac{\partial G}{\partial e^{V_{in}}}\left(e^{V_{1n}},\dots,e^{V_{J_n n}}\right)$$

The **Multinomial Logit Model** is an instance of the GEV family, with

$$G(x) = \sum_{i=1}^{J_n} x_i^{\mu},$$

[1] McFadden's original formulation with μ=1 was generalized to μ>0 by Ben-Akiva and François (1983).

yielding to the following probability model :

$$P(i \mid C_n) = \frac{e^{V_i}}{\sum_{j \in C_n} e^{V_j}}.$$

An important property of the Multinomial Logit Model is Independence from Irrelevant Alternatives (IIA). This property can be stated as follows: *The ratio of the probabilities of any two alternatives is independent of the choice set.* That is, for any choice sets C_1 and C_2 such that $C_1 \subseteq C_n$ and $C_2 \subseteq C_n$, and for any alternatives i and j in both C_1 and C_2, we have

$$\frac{\mathrm{P}(i|C_1)}{\mathrm{P}(j|C_1)} = \frac{\mathrm{P}(i|C_2)}{\mathrm{P}(j|C_2)}.$$

An equivalent definition of the IIA property is: *The ratio of the choice probabilities of any two alternatives is unaffected by the systematic utilities of any other alternatives.*

The IIA property of Multinomial Logit Models is a limitation for some practical applications. This limitation is often illustrated by the red bus/blue bus paradox in the modal choice context. We use here instead the following path choice example.

Consider a commuter traveling from origin O to destination D. He/she is confronted with the path choice problem described in Figure 2-1, where the choice set is {1,2a,2b} and the only attribute considered for the choice is travel time. We assume furthermore that the travel time for any alternative is the same, that is $V(1) = V(2\mathrm{a}) = V(2\mathrm{b}) = T$, and that the travel time on the small sections a and b is δ.

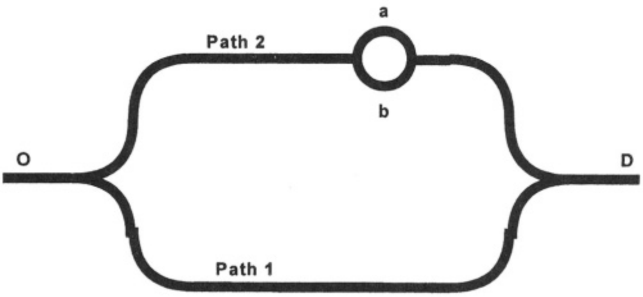


Figure 2-1. Path Choice Problem

The probability of each alternative provided by the Multinomial Logit Model for this example is

$$P(1|\{1,2a,2b\}) = P(2a|\{1,2a,2b\}) = P(2b|\{1,2a,2b\}) = \frac{e^{\mu T}}{\sum_{j\in\{1,2a,2b\}} e^{\mu T}} = \frac{1}{3}$$

Clearly, this result is independent of the value of δ. However, when δ is significantly smaller than the total travel time T, we expect the probabilities to be close to 50%/25%/25%. The Multinomial Logit Model is not consistent with this intuitive result. This situation appears in choice problems with significantly correlated random utilities, as it is clearly the case in the path choice example. Indeed, alternatives 2a and 2b are so similar that their utilities share many unobserved attributes of the path and, therefore, the assumption of independence of the random parts is not valid in this context.

The **Nested Logit Model**, first proposed by Ben-Akiva (1973 and 1974) and derived as a random utility model and a special case of GEV by McFadden (1978), is an extension of the Multinomial Logit Model designed to capture some correlations among alternatives. It is based on the partitioning of the choice set C_n into M nests C_{mn} such that

$$C_n = \bigcup_{m=1}^{M} C_{mn}$$

and $C_{mn} \cap C_{m'n} = \varnothing \ \forall m \neq m'$. It is also an instance of the GEV family, with

$$G(x) = \sum_{m=1}^{M}\left(\sum_{i\in C_{mn}} x_i^{\mu_m}\right)^{\frac{\mu}{\mu_m}}$$

where $\mu > 0$, $\mu_m > 0$ and $\mu \le \mu_m$. Each nest within the choice set is associated with a composite utility

$$V_{C_{mn}} = \tilde{V}_{C_{mn}} + \frac{1}{\mu_m} \ln \sum_{j\in C_{mn}} e^{\mu_m \tilde{V}_{jn}} ,$$

where $\tilde{V}$ denotes the partial utility common to all alternatives in the nest. The second term is called expected maximum utility, LOGSUM, inclusive value or accessibility in the literature. The probability for individual n to choose alternative i within nest C_{mn} is given by

$$P(i|C_n) = P(C_{mn}|C_n)P(i|C_{mn})$$

where

$$P(C_{mn}\,|\,C_n) = \frac{e^{\mu V_{C_{mn}}}}{\sum_{l=1}^{M} e^{\mu V_{C_{ln}}}} ,$$

and

$$P(i|C_{mn}) = \frac{e^{\mu_m \tilde{V}_{in}}}{\sum_{j \in C_{mn}} e^{\mu_m \tilde{V}_{jn}}}.$$

Parameters μ and μ_m reflect the correlation among alternatives within the nest C_{mn}. The correlation between the utility of two alternatives i and j in nest C_{mn} can be derived (see Ben-Akiva and Lerman, 1985) as

$$\text{Corr}(U_{in}, U_{jn}) = \begin{cases} 1 - \frac{\mu^2}{\mu_m^2} & \text{if } i \text{ and } j \in C_{mn} \\ 0 & \text{otherwise} \end{cases}.$$

Therefore,

$$\frac{\mu}{\mu_m} = 1 \Leftrightarrow \text{corr}(U_{in}, U_{jn}) = 0\cdot$$

The parameters μ and μ_m are closely related in the model. Actually, only their ratio is meaningful. It is not possible to identify them separately. A common practice is to arbitrarily constrain one of them to a specific value (usually 1). If $\mu=\mu_m$, the Nested Logit Model collapses to a Multinomial Logit Model.

This is illustrated by the following example (Bierlaire, 1998). We apply the Nested Logit Model to the route choice problem described in Figure 1. We partition the choice set C_n={1,2a,2b} into C_{1n}={1} and C_{2n}={2a,2b}. The probability of choosing path 1 is given by

$$P(1 \mid \{1,2a,2b\}) = \frac{1}{1+2^{\frac{\mu}{\mu_2}}},$$

where μ_2 is the scale parameter of the random term associated with C_{2n}, and μ is the scale parameter of the choice between C_{1n} and C_{2n}. Note that we require $0 \leq \mu/\mu_2 \leq 1$. The probability of the two other paths is

$$P(2a \mid C_n) = P(2b \mid C_n) = \frac{1}{2}\frac{2^{\frac{\mu}{\mu_2}}}{1+2^{\frac{\mu}{\mu_2}}}\cdot$$

In this example, we need to normalize either μ or μ_2 to 1. In the latter case we have

$$P(1|\{1,2a,2b\}) = \frac{1}{1+2^{\mu}}$$

and

$$P(2a|C_n) = P(2b|C_n) = \frac{1}{2}\left(\frac{2^{\mu}}{1+2^{\mu}}\right)$$

and we require that $0 \leq \mu \leq 1$. Note that for $\mu=1$ we obtain the MNL result. For μ approaching zero, we obtain the expected result when paths 2a and 2b fully overlap. A model where the scale parameter μ is normalized to 1 is said to be "normalized from the top." A model where one of the parameters μ_m is normalized to 1 is said to be "normalized from the bottom." The latter may produce a simpler formulation of the model. We illustrate it using the following example.

In the context of a mode choice with C_n={bus, metro, car, bike}, we consider a model with two nests: C_{1n}={bus,metro} contains the public transportation modes and C_{2n}={car,bike} contains the private transportation modes. For the example's sake, we consider the following deterministic terms of the utility functions:

$$V_{\text{bus}}=\beta_1\, t_{\text{bus}};\; V_{\text{metro}}=\beta_1\, t_{\text{metro}};\; V_{\text{car}}=\beta_2\, t_{\text{car}};\; V_{\text{bike}}=\beta_2\, t_{\text{bike}}$$

where t_i is the travel time using mode i and β_1 and β_2 are parameters to be estimated. Note that we have one parameter for private and one for public transportation, and we have not included the alternative specific constants in order to keep the example simple.

Applying the Nested Logit Model, we obtain

$$P(\text{bus}) = \left(\frac{e^{\mu_1\beta_1 t_{\text{bus}}}}{e^{\mu_1\beta_1 t_{\text{bus}}}+e^{\mu_1\beta_1 t_{\text{metro}}}}\right)\frac{e^{\frac{\mu}{\mu_1}\ln\left(e^{\mu_1\beta_1 t_{\text{bus}}}+e^{\mu_1\beta_1 t_{\text{metro}}}\right)}}{e^{\frac{\mu}{\mu_1}\ln\left(e^{\mu_1\beta_1 t_{\text{bus}}}+e^{\mu_1\beta_1 t_{\text{metro}}}\right)}+e^{\frac{\mu}{\mu_2}\ln\left(e^{\mu_2\beta_2 t_{\text{car}}}+e^{\mu_2\beta_2 t_{\text{bike}}}\right)}}$$

The normalization from the bottom is obtained by defining $\theta_1= \mu/\mu_1$, $\theta_2= \mu/\mu_2$, $\beta_1^*=\mu_1\beta_1$ and $\beta_2^*=\mu_2\beta_2$. Consequently,

$$P(\text{bus}) = \frac{e^{\beta_1^* t_{\text{bus}}}}{e^{\beta_1^* t_{\text{bus}}}+e^{\beta_1^* t_{\text{metro}}}}\frac{e^{\theta_1\ln\left(e^{\beta_1^* t_{\text{bus}}}+e^{\beta_1^* t_{\text{metro}}}\right)}}{e^{\theta_1\ln\left(e^{\beta_1^* t_{\text{bus}}}+e^{\beta_1^* t_{\text{metro}}}\right)}+e^{\theta_2\ln\left(e^{\beta_2^* t_{\text{car}}}+e^{\beta_2^* t_{\text{bike}}}\right)}},$$

with $0\leq\theta_1,\theta_2\leq 1$.

This formulation simplifies the computation of the derivatives, needed by parameter estimation procedure (see Daly, 1987). For this reason, it has been adopted by the Ben-Akiva and Lerman (1985) textbook and in estimation packages like ALOGIT (Daly, 1987) and HieLoW (Bierlaire, 1995, Bierlaire and Vandevyvere, 1995). We emphasize here that these packages should be used with caution when the same parameters are present in more than one nest. Specific techniques inspired from artificial trees proposed by Bradley and Daly (1991) must be used to obtain a correct specification of the model. In the above example, if $\mu_1=\mu_2$, then imposing the restriction $\beta_1=\beta_2$ is straightforward. However, for the case of $\mu_1\neq\mu_2$ and $\beta_1=\beta_2=\beta$, we define $\beta^*=\mu_1\mu_2\beta$ and create artificial nodes below each alternative, with a scale μ_2 for the first nest and scale μ_1 for the second. We refer the reader to Koppelman and Wen

(1998) and Hensher and Greene (2002) for further discussion. Note that the new package BIOGEME (Bierlaire, 2001b) for GEV model estimation does not impose a specific normalization for the Nested Logit model and therefore does not require such techniques.

A direct extension of the Nested Logit Model consists in partitioning some or all nests into sub-nests, which can in turn, be divided into sub-nests. The model described above is valid at every layer of the nesting, and the whole model is generated recursively. Therefore, a tree structure is a convenient representation of Nested Logit models. Clearly, the number of potential structures reflecting the correlation among alternatives can be very large. No technique has been proposed thus far to identify the most appropriate correlation structure directly from the data.

The Nested Logit Model is designed to capture choice problems where alternatives within each nest are correlated. No correlation across nests can be captured by the Nested Logit Model. When alternatives cannot be partitioned into well separated nests to reflect their correlation, the Nested Logit Model is not appropriate.

The **Cross-Nested Logit Model** is a direct extension of the Nested Logit Model, where each alternative may belong to more than one nest. It is also an instance of the GEV family, with

$$G(x) = \sum_{m=1}^{M} \left(\sum_{j \in C_n} \alpha_{jm} x_j^{\mu_m} \right)^{\frac{\mu}{\mu_m}},$$

where $\mu > 0$, $\mu_m > 0$, $\mu \leq \mu_m$, $\alpha_{jm} \geq 0$ and $\sum_m \alpha_{jm} > 0$.

This model was first presented by McFadden (1978) as a special case of the GEV model. It was applied by Small (1987) for departure time choice, by Vovsha (1997) for mode choice, and by Vovsha and Bekhor (1998) for route choice. Swait (2001) proposes a Cross-Nested formulation for a model including choice set generation. The general formulation proposed above has been introduced by Ben-Akiva and Bierlaire (1999). The proof that it is indeed a GEV model is detailed by Bierlaire (200la). Wen and Koppelman (2001) provide an analysis of the model elasticities. They use the name "Generalized Nested Logit" model for Cross-Nested. Papola (2000) describes a technique to design a specific Cross-Nested logit model for any given homoscedastic variance-covariance structure.

The parameter α_{jm} is usually interpreted as the degree at which alternative *j* belongs to nest *m*. Therefore, a common normalization of the model imposes that $\sum_m \alpha_{jm} = 1$. We emphasize that this condition is a convenient normalization condition, but is not necessary for the model to comply with random utility theory. The Recursive Nested Extreme Value Model (RNEV), proposed by Daly (2001), generalizes the Cross-Nested model by allowing several levels of nests in the formulation.

The **Network GEV model** is a class of models within the GEV family proposed by Bierlaire (2002) and based on the same idea as Daly's RNEV. Each instance is defined by a network where each edge (m,k) is associated with a non-negative parameter $\alpha_{(m,k)}$. The network must have the following properties.

1. It does not contain any circuit.
2. It has one special node with no predecessor, called the root.
3. It has J special nodes with no successor, called the alternatives.
4. For each alternative i, there exists a path between the root and i such that all α parameters on the path are non-zero.

Each node m of the network is associated with an homogeneous function $\mathbf{G}_m$, with homogeneity parameter μ_m, such that

$$G_m(x) = \sum_{k \in \mathrm{succ}(m)} \alpha_{(m,k)} G_k(x)^{\frac{\mu_m}{\mu_k}} .$$

If each alternative i is associated with the trivial function

$$G_i(x) = x_i^{\mu_i}$$

then the G function associated with each node of the network generates a GEV model. In general, only the GEV model associated with the root is considered. This result, formally proven by Bierlaire (2001c), provides an intuitive and general way of generating new GEV models. Namely, all GEV models mentioned above fit in that framework.

Multinomial Probit Model

The Probability Unit (or Probit) model is derived from the assumption that the error terms of the utility functions are normally distributed. The Probit model captures explicitly the correlation among all alternatives. Therefore, we adopt a vector notation for the utility functions:

$$\boldsymbol{U}_n = \boldsymbol{V}_n + \boldsymbol{\varepsilon}_{n,}$$

where $\boldsymbol{U}_n$, $\boldsymbol{V}_n$ and $\boldsymbol{\varepsilon}_n$ are $(J_n \times 1)$ vectors. The vector of error terms $\varepsilon_n=[\varepsilon_{1n},\varepsilon_{2n},\dots,\varepsilon_{Jn}]^T$ is multivariate normal distributed with a vector of means $\mathbf{0}$ and a J_nxJ_n variance-covariance matrix Σ_n.

The probability that a given individual *n* chooses alternative *i* from the choice set C_n is given by

$$P(i|C_n) = P(U_{jn} - U_{in} \leq 0 \ \ \forall j \in C_n).$$

Denoting Δ_i the $(J_n\text{-}1 \times J_n)$ matrix such that

$$\Delta_i \boldsymbol{U}_n = [U_{1n}\text{-}U_{in},\dots,U_{(i\text{-}1)n}\text{-}U_{in},U_{(i+1)n}\text{-}U_{in},\dots,U_{J_n n}\text{ -}U_{in}]^T,$$

The matrix Δ_i is such that the i^{th} column contains -1 everywhere. If the i^{th} column is removed, the remaining $(J_n-1\times J_n-1)$ matrix is the identity matrix. For example, in the case of a trinomial choice model, we have

$$\Delta_2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

Given this transformation, we have that

$$\Delta_i U_n \sim \mathrm{N}(\Delta_i V_n, \Delta_i \Sigma_n \Delta_i^T).$$

The density function is given by

$$f_i(x) = \lambda \exp\left[-\frac{1}{2}(x - \Delta_i V_n)^T (\Delta_i \Sigma_n \Delta_i^T)^{-1} (x - \Delta_i V_n)\right]$$

where

$$\lambda = (2\pi)^{-\frac{J_n - 1}{2}} \mid \Delta_i \Sigma_n \Delta_i^T \mid^{-1/2}$$

and

$$P(i \mid C_n) = P(\Delta_i U_n \leq 0) = \int_{-\infty}^{0} \ldots \int_{-\infty}^{0} f_i(x) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_{J_n} \cdot$$

We note that the multifold integral becomes intractable even for a relatively low number of alternatives. Moreover, the number of unknown parameters in the variance-covariance matrix grows with the square of the number of alternatives. We refer the reader to McFadden (1989) for a detailed discussion of multinomial Probit models. The complexity of Probit models can be reduced using a Factor Analytic form of the model, as described in the next section.

Generalized Factor Analytic Specification of the Random Utility

The general formulation of the factor analytic formulation is

$$U_n = V_n + \varepsilon_n = V_n + \mathrm{F}_n \mathrm{T} \zeta_n,$$

where U_n is a $(J_n\times 1)$ vector of utilities, V_n is a $(J_n\times 1)$ vector of deterministic utilities, ε_n is a $(J_n\times 1)$ vector of random terms, ζ_n is an $(M\times 1)$ vector of factors which are IID standard normal distributed, F_n is a $(J_n\times M)$ matrix of loadings that map the factors to the random utility vector and T is a *MxM* lower triangular matrix, capturing the Cholesky factor of the variance-covariance matrix. This specification is very general, and allows explicitly specifying some structure in the model and, consequently, decreasing the complexity. We describe here special cases of factor analytic representations. They are discussed in more detail by Ben-Akiva, Bolduc, and Walker (2001) and Walker (2001).

Heteroscedasticity A heteroscedastic[2] model is obtained when $\mathbf{F}_n$ is the identity matrix, and T is a diagonal matrix containing the alternative specific standard deviations $\sigma_{i,}$.

Error components The error component formulation is based on fixed factor loadings equal to 0 or 1. Entry in row *i* and column *j* of $\mathbf{F}_n$ is 1 if error term ζ_{jn} applies to alternative *i*, and 0 otherwise. A typical specification of $\mathbf{F}_n$ is based on a nested structure, where each alternative belongs to exactly one next. In that case, a factor is associated with each nest, and the entry (i,j) of $\mathbf{F}_n$ is 1 if alternative *i* belongs to nest *j*. A cross-nested specification is also possible, by allowing an alternative to belong to more than one nest. Finally, we note that the matrix T is usually diagonal and must be estimated.

Factor analytic The term "factor analytic" usually refers to the formulation where the loading factor F_n is not imposed a priori and must be estimated. In general, the matrix T is usually diagonal in that case.

General autoregressive process Assuming that the disturbances follow an autoregressive process allows decreasing the model complexity while keeping a reasonable level of generality. Interestingly, such an assumption fits in the Generalized Factor Analytic Specification. We consider the case where the error term ε_n is generated from a first-order autoregressive process:

$$\varepsilon_n = \rho W_n \varepsilon_n + T\,\zeta_n,$$

where W_n is a $(J_n \times J_n)$ matrix of weights describing the influence of each component of the error terms on the others, and ζ_n is an $(M\times 1)$ vector of error terms which are IID standard normal distributed. We can write the process as

$$\varepsilon_n = (I-\rho W_n)^{-1}\, T\,\zeta_n,$$

which is a special case of the Generalized Factor Analytic Specification with

$$F_n = (I-\rho W_n)^{-1}.$$

Random parameters We consider a utility function $X_n\beta$, and we assume that the parameters β are normally distributed with mean $\hat{\beta}$ and variance-covariance matrix TT'. Therefore, $\beta = \hat{\beta} + T\zeta_n$ where ζ_n are IID normal distribution. The utility function can be written as

$$X_n\hat{\beta} + X_n T\zeta_n,$$

[2] Heteroscedasticity here refers to different variances among the alternatives. We use it in this context to refer to a diagonal variance-covariance matrix with potentially different terms on the diagonal.

which is a Generalized Factor Analytic formulation, with $\mathbf{F}_n=\mathbf{X}_n$. Using such a utility function in a probit model does not cause any difficulty, as all random terms are normally distributed. If a Multinomial Logit model is preferred, the formulation contains both Gumbel and normal error terms, and the model becomes a Hybrid Logit model, or Mixed Logit model, which is described in the next section.

Hybrid Logit Model

The Multinomial Probit with a Logit kernel model, called Hybrid Logit or Mixed Logit, has been introduced by Bolduc and Ben-Akiva (1991). It is intended to bridge the gap between Logit and Probit models by combining the advantages of both. It is based on the following utility functions:

$$U_{in}= V_{in}+ \xi_{in}+ \upsilon_{in},$$

where ξ_{in} are normally distributed and capture correlation between alternatives, and υ_{in} are independent and identically distributed Gumbel variables. If the ξ_{in} are given, the model corresponds to a Multinomial Logit formulation:

$$P(i|C_n,\xi_n) = \frac{e^{V_{in}+\xi_{in}}}{\sum_{j\in C_n} e^{V_{jn}+\xi_{jn}}},$$

where $\xi_n=[\xi_1,..., \xi_J]^T$ is the vector of unobserved random terms. Therefore, the probability to choose alternative i is given by

$$P(i\,|\,C_n) = \int_{\xi_n} P(i\,|\,C_n,\xi_n) f(\xi_n) d\xi_n$$

where $f(\xi_n)$ is the probability density function of ξ_n. This model is a generalization of the Multinomial Probit Model when the distribution $f(\xi_n)$ is a multivariate normal. Other distributions may also be used. The earliest application of this model to capture random coefficients in the Logit Model (see below) was by Cardell and Dunbar (1980). More recent results highlighted the robustness of Hybrid Logit (see McFadden and Train, 2000). We note that the Hybrid Logit model can be combined with any Generalized Factor Analytic formulation. The random parameters model presented above is an example of such a combination. We refer the reader to Ben-Akiva et al. (2002) for a review of Hybrid Logit models.

Latent Class Choice Model

Latent class choice models are also designed to capture unobserved heterogeneity (see Everitt, 1984, for an introduction to latent variable models). The underlying assumption is that the heterogeneity is generated by discrete constructs. These constructs are not directly observable and therefore are represented by latent classes. For example, heterogeneity may be produced by taste variations across segments of the population, or when choice sets considered by individuals vary (latent choice set).

The latent class choice model is given by:

$$P(i|X_n) = \sum_{s=1}^{S} P(i|X_n;\beta_s,C_s)P(s|X_n;\theta)$$

where S is the number of latent classes, X_n is the vector of attributes of alternatives and characteristics of decision-maker n, β_s *are* the choice model parameters specific to class s, C_s is the choice set specific to class s, and θ is an unknown parameter vector.

The model

$$P(s|X_n;\theta)$$

is the class membership model, and

$$P(i|X_n;\beta_s,C_s)$$

is the class-specific choice model (Kamakura and Russell, 1989, Gopinath, 1995).
Special case: latent choice sets A special case is the choice model with latent choice sets:

$$P(i\,|\,X_n) = \sum_{C_n \in G} P(i\,|\,X_n,C_n)P(C_n\,|\,X_n)$$

where G is the set of all non-empty subsets of the universal choice set M, and $P(i|X_n,C_n)$ is a choice model. We note here that the size of G grows exponentially with the size of the universal choice set.

The latent choice set can be modeled using the concept of alternative availability. For such a model, a list of constraints or criteria is used to characterize the availability of alternatives. For each alternative i, a binary random variable A_{in} is defined such that $A_{in}=1$ if alternative i is available to individual n, and 0 otherwise. A list of K_{in} constraints is defined as follows:

$$A_{in} = 1 \text{ if } H_{ink} \geq 0,\ \forall k=1,\ldots,K_{in}.$$

For example, in a path choice context, one may consider that a path is not available if the ratio between its length and the shortest path length is above some threshold, represented by a random variable. For example, the associated constraint for path i could be:

$$L_i\,/\,L^* \geq 2+\varepsilon$$

where L^* is the length of the shortest path, L_i is the length of path i and ε a random variable with zero mean. It means that, on average, paths longer than twice the length of the shortest path are rejected.

The probability for an alternative to be available is given by

$$\mathrm{P}(A_{in} = 1) = \mathrm{P}(H_{ink} \geq 0\ \forall k=1,\ldots,K_{in}).$$

The latent choice set probability is then:

$$P(C_n) = \frac{P(A_{in} = 1, \forall i \in C_n \text{ and } A_{jn} = 0, \forall j \notin C_n)}{1 - P(A_{\mathrm{ln}} = 0, \forall l \in M)}.$$

If the availability criteria are assumed to be independent, we have

$$P(C_n) = \frac{\prod_{i \in C} P(A_{in} = 1) \prod_{j \notin C} P(A_{in} = 0)}{1 - \prod_{l \in M} P(A_l = 0)}.$$

Swait and Ben-Akiva (1987) estimate a latent choice set model of mode choice in a Brazilian city. See also Ben-Akiva and Boccara (1995) for a more detail analysis of discrete choice models with latent choice sets.

2.3 Model estimation

The estimation of discrete choice models from sample data is a difficult and important task. Most statistical packages provide estimation capabilities for simple models, like the Multinomial Logit models. Dedicated commercial software packages are available for the estimation of Nested-Logit models. Free software for estimation of Hybrid, or Mixed, Logit models (emlab.berkeley.edu/users/train/software.html) and GEV models (rosowww.epfl.ch/mbi/biogeme) is also available. However, these packages do not cover the entire range of models, and a specific implementation of an estimation procedure is sometimes necessary. We discuss here some issues related to such implementation.

Maximum likelihood estimation is the most widely used technique for discrete choice model estimation (see statistical textbooks, such as Sprott, 2000, and Severini, 2000). It aims at identifying the set of parameters maximizing the probability that a given model perfectly reproduces the observations. It is a nonlinear programming problem. The nature of the objective function and of the constraints determines the type of solution algorithm that must be used.

The objective function of the maximum likelihood estimation problem for GEV models is a nonlinear analytical function, as the probability density function has a closed form. In general, the function is not concave (except for the Multinomial Logit Model) and, therefore, significantly complicates the identification of a (global) maximum. Most nonlinear programming algorithms (see Dennis and Schabel, 1983, or Bertsekas, 1995) are designed to identify local optima of the objective function. There exists some meta-heuristics designed to identify global optima (like genetic algorithms, and simulated annealing) but none of them can guarantee that the provided solution is a global optimum. Therefore, whatever algorithm is preferred, starting it from different initial solutions is a good practice.

For the Probit or Hybrid-logit models, the objective function does not have an analytical form and must be evaluated based on Monte Carlo (Metropolis and Ulam, 1949) or Quasi-Monte Carlo methods (Morokoff and Caflish, 1995). Contrarily to MonteCarlo, Quasi-Monte Carlo techniques are deterministic. They require fewer "draws" than Monte-Carlo simulation to reach the same level of accuracy (see Spanier and Maize, 1994).

Not all parameters of a model can be identified from the data. Parameter identification and model normalization issues are important to analyze before performing an actual estimation. We refer the reader to Ben-Akiva and Lerman (1985) for a general discussion on such issues. Bunch (1991) and Bolduc (1992) address the case of the Probit model. Walker (2001) provides a detailed analysis of identification issues for the Hybrid Logit model.

The parameters to be estimated must verify some constraints. First, most of them must lie within bounds in order for the model to be consistent with the theory (e.g. the homogeneous parameters of GEV functions must be non-negative) or with their intuitive interpretation (e.g. the coefficient for cost or travel time in a utility function is usually non-positive). Moreover, some constraints have to be verified in order for the model to be estimable (e.g. the sum of α parameters must sum up to one in a Cross-Nested Logit model). In the past, it was usually advised to ignore the bound constraints, to eliminate other constraints by incorporating them in the objective function, and to use unconstrained optimization algorithms. The increasing complexity of the models, combined with the availability of efficient software packages for constrained optimization motivate now the explicit management of constraints in the estimation process.

2.4 Route Choice Applications

The route choice problem plays an important role in many transportation related applications. In this section, we analyze its specific assumptions, and present some models designed to capture this complex behavioral problem.

Given a transportation network composed of nodes, links, origins and destinations; and given an origin *o,* a destination *d* and a transportation mode *m*, what is the chosen route between *o* and *d* on mode *m.* This discrete choice problem has specific characteristics. First, the universal choice set is usually very large. Second, the decision-maker considers not all physically feasible alternatives. Third, the alternatives are usually correlated, due to overlapping paths.

We now describe typical assumptions associated with route choice models.

Decision-Maker

The traveler's characteristics most often used for route choice applications are:

- <u>Value-of-time.</u> Obviously, travel time is a key attribute of alternative routes. Its influence on behavior, however, may vary across individuals. The sensitivity of an individual to travel time is usually referred to as the value-of-time. It can be represented by a continuous variable (e.g., the dollar-value equivalent of a minute spent traveling) or by a discrete variable identifying the decision-maker's value-of-time as low, medium or high.

- Access to information. Information about network conditions may significantly influence route choice behavior. Therefore, it may be important that a route choice model explicitly differentiates travelers with access to such information from those without access. It is also an important policy variable. It may be modeled by a single binary attribute (access/no access) or by several binary variables identifying the type of information available to the traveler (pre-trip information, on-board computer, etc.)
- Trip purpose. The purpose of the trip may significantly influence the route choice behavior. For example, a trip to work may be associated with a penalty for late arrival, while a shopping trip would usually have no such penalty.

Alternatives

Identifying the choice set in a route choice context is a difficult task. Two main approaches can be considered.

First, it may be assumed that each individual can potentially choose any path between her/his origin and destination. The choice set is easy to identify, but the number of alternatives can be very large, causing operational problems in estimating and applying the model. Moreover, this assumption is behaviorally unrealistic.

Second, a restricted number of paths may be considered in the choice set. The choice set generation can be deterministic or stochastic, depending on the analyst's knowledge of the problem.

Dial (1971) proposes to include in the choice set "reasonable" paths composed of links that would not move the traveler farther away from her/his destination. The labeling approach (proposed by Ben-Akiva *et al.,* 1984) includes paths meeting specific criteria, such as shortest paths, fastest paths, most scenic paths, paths with fewest stop lights, paths with least congestion, paths with greatest portion of freeways, paths with no left turns, etc.

Azevedo *et al.* (1993) propose the link elimination approach, where the shortest path (according to a given impedance) is first calculated and introduced in the choice set. Then, some links belonging to the shortest path are removed, and a the shortest path in the modified network is computed and introduced in the choice set.

Cascetta and Papola (1998) propose an implicit probabilistic choice set generation model, where the utility function associated with path *i* by individual *n* is defined as

$$U_{in} = V_{in} + \ln q_{in} + \varepsilon_{in},$$

where q_{in} is a random variable with mean

$$\bar{q}_{in} = \frac{1}{1 + e^{\sum_k -\gamma_k X^A_{ink}}}.$$

X^A_{ink} are the attributes for availability and perception of the path and γ_k are parameters to be estimated.

Swait (2001) combines the probabilistic choice set generation with the route choice model within a Cross-Nested structure.

Some recent models (Nguyen and Pallottino, 1987, Nguyen, Pallottino and Gendreau, 1988) consider hyperpaths instead of paths as alternatives. A hyperpath is a collection of paths with associated strategies at decision nodes. This technique is particularly appropriate for a public transportation network.

Attributes

In describing the attributes of the alternatives to be included in the utility function, we need to distinguish between link-additive and non-link-additive attributes.

If *i* is a path composed of links $a \in \Gamma_i$, x_i is a link-additive attribute of *i* if

$$x_i = \sum_{a \in \Gamma_i} x_a \text{ ,}$$

where x_a is the corresponding attribute of link *a*. For example, the travel time on a path is the sum of the travel times on links composing the path. Qualitative attributes are in general non-link-additive. For example, a binary variable x_i equal to one if the path is a habitual path and 0 otherwise, is non link-additive. In the context of public transportation, variables like transfers and fares are usually not link-additive. The distinction is important because some models, designed to avoid path enumeration, use link attributes and not path attributes.

Among the many attributes that can potentially be included in a utility function, travel time is probably the most important. But what does travel time mean for the decision-maker? How does she/he perceive travel time? Many models are based on the assumption that most travelers are sufficiently experienced and knowledgeable about usual network conditions and, therefore, are able to estimate travel times accurately. This assumption may be satisfactory for planning applications using static models. With the emergence of Intelligent Transportation Systems, models that are able to predict the impact of real-time information have been developed. In this context, the "perfect knowledge" assumption is contradictory with the ITS services that provide information. Several approaches can be used to capture perceptions of travel times. One approach represents travel time as a random variable in the utility function. This idea was introduced by Burrell (1968) and is captured by a random utility model. Also, the uncertainty or the variability of travel time along a given path can be explicitly included as an attribute of the path.

In addition to travel time, the following attributes are usually included.

- Path length. The length of the path is likely to influence the decision maker's choice. Also, this attribute is easy to measure. Note that it may be highly correlated with travel time, especially in uncongested networks.

- Travel cost. In addition to the obvious behavioral motivation, including travel cost in the utility function is necessary to forecast the impact of tolls and congestion pricing, for example. It is common practice to distinguish the so-called out-of-pocket costs (like tolls), which are directly associated with a specific trip, from other general costs (like car operating costs).

- Transit specific. Attributes specific to route choice in transit networks include number of transfers, waiting and walking time and service frequency.
- Others. Traffic conditions (e.g. level of congestion, volume of conflicting traffic streams or pedestrian movements), obstacles (e.g. number of stop signs, number of traffic lights, number of left turns against traffic), road types (e.g. dummy variable capturing preference for freeways) and road condition (e.g. surface quality, number of lanes, safety, scenery) are some of the other attributes that may be considered. Whether to include them in the utility function depends on their behavioral pertinence in a specific context, and on data availability.

Finally, the level of path overlapping can also be included in the utility function of a path. It is not one of its attributes per se. It is more a measure of how the alternative is perceived within a choice set. Several formulations have been proposed in the literature.

Commonality Factor. Cascetta *et al.* (1996) propose the following specification for the commonality factor

$$\mathrm{CF}_{in} = \beta_{\mathrm{CF}} \ln \sum_{j \in C_n} \left(\frac{L_{ij}}{\sqrt{L_i L_j}} \right)^{\gamma}$$

where L_{ij} is the length[3] of links common to paths i and j, and L_i and L_j are the overall length of paths i and j, respectively. β_{CF} is a coefficient to be estimated. The parameter γ may be estimated or constrained to a convenient value, often 1 or 2.

Considering the path choice example in Figure 1, the commonality factor for path 1 is zero because it does not overlap with any other path. The commonality factor for paths 2a and 2b is

$$\beta_{CF} \ln(1 + [(T\text{-}\delta)/T]^{\gamma}).$$

Path Size. The Path Size model, first proposed by Ben-Akiva and Bierlaire (1999), is an application of the notion of elemental alternatives and size variables. In the route choice context, we assume that an overlapping path may not be perceived as a distinct alternative. Indeed, a path contains links that may be shared by several paths. Hence, the size of a path with one or more shared links may be less than one. We include in the utility function of path i for individual n a size variable defined by

[3] or any other link-additive attribute

$$PS_{in} = \ln \sum_{a\in\Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j\in C_n} \delta_{aj} \frac{L^*_{C_n}}{L_j}}$$

and Γ_i is the set of links in path i; l_a and L_i are the length of link a and path i, respectively; δ_{aj} is the link-path incidence variable that is one if link a is on path j and 0 otherwise; and $L^*_{C_n}$ is the length of the shortest path in C_n. Considering again the path choice problem from Figure 1, the size of path 1 is 1, and the size of paths 2a and 2b is $(T+\delta)/2T$.

Generalized Path Size Ramming (2001) proposed a generalized formulation of the Path Size where

$$PS_{in} = \ln \sum_{a\in\Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j\in C_n} \frac{G(L_i,\gamma)}{G(L_j,\gamma)} \delta_{aj}}$$

and G is a function with paratemer g. The Exponential Path-Size formulation is obtained with $G(L_i,\gamma) = L_i^{\gamma}$. Note that $\gamma=0$ corresponds to a normalized version of the original Path Size model. Based on experiments on a case study in Boston, Ramming (2001) observes a better behavior of the Path Size correction in terms of its capability of reproducing observed data, compared to the Commonality Factor. Also, it is observed that low values of γ may lead to counter-intuitive results, motivating the generalized version.

Decision Rules

Shortest path. The simplest possible decision rule in the route choice context assumes that each individual chooses the path with the highest utility. Models based on deterministic utility maximization are supported by efficient algorithms to compute shortest paths in a graph (e.g. Dijkstra, 1959, and Dial, 1969). However, the behavioral limitations of this approach have motivated the development of stochastic models based on the random utility model.

Logit route choice. A Multinomial Logit Model with an efficient algorithm for route choice has been proposed by Dial (1971). Using the concept of "reasonable paths" to define the choice set and assuming the paths attributes to be link-additive, this algorithm avoids explicit path enumeration.

As described earlier, the IIA property of the Multinomial Logit Model is the major weakness of Dial's algorithm in the context of highly overlapping routes. Therefore, its use is limited to networks with specific topologies. A Logit model may also be used with a choice set generation model, such as the Labeling approach, that results in a small size choice set with limited overlap.

Probit route choice. Given the shortcomings of the Logit route choice model, Probit models have been proposed in the context of stochastic network loading by Burrell (1968), Daganzo and Sheffi (1977) and Yai *et al.* (1997). The two problems in this case are the complexity of the variance-covariance matrix and the lack of an analytical formulation for the probabilities. The covariance structure can be simplified when path utilities are link-additive, the variance of link utility is proportional to the utility itself, and the covariance of utilities of two different links is zero. The use of a factor analytic formulation, where the matrix F_n is the link-path incidence matrix, enables to reduce the complexity of the model from the number of paths in the network down to the number of links. A Monte Carlo or Quasi-Monte Carlo simulation is often used to circumvent the absence of a closed analytical form.

Cross-Nested Logit route choice. Vovsha and Bekhor (1998) have proposed an interesting Cross-Nested formulation, where each link of the network corresponds to a nest, and each path to an alternative. The α parameters of the Cross-Nested logit are not estimated. They capture the network topology and, consequently, the path overlapping. Note that this approach can be combined with attribute-based path overlapping measures, like the commonality factor and the path-size.

Hybrid Logit route choice. Ramming (2001) has estimated a route choice model based on Hybrid Logit. Although there are some issues regarding the number of draws for the Simulated Maximum Likelihood Estimation, experiments on a case study in Boston report a good behavior of the Hybrid Logit approach, especially when it is combined with the Path Size overlapping attribute.

2.5 Departure Choice Applications

Modeling the choice of departure time appears in the context of dynamic traffic assignment as an extension of the route choice problem. It is important to distinguish the departure time choice itself and the choice of changing departure time. The latter appears usually in the context of Traveler Information Systems, where individuals may revisit a previous choice using additional information. We now describe typical modeling assumptions associated with the departure time choice model.

Decision-Maker

The central traveler's characteristic of departure time choice models is the preferred arrival time at the destination. It is often presented as a time interval or window with variable length reflecting schedule flexibility. Other relevant traveler's characteristics are the (monetary and psychological) penalties for early and late arrivals. In the context of a departure time change, the individual's "habitual" departure time must also be known.

Travelers generally use their expectations of travel time and "subtract" this from their intended arrival time to determine what the departure time should be, with a safety margin factored in. The magnitude of the safety margin depends on the travel time variability and the penalties for late or early arrival. In this context, an intended arrival time is the outcome of a choice and is not necessarily the same as the preferred arrival time. Within some constraints, travelers may change their intended arrival time in the process of making a departure time choice.

From travel diaries one can obtain data on actual departure and arrival times. Preferred arrival times can be obtained only by a direct question. Such questions are usually not included in travel diary surveys. Moreover, even when these questions are included, the answers may be biased towards the actual or the intended arrival time, since respondents may try to justify (to themselves and/or the interviewer) their actual behavior when questioned about their preferences. Therefore, the preferred arrival time characteristic may be measured with significant errors. The modeling implications are considered below.

Alternatives

The choice set specification for departure time models is an intricate problem. First, the continuous time must be discretized. A reasonable compromise must be found between a fine temporal resolution and the model complexity. Indeed, there is a potentially large number of alternatives, particularly for realistic dynamic traffic applications. Second, the correlation among alternatives cannot be ignored, especially when time intervals are short. Choosing between the 7:45-7:50 and 7:50-7:55 time intervals differs from choosing between 7:45-7:50 and 8:45-8:50. In the first case, the two alternatives are likely to share unobserved attributes. Third, the perception of the alternatives depends on trip travel time. Most individuals round time and the rounding may depend on the travel time and travel time variability. For short trips, 7:52 may be rounded to 7:50, whereas for long trips it may be approximated by 8:00.

The choice set generation consists of defining an acceptable range of departure time intervals considered by an individual *n*. A common procedure is based on the preferred arrival time PAT^*_n. Let $[\mathrm{PAT}_{n,\min}; \mathrm{PAT}_{n,\max}]$ be the feasible arrival time interval, and let $[\mathrm{TT}_{n,\min}; \mathrm{TT}_{n,\max}]$ be the range of travel times. Then the interval of acceptable departure times is $[\mathrm{DT}_{n,\min}; \mathrm{DT}_{n,\max}] = [\mathrm{PAT}_{n,\min}\text{-}\mathrm{TT}_{n,\max}; \mathrm{PAT}_{n,\max}\text{-}\mathrm{TT}_{n,\min}]$. Overestimating the length of the acceptable or feasible departure time interval should not cause errors if the model is otherwise well specified. The tendency may be to attempt to reduce the number of alternatives in the model by understating this interval. This, unfortunately, may cause significant errors. Small (1987) analyzed the impact of truncating the departure time choice set. He concluded that there is no problem if the true model is a Multinomial Logit Model. Some adjustments are needed if a Cross-Nested Logit with ordered alternatives is assumed.

In the context of departure time change, the alternatives may be described in a relative way. Antoniou *et al.* (1997) propose a choice set with five alternatives: do not change, switch to an earlier or a later departure, by one or two time intervals.

Attributes

Travel time is a key attribute of departure time alternatives. Travel time variability also affects departure time choice through the above mentioned safety margin. However, data on this attribute are rarely available and it is conveniently assumed to be constant across the peak period. Other important attributes are the early and late schedule delays. These are interactions of travel time and preferred arrival time, as follows. Given a preferred arrival time PAT^*_n, a penalty-free interval is defined: $[\mathrm{PAT}^*_{n,min};\mathrm{PAT}^*_{n,max}]$. It is assumed that the individual suffers no penalty if the arrival times lies within the interval. The actual arrival time AT_n is equal to $\mathrm{DT}_n+\mathrm{TT}(\mathrm{DT}_n)$, where $\mathrm{TT}(\mathrm{DT}_n)$ is the travel time if the trip starts at time DT_n. The early schedule delay is defined as

$$\mathrm{Max}\ [\mathrm{PAT}^*_{n.min}-\mathrm{AT}_n\ ,\ 0]$$

and the late schedule delay is defined as

$$\mathrm{Max}\ [\mathrm{AT}_n-\mathrm{PAT}^*_{n.max}\ ,\ 0].$$

In the context of a change in departure time, individuals may also assign a penalty to departure times that are significantly different from their habitual departure times, due to the inertia associated with habits.

The key data requirement concerns the travel times for the alternative departure times, denoted by TT(DT). A calibrated dynamic traffic assignment or a traffic simulation model with sufficiently high temporal resolution could be used to predict these travel times.

Decision Rules

Deterministic models assume that there are no unobserved effects. The chosen departure time is determined by maximizing a utility (or minimizing a generalized cost) function that is a linear combination of the above-mentioned attributes. A deterministic departure time choice model, when applied as part of an equilibrium model with congested networks, may result in a spread of demand across a time interval with a constant utility (or generalized cost). Across this constant utility interval the change in travel time disutility is equal to the negative of the change in schedule delay disutility. If schedule delay disutility changes linearly, then the equilibrium travel time also changes linearly.

It is clearly unreasonable that such a constant utility condition and a linear trend of travel time will hold over long time intervals. Thus, to achieve a reasonable spread of departure times, an application of a deterministic model must account for a high degree of heterogeneity. To a limited extent, heterogeneity may be captured by market segmentation (as in multiple classes of travelers). However, for the same considerations that apply to mode choice models, there are significant unobserved effects that can only be captured by continuous distributions using probabilistic departure time choice models.

Probabilistic Choice Models. Departure time choice models have been estimated using both revealed preferences (RP) and stated preferences (SP) data. Availability of suitable RP data sets for departure time choice is limited because data on travel times for the different departure time alternatives are usually unavailable at sufficient detail. SP data have been used to estimate models of the choice of changing departure time.

The cumulative experience in estimating RP departure time choice models is very limited. Small (1982), Hendrickson and Plank (1984) and Cascetta et al. (1992) used RP data to estimate Logit models of departure time choice. Small (1987) extended his earlier work with a Cross-Nested Logit model, where m adjacent departure time intervals are nested together, capturing their intrinsic correlation. A single departure time interval belongs to m different nests, source of the cross-nested structure.

de Palma, Fontan and Mekkaoui (2000) estimate the distribution of desired departure times for public transportation users, based on traffic counts, using a non parametric regression approach.

de Palma and Fontan (2001) calibrate a Hybrid Logit model with SP data, collected from a computer-assisted survey, with personalized scenarios. In the context of departure time change, Antoniou et al. (1997) proposed a Nested Logit Model based on SP data for joint choice of departure time and route. Liu and Mahmassani (1998) used SP data to estimate a Probit model where day-to-day correlation is assumed.

Choice Models with Latent Variables. The departure time choice model considered so far is the conditional probability of departure time given a known preferred arrival time, or preferred arrival time window. It was argued above that preferred arrival times cannot be observed from travel diaries and are measured with significant errors. Thus, the proper modeling framework requires that the preferred arrival time be treated as a latent variable. A marginal departure choice probability is then calculated by integrating the model over a distribution of preferred arrival time. The parameters of this preferred arrival time distribution are unknown and need to be estimated jointly with the unknown parameters of the conditional departure time probability. Responses to survey questions about preferred arrival times can be used by adding the corresponding measurement equations or by estimating a model of the joint probability of the departure time choice and the reported preferred arrival time. See Ben-Akiva et al. (1997) for more detail concerning this modeling framework.

2.5 Conclusion

Discrete choice methods are constantly evolving to accommodate the requirements of specific applications. This is an exciting field of research, where a deep understanding of the underlying theoretical assumptions is necessary both to apply the models and develop new ones. In this Chapter, we have summarized the fundamental aspects of discrete choice theory, and we have introduced recent model developments, illustrating their richness. A discussion on route choice and departure time choice applications have shown how specific aspects of real applications must be addressed.

References

Antoniou, C., Ben-Akiva, M., Bierlaire, M. and Mishalani, R. (1997). Demand simulation for dynamic traffic assignment, Proceedings of the 8th IFAC Symposium on Transportation Systems, Chania, Greece.

Azevedo J.A., Santos Costa M.E.O., Silvestre Madeira J.J.E.R. and Vieira Martins, E.Q. (1993) "An algorithm for the ranking of shotest paths". *European Journal of Operational Research* **69** 97-106.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions, in R. Hall (ed.), Handbook of Transportation Science, Kluwer, pp. 5-34.

Ben-Akiva, M. and Bolduc, D. (1996). Multinomial probit with a logit kernel and a general parametric specification of the covariance structure. Paper presented at the 3rd Invitational Choice Symposium, Columbia University.

Ben-Akiva, M., Bolduc, D. and Walker, J. (2001). Specification, Identification, & Estimation of the Logit Kernel (or Continuous Mixed Logit) Model, Working paper, Department of Civil Engineering, MIT, Camridge, MA.

Ben-Akiva, M. E. (1973). Structure of passenger travel demand models, PhD thesis, Department of Civil Engineering, MIT, Cambridge, Ma.

Ben-Akiva, M. E. (1974). Structure of passenger travel demand models, Transportation Research Record 526.

Ben-Akiva, M.E. and Boccara, B. (1995) Discrete Choice Models with Latent Choice sets, International Journal of Research in Marketing 12, pp. 9-24.

Ben-Akiva, M. E., Bergman, M. J., Daly, A. J. and Ramaswamy, R. (1984). Modeling inter-urban route choice behaviour, in J. Volmuller and R. Hamerslag (eds), Proceedings from the ninth international symposium on transportation and traffic theory, VNU Science Press, Utrecht, Netherlands, pp. 299-330.

Ben-Akiva, M. E. and Lerman, S. R. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, Cambridge, Ma.

Ben-Akiva, M. and Francois, B. (1983). μ homogeneous generalized extreme value model, Working paper, Department of Civil Engineering, MIT, Cambridge, Ma.

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D., Daly, A., de Palma, A., Gopinath, D., Karlstrom, A. and Munizaga, M. A. (2002). Hybrid choice models: Progress and challenges. Accepted for publication in *Marketing Letters.*

Bertsekas, D. P. (1995). Nonlinear Programming, Athena Scientific, Belmont.

Bhat, C. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model, TRB 35(7): 677-693.

Bierlaire, M. (1995). A robust algorithm for the simultaneous estimation of hierarchical logit models, GRT Report 95/3, Department of Mathematics, FUNDP.

Bierlaire, M. (1998). Discrete choice models, in M. Labbé, G. Laporte, K. Tanczos and P. Toint (eds), Operations Research and Decision Aid Methodologies in Traffic and Transportation Management, Vol. 166 of NATO ASI Series, Series F: Computer and Systems Sciences, Springer Verlag, pp. 203-227.

Bierlaire, M. (2001a). A general formulation of the cross-nested logit model, Proceedings of the 1st Swiss Transportation Research Conference, Ascona, Switzerland, http ://www.strc.ch

Bierlaire, M. (2001b). An introductory tutorial to BIOGEME. URL: http://rosowww.epfl.ch/mbi/biogeme

Bierlaire, M. (2002). The network GEV model, Proceedings of the 2nd Swiss Transportation Research Conference, Ascona, Switzerland. http://www.strc.ch

Bierlaire, M., Lotan, T. and Toint, P. L. (1997). On the overspecification of multinomial and nested logit models due to alternative specific constants, Transportation Science 31(4): 363-371.

Bierlaire, M. and Vandevyvere, Y. (1995). HieLoW: the interactive user's guide, Transportation Research Group - FUNDP, Namur.

Bolduc, D. (1992). Generalized Autoregressive Errors in the Multinomial Probit Model, Transportation Research B 26(2), 155-170

Bolduc, D. (1999). A practical technique to estimate multinomial probit models in transportation, Transportation Research B 33(1): 63-79.

Bolduc, D. and Ben-Akiva, M. (1991) A Multinomial Probit Formulation for Large Choice Sets. *Proceedings of the 6th International Conference on Travel Behaviour* **2**, 243-258.

Bradley, M. A. and Daly, A. (1991). Estimation of logit choice models using mixed stated preferences and revealed preferences information, Methods for understanding travel behaviour in the 1990's, International Association for Travel Behaviour, Qu'ebec, pp. 116-133. 6th international conference on travel behaviour.

Bunch, D.A. (1991) Estimability in the Multinomial Probit Model. *Transportation Research B* **25** 1-12.

Burrell, J. E. (1968). Multipath route assignment and its application to capacity restraint, Proceedings of the 4th International Symposium on the Theory of Road Traffic Flow (Karlsruhe).

Cardell and Dunbar (1980). Measuring the societal impacts of automobile downsizing, Transportation Research A 14(5-6): 423-434.

Cascetta, E., Nuzzolo, A. and Biggiero, L. (1992). Analysis and modeling of commuters' departure time and route choice in urban networks, Proceedings of the second international Capri seminar on urban traffic networks.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks, Proceedings of the 13th International Symposium on the Theory of Road Traffic Flow (Lyon, France).

Cascetta, E. and Papola, A. (1998). Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand, Technical report, Universita degli Studi di Napoli Federico II.

Conn, A., Gould, N. and Toint, P. (2000). Trust region methods, MPS-SIAM Series on Optimization, SIAM.

Daganzo, C. F. and Sheffi, Y. (1977). On stochastic models of traffic assignment, Transportation Science 11(3): 253-274.

Daly, A. (1987). Estimating "tree" logit models, Transportation Research B 21(4): 251-268.

Daly, A. (2001). Recursive nested EV model. Paper presented at the Invitational Choice Symposium, Asilomar, Ca.

de Palma, A. and Fontan, C. (2001). Departure time choice and heterogeneity of commuters, Proceedings of the 9th WCTR. Paper 4149.

de Palma, A., Fontan, C. and Mekkaoui, O. (2000). Trip timing for public transportation: an empirical application, Technical Report 2000- 19, Université de Cergy-Pontoise - Univcersité de Paris X - Nanterre, Théorie Economique, Modélisation et Application. UMR 7536 CNRS.

Dennis, J. E. and Schnabel, R. B. (1983). Numerical methods for unconstrained optimization and nonlinear equations, Prentice-Hall, Englewood Cliffs, USA.

Dial, R. B. (1969). Algorithm 360: shortest path forest with topological ordering., Communications of ACM 12: 632-633.

Dial, R. B. (1971). A probabilistic multipath traffic assignment algorithm which obviates path enumeration, Transportation Research 5(2): 83-111.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs, Numerische Mathematik 1: 269-271.

Everitt,B.S. (1984). An Introduction to Latent Variable Models, Monographs on Statistical and Applied Probability, Chapman and Hall.

Gopinath, D.A. (1995) *Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand,* Ph.D. Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Gumbel, E. J. (1958). Statistics of Extremes, Columbia University Press, New York.

Hendrickson, C. and Plank, E. (1984). The flexibility of departure times for work trips, Transportation Research A 18: 25-36.

Hensher D.A. and Greene W.H. (2002). Specification and estimation of the nested logit model: alternative normalizations, Transportation Research B 36(1): 1-18.

Kamakura, W.A. and G.J. Russell (1989). A Probabilistic Choice Model for Market Segmentation and Elasticity Structure, *Journal of Marketing Research* 25, 379-390.

Koppelman, F. S. and Wen, C.-H. (1998). Alternative nested logit models: Structure, properties and estimation, Transportation Research B 32(5): 289-298.

Liu, Y.-H. and Mahmassani, H. (1998). Dynamic aspects of departure time and route decision behavior under ATIS: modeling framework and experimental results, presented at the 77th annual meeting of the Transportation Research Board, Washington DC.

Luce, R. (1959). Individual choice behavior: a theoretical analysis, J. Wiley and Sons, New York.

Manski, C. (1977). The structure of random utility models, Theory and Decision 8: 229-254.

McFadden, D. (1978). Modelling the choice of residential location, in A. Karlquist et al. (ed.), Spatial interaction theory and residential location, North-Holland, Amsterdam, pp. 75-96.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration, Econometrica 57(5): 995-1026.

McFadden, D. and Train, K. (2000). Mixed multinomial logit models for discrete response, Journal of Applied Econometrics, Vol. 15, No. 5, pp. 447-470.

Morokoff, W.J. and Caflish R.E. (1995) Quasi-Monte Carlo integration. *Journal of Computational Physics* **122**.

Nguyen, S. and Pallottino, S. (1987). Traffic assignment for large scale transit networks, in A. Odoni (ed.), Flow control of congested networks, Springer Verlag.

Nguyen, S., Pallottino, S. and Gendreau, M. (1998). Implicit enumeration of hyperpaths in a logit model for transit networks, Transportation Science 32(1).

Panier, E. and Tits, A. L. (1993). On combining feasibility, descent and superlinear convergence in inequality constrained optimization, MP 59: 261-276.

Papola, A. (2000). Some development of the cross-nested logit model, Proceedings of the 9th IATBR Conference.

Ramming, M. S. (2001). Network Knowledge and Route Choice, PhD thesis, Massachusetts Institute of Technology.

Severini, T.A. (2000) Likelihood methods in statistics, Oxford statistical science series, Oxford University Press.

Small, K. (1982). The scheduling of consumer activities: work trips, The American Economic Review pp. 467-479.

Small, K. (1987). A discrete choice model for ordered alternatives, Econometrica 55(2): 409-424.

Spanier, J. and Maize, H. (1994). Quasi-random methods for estimating integrals using relatively small samples, SIREV 36(1): 18-44.

Sprott D. A. (2000) Statistical inference in science, Springer series in statistics, Springer, New York.

Swait, J. (2001). Choice set generation within the generalized extreme value family of discrete choice models, TRB 35(7): 643-666.

Swait, J. and Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation, TRB 21(2).

Tversky, A. (1972). Elimination by aspects: a theory of choice, Psychological Review 79: 281-299.

Vovsha, P. (1997). Cross-nested logit model: an application to mode choice in the Tel-Aviv metropolitan area, Transportation Research Board, 76th Annual Meeting, Washington DC. Paper #970387.

Vovsha, P. and Bekhor, S. (1998). The link-nested logit model of route choice: overcoming the route overlapping problem, Transportation Research Record 1645: 133-142.

Walker, J. L. (2001). Extended discrete choice models: integrated framework, flexible error structures, and latent variables, PhD thesis, Massachusetts Institute of Technology.

Wen, C.-H. and Koppelman, F. S. (2001). The generalized nested logit model, Transportation Research B 35(7): 627-641.

Yai T., Iwakura, S. and Morichi S. (1997). Multinomial Probit with Structured Covariance for Route Choice Behavior. *Transportation Research B.* **31**(3) 195-207.

This page intentionally left blank

3 ACTIVITY-BASED MODELING OF TRAVEL DEMAND

Chandra R. Bhat and Frank S. Koppelman

3.1 Introduction and Scope

Since the beginning of civilization, the viability and economic success of communities have been, to a major extent, determined by the efficiency of the transportation infrastructure. To make informed transportation infrastructure planning decisions, planners and engineers have to be able to forecast the response of transportation demand to changes in the attributes of the transportation system and changes in the attributes of the people using the transportation system. Travel demand models are used for this purpose; specifically, travel demand models are used to predict travel characteristics and usage of transport services under alternative socio-economic scenarios, and for alternative transport service and land-use configurations.

The need for realistic representations of behavior in travel demand modeling is well acknowledged in the literature. This need is particularly acute today as emphasis shifts from evaluating long-term investment-based capital improvement strategies to understanding travel behavior responses to shorter-term congestion management policies such as alternate work schedules, telecommuting, and congestion-pricing. The result has been an increasing realization in the field that the traditional *statistically-oriented* trip-based modeling approach to travel demand analysis needs to be replaced by a more *behaviorally-oriented* activity-based modeling approach. The next two sections discuss the basic concepts of the trip-based and the activity-based approaches to travel demand analysis.

The Trip-Based Approach

The trip based approach uses individual trips as the unit of analysis and usually includes four sequential steps. The first, trip generation, step involves the estimation of the number of home-based and non-home based person-trips produced from, and attracted to, each zone in the study area. The second, trip distribution, step determines the trip-interchanges (*i.e.,* number of trips from each zone to each other zone). The third, mode choice, step splits the person-trips between each pair of zones by travel mode obtaining both the number of vehicle trips and number of transit trips between zones. The fourth, assignment, step assigns the vehicle trips to the roadway network to obtain link volumes and travel times and the person trips to the transit network. Time-of-day of trips is either not modeled or is modeled in only a limited way, in the trip-based approach. Most commonly, time is introduced by applying time-of-day factors to 24-hour travel volumes at the end of the traffic assignment step or at the end of the trip generation step.

A fundamental conceptual problem with the trip-based approach is the use of trips as the unit of analysis. Separate models are developed for home-based trips and non-home based trips, without consideration of dependence among such trips. Further, the organization (scheduling) of trips is not considered; that is, there is no distinction between home-based trips made as part of a single-stop sojourn from home and those made as part of a multiple-stop sojourn from home. Similarly, there is no distinction between non-home based trips made during the morning commute, evening commute, from work, and as part of pursuing multiple stops in a single sojourn from home. Thus, the organization of trips and the resulting inter-relationship in the attributes of multiple trips is ignored in all steps of the trip-based method. This is difficult to justify from a behavioral standpoint. It is unlikely that households will determine the number of home-based trips and the number of non-home based trips separately. Rather, the needs of the households are likely to be translated into a certain number of total activity stops by purpose followed by (or jointly with) decisions regarding how the stops are best organized. Similarly, the location of a stop in a multistop sojourn (or tour) is likely to be affected by the location of other stops on the tour. Such multistop tours are becoming increasingly prevalent (see Gordon *et al.,* 1988; Lockwood and Demetsky, 1994) and ignoring them in travel analysis means *"discarding an element that is doubtless important in the individual's organization of time and space"* (Hanson, 1980). Also, in a multistop tour from home consisting of, say, a grocery shopping stop and a social visit, the trip-based approach fails to recognize that the travel mode for all three trips (home to shop, shop to visit, and visit to home) will be the same. The travel mode chosen will depend on various characteristics of all three trips (and not any one single trip) and, consequently, these trips cannot be studied independently.

The behavioral inadequacy of the trip-based approach, and the consequent limitations of the approach in evaluating demand management policies, has led to the emergence of the activity-based approach to demand analysis.

The Activity-Based Approach

The activity-based approach to travel demand analysis views travel as a derived demand; derived from the need to pursue activities distributed in space (see Jones *et al.,* 1990 or Axhausen and Gärling, 1992). The approach adopts a holistic framework that recognizes the complex interactions in activity and travel behavior. The conceptual appeal of this approach originates from the realization that the need and desire to participate in activities is more basic than the travel that some of these participations may entail. By placing primary emphasis on activity participation and focusing on sequences or patterns of activity behavior (using the whole day or longer periods of time as the unit of analysis), such an approach can address congestion-management issues through an examination of how people modify their activity participations (for example, will individuals substitute more out-of-home activities for in-home activities in the evening if they arrived early from work due to a work-schedule change?).

The shift to an activity-based paradigm has also received an impetus because of the increased information demands placed on travel models by the 1990 Clean Air Act Amendments (CAAAs). These amendments require the inclusion of transportation control measures (TCMs) in transportation improvement programs for MPOs in heavily polluted non-attainment areas and, by state law, for all non-attainment areas in California. Some TCMs, such as HOV lanes and transit extensions, can be represented in the existing modeling framework; however, non-capital improvement measures such as ridesharing incentives, congestion pricing and employer-based demand management schemes can not be so readily represented (Deakin, Harvey and Skabardonis, Inc. 1993, Chapter 2). The ability to model both individual activity behavior and interpersonal linkages between individuals, a core element of activity modeling, is required for the analysis of such TCM proposals. The CAAAs also require travel demand models to provide (for the purpose of forecasting mobile emission levels) link flows at a high level of resolution along the time dimension (for example, every 30 minutes or an hour as opposed to peak-period and off-peak period link flows) and also to provide the number of new vehicle trips (*i.e.,* cold starts) which begin during each time period. Because of the simplistic, "individual-trip" focus of the trip-based models, they are not well-equipped to respond to these new requirements (see Cambridge Systematics, Inc., 1994; Chapter 5). Since the activity-based approach adopts a richer, more holistic approach with detailed representation of the temporal dimension, it is better suited to respond to the new requirements.

The activity-based approach requires time-use survey data for analysis and estimation. A time-use survey entails the collection of data regarding all activities (in-home and out-of-home) pursued by individuals over the course of a day (or multiple days). Travel constitutes the medium for transporting oneself between spatially dislocated activity participations. The examination of both in-home and out-of-home activities facilitates an understanding of how individuals substitute out-of-home activities for in-home activities (or vice-versa) in response to changing travel conditions. This, in turn, translates to an understanding of when trips are generated or suppressed.

It is important to note that administrating time-use surveys is similar to administrating household travel surveys, except for collection of in-home as well as out-of-home activities. The information elicited from respondents is a little more extensive in time-use surveys compared to travel surveys, but experience suggests that the respondent burden or response rates are not significantly different between time-use and travel surveys (see Lawton and Pas, 1996 for an extensive discussion).

The activity-based approach does require more careful and extensive preparation of data to construct entire "sequences" of activities and travel. On the other hand, such intensive scrutiny of data helps identify data inconsistencies which might go unchecked in the trip-based approach (for example, there might be "gaps" in an individual's travel diary because of non-reporting of several trips; these will be identified during data preparation for activity analysis, but may not be identified in the trip-based approach since it highlights individual trips and not the sequence between trips and activities).

The rest of this chapter focuses on the activity-based approach to travel demand forecasting. The next section traces the history of research on activity analysis. Section 3.3 presents an overview of the modeling methods being used in activity-travel analysis. Section 3.4 discusses how activity-based travel research has been influencing travel demand analysis. Section 3.5 concludes the chapter by identifying important future research topics in the activity analysis area.

3.2 History of Research on Activity Analysis

The seminal works by Chapin (1971), Hagerstrand (1970) and Cullen and Godson (1975) form the basis for much of the research on activity analysis. Chapin (1971) proposed a motivational framework in which societal constraints and inherent individual motivations interact to shape revealed activity participation patterns. Hagerstrand (1970), on the other hand, emphasized the constraints imposed by the spatial distribution of opportunities for activity participation and temporal considerations on individual activity participation decisions, thus laying the foundation for what is now commonly referred to as the space-time "prism". Cullen and Godson (1975) argued that the spatial and temporal constraints identified by Hagerstrand are fundamentally characterized by varying degrees of rigidity (or flexibility). They undertook extensive empirical analysis to indicate that temporal constraints are more rigid than spatial constraints and that the rigidity of temporal constraints is closely related to activity type of participation (with more temporal rigidity associated with work-related activities compared to leisure activities).

Activity-based travel research has received much attention and seen considerable progress since these early studies. In the following review, we will use the term "activity episode" to refer to a discrete activity participation. The term "activity" refers to a collection of episodes of the same type or purpose over some time unit (say a day or a week). The review is undertaken in two categories. The first category focuses on participation decisions associated with a single activity episode. The second category examines individual decisions regarding activity episode patterns (that is, multiple activity episodes and their sequencing).

Single Activity Episode Participation

The studies in this section focus on the participation of individuals in single activity episodes, along with one or more accompanying characteristics of the episode such as duration, location, or time window of participation. The effect of household interdependencies on individual activity choice is represented in these models in the form of simple measures such as presence of working spouse, number of adults, and household structure.

Damm (1980) developed a multivariate daily model of participation and duration in out-of-home non-work activities (no distinction between activity types is made). He partitions the day into five periods based on the work schedule and introduces interdependence in activity participation and duration among time periods using variables which measure the "time spent in other periods". Temporal constraints are represented in the model in the form of variables like duration of work, flexibility of work hours and time spent in other periods. Spatial fixity of work place (an indicator of whether the individual has a fixed work location or not), accessibility, years lived at residence and presence of driver's license are defined to represent spatial constraints. Other socio-economic variables are included to represent the influence of lifecycle (eg., number of children), potential allocation process (eg., work status of spouse), and other familial responsibilities on individual activity participation.

Van der Hoorn (1983) developed an activity episode model for the choice of activity type and location of the episode. The three available locations in his analysis are "at home", "in town" and "outside town" (the term "town" representing the area of residence). Separate logit models are proposed for each previous location, each of five person groups, and for the workweek and weekend. Activity episode choice is regarded as being conditional on the location of the previous activity episode, but not on the activity type performed at the previous location. Location choice is dependent on the previous location and on the next activity episode. The restrictions imposed by external constraints and mandatory activities are taken into account while defining the choice set of available activities and locations.

Hirsh et al. (1986) developed a dynamic theory of weekly activity behavior and modified it suitably to model shopping activity in Israel. They recognized the benefit of studying activities on the basis of a weekly cycle rather than on a daily period. The attributes used in the model are similar to the ones used by Damm and van der Hoorn.

Mannering and his colleagues (Mannering *et al.,* 1994, Kim *et al.,* 1993) analyzed home-stay duration between successive participations in out-of-home activity episodes. Bhat (1996a) and Neimeier and Morita (1996), on the other hand, formulated and estimated models for the duration of out-of-home activity episodes. The results from these studies suggest that the socio-demographics of the individual's household and the individual (such as household size, income earnings, age, sex, *etc.*), and the work schedule characteristics of the individual, have a substantial effect on the duration of home-stay and out-of-home activity episodes. All the duration studies listed above use a hazard-based duration structure in their analyses.

Activity Episode Pattern Analysis

In this section, we review studies which examine activity episode patterns (*i.e.,* multiple activity episodes and their sequence). Some of these studies focus only on activity episode scheduling and consider the generation of activity episodes and their attributes as exogenous inputs. Such studies are reviewed in the next section. Other studies analyze both activity episode generation and scheduling, and these are reviewed in the subsequent section.

Activity Episode Scheduling A fundamental tenet of the activity episode scheduling approach to the analysis of activity/travel patterns is that travel decisions are driven by the collection of activities that form an agenda for individual participation. Travel patterns are viewed as arising from a more fundamental activity scheduling process. Activity scheduling is affected by spatial/temporal constraints of travel, specifications of precedence among activities, requirements to be with other family members at particular times and places (coupling constraints), and available individual transportation supply environment (the allocation of activities between household members that shapes the activity agenda of each individual and the allocation of household transportation supply between members is presumed to be exogenous in these studies).

Activity episode scheduling models generally take the structure of a computerized production system which comprises a set of rules in the form of condition-action pairs (see Newell and Simon, 1972). Studies in the psychology field suggest that a production system is consistent with the way in which humans perceive, appraise, and respond to spatial and aspatial information within the context of limited-information processing ability (Gärling *et al.,* 1994).

One of the earliest scheduling models was CARLA, developed by the Oxford University Transport Studies Unit (Clarke 1986). This model uses the list of activities to be scheduled and their durations to produce all feasible activity patterns in response to a change in the travel environment (for example, transit service improvements or cuts). It does so through the use of a branch-and-bound based combinatorial algorithm which reorganizes a given activity program and selects only those patterns which are feasible in terms of spatio-temporal and inter-personal constraints.

Recker et al. (1986a, 1986b) developed another scheduling model called STARCHILD. Their model partitions the daily scheduling process into two stages. In the first stage (also referred to as the pre-travel stage), the individual decides on a planned activity episode schedule based on a pre-determined directory of activities and their duration, location and time window for participation. STARCHILD models the selection of a planned activity program by generating distinct non-inferior patterns using combinatorics and then applying a logit choice model to establish the pattern choice with highest utility. The assignment of a utility value to each pattern is a function of the amount of time in the pattern associated with activity participation, wait time and travel time. The planned activity episode schedule is continuously revised and updated in the second dynamic scheduling stage circumstances or new activity demands. More recently, Recker (1995a) has extended the STARCHILD approach to include a mathematical programming formulation for the choice of an activity-travel pattern from several possible patterns.

Gärling *et al.* (1989) proposed yet another activity scheduling model labeled SCHEDULER. This computational model assumes the presence of a long term calendar (an agenda of activity episodes with duration, appointment details and preference) at the start of any time period. A small set of episodes with high priority are selected from this long term "calendar" and stored in a short term calendar as the subset of episodes to be executed in the short-run. This activity subset is sequenced, and activity locations determined based on a "distance-minimizing" heuristic procedure (see Axhausen and Gärling for a detailed review). SMASH (Ettema *et al.,* 1993) is a development of the SCHEDULER framework in which heuristic scheduling rules are specified and tested.

A more recent scheduling model is the adaptation simulation system labeled AMOS (for Activity-MObility Simulator) developed by Kitamura *et al.* (1996) to examine the short-term responses to Transportation Control Measures (TCMs). The model takes an observed daily activity-travel pattern of an individual (baseline pattern) and determines an adaption choice (for example, do nothing, change mode, change departure time, *etc.*) to a TCM using a response option generator.

Activity Episode Generation and Scheduling The studies reviewed in this section attempt to capture individual activity/travel patterns by focusing on the mechanism by which individual activities are generated and sequenced.

Kitamura (1983) studied episode sequencing and the tendencies or preferences in the formation of the set of activities to be pursued. A sequential history dependent approach is taken (sequential in that the probability of a given set of activities being chosen and pursued in a particular order is represented by a set of sequential and conditional probabilities). He found a consistent hierarchical order in sequencing episodes (with the less-flexible activities being pursued earlier). Kitamura and Kermanshah (1983) adopted the same sequential view in their extension of the above study to include the time dimension of activity choice. Adler and Ben Akiva (1979) examined inter-trip linkages from a simultaneous decision perspective, *i.e,* on the premise that the individual plans and pre-determines her/his daily travel schedule. The choice alternatives in this approach are entire daily patterns. However, the daily patterns are described by rather simple aggregate measures such as the mode used in travel and number of tours in the pattern. Golob (1986) also used a simultaneous decision approach, though his focus was on trip-chains or tours rather than daily patterns. The spatial and temporal dimensions are suppressed in this analysis. A set of different types of trip-chains are identified and modeled as dependent variables. A multivariate statistical technique (non-linear canonical correlation analysis) is employed for the analysis. Other studies of inter-trip linkage are Kitamura (1984), Nishii et al (1988) and O'Kelly & Miller (1984).

More recently, Ben-Akiva and Bowman (1994) have estimated a utility-based choice model of daily activity schedule of individuals that comprises a nested logit model of activity pattern choices (*i.e.,* purposes, priorities and structure of the day's activities and travel) and tour choices (mode choice, destination choice of stops in tours, and departure time from home and from the "primary" activity in tour). Similar efforts by Wen and Koppelman (1997, 1999) include generation and allocation of maintenance stops and automobiles to household members but excludes mode and destination choice. In contrast to the utility-maximizing discrete choice formulations of Ben-Akiva and Bowman and Wen and Koppelman, Vause (1997) proposes the use of a rule-based mechanism to restrict the number of activity-related choices available to an individual as well as for choice selection from the restricted choice set. Vause emphasizes the need to avoid the use of a single choice strategy in modeling and advances the use ofthe rule-based mechanism as a method to simulate different choice strategies (such as satisfaction, dominance, lexicographic and utility) within the same operational framework.

Vaughn, Speckman and Pas (Vaughn *et al,* 1997, and Speckman *et al,* 1997) developed a statistical approach to generate a set of baseline household activity patterns including the number and type of each activity episode and its duration, the number of home-based and work/school-based tours and start and end times for tours for a synthetic population represented by a continuous path through space and time. The statistical (as opposed to behavioral) basis of this approach raises questions about its use in prediction. However, it could provide initial travel-activity patterns for input to adaptive modeling systems such as SMASH and AMOS.

The studies ofepisode patterns discussed thus far either do not model the temporal dimension ofepisodes or assume broad time periods in the analysis. More recently, two approaches have been proposed to model activity episode generation and scheduling within the context of a continuous time domain. The first is the Prism-Constrained Activity-Travel Simulator proposed by Kitamura and Fujii, 1998 and the other is the Comprehensive Activity-Travel Generation for Workers (CATGW) model system proposed by Bhat and Singh (1999). These two studies are discussed in section 4.2 under the heading of "Emergence of Comprehensive Activity-based Travel Demand Models".

3.3 Modeling Methods in Activity-Travel Analysis

The methods used in activity-based travel analysis include discrete choice models as well as other methods that accommodate non-discrete variables in activity modeling. The latter methods have emerged more recently because ofthe need to model travel as part of a larger (and holistic) activity-travel pattern and involve relatively non-traditional (in the travel analysis field) methodologies such as duration analysis and limited-dependent variable models. In this section, we discuss these various methods. The material here is drawn liberally from Bhat (1997a), though in a substantially condensed form.

Discrete Choice Models

The multinomial logit (MNL) model has been the most widely used structure for modeling discrete choices in travel behavior analysis. The random components ofthe utilities ofthe different alternatives in the MNL model are assumed to be independent and identically distributed (IID) with a type I extreme-value (or Gumbel) distribution (McFadden, 1973). The MNL model also maintains an assumption of homogeneity in responsiveness to attributes of alternatives across individuals (*i.e.,* an assumption of response homogeneity). Finally, the MNL model also maintains an assumption that the error variance-covariance structure of the alternatives is identical across individuals (*i.e.,* an assumption of error variance-covariance homogeneity). The three assumptions together lead to the simple and elegant closed-form mathematical structure of the MNL. However, these assumptions also leave the MNL model saddled with the "independence of irrelevant alternatives" (IIA) property at the individual level (Ben-Akiva and Lerman, 1985). In the next three sections, we will discuss generalizations of the MNL structure along each of the three dimensions mentioned above: a) Relaxation of the IID (across alternatives) error structure, b) Relaxation of response homogeneity, and c) Relaxation of the error variance-covariance structure homogeneity. While we discuss each of the dimensions separately, one can combine extensions across different dimensions to formulate several more generalized and richer structures.

Relaxation of the IID (Across Alternatives) Error Structure The rigid inter-alternative substitution pattern of the multinomial logit model can be relaxed by removing, fully or partially, the IID assumption on the random components of the utilities of the different alternatives. The IID assumption can be relaxed in one of three ways: a) allowing the random components to be correlated while maintaining the assumption that they are identically distributed (identical, but non-independent random components), b) allowing the random components to be non-identically distributed (different variances), but maintaining the independence assumption (non-identical, but independent random components), and c) allowing the random components to be non-identical and non-independent (non-identical, non-independent random components). Each of these alternatives is discussed below.

Identical, Non-Independent Random Components The distribution of the random components in models which use identical, non-independent random components can be specified to be either normal or type I extreme value. Discrete choice literature has mostly used the type I extreme value distribution since it nests the multinomial logit and results in closed-form expressions for the choice probabilities.

The models with the type I extreme value error distribution belong to the Generalized Extreme Value (GEV) class of random utility-maximizing models. Five model structures have been formulated and applied within the GEV class. These are: the Nested Logit (NL) model, the Paired Combinatorial Logit (PCL) model, the cross-nested logit (CNL) model, the Ordered GEV (OGEV) model, and the Multinomial Logit-Ordered GEV (MNL-OGEV) model.

The nested logit (NL) model permits covariance in random components among subsets (or nests) of alternatives (each alternative can be assigned to one and only one nest). Alternatives in a nest exhibit an identical degree of increased sensitivity relative to alternatives not in the nest (Williams, 1977, Daly and Zachary, 1978, Daganzo and Kusnic, 1993).

The paired combinatorial logit (PCL) model initially proposed by Chu (1989) and recently examined in detail by Koppelman and Wen (1996) generalizes, in concept, the nested logit model by allowing differential correlation between each pair of alternatives. While the nested logit model is not nested within the PCL structure, an appropriate constrained PCL closely approximates the nested logit model.

Another generalization of the nested logit model is the cross-nested logit (CNL) model of Vovsha (1996). In this model, an alternative need not be exclusively assigned to one nest as in the nested logit structure. Instead, each alternative can be probabilistically assigned to multiple nests. Vovsha proposes a heuristic procedure for estimation of the CNL model.

The ordered GEV model was developed by Small (1987) to accommodate correlation among the unobserved random utility components of alternatives close together along a natural ordering implied by the choice variable (examples of such ordered choice variables might include car ownership, departure time of trips, *etc.*).

The MNL-OGEV model formulated by Bhat (1998b) generalizes the nested logit model by allowing adjacent alternatives within a nest to be correlated in their unobserved components.

The advantage of all the GEV models discussed above is that they allow partial relaxations of the independence assumption among alternative error terms while maintaining closed-form expressions for the choice probabilities. The problem with these models is that they are consistent with utility maximization only under rather strict (and often empirically violated) restrictions on the dissimilarity parameters. The origin of these restrictions can be traced back to the requirement that the variance of the joint alternatives be identical.

Non-Identical, Independently Distributed Random Components The concept that heteroscedasticity in alternative error terms (i.e., independent, but not identically distributed error terms) relaxes the IIA assumption is not new (see Daganzo, 1979), but has received little (if any) attention in travel demand modeling and other fields. Four models have been proposed which allow non-identical random components. The first is the negative exponential model of Daganzo (1979), the second is the heteroscedastic multinomial logit (HMNL) model of Swait and Stacey (1996), the third is the oddball alternative model of Recker (1995b) and the fourth is the heteroscedastic extreme-value (HEV) model of Bhat (1995).

Daganzo (1979) used independent negative exponential distributions with different variances for the random error components to develop a closed-form discrete choice model which does not have the IIA property. His model has not seen much application since it requires that the perceived utility of any alternative not exceed an upper bound.

Swait and Stacey (1996) allowed heteroscedasticity by specifying the variance of the alternative error terms to be functions of observed alternative characteristics. The error terms themselves are assumed to be type I extreme-value. The scale parameter θ_i characterizing the variance of each alternative i is written as $\theta_i = \exp(\beta' z_i)$, where z_i

is a vector of attributes associated with alternative i and β is a corresponding vector of parameters to be estimated. The resulting model has a closed-form structure.

Recker (1995b) proposed the oddball alternative model which permits the random utility variance of one "oddball" alternative to be larger than the random utility variances of other alternatives. This situation might occur because of attributes which define the utility of the oddball alternative, but are undefined for other alternatives. Then, random variation in the attributes that are defined only for the oddball alternative will generate increased variance in the overall random component of the oddball alternative relative to others.

Bhat (1995) formulated the heteroscedastic extreme-value (HEV) model which assumes that the alternative error terms are distributed with a type I extreme value distribution. The variance of the alternative error terms are allowed to be different across all alternatives (with the normalization that the error terms of one of the alternatives has a scale parameter of one for identification). Bhat develops an efficient Gauss-Laguerre quadrature technique to approximate the one-dimensional integral in the choice probabilities of the HEV model. The reader is referred to Hensher (1998a; 1998b) and Hensher *et al.* (1999) for applications of the HEV model to estimation from revealed and stated preference data.

The advantage of the heteroscedastic class of models discussed above is that they allow a flexible cross-elasticity structure among alternatives than many of the GEV models discussed earlier. Specifically, the models (except the oddball model) permit differential cross-elasticities among all pairs of alternatives. The limitation (relative to the GEV models) is that the choice probabilities do not have a closed-form analytical expression in the HEV model.

Non-Identical, Non-Independent Random Components Models with non-identical, non-independent random components use one of two general structures: the first is an error-components structure and the second is the general multinomial probit (MNP) structure.

The error-components structure partitions the overall error into two components: one component which allows the random components to be non-identical and non-independent, and the other component which is specified to be independent and identically distributed across alternatives. In particular, consider the following utility function for alternative i:

$$\begin{aligned} U_i &= V_i + \zeta_i \\ &= V_i + \mu' z_i + \epsilon_i \end{aligned} \qquad (1)$$

where V_i and ζ_i are the systematic and random components of utility, and ζ_i is further partitioned into two components, $\mu' z_i$ and ϵ_i. z_i is a vector of observed data associated with alternative i, μ is a random vector with zero mean and density $g(\mu \mid \Sigma)$, Σ is the variance-covariance matrix of the vector μ, and ϵ_i is independently and identically standard distributed across alternatives with density function f(.). The component $\mu' z_i$ induces heteroscedasticity and correlation across unobserved utility components of the alternatives (see Train, 1995). While different distributional assumptions might be made regarding f(.) and g(.), it is typical to assume a standard type I extreme value for f(.), and a normal distribution for g(.). This results in a error-components model with a logit kernel. On the other hand, if a standard normal distribution is used for f(.), the result is a error-components probit model. Both these structures will involve integrals in the choice probability expressions which do not have a closed-form solution. The estimation of these models is achieved using logit simulators (in the first case) or probit simulators (in the second case). Different and very general patterns of heteroscedasticity and correlation in unobserved components among alternatives can be generated by appropriate specification of the μ and z_i vectors (see Bhat, 1998c, Ben-Akiva and Bolduc, 1996 and Brownstone and Train, 1999).

The general multinomial probit (MNP) structure does not partition the error terms, and estimates (subject to certain identification considerations) the variance-covariance matrix of the overall random components among the different alternatives (see Bunch and Kitamura, 1990; Lam, 1991; and Lam and Mahmassani, 1991). However, McFadden and Train (1996) have shown that the error-components formulation can approximate a multinomial probit formulation as closely as one needs it to. Further, the error-components models can be estimated using simulators which are conceptually simple, easy to program and inherently faster than simulators for the MNP model (see Brownstone and Train, 1999).

Relaxation of Response Homogeneity The standard multinomial logit, and other models which relax the IID assumption across alternatives, typically assume that the response parameters determining the sensitivity to attributes of the alternatives are the same across individuals in the population. However, if such an assumption is imposed when there is response heterogeneity, the result is biased and inconsistent parameter and choice probability estimates (see Chamberlain, 1980).

Response heterogeneity may be accommodated in one of two ways. In the first approach, the varying coefficients approach, the coefficients on alternative attributes are allowed to vary across individuals while maintaining a single utility function. In the second approach, the segmentation approach, individuals are assigned to segments based on their personal/trip characteristics, and a separate utility function is estimated for each segment. Each of these approaches is discussed next.

Varying Coefficients Approach Consider the utility U_{qi} that an individual q associates with alternative i and write it as:

$$U_{qi} = \alpha_i + \delta_i' z_q + \epsilon_{qi} + \eta_q' x_{qi} \tag{2}$$

where α_i is an individual-invariant bias constant, z_q is a vector of observed individual characteristics, δ_i is a vector of parameters to be estimated, ϵ_{qi} is a random term representing idiosyncracies in preferences, and η_q is a vector representing the responsiveness of individual q to a corresponding vector of alternative-associated variables x_{qi}.The ϵ_{qi} terms may be specified to have any of the structures discussed in Section 3.2. Conditional on η_q and the assumption regarding the ϵ_{qi} terms, the form of the conditional choice probabilities can be developed. The unconditional choice probabilities corresponding to the conditional choice probabilities will depend on the response heterogeneity specification adopted for the η_q vector. A general heterogeneity specification involves allowing each element η_{qk} of the vector η_q to vary across individuals based on observed as well as unobserved individual characteristics: $\eta_{qk} = \pm\exp(\gamma_k + \beta_k' w_{qk} + v_{qk})$, where w_{qk} is a vector of relevant observed individual characteristics and v_{qk} is a term representing random taste variation across individuals with the same observed characteristics w_{qk}. The exponential form is used to ensure the appropriate sign on the response coefficients: a '+' sign is applied for a non-negative response coefficient and the '-' sign is applied for a non-positive response coefficient. v_{qk} is typically assumed to be normally distributed. The random response specification does not exhibit the restrictive independence from irrelevant alternatives (IIA) property even if the IID error assumption across alternatives of the MNL is maintained (see Bhat, 1998d).

Segmentation Approaches Two segmentation approaches may be identified depending on whether the assignment of individuals to segments is exogenous (deterministic) or endogenous (probabilistic).

The exogenous segmentation approach to capturing heterogeneity assumes the existence of a fixed, finite number of mutually-exclusive market segments (each individual can belong to one and only one segment). The segmentation is based on one or two key socio-demographic variables (sex, income, *etc.*). Within each segment, all individuals are assumed to have *identical* preferences and identical sensitivities to level-of-service variables (*i.e.,* the same utility function). Typically, very few (one or two) demographic variables are used for segmentation. The advantage of the exogenous segmentation approach is that it is easy to implement. The disadvantage is that its practicality comes at the expense of suppressing potentially higher-order interaction effects of the segmentation variables on response to alternative attributes.

The endogenous market segmentation approach attempts to accommodate heterogeneity in a practical manner not by suppressing higher-order interaction effects of segmentation variables (on response to alternative attributes), but by reducing the dimensionality of the segment-space. Each segment, however, is allowed to be characterized by a large number of segmentation variables. Individuals are assigned to segments in a probabilistic fashion based on the segmentation variables. Since this approach identifies segments without requiring a multi-way partition of data as in the exogenous market segmentation method, it allows the use of many segmentation variables in practice and, therefore, facilitates incorporation of the full order of interaction effects of the segmentation variables on preference and sensitivity to alternative attributes (see Bhat, 1997b and Gopinath and Ben-Akiva, 1995).

Relaxation of Error Variance-Covariance Structure Homogeneity The assumption of error variance-covariance structure homogeneity across individuals can be relaxed either by a) allowing the variance components to vary across individuals (variance relaxation), b) allowing the covariance components to vary across individuals (covariance relaxation), or c) allowing both variance and covariance components to vary across individuals (variance-covariance relaxation).

Variance Relaxation Swait and Adamowicz (1996) formulate a heteroscedastic multinomial logit (HMNL) model that allows the variance of alternatives to vary across individuals based on attributes characterizing the individual and her/his environment (the variance, however, does not vary across alternatives). The motivation for such a model is that individuals with the same deterministic utility for an alternative may have different abilities to accurately perceive the overall utility offered by the alternative. The HMNL model has exactly the same structure as the heteroscedastic model described earlier in this section, though the motivations for their development are different. McMillen (1995) also proposes a heteroscedastic model in the context of spatial choice and Gliebe *et al* (1998) incorporated heteroscedastic scaling into the PCL model for stochastic route choice.

Covariance Relaxation Bhat (1997c) develops a nested logit model that allows heterogeneity across individuals in the magnitude of covariance among alternatives in a nest. The heterogeneity is incorporated by specifying the logsum (dissimilarity) parameter(s) in the nested logit model to be a deterministic function of individual-related characteristics. The model is applied to intercity mode choice analysis, where such heterogeneity may be likely to occur.

The author is not aware of any study that allows both variance and covariance components to vary across individuals (variance-covariance relaxation), though in concept the extension involves combining the variance and covariance relaxations discussed earlier.

Hazard Duration Models

Hazard-based duration models are ideally suited to modeling duration data. Such models focus on an end-of-duration occurrence (such as end of shopping activity participation) given that the duration has lasted to some specified time (Hensher and Mannering, 1994). This concept of conditional probability of "failure" or termination of activity duration recognizes the dynamics of duration; that is, it recognizes that the likelihood of ending a shopping activity participation depends on the length of elapsed time since start of the activity.

Hazard-based duration models are being increasingly used to model duration time in activity analysis. To include an examination of covariates which affect duration time, most studies use a proportional hazard model which operates on the assumption that covariates act multiplicatively on some underlying or baseline hazard.

Two important methodological issues in the proportional hazard model are a) the distributional assumptions regarding duration (equivalently, the distributional assumptions regarding the baseline hazard) and b) the assumptions about unobserved heterogeneity (*i.e.,* unobserved differences in duration across people). We discuss each of these issues in next two sections. A comprehensive review of the extension of the simple univariate duration model to include multiple duration processes, multiple spells from the same individual, and related issues may be found in Bhat (1997a).

Baseline Hazard Distribution The distribution of the hazard may be assumed to be one of many parametric forms or may be assumed to be nonparametric. Common parametric forms include the exponential, Weibull, log-logistic, gamma, and log-normal distributions. Different parametric forms imply different assumptions regarding duration dependence. For example, the *exponential* distribution implies no duration dependence; that is, the time to "failure" is not related to the time elapsed. The *Weibull* distribution generalizes the exponential distribution and allows for monotonically increasing or decreasing duration dependence. The form of the duration dependence is based on a parameter that indicates whether there is positive duration dependence (implying that the longer the time has elapsed since start of the duration, the more likely it is to exit the duration soon), negative duration dependence (implying that the longer the time has elapsed since start of the duration, the less likely it is to exit the duration soon), or no duration dependence (which is the exponential case). The *log-logistic* distribution allows a non-monotonic hazard function.

The choice of the distributional form for the hazard function may be made on theoretical grounds. However, a serious problem with the parametric approach is that it inconsistently estimates the baseline hazard and the covariate effects when the assumed parametric form is incorrect (Meyer, 1990). The advantage of using a nonparametric form is that even when a particular parametric form is appropriate, the resulting estimates are consistent and the loss of efficiency (resulting from disregarding information about the hazard's distribution) may not be substantial.

Most studies of duration to date have made an *a priori* assumption of a parametric hazard. The most relevant duration studies for activity-travel modeling include a) the homestay duration models for commuters (*i.e.,* the time between coming home from work and leaving home for another out-of-home activity participation) of Mannering *et al.* (1992) and Hamed and Mannering (1993), b) the sex-differentiated shopping duration models of Niemeier and Morita (1996), c) the shopping activity duration during the evening work-to-home commute of Bhat (1996a), and d) the delay duration model for border crossings by Paselk and Mannering (1993). These studies have been reviewed in greater detail by Pas (1997).

Unobserved Heterogeneity Unobserved heterogeneity arises when unobserved factors (*i.e.,* those not captured by the covariate effects) influence durations. It is well-established now that failure to control for unobserved heterogeneity can produce severe bias in the nature of duration dependence and the estimates of the covariate effects (Heckman and Singer, 1984).

The standard procedure used to control for unobserved heterogeneity is the random effects estimator. This involves specification of a distribution for the unobserved heterogeneity (across individuals) in the population. Two general approaches may be used to specify the distribution of unobserved heterogeneity. One approach is to use a parametric distribution such as a gamma distribution or a normal distribution (most earlier research has used a gamma distribution). The problem with the parametric approach is that there is seldom any justification for choosing a particular distribution; further, the consequence of a choice of an incorrect distribution on the consistency of the model estimates can be severe (see Heckman and Singer, 1984). A second approach to specifying the distribution of unobserved heterogeneity is to use a nonparametric representation for the distribution and to estimate the distribution empirically from the data. This is achieved by approximating the underlying unknown heterogeneity distribution by a finite number of support points and estimating the location and associated probability masses of these support points. The nonparametric approach enables consistent estimation since it does not impose a prior probability distribution.

Application of duration models in the transportation field have, for the most part, ignored unobserved heterogeneity (but see Bhat, 1996a and Hensher, 1994).

Limited-Dependent Variable Models

Limited-dependent variable models encompass a wide variety of structures. In this section, we will focus on inter-related discrete and non-discrete variable systems. The non-discrete variable can take several forms. However, the three most interesting cases in the context of travel and activity modeling are the continuous, ordinal, and grouped forms. Further, the structure for the discrete/ordinal and discrete/grouped variable systems are very similar; so we will examine limited-dependent variable systems under two headings: discrete/continuous and discrete/ordinal models.

Discrete/Continuous Models Hamed and Mannering (1993) use the discrete/continuous model framework to model activity type choice, travel time duration to the activity, and activity duration. Barnard and Hensher (1992) estimate a discrete/continuous model of shopping destination choice and retail expenditure. They use Lee's (1983) transformation method for polychotomous choice situations with non-normal error distributions in the choice model. Bhat (1998e) has also used Lee's method for discrete/continuous models, but extends the method to jointly estimate a polychotomous discrete choice and two continuous choices.

Discrete/Ordinal Models Bhat and Koppelman (1993) estimate a discrete/grouped system of employment status (represented by a binary flag indicating whether or not an individual is employed) and annual income earnings. Observed income earnings in their data is in grouped form (*i.e.,* observed only in grouped categories such as < 20,000, 20,000-39,999, 40,000-59,999, *etc.*). Since it is likely that people who are employed are also likely to be the people who can earn higher incomes, the two variables are modeled jointly.

Bhat (1997d) has recently developed a joint model of polychotomous work mode choice and number of non-work activity stops during the work commute (*i.e.,* the total number of non-work stops made during the morning home-to-work commute and evening work-to-home commute). The joint model provides an improved basis to evaluate the effect on peak-period traffic congestion of conventional policy measures such as ridesharing improvements and solo-auto use dis-incentives.

3.4 Results of Activity-Travel Analysis

The substantial literature on activity-travel studies precludes a discussion of the results of individual studies. Instead, in this section, we discuss how activity-based travel research has and is influencing travel demand modeling.

Better Specification of Travel Demand Models

The insights obtained from activity-based research has enabled the incorporation of measures of complex behavior in a simple, albeit valuable way in travel choice models. Beggan (1988) used simple descriptors of travel-activity behavior such as the number of stops made during the work tour and the number of tours made during the work day as independent variables and found that even these simplified descriptors had a significant influence on mode-choice to work. Damm (1980) used various descriptors of lifecycle, temporal constraints, spatial constraints, interaction between time periods and interaction between household members in a nested logit model to estimate the participation and duration in discretionary activities. Goulias *et al.* (1989), Bhat *et al.* (1999) and Felendorf *et al.* (1997) recognize the inter-relationships among home-based and non-home based trips in a sojourn from home or from work and develop methods that can be used not only to generate trips but also to determine their placement within the larger daily activity-travel pattern of individuals. Purvis and his colleagues (Purvis *et al.,* 1996) at the Metropolitan Transportation Commission (MTC) of the San Francisco Bay area introduced the notion of time constraints by using work travel time as an explanatory variable in their traditional non-work trip generation model.

Clearly, one way that activity-based research is influencing (and has influenced) travel demand modeling is through incremental improvements to trip-based planning methods.

Emergence of comprehensive Activity-Based Travel Forecasting Models

As indicated earlier in the section on activity episode generation and scheduling, two approaches have been recently proposed to model the entire diary activity-travel pattern of individuals within the context of a continuous time domain. The first is the Prism-Constrained Activity-Travel Simulator proposed by Kitamura and Fujii, 1998 and the other is the Comprehensive Activity-Travel Generation for Workers (CATGW) model system proposed by Bhat and Singh (1999).

PCATS divides the day (or any other unit of time) into two types of periods: "open" periods and "blocked" periods. "Open" periods represent times of day when an individual has the option of traveling and engaging in "flexible" activities. "Blocked" periods represent times when an individual is committed to performing "fixed" activities. PCATS then attempts to "fill" the open periods based on a space-time prism of activities contained within the open period. PCATS uses a sequential structure for generation of the activity episodes and associated attributes (activity type, activity duration, activity location, and mode choice) within the "open" period (thus, the unit of analysis in PCATS is the individual activity).

The CATGW framework is based on the fixity of two temporal points in a worker's continuous daily time domain. The two fixed points correspond to the arrival time of an individual at work and the departure time of an individual from work. The day is divided into four different patterns: before morning commute pattern, work commute pattern, midday pattern, and post home-arrival pattern. Within each of the before work, midday and post home-arrival patterns, several tours may be present. A tour is a circuit that begins at home and ends at home for the before work and post home-arrival patterns and is a circuit that begins at work and ends at work for the midday pattern. Further, each tour within the before work, midday and post home-arrival patterns may comprise several activity episodes. Similarly, the morning commute and evening commute components of the work commute pattern may also comprise several activity episodes. The modeling representation for the entire daily activity-travel pattern is based on a descriptive analysis of actual survey data from two metropolitan areas in the U.S. The suite of models in the modeling representation can be used for generation of synthetic baseline patterns as well as to evaluate the effect of Transportation Control Measures (TCMs). The models have been applied to evaluate the potential effect of TCMs on stop-making and cold starts in the Boston Metropolitan area.

Study of Important Policy Issues

The study of policy issues is improved and/or made possible by the activity-based approach. Demand management strategies that attempt to suppress or spread traffic peaks need to be designed based on the effect of these measures on re-scheduling of activities and household interactions. For example, a change in work schedule to an early departure from work may lead to increased trip-making at the evening because of the additional time available to participate in out-of-home activities. If some of this travel is undertaken during the same time as the PM peak-period travel, the extent of congestion alleviation projected by traditional models will not be realized (see Jones et al., 1990). In fact, from an air quality standpoint, Bhat (1998a) illustrates that an early departure from work would lead to more cold starts because of the increased activity durations of evening commute stops resulting from more time availability. Similarly, improvements in high-occupancy vehicle modes or peak period pricing measures are likely to have a rather small impact on the mode choice of individuals who make stops during the commute. The activity-based approach would recognize this association, while traditional mode choice models will overestimate the shift to high-occupancy modes, as clearly demonstrated by Bhat (1997d) using actual empirical data.

Another example of the advantage of activity-based analysis relative to traditional methods is in the evaluation of the travel impacts of telecommuting. Specifically, displacements of travel (and its associated consequences) to other times of day due to a change in activity patterns caused by adoption of work telecommuting strategies cannot be examined by the narrow trip-based models, but can be examined using activity-based models (see Mokhtarian, 1993).

Improvements in Data Collection Procedures

Activity research has and continues to provide insights into cost-effective methods of collecting data and improving the accuracy of data collection procedures. It also facilitates the development of new data collection techniques that are responsive to current needs. Improvements in the accuracy ofconventional data collection procedures due to activity-based research include the employment of a verbal activity recall framework, stated preference techniques, multi-day surveys, longitudinal data collection, pattern reconstruction techniques, and interactive measurement and gaming simulation techniques (see Lawton and Pas, 1996, for a comprehensive resource paper on survey methods associated with activity analysis).

Contributions to Regional and Community Planning

Models with a sound behavioral casual linkage between individual activity patterns and the travel environment will be critical to good regional and community planning. The activity perspective of travel provides a clear picture of the functioning of urban areas (for example, the spatial characteristics of intra-urban labor markets) and has the potential to identify the differential quality of life associated with different segments of the population. For example, some researchers (see Johnston-Anumonwo, 1995; Hanson and Pratt, 1988,1992; Preston *et al.*, 1993; and MacDonald and Peter, 1994) have used the activity analysis framework to study the social and spatial context of information exchange with regard to employment-related decisions. Ferguson and Jones (1990), on the other hand, used the activity-based perspective to identify the special needs of the elderly and disabled in Adelaide and were able to make specific recommendations to improve the mobility ofthese population groups by identifying the rhythms and timing under which such individuals live.

3.5 Future directions in Activity-Based Travel Research

The review of activity-based studies in section 3.2 indicates the substantial progress that has been made in recent years. There is no question that there is an increasing realization and awareness of the need to model travel as part of a holistic (and temporally continuous) activity-travel pattern. However, there is still a long way to go in understanding how households and individuals make choices that drive their activity and travel patterns. The objective of this section is to highlight some of the directions that we consider important in activity-based travel analysis.

Inter-Individual Interactions in Activity Behavior

An area that has received limited attention thus far in the activity analysis literature is the interactions among individuals in a household and the effect of such interactions on individual activity episode patterns. Interactions among individuals might take the form of joint participation in certain activities (such as shopping together or engaging in recreational/social activities together), "serve-passenger" and "escort" activities where one individual facilitates and oversees the participation of another in activities (for example, the "soccer mom" phenomenon), and allocation of autos and activities among individuals (especially in multi-adult, one-car households). Such interactions can lead to constraints that may be very important in individual activity/travel responses to changes in the transportation or land-use environment. However, the comprehensive activity analysis frameworks today that model individual activity patterns within a continuous time domain (such as those discussed in section 3.2) do not consider inter-individual interactions. On the other hand, some recent efforts (for example, see Wen and Koppelman, 1999) have focused on inter-individual interactions in activity decisions but have not examined individual activity-travel patterns at a fine level of temporal resolution. Integration of efforts which accommodate inter-individual interactions in activity patterns with efforts that use a continuous time domain is, therefore, likely to be a very fruitful area for further research.

Time-Space Interactions in Activity Behavior

Another area that needs substantial attention in the future is the explicit accommodation of time and space interactions. Most early research in the activity analysis area emphasized the dependence in spatial choices among activities using either semi-Markov processes or discrete-choice models (Horowitz, 1980; Kitamura, 1984; O'Kelly and Miller, 1984; Lerman, 1979). These studies ignored the temporal aspects of activity participation. More recently, some studies have focused on the timing and duration of activities (Ettema *et al.,* 1995; Hamed and Mannering, 1993; Bhat, 1996a). But these studies have not examined spatial issues. Thus, though one of the key concepts of the activity-based approach is the time-space interaction, little work has been done toward developing such an integrated modeling approach. Thill and Thomas, 1987, indicated the following in their review of travel behavior research: *"In spite of various devices to account for links between decisions, no study has thus far appropriately restored the simultaneity of intended choices....It is necessary to conceive a framework that combines both temporal and spatial aspects of travel choice and that considers multipurpose multistop behavior as a multidimensional whole".* This statement remains valid even today. Recent work by Thill and Horowitz (1997a,b), Dijst and Vidakovic (1997), and Bhat (1998a) starts to address this concern, but there is still much work to be done in this area.

In-Home and Out-of-Home Activity Substitution

In-home and out-of-home activities have quite diiferent implications for travel; an in-home episode does not involve travel (for a person already at home), while an out-of-home episode requires travel. Thus, the in-home/out-of-home participation decision has an impact on the generation of trips (see Jones *et al.,* 1993). Understanding this substitution is important, particularly at a time when opportunities for entertainment at home are increasing because of the increasing accessibility of households to computers, theater quality audio and video systems, and an almost unlimited choices of movies to view from home. Despite the importance of understanding in-home and out-of-home substitution effects, very few studies have examined this issue (see Kitamura *et al,* 1996, Kraan, 1996 and Bhat, 1998e). And even these studies have examined substitution only in the context of broad activity types (such as discretionary activities, maintenance activities, *etc.*) rather than the more relevant substitution in specific activity types.

One of the impediments to a detailed analysis of in-home and out-of-home substitution has been (until recently) the unavailability of data on in-home activities. From a data collection standpoint, a related complication is the participation of individuals in multiple activities at the same time at home (for example, eating and watching television at the same time). Thus, research is required into how we might collect detailed data on activity type of participation at home and how we might be able to elicit information on multi-activity participation.

Unit of Analysis

The unit of analysis typically used in the activity-based travel models is the weekday. The implicit assumption is that there is little variation in activity-travel patterns across different days of the week. Research focusing even on simple aggregate measures of activity-travel behavior (such as trip frequency, and number and type of stops made during the morning/evening commutes) has indicated quite substantial intrapersonal variability across weekdays (see Pas and Koppelman, 1986; Jou and Mahmassani, 1997). One can therefore expect substantial day-to-day variations when considering entire activity-travel patterns. In addition, the focus on a single weekday does not allow the examination of the interaction in activity participation between weekends and weekdays. Of course, the use of an entire week as the unit of analysis will require the collection of time-use diary data over at least one week. This offers research opportunities for the development of data techniques that can collect time-use data over a week without being prohibitively expensive or appearing excessively intrusive.

The Decision Mechanism

As described earlier in the paper, there have been several previous modeling efforts to generate activity episode patterns. However, we still lack a good understanding of the decision mechanism underlying revealed activity episode patterns. For example, how do households and individuals acquire and assimilate information about their activity/travel environment, is activity-travel behavior pre-planned or is it subject to dynamic adjustment or is there a mixture of these processes, are attributes of activity episodes determined jointly or sequentially, and what objective do individuals follow while determining their scheduling decisions? The main challenge to studying these issues is that the generation and scheduling process that determines the revealed episode patterns can only be understood if additional data on the internal mechanism leading up to revealed episode patterns is collected. Such data are not currently available and again this offers another research opportunity in the area of data collection.

Clearly, there are important theoretical and methodological advances still to be made in the activity-based travel research field. As progress is made on these fronts, we are bound to see more applications of the activity paradigm in travel demand modeling. Some metropolitan planning organizations (MPOs) are already embracing this new paradigm and pursuing efforts to develop comprehensive activity model systems to replace the traditional four-step trip-based methods. Many other MPOs realize the need to switch to an activity-based modeling system in the near future. To conclude, the activity-based approach to travel demand modeling is slowly, but steadily, finding its way into actual practice.

Acknowledgements

This research was supported by National Science Foundation grants DMS 9208758 and DMS 9313013 to the National Institute of Statistical Sciences (NISS).

References

Adler, T. and M. Ben-Akiva (1979) Atheoretical and empirical model of trip-chaining behavior, *Transportation Research,* 13B, 243-257.

Axhausen, K. and T. Garling (1992) Activity-based approaches to travel analysis: conceptual frameworks, models and research problems, *Transport Reviews,* 12, 324-341.

Barnard and Hensher (1992) The spatial distribution of retail expenditures, *Journal of Transport Economics and Policy,* September, 299-311.

Beggan, J.G. (1988) The relationship between travel/activity behavior and mode choice for the work trip, unpublished Master's Thesis, Northwestern University.

Ben-Akiva, M. and S.R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand,* The MIT Press, Cambridge.

Ben-Akiva, M. and D. Bolduc (1996) Multinomial probit with a logit kernel and a general parametric specification of the covariance structure, working paper, Department of Civil and Environmental engineering, Massachusetts Institute of Technology, Cambridge, MA and Département d'économique, Université Laval, Sainte-Foy, Qc, Canada.

Ben-Akiva, M.E. and J.L. Bowman (1995) Activity-based disaggregate travel demand model system with daily activity schedules, Paper presented at the EIRASS Conference on Activity-Based Approaches: Activity Scheduling and the Analysis of Activity patterns, Eindhoven, The Netherlands.

Bhat, C.R. and F.S. Koppelman (1993) An endogenous switching simultaneous equation system of employment, income, and car ownership, *Transportation Research,* 27A, 6, 447-459.

Bhat, C.R. (1995) A heteroscedastic extreme-value model of intercity mode choice, *Transportation Research,* 29B, 6, 471-483.

Bhat, C.R. (1996a) A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogenetty, *Transporation Research*, 30B, 3, 189-207.

Bhat, C.R. (1996b) Incorporating observed and unobserved heterogeneity in work mode choice modeling, *forthcoming* in *Transportation Science.*

Bhat, C.R. (1997a) Recent methodological advances relevant to activity and travel behavior analysis, invitational resource paper prepared for presentation at the International Association of Travel Behavior Research Conference to be held in Austin, Texas, September 1997.

Bhat, C.R. (1997b) An endogenous segmentation mode choice model with an application to intercity travel, *Transportation Science,* 31, 34-48.

Bhat, C.R. (1997c) A nested logit model with covariance heterogeneity, *Transportation Research,* 31B, 11 -21.

Bhat, C.R. (1997d) Work mode choice and number of non-work commute stops, *Transportation Research,* 31 B, 41-54.

Bhat, C.R. (1998a) Modeling the commute activity-travel pattern of workers: formulation and empirical analysis, Technical Paper, Department of Civil Engineering, University of Texas at Austin.

Bhat, C.R. (1998b) An analysis of travel mode and departure time choice for urban shopping trips, *Transportation Research,* 32B, 387-400.

Bhat, C.R. (1998c) Accommodating flexible substitution patterns in multidimensional choice modeling: formulation and application to travel mode and departure time choice, *Transportation Research,* 32B, 425-440.

Bhat, C.R. (1998d) Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling, *Transportation Research,* 32A, 495-507.

Bhat, C.R. (1998e) A post-home arrival model of activity participation behavior, *Transportation Research,* 32B, 361-371.

Bhat, C.R. and S.K. Singh (1999) A comprehensive daily activity-travel generation model system for workers, *forthcoming, Transportation Research.*

Bhat, C.R. and F.S. Koppelman (1998) A retrospective and prospective survey of time-use research, under consideration for possible publication in a special time-use issue in *Transportation.*

Bhat, C.R., J.P. Carini and R. Misra (1999) On the generation and organization of household activity stops, *forthcoming,* Transportation Research Record.

Brownstone, D. and K. Train (1999) Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics,* Vol. 89, pp. 109-129.

Bunch, D.S. and R. Kitamura (1990) Multinomial probit model estimation revisited: testing estimable model specifications, maximum likelihood algorithms, and probit integral approximations for trinomial models of household car ownership, UCD-ITS-RP-90-1, Institute of Transportation Studies, University of California, Cambridge Systematics, Inc. (1994) *Short-Term Travel Model Improvements,* Final report (DOT-95-05), Prepared for the U.S. Department ofTransportation and US Environmental Protection Agency.

Chamberlain, G. (1980) Analysis of covariance with qualitative data, *Review of Economic Studies,* 47, 225-238.

Clarke, M.I. (1986) Activity modelling - a research tool or a practical planning technique ?, in *Behavioral Research for Transport Policy,* 3-15, VNU Science Press, Utrecht, The Netherlands.

Chapin, F.S., Jr. (1971) Free-time activities and the quality of urban life, *Journal of the American Institute of Planners,* 37, 411-417.

Chu, C. (1989) A paired combinatorial logit model for travel demand analysis, Proceedings of the Fifth World Conference on Transportation Research, 295-309, Ventura, CA.

Cox, D. R. (1972) Regression models with life tables, *Journal of the Royal Statistical Society Series B,34,187-220.*

Cullen, I. and V. Godson (1975) Urban networks: the structure of activity patterns, *Progress in Planning,* 4, 1-96.

Cullen, I. and E. Phelps (1975) Diary techniques and the problems ofurban life, Final Report submitted to the Social Science Research Council, Grant HR 2336, London.

Damm, D. (1980) Interdependencies in activity behavior", *Transportation Research Record,* 750, 33-40.

Daganzo, C. (1979) *Multinomial Probit: The Theory and its Application to Demand Forecasting,* Academic Press, New York.

Daganzo, C.F. and M. Kusnic (1993) Two properties of the nested logit model, *Transportation Science,* 27, 395-400.

Daly, A.J. and S. Zachary (1978) Improved multiple choice models, in D.A. Hensher and M.Q. Dalvi (eds.) *Determinants of Travel Choice,* Saxon House, Westmead.

Deakin, Harvey, Skabardonis, Inc. (1993) Manual of Regional Transportation Modeling Practice for Air Quality Analysis, The National Association ofRegional Councils, Washington, D.C.

Dijst, M. and V. Vidakovic. (1997) Individual action space in the city, in *Activity-Based Approaches to Travel Analysis,* ed. Dick Ettema and Harry Timmermans. Elsevier Science, Ltd., pp. 117-134.

Ettema, D., A. W.J. Borgers and H.J.P. Timmermans (1993) Simulation model ofactivity scheduling behavior, *Transportation Research Record,* 1413, 1-11.

Fellendorf, M., T. Haupt, U. Heidi and W. Scherr (1997) PTV vision: activity-based demand forecasting in daily practice, in *Activity-Based Approaches to Travel Analysis,* editors: Ettema, D.F. and H.J.P. Timmermans, 55-71, Elsevier Science Ltd. New York.

Ferguson, D. and P.M. Jones (1990) HATS study of wheelchair disability in Adelaide, Report to the Director General of Transport, South Australia.

Gärling, T..K. Brannas, J. Garvill, R.G. Golledge, S.Gopal, E.Holm and E. Lindberg(1989). Household activity scheduling. In Transport policy, management & technology towards 2001: Selected proceedings of the fifth world conference on transport research (Vol. IV, pp. 235-248). Ventura, CA: Western Periodicals.

Gärling, T., M.P. Kwan and R. G. Golledge (1994) Computational-process modeling of household travel activity scheduling, *Transportation Research,* 25B, 355-364.

Gliebe, J.P., F.S. Koppelman and A. Ziliaskopoulos (1998) Route choice using a paired combinatorial logit model, prepared for presentation at the 78th meeting of the Transportation Research Board, Washington, D.C., January 1999.

Golledge, R.G., M.P. Kwan and T. Gärling (1994) Computational process modeling of travel decisions using geographical information systems, *Papers in Regional Science,* 73-99.

Golob, T.F. (1986) A non-linear canonical correlation analysis of weekly chaining behavior, *Transportation Research,* 20A, 385-399.

Gopinath, D. and M. Ben-Akiva (1995) Estimation of randomly distributed value of time, working paper, Department of Civil Engineering, Massachusetts Institute of Technology.

Gordon, P., A Kumar and H. W. Richardson (1988) Beyond the journey to work, *Transportation Research,* 21A, 6, 419-426.

Goulias, K.G., R.M. Pendyala and R. Kitamura (1989) Practical method for the estimation oftrip generation and trip chaining, *Transportation Research Record,* 1285, 47-56.

Hägerstrand, T. (1970) What about people in regional science?, *Papers and Proceedings of the Regional Science Association,* 24, 7-24.

Hamed, M.M and F.L. Mannering (1993) Modeling travelers' postwork activity involvement: toward a new methodology, *Transportation Science,* 27, 4, 381-394.

Han, A. and J.A. Hausman (1990) Flexible parametric estimation of duration and competing risk models, *Journal of Applied Econometrics,* 5, 1-28.

Hanson, S. (1980) Spatial diversification and multipurpose travel: implications for choice theory, *Geographical Analysis,* 12, 245-257.

Hanson, S. and G. Pratt (1988) Reconceptualizng the links between home and work in urban geography, *Economic Geography,* 64, 299-321.

Hanson, S. and G. Pratt (1992) Dynamic dependencies: a geographic investigation of local labor markets, *Economic Geography,* 68, 373-405.

Heckman, J. and B. Singer (1984) A method for minimizing the distributional assumptions in econometric models for duration data, *Econometrica,* 52, 271-320.

Hensher, D.A. and F.L. Mannering (1994) Hazard-based duration models and their application to transport analysis, *Transport Reviews,* 14, 1, 63-82.

Hensher, D.A. (1998a) Establishing a fare elasticity regime for urban passenger transport, *Journal of Transport Economics and Policy,* 32 (2), 221-246.

Hensher, D.A. (1998b) Extending valuation to controlled value functions and non-uniform scaling with generalized unobserved variances. In Gärling, T., Laitila, T., & Westin, K. *Theoretical Foundations of Travel Choice Modeling*. Oxford: Pergamon, 75-102

Hensher, D. A., J. Louviere, and J. Swait (1999) Combining sources of preference data, *Journal of Econometrics,* 89, 197-221.

Hirsh, M., J.N. Prashker and M.E. Ben-Akiva (1986). "Dynamic Model of Weekly Activity Pattern, *Transportation Science,* 20,, 24-36.

Horowitz, J.L. (1980). A utility maximizing model of the demand for multi-destination non-work travel, *Transportation Research,* 14B, 369-386.

Johnson-Anumonwo, I. (1995) Racial differences in the commuting behavior of women in Buffalo, 1980-1990, *Urban Geography,* 16, 23-45.

Jones, P. M., F.S. Koppelman, and J.P. Orfeuil (1990) Activity analysis: state of the art and future directions, in *Developments in Dynamic and Activity-Based Approaches to Travel Analysis,* 34-55, Gower, Aldershot, England.

Jou, R-C and H.S. Mahmassani (1997) Comparative analysis of day-to-day trip chaining behavior of urban commuters in two cities, *forthcoming, Transportation Research Record.*

Kim, S.G., M. Hamed and F. Mannering (1993) A note on modeling travelers' home-stay duration and the efficiency of proportional hazards model, working paper, Department of Civil Engineering, University of Washington.

Kitamura, R. (1983) A sequential, history dependent approach to trip chaining behavior, *Transportation Research Record,* 944, 13-22.

Kitamura, R. and M. Kermanshah (1983) Identifying time and history dependencies of activity choice, *Transportation Research Record,* 944, 22-30.

Kitamura, R. (1984) Incorporating trip chaining into analysis of destination choice, *Transportation Research,* 18B, 67-81.

Kitamura, R., E.I. Pas, C.V. Lula, T.K. Lawton, and P.E. Benson (1996) The sequenced activity mobility simulator (SAMS): an integrated approach to modeling transportation, land use and air quality, *Transportation,* 23, 267-291.

Kitamura, R. and S. Fujii (1998) Two computational process models of activity-travel behavior, In T. Garling, T. Laitila and K. Westin (eds.) Theoretical Foundations of Travel Choice Modeling, Oxford: Elsevier Science, pp. 251-279.

Kitamura, R., T. Yamamoto, S. Fujii and S. Sampath (1996). A discrete-continuous analysis of time allocation to two types of discretionary activities which accounts for unobserved heterogeneity", in J.B. Lesort (editor) Transportation and Traffic Theory, 431-453, Elsevier, Oxford.

Koppelman, F.S. and C-H Wen (1996) The paired combinatorial logit model: properties, estimation and application, Department of Civil Engineering, Northwestern University, Evanston, Illinois.

Kraan, M. (1996) Time to travel?; a model for the allocation of time and money, Unpublished Ph.D. dissertation, Department of Civil Engineering, University of Twente, The Netherlands.

Lam, S-H (1991) Multinomial probit model estimation: computational procedures and applications, Unpublished Ph.D. dissertation, Department of Civil Engineering, The University of Texas at Austin.

Lam, S-H. and H.S. Mahmassani (1991) Multinomial probit model estimation: computational procedures and applications, in *Methods for Understanding Travel Behavior in the 1990s,* Proceedings of the International Association of Travel Behavior, 229-242.

Lawton, T.K. and E.I. Pas (1996) Resource paper for survey methodologies workshop, Conference Proceedings on Household Travel Surveys: New Concepts and Research Needs, Transportation Research Board, National Research Council, Washington, D.C.

Lee, M. (1996). Analysis of accessibility and travel behavior using GIS, Unpublished Master of Science Thesis, Department of Civil and Environmental Engineering, The Pennsylvania State University, May.

Lerman, S. (1979). "The Use of Disaggregate Choice Models in Semi-Markov Process Models of Trip Chaining Behavior", *Transportation Science,* 13, 273-291.

Lockwood, P.B. and M.J. Demetsky (1994) Nonwork travel - A study of changing behavior", presented at the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., January.

Lee, L.F. (1983) Generalized econometric models with selectivity, *Econometrica,* 51, 2, 507-512.

MacDonald, H. and A. Peters (1994) Spatial constraints on rural women workers, *Urban Geography,* 15, 720-740.

Mannering, F., E. Murakami and S.G. Kim (1994) Temporal stability of travelers' activity choice and home-stay duration: some empirical evidence, *Transportation,* 21, 371-392.

McFadden, D. (1973) Conditional logit analysis of qualitative choice behavior in Zaremmbka, P. (ed.) *Frontiers in Econometrics,* Academic press, New York.

McFadden, D. and K. Train (1996) Mixed MNL models for discrete response, working paper, Department of economics, University of California, Berkeley.

McMillen, D.P. (1995) Spatial effects in probit models: a monte carlo investigation, in L. Anselin and R.J.G.M. Florax (editors) *New Directions in Spatial Econometrics,* Springer-Verlag, New York.

Meyer, B.D. (1990) Unemployment insurance and unemployment spells, *Econometrica,* 58, 4, 757-782.

Mokhtarian, P.L. (1993) The travel and urban form implications of telecommunications technology, Discussion Paper for FHWA/LILP Workshop on Metropolitan America in Transition: Implications for Land Use and Transportation Planning, Washington, D.C..

Niemeier, D.A. and J.G. Morita (1996) Duration of trip-making activities by men and women, *Transportation,* 23, 353-371.

Newell, A. and H.A. Simon (1972) *Human Problem Solving,* Prentice-Hall, Englewood Cliffs, NJ.

Nishii, K., K. Kondo and R. Kitamura (1988) Empirical analysis of trip-chaining behavior, *Transportation Research Record,* 1203, 48-59.

O'Kelly, M.E. and E.J. Miller (1984) Characteristics of multistop multipurpose travel: an empirical study of trip length, *Transportation Research Record,* 976,33-39.

Pas, E.I. (1997) Recent advances in activity-based travel demand modeling, *Activity-Based Travel Forecasting ConferenceProceedings,* June 2-5,1996, Summary, Recommendations and Compendium of Papers, February 1997

Pas, E.I. and F.S. Koppelman (1986). "An Examination of the Determinants of Day-to-Day Variability in Individuals' Urban Travel Behavior, *Transportation,* 13, 183-200.

Paselk, T. and F. Mannering (1993) Use of duration models to predict vehicular delay at US/Canadian border crossings, paper presented at the 1993 Annual Meeting of the Transportation Research Board, Washington, D.C.

Preston, V., S. McLafferty and E. Hamilton (1993) The impact of family status on black, white, and hispanic women's commuting, *Urban Geography,* 14, 228-250.

Purvis, C.L., M. Iglesias and V.A. Eisen (1996) Incorporating work trip accessibility in non-work trip generation models in the San Francisco Bay area, presented at the 75th Annual Transportation Research Board Meeting, Washington, D.C., January.

Recker, W.W. (1995a) The household activity pattern problem: general formulation and solution, *Transportation Research,* 29B, 61-77.

Recker, W.W. (1995b) Discrete choice with an oddball alternative, *Transportation Research,* 29B, 201-212.

Recker, W.W., M.G. McNally and G.S. Root (1986a) A model of complex travel behavior: part i. Theoretical development, *Transportation Research,* 20A, 307-318.

Recker, W.W., M.G. McNally and G.S. Root (1986b) A model of complex travel behavior: part ii. An operational model, *Transportation Research,* 20A, 319-330.

Small, K.A. (1987) A discrete choice model for ordered alternatives, *Econometrica,* 55(2), 409-424.

Spear, B.D. (1994) New approaches to travel forecasting models: a synthesis of four research proposals, Final Report (DOT-T-94-15) submitted to the U.S. Department of Transportation and US Environmental Protection Agency.

Speckman, P., K.M. Vaughn and E.I. Pas (1997) A Continuous Spatial Interaction Model: Application to-Home-Work Travel in Portland, Oregon, presented at the 1997 Annual Transportation Research Board Meeting, Washington, D.C.

Steckel, J.H. and W.R. Vanhonacker (1988) A heterogenous conditional logit model of choice, *Journal of Business and Economic Statistics,* 6, 391-398.

Swait, J. and E.C. Stacey (1996) Consumer brand assessment and assessment confidence in models of longitudinal choice behavior, presented at the 1996 INFORMS Marketing Science Conference, Gainesville, FL, March 7-10.

Swait, J. and W. Adamowicz (1996) The effect of choice environment and task demands on consumer behavior: discriminating between contribution and confusion, working paper, Department of Rural Economy, University of Alberta.

Thill, J-C. and J.L. Horowitz (1997a), Travel time constraints on destination-choice sets, *Geographic Analysis,* 29, 108-123.

Thill, J-C. and J.L. Horowitz (1997b), Modeling nonwork destination choices with choice sets defined by travel time constraints, Department of Geography, State University of New York at Buffalo, Working Paper.

Train, K. (1995) Simulation methods for probit and related models based on convenient error partitioning, working paper, Department of economics, University of California, Berkeley.

Van der Hoorn, T. (1983) Development of an activity model using a one-week activity-diary data base, in S. Carpenter & P. Jones (eds.), *Recent Advances in Travel Demand Analysis,* 335-349, Gower, Aldershot, England.

Vause, M. (1997) A rule-based model of activity scheduling behavior, in *Activity-Based Approaches to Travel Analysis,* editors: Ettema, D.F. and H.J.P. Timmermans, 73-88, Elsevier Science Ltd. New York.

Vovsha, P. (1997) The cross-nested logit model: application to mode choice in the Tel-Aviv metropolitan area, presented at the 1997 Annual Transportation Research Board Meeting, Washington, D.C.

Vaughn, K.M., P. Speckman and E.I. Pas. (1997) Generating household activity-travel patterns (HATPs) for synthetic populations, presented at the 1997 Annual Transportation Research Board Meeting, Washington, D.C.

Wen, C. (1998) Development of stop generation and tour formation models for the analysis of travel/activity behavior, Unpublished Ph.D. dissertation, Dept. of Civil Engineering, Northwestern University, Evanston, Illinois.

Wen, C. and F.S. Koppelman (1997) A conceptual and methodological framework for the generation of activity-travel patterns, Proceedings, The International Association of Travel Behavior Research Conference, Austin, Texas, September 1997.

Wen, C. and F.S. Koppelman (1999) An Integrated Model System of Stop Generation and Tour Formation for the Analysis of Activity and Travel Patterns, *forthcoming,* Transportation Research Record.

Williams, H.C. W.L (1977) On the formation of travel demand models and economic evaluation measures of user benefit, *Environment and Planning,* 9 A, 285-344.

This page intentionally left blank

4 TRANSPORTATION SAFETY

Leonard Evans

4.1 Introduction

Specialists and the public widely use the term *safety.* Such use rarely generates serious misunderstanding even though there is no precise, let alone quantitative, definition of *safety.* The general concept is the absence of unintended harm to living creatures or inanimate objects. Quantitative safety measures nearly always focus on the magnitudes of departures from perfect safety, rather than directly on safety as such. Depending on the specific subject and on available data, many measures have been used.

A feature that measures of safety have in common is that they are, in essentially all cases, rates. That is, some measure of harm (deaths, injuries, or property damage) divided by some indicator of exposure to the risk of this harm. For example, rates related to driver deaths include the number of driver deaths per kilometer of travel, per vehicle, per licensed driver, and per year. Note that the number of driver deaths per year is just as much a rate as any of the other examples.

Even within a narrow portion of transportation (say, scheduled airlines or motorcycles), there is no one rate that is superior to others in any general sense. Which rate is appropriate depends on what question is asked (and also on what data are available).

While safety is an important consideration in many human activities, it has a particularly prominent role in transportation. Every type of transportation system involves some risk of harm, as has been the case since antiquity, and seems likely to remain the case in the future. The primary goal of transportation, the effective movement of people and goods, is better served by ever increasing speeds. A substantial proportion of technological innovation for the last few thousand years has focused on increasing transportation speeds, from animal-powered to supersonic flight. In general, as speed increases so does risk.

The Sinking of the Titanic

Some safety concepts can be illustrated by the best known of all unintended events in transportation safety -- the sinking of the *Titanic.* (We have no way to know whether in 90 years the intrinsically more important intentional events of 11 September 2001 will have left as indelible an impression on the world's consciousness). On Sunday 14 April 1912 the 47,000-ton liner *Titanic* maintained its top speed of 22.5 knots (42 km/h) despite receiving nine ice warnings. At 11:40 p.m. the crew reported an iceberg directly ahead. Despite vigorous evasive action, a glancing impact ripped a 90 m gash in the starboard side. The *Titanic* sank at 2:20 a.m. on Monday 15 April, 2 hours and 40 minutes after the impact, with the loss of over 1500 lives, including that of the 62-year-old captain, Edward J. Smith, on his scheduled last voyage (Company captain, 1998).

What if? Any unintended incident leading to harm begs a series of "what if" questions. What if, by chance, the *Titanic* had been a few dozen meters north or south of its actual position? What if the lookout had spotted the iceberg a few seconds earlier? What if there had been more effective procedures for deploying the available lifeboats? What if there had been more lifeboats? US law prohibits 62-year-olds from piloting passenger-carrying aircraft. So, was it an older driver problem? It is generally concluded that if the ship had maintained its initial high speed, the resulting increase in rudder effectiveness would have prevented contact with the iceberg. It is also claimed that cutting the speed to half rather than stopping completely after impact forced additional water into the vessel. Another hour afloat could have substantially reduced casualties as the liner *Carpathia* arrived less than two hours after the *Titanic* sank.

What if impact had been head-on? One "what if" given less attention than others is: What if no one had detected the iceberg and the *Titanic* had crashed head-on into it at 42 km/h? When a car traveling at 42 km/h strikes an immovable barrier, about 8% of its total length (or about 0.4 m) is crushed (Wood, 1997). The uncrushed portion of the car experiences an average deceleration of 167 m/s^2, equivalent to 17 times the acceleration due to gravity, or 17 g. The associated forces of the occupants against their safety belts are likely to produce some injuries (unbelted occupants would sustain greater levels of injury as they continued to travel at 42 km until abruptly stopped by striking the near stationery interior of the vehicle). Assume, as a very rough approximation, that 8% of the *Titanic's* 269 m length would have been crushed by the head-on impact. This 21.5 m of crush would generate an average deceleration of 3 m/s^2, or about 0.3 g. The energy dissipated, equivalent to 30,000 cars crashing (in the 4 seconds during which crush occurred) would have made an enormous noise. Those in the 92% of the liner that was not crushed by the impact would have experienced a mild deceleration, not too unlike that of a car or train coming to a gentle stop at a traffic light or station. Anyone in the portion that was crushed would likely have been killed or seriously injured. As few crew members, and even fewer passengers, would be close to the front of the ship at near midnight, casualties would have been light. The ship would have been in no danger of sinking because of its watertight compartment structure. It would likely have returned to its maker in Belfast for repairs, and today almost nobody would have heard of it.

Number of fatalities – reliability of data. Immediately after the sinking, official inquiries were conducted by a special committee of the U.S. Senate (because American lives were lost) and the British Board of Trade (under whose regulations the *Titanic* operated). The total numbers of deaths established by these hearings were:

U.S. Senate committee:	1,517 lives lost
British Board of Trade:	1,503 lives lost

Confusion over the number of fatalities was exacerbated by the official reports to the U.S. Senate and the British Parliament, which revised the numbers to 1,500 and 1,490, respectively. Press reports included numbers as high as 1,522. Additional revisions cement the conclusion that we will never how many people died on the *Titanic*. (We do know that there were 705 survivors). Likewise, we will never know how many people were killed in the 11 September 2001 terrorist attacks.

The uncertainty regarding the number of deaths in exhaustively investigated prominent events alerts us to the likelihood of uncertainties in even the most seemingly reliable data. At some intuitive level, one might expect the number of deaths to be generally determinable without mistake. For various reasons, this is rarely the case. Arbitrary criteria are often necessary even to classify whether a death should be counted as a transportation death. Drivers may have fatal heart attacks at the wheel prior to crashing; vehicle occupants may be transported to hospital after a crash and die later for reasons, such as pneumonia, that may not be strongly linked to the crash. While there is uncertainty associated with fatality data, such data constitute, by far, the most reliable safety data available. Hence, much of the scientific study in safety focuses on fatalities.

Crashworthiness and crash avoidance. Neither builder nor owner ever used the term "unsinkable." However, the claim of a high level of design safety was well justified, notwithstanding many later questions about the quality of the steel sheeting, the absence of tops on the watertight compartments, and the number of lifeboats. The *Titanic* contained the best *crashworthiness* that had ever been engineered into a ship. However, engineering safety must be viewed in the context of the way it is used. Interactions between *crashworthiness* and crash *avoidance* are examples of more general behavior feedback effects (or technology/human interface effects) that are important in safety (Evans, 1991; 1996). If the *Titanic* had not processed such superior crashworthiness, it would have sunk in minutes rather than hours, with the near-certain loss of all on board. Indeed, its fate may have remained a mystery to this day. Less confidence in *Titanic's* crashworthiness would likely have led to more caution on the bridge. Shakespeare writes, "Best safety lies in fear" (Hamlet, Act I, Scene 3). Because of the ice conditions less safe vessels were waiting for dawn before proceeding. The sinking of the *Titanic* raises a fundamental safety question with parallels in other areas, such as the effect of airbags on fatalities: "Did the *Titanic's* superior crashworthiness save 705 lives or cause over 1,500 deaths?"

Terminology

The above discussion has already introduced a number of terms, which we now discuss more fully.

A vehicle striking anything is referred to as a *crash.* The widely used term *accident* is unsuitable for technical use (Pless and Davis 2001, Evans, 1994; Langley, 1988; Doege, 1978). *Accident* conveys a sense that the losses incurred are due exclusively to fate. Perhaps this is what gives *accident* its most potent appeal -- the sense that it exonerates participants from responsibility. *Accident* also conveys a sense that losses are devoid of predictability. Yet the purpose of studying safety is to examine factors that influence the likelihood of occurrence and resulting harm from crashes. Some crashes are purposeful acts for which the term *accident* would be inappropriate even in popular use. There can be little doubt that at least a few percent (perhaps as much as 5%) of driver fatalities are suicides (Hernetkoski and Keskinen, 1998; Ohberg et al., 1997; Bollen and Philipps, 1981; Philipps, 1979). Although the use of vehicles for homicide may be less common than in the movies, such use is certainly not zero. Popular usage refers to *the crash of Pan Am flight 103,* now known to be no *accident,* in any sense of the word. Even more so, the events of 11 September 2001, known to be intentional acts immediately after the second plane crashed into the World Trade Center.

Generally the term *cause* is avoided, in large measure because it all too often invokes the inappropriate notion of a single cause. Crashes result from many factors operating together. To say that the loss of life on *Titanic* was caused by the absence of a mandatory retirement age for captains, the owner being on board, the look-out being too alert or not alert enough, by climate conditions, or by poor quality steel may generate more confusion than clarity. Instead of focusing on a single cause, we generally think in terms of a list of *factors,* which, if different, would have led to a different outcome. The goal in safety analysis is to examine factors associated with crashes with the aim of identifying those which can be changed by countermeasures, or interventions, to enhance future safety.

Collections of observed numbers are referred to as *data* and not *statistics.* Since *statistics* is the name of a branch of mathematics dealing with hypothesis testing and confidence limits, using it to also mean data invites needless ambiguity.

We follow common usage in referring to ages; age 20 means people with ages equal to or greater than 20 years, but less than 21 years. This is plotted at 20.5 years, very close to the average age of 20-year-olds; 40-year-olds are not quite twice as old as 20-year-olds, which might come as good news to some!

The consequences of crashes include fatalities, injuries and property damage. Useful terms encompassing all of these are *harm* and *losses.* Measures that reduce harm can be placed into two distinct categories.

Crashworthiness refers to engineering features aimed at reducing losses, given that a specific crash occurs. Examples are improved occupant protection by making the structure close to the occupant less likely to crush, and devices such as collapsible steering columns; other examples of crashworthiness include reducing risks of post-crash fires, or of ships sinking from crash impact.

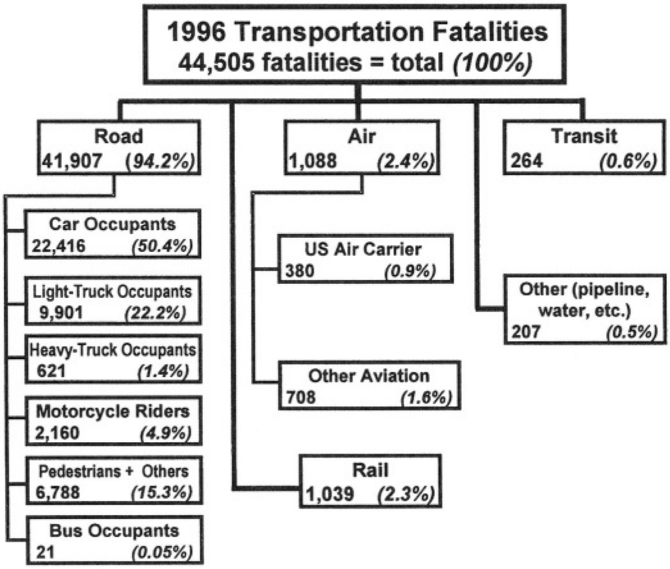
Crash prevention refers to measures aimed at preventing the crash from occurring. Such measures may be either of an engineering nature (making vehicles easier to see, better braking, radar, etc.) or of a behavioral nature (driver selection, training, motivating and licensing, traffic law enforcement, etc.).

A fundamental difference between *crashworthiness* and *crash prevention* is that when a crash is prevented all harm is reduced to zero. Improved *crashworthiness* rarely eliminates harm in a severe crash, but does reduce the level of harm (say, converting a fatality into a severe injury, or a severe injury into a less severe injury, or an expensive vehicle-repair into a less expensive repair). The finding that safety belts reduce car-driver fatality risk by 42% means that out of 100 drivers who would have been killed without belts, 42 would have survived if all had worn belts. However, the 42 survivors would sustain injuries, in many cases very severe injuries. *Crashworthiness* is measured by the percent reduction in risk for some specific level of injury, such as fatality or minor injury. A *crash prevention* measure that reduces crash risk by some percent is necessarily a far more effective intervention than a *crashworthiness* measure with the same percent effectiveness.

4.2. Overview of Transportation Fatalities

The US Department of Transportation (1998) estimates that 44,505 people lost their lives in connection with transportation in the United States in 1996. The distribution of these by transportation mode is presented in Table 4-1. The numbers in Table 4-1 in a few cases differ slightly from those in the original source because of minor

Table 4-1. Distribution by mode of transportation fatalities in the United States in 1996. (Based on US Department of Transportation, 1998).



corrections to achieve consistent totals.

These 44,505 deaths occurred in a system in which vehicles traveled over 4 billion km in 1996, as detailed in Table 4-2. As there is, on average, more than one person per vehicle, the distance traveled by all people will exceed the distance traveled by all vehicles (details in Table 4-3.) The unfortunate term *passenger miles* (or *passenger km)* appears often in the literature, even though most travel is in vehicles containing no passengers. People traveling in (or on) road transportation vehicles are more appropriately referred to as *occupants.* Occupants are either *drivers* or *passengers.* Vehicles referred to in the literature as "passenger cars" will here be called simply "cars." Because different situations arise for different modes, additional categories (such as *crew*) are also used.

Table 4-4 shows the number of deaths per billion vehicle miles derived by dividing the estimates in Table 4-1 by those in Table 4-2. No estimates are given in Table 4-4 if the definitions for the categories of distance of vehicle travel and the fatalities were substantially different, or the estimates of travel are too uncertain. Even without problems of data availability and reliability, it is surprisingly difficult to define categories that apply across all modes. For US road traffic, a fatality is counted if the crash occurs on a US public road, without regard to the origin of the vehicle or its occupants, whereas for air traffic factors such as the home base of the airline are relevant while the location of the crash may not be. Rail rates are not given as most fatalities occur to people outside the train (at grade crossings), and passenger and freight-train data are collected in different ways. A car driver killed in a car-train crash is likely to be added to both the road traffic and the train totals. A worker killed in a fire unrelated to transportation in a railroad facility is counted as a railroad fatality. Tables 4-4 (and 4-5) should be interpreted in the context of these uncertainties.

The overall national rate for all modes of transportation is 11.1 fatalities per billion km of vehicle travel. The road transportation rate of 10.5 fatalities per billion km is equivalent to 1.7 fatalities per hundred million miles (conversion factor is 1.609334/10 exactly). As vehicles with high occupancy travel with more people at risk, it is appropriate to examine the deaths for the same distance of occupant travel.

Table 4-5 shows the number of deaths per billion km of travel, derived by dividing the estimates in Table 4-1 by those in Table 4-3. The cross-modal comparisons in Tables 4-4 and 4-5 are sufficiently unreliable that they should be interpreted as little more than suggestive. Scheduled airline rates are much lower than the average for all airline travel. As one or two major airline crashes have a large influence on this rate, it is highly unstable from year to year. The rate averaged over 1990-1996 is 0.2 deaths per billion aircraft km. The overwhelming majority of those killed in airline crashes have minimal control over events. All are at similar risk, regardless of behavior or personal characteristics. While the average rate for road-vehicle occupants is much higher, this rate varies greatly according to such characteristics as driver age, use of alcohol, safety-belt use, conformance with traffic law, etc. A car driver with many characteristics associated with lower crash risk can drive a 1,000 km trip with no more risk of death than taking a plane for the

Table 4-2. Distribution by types of vehicles of the total distance traveled by all vehicles in the United States in 1996. (Based on US Department of Transportation, 1998).

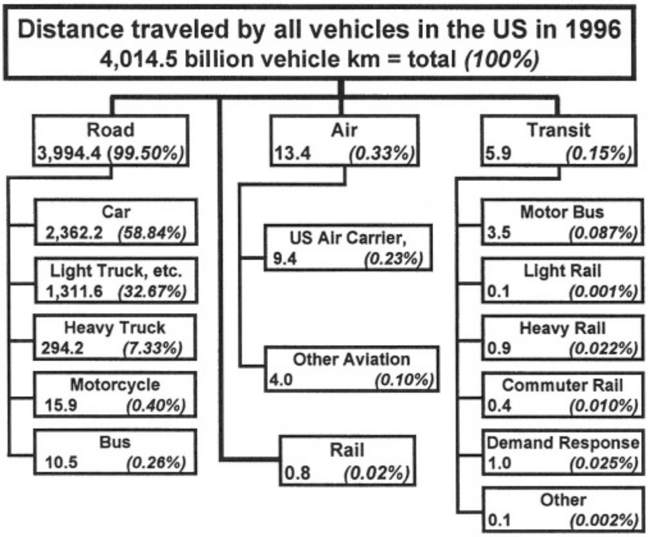


Table 4-3 Distribution by vehicle types of travel by all vehicle occupants in the United States in 1996. (Based on US Dept. of Transportation, 1998).

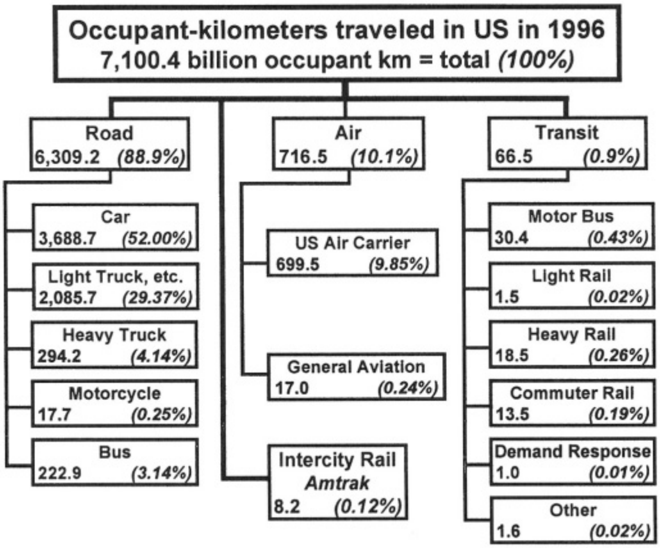


Table 4-4. Death rates for the same distance of vehicle travel, computed from *Tables 4-1 and 4-2*.

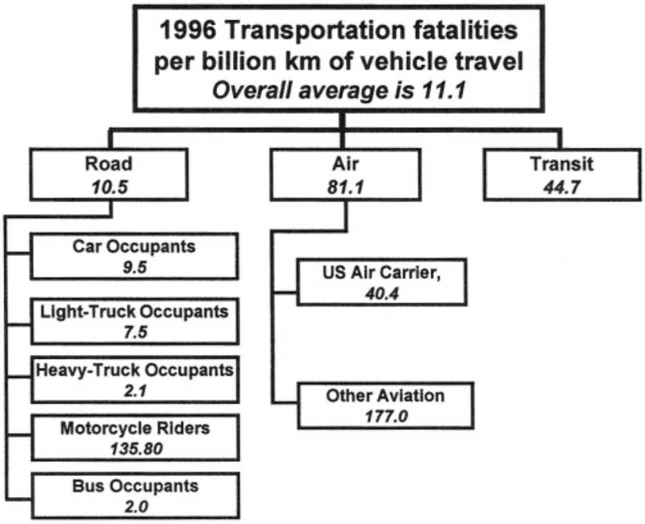
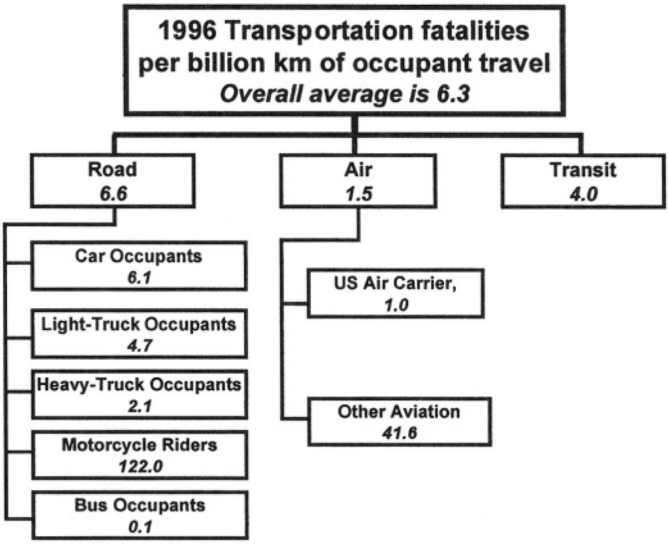


Table 4-5. Death rates for the same distance of occupant travel, computed from Tables 4-1 and 4-3.



same trip. The longer the trip the greater is the safety advantage of air travel, because nearly all the risk is concentrated in the take-off and landing phases, whereas the ground vehicle risk is approximately proportional to the distance traveled.

Within the road transportation mode, the comparisons are more reliable. The risk of occupant death depends systematically, and very strongly, on the mass and size of the vehicle (Evans, 2001a). For a same distance journey, a motorcycle rider is about 20 times as likely to be killed as a car occupant, and about a thousand times as likely to be killed as a bus occupant.

Tables 4-1 through 4-5 underline the dominance of road transportation over all other modes combined in the US. Road transportation accounts for over 99% of all the distance traveled by vehicles, and almost 90% of all the distance traveled by people. It accounts for 94% of all transportation deaths, and for an even higher percent of injuries and property damage. Because of its dominant role, most of the rest of this chapter is devoted to road transportation. Unless otherwise apparent, the term *vehicle* denotes an engine-powered vehicle designed to travel on a road, and the term *traffic crash* denotes a crash involving at least one such a vehicle. *Traffic crashes* also generally involve non-vehicles (pedestrians, bicycles, animal-powered vehicles, and fixed objects –trees being the most common).

4.3 Introduction To Road Traffic Fatalities

Road traffic deaths and injuries constitute one of the largest public health problems in industrialized countries. In the US, traffic crashes account for half of all injury deaths (National Safety Council 1997), and 94% of all transportation deaths. In a typical two-week period, more people are killed on US roads than the 1500 lost on the *Titanic*. In a typical month, more Americans die on US roads than were killed in the terrorist attacks.

In the US, traffic crashes account for half of 19-year-old female deaths and a third of 19-year-old male deaths (Evans, 2000). The fraction is lower for males because of so many male deaths from firearms. The total number of pre-retirement years of life lost due to traffic crashes is similar to that due to the combined effects of the two leading diseases, cancer and heart disease. Worldwide, about a million people are killed annually in traffic crashes (WHO, 2001), with injuries about 70 times this number. The victims are predominantly young, and about 65% are male. As motorization increases, totals are expected to increase.

Analysis of road safety differs from that for the other modes in that enormous quantities of relevant data are available, most commonly based on police reports. The *Fatality Analysis Reporting System* (FARS – before 1998 called *Fatal Accident Reporting System*) documents over a million people killed on US roads since 1975. The availability of large quantities of data lead to safety for roads being better understood than safety for any other transportation mode.

Variables coded in large data sets generally include gender and age of crash participants, weather, make and model of vehicle, etc. Variables not known include vehicle speed at onset of crash event, vehicle speed just prior to impact, amount of vehicle crush, and medical details of injuries. Such details can be estimated only after expensive post-crash investigations, which are not routinely performed. For other transportation modes, nearly all information comes from intensive in-depth analysis of the few crashes that occur.

Historical Trends

In the early decades of the twentieth century few people were killed on US roads because there were few motorized vehicles (Figure 4-1). As vehicle ownership increased rapidly, so did traffic deaths, peaking in 1972 at 54,589, and declining later to a present fairly stable rate of just over 40,000 per year. The rate in China and other rapidly industrializing countries continues to increase rapidly.

The number of traffic deaths per year shows little in the way of a pattern. However, if we instead examine the number of traffic deaths in the US for the same distance of vehicle travel, a clear trend emerges (Figure 4-2). Ever since 1921 when data on the total distance traveled by all vehicles were first collected, the number of traffic deaths for the same distance of travel has trended downwards at an average decrease of about 3.5% per year. The 2000 rate of 9.7 traffic deaths per billion km of travel is 94% below the 1921 rate of 150. If the 1921 rate were to apply today, the number of US traffic fatalities would exceed half a million. The downward trend in the number of deaths for the same distance of travel is observed in all countries for which data are available (Evans, 1997). As motorization continues, the fraction of all deaths that are pedestrians trends downwards (Figure 4-3).

The number of traffic deaths for the same distance of travel can be measured only after a nation instigates a procedure to estimate the distance all vehicles are driven. Even when available, estimates of distance of travel differ greatly in reliability from country to country. A useful universally available measure is the number of traffic deaths per thousand registered vehicles. The registration, and thereby counting, of vehicles is routinely performed by nearly all jurisdictions. The number of deaths per thousand vehicles varies greatly between countries -- by more than a factor of one hundred (Table 4-6 and Figure 4-4). In general, the higher the degree of motorization (as indicated by the number of vehicles per 1000 people), the lower is the number of traffic fatalities per thousand vehicles. Another key factor is mix between rural and urban driving. Fatality risk tends to be lower in urban areas where speeds are lower. Within the US, the fairly urban states of Rhode Island, Massachusetts and Connecticut have 0.09, 0.11 and 0.12 deaths per thousand vehicles, respectively, whereas the more rural states of Mississippi and Arkansas have 0.39 deaths per thousand vehicles. While the rate for China, the world's most populous nation, is substantially higher than that for more motorized countries, it is dropping at a much faster rate than in the US and other more motorized countries (Figure 4-5). Although the rate is dropping, the dramatic growth of vehicle ownership in China (Figure 4-6) and in other countries that are rapidly industrializing will inexorably increase casualties.

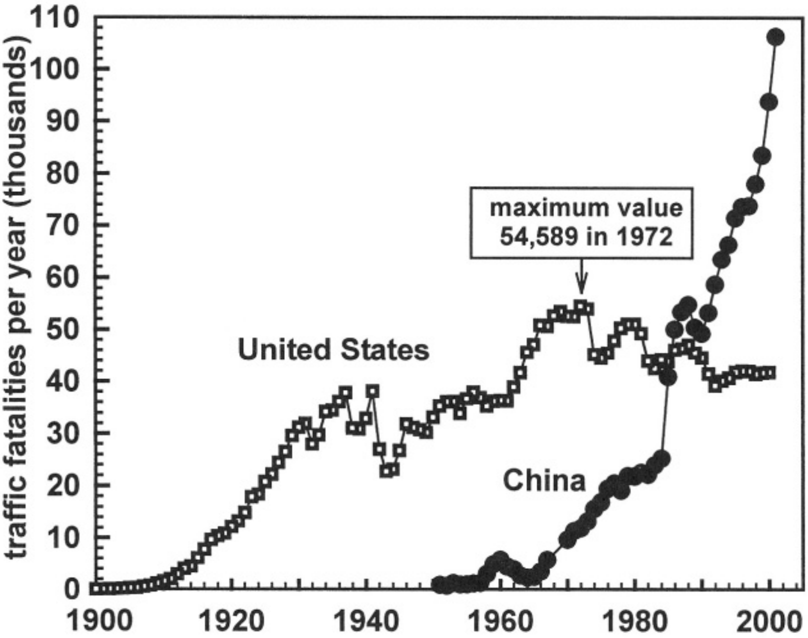


Figure 4-1. *Total annual traffic fatalities in the US and China.*

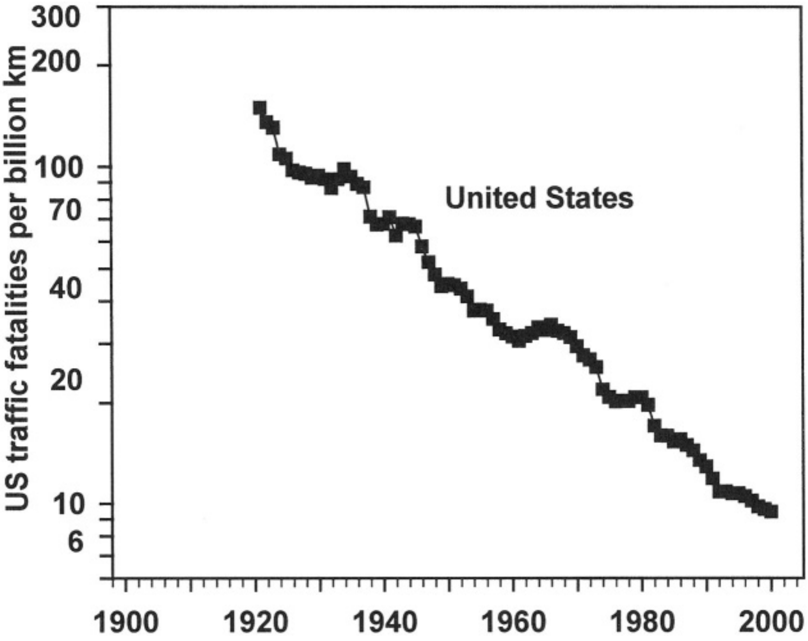


Figure 4-2. Total annual traffic fatalities per billion kilometers of vehicle travel in the US.

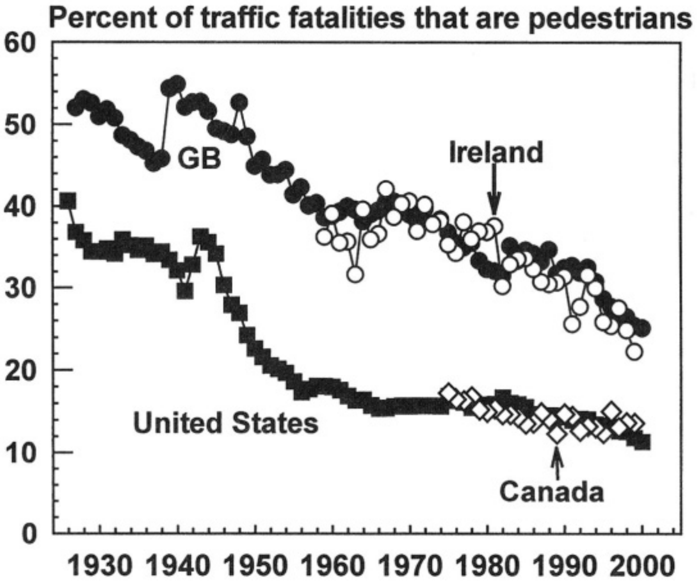


Figure 4-3. The percent of all traffic fatalities that are pedestrian fatalities.

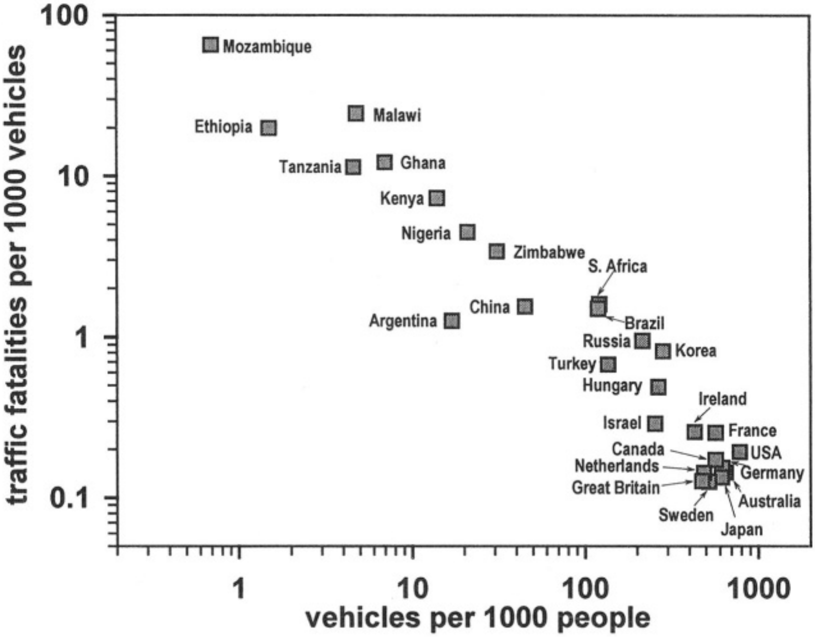


Figure 4-4. Safety related to degree of motorization

Table 4-6. Various rates for a number of countries.

Country	Vehicles per 1000 people	Fatalities per 1000 vehicles	Fatalities per million people	Fatalities per year	Data year
USA	790	0.19	153	41,821	2000
Australia	647	0.14	93	1,763	1999
Germany	617	0.15	95	7,772	1999
Japan	614	0.13	82	10,372	1999
Canada	567	0.17	97	2,972	1999
France	567	0.25	144	8,487	1999
Sweden	520	0.13	66	580	1999
Netherlands	485	0.14	69	1,090	1999
UK	473	0.13	60	3,564	1999
Ireland	429	0.26	110	413	1999
Korea	282	0.82	232	10,756	1999
Hungary	265	0.49	129	1,306	1999
Israel	254	0.29	74	469	1999
Russia	215	0.95	204	29,600	2000
Turkey	136	0.68	92	5,975	1999
South Africa	121	1.60	193	9,068	1998
Brazil	119	1.50	179	30,000	1998
China	45	1.55	70	83,529	1999
Zimbabwe	31	3.39	106	1,205	1996
Nigeria	21	4.49	94	6,185	1995
Argentina	17	1.26	210	7,545	2000
Kenya	14	7.29	103	2,617	1995
Ghana	7.0	12.19	86	1,646	1998
Malawi	4.8	24.49	119	1,382	1996
Tanzania	4.6	11.39	53	1,583	1998
Ethiopia	1.5	19.91	29	1,693	1998
Mozambique	0.7	65.18	43	805	1997

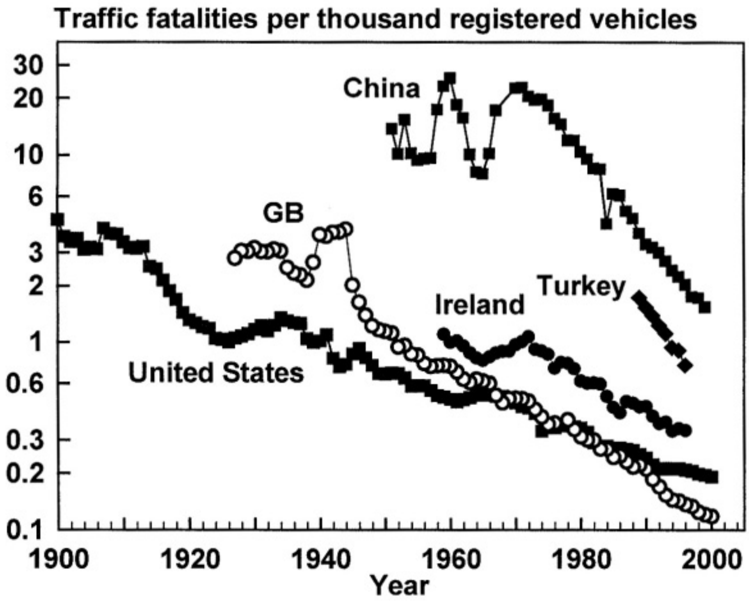


Figure 4- 5. Total annual traffic fatalities per thousand registered vehicles in some countries.

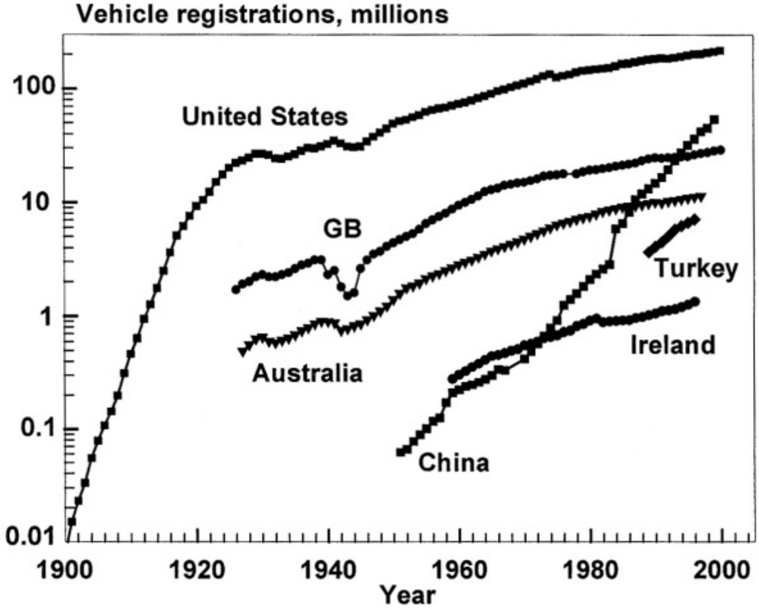
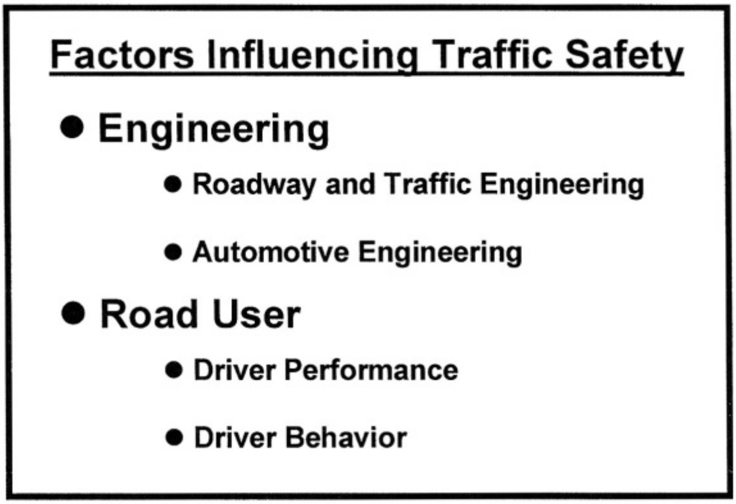


Figure 4- 6. Vehicle growth in some countries.

Approaches To Reducing Harm From Traffic Crashes

Why do fatality rates decline in time and vary so much from country to country? This question is somewhat akin to asking why average longevity increases in time and varies so much from country to country. Such effects are due to many factors – public health policy and implementation, availability of and advancements in drugs, surgery, preventative medicine, plumbing, nutrition, hygiene, etc. In the traffic crash and longevity cases it is difficult to assign in any quantitative way the relative contributions of the different factors. The structure in Table 4-7, which is one of a number of possible categorizations, is aimed at clarifying some of the main factors that contribute to traffic safety. Not reflected, because it is somewhat outside the scope of this chapter, are the important contributions from improved medicine, which reduce average harm from all sources. As medical science continues to advance, those injured in any transportation crash are less likely to die. Indeed, it is often claimed that if a victim can be transported alive to a modern well-equipped emergency trauma center, the probability of survival is extremely high. This places high value on rapid transportation from the crash site to the hospital. Here the infrastructure of vehicular transportation contributes in a fairly direct way to reducing the severity of the harm from the crashes that occur on it.



Factors Influencing Traffic Safety

- **Engineering**
  - **Roadway and Traffic Engineering**
  - **Automotive Engineering**
- **Road User**
  - **Driver Performance**
  - **Driver Behavior**

Table 4-7. One way to categorize the main factors central to road traffic safety.

4.4 Engineering Factors

Roadway Engineering

On rural two-lane roads, vehicles traveling in opposite directions pass each other only a meter or so apart. Even if speed limits are obeyed, the combined relative speed may still far exceed 150 km/h. A head-on crash at such relative speeds will likely prove fatal, yet such crashes occur due to, for example, improper overtaking or loss of control on curves. On freeways where there is physical separation between traffic traveling in opposite directions, the only vehicles permitted to drive close to each other are traveling in the same direction at similar speeds. Fixed objects, such as trees, are far removed from the path of vehicles. Risk of side-impact at intersections is eliminated through the replacement of intersections by under- or over-passes.

The roadway engineering improvements typified by the differences between freeways and rural two-lane roads constitute one of the most effective engineering countermeasures available. In the US, fatality risk on interstate rural freeways is 55% lower than the average for all non-interstate rural roads (Table 4-8). The lowest rate in Table 4-8 is 85% lower than the highest rate. Such dramatic safety effects dependant on roads and road use bring one face to face with the types of tradeoffs that often arise in traffic safety decisions. Freeways are expensive undertakings that are justified mainly to produce improved mobility. They can are rarely installed primarily to improve safety. Additional considerations may argue against building freeways, including their effect on city neighborhoods, landscape aesthetics, and wild life. Better roads generate more traffic and stimulate urban sprawl, increasing pressure on resources and the environment. The additional travel that freeways stimulate generates additional travel risk, but this effect is small compared to the risk reduction resulting from replacing rural two-lane roadways by freeways.

Table 4-8. Fatality rates on different types of US roads for 1998 (Bureau of Transportation Statistics, 2000)

Roadway category	Fatalities per billion km	
	Rural	**Urban**
NON-INTERSTATE:		
Arterial	14.8	7.1
Collector	18.3	4.9
Local	23.0	7.9
NON-INTERSTATE AVERAGE	**17.2**	**7.0**
INTERSTATE	7.7	**3.8**

Vehicle Engineering

In the earliest days of the auto industry, crashes often resulted from the mechanical failure of such key components as wheels, tires or brakes. As component reliability increased, focus shifted towards fundamental understanding of injury mechanisms and on technologies aimed at protecting occupants of vehicles when crashes occur.

Biomechanics – the science of relating injury to mechanical force. Biomechanics is the bridge that links engineering and medicine. Trauma surgeons distinguish between penetrating trauma and blunt trauma. Penetrating trauma occurs when small objects exert sufficient localized force to penetrate the human body, an obvious example being a bullet. Blunt trauma occurs when an object of larger area applies sufficient force on the body to damage its structure, such as occurs when someone falls from a building. Nearly all traffic injuries, whether to vehicle occupants or to pedestrians, involve blunt trauma. Consider a vehicle traveling at, say, 50 km/h and crashing into a perfectly rigid horizontal barrier. An unbelted driver would, in accord with elementary physics, continue to travel at 50 km/h until stopped by a force. Such a force occurs when the driver impacts, at a speed of 50 km/h, the interior of the now stationary vehicle. It is this so-called *second collision* that causes injuries, not the first collision of the vehicle striking the barrier. A person falling from a fourth floor window would strike the ground at a similar speed and be subject to similar injury forces. While evolution has provided humans with a protective fear of heights, no corresponding fear exists for the relatively new experience of traveling at speeds faster than can be produced by muscle power.

Goal of occupant protection. The theoretical best protection would be for the occupant to slow down from the initial speed of 50 km/h to zero at a constant deceleration using the entire distance between his or her body and the barrier. The engine and other rigid components make it impossible to achieve this ideal goal. The practical goal is for the vehicle structure to crumble in such a way as to provide as much ride-down distance as possible, and for the occupant to travel this distance at as uniform a deceleration as possible. In addition, a strong "safety cage" that does not crumple reduces the risk of occupants being crushed.

Engineering changes that have contributed to reductions in driver risk include collapsible steering columns, lap/shoulder safety belts and design changes in the structure surrounding the occupants to reduce intrusion. When a driver's chest strikes a steering wheel, the collapsible steering column allows the steering column to compress and thereby reduce the maximum force on the chest. This simple device is estimated to reduce overall driver fatality risk in a crash by about 6%.

Estimates of the effectiveness of occupant protection devices are summarized in Table 4-9. The interpretation is that if 100 fatally injured drivers not wearing belts had been wearing belts, 42 would have survived. This is equivalent to saying that wearing a belt reduces a driver's risk of being killed in a crash by 42%.

Table 4-9. Effectiveness of safety belts and airbags in reducing driver fatality risk.

Occupant protection device	*Effectiveness in preventing driver fatalities*
Lap/shoulder belt alone	**42 %**
Lap/shoulder belt plus airbag	**47 %**
Airbag alone*	**13 %**

* No manufacturer offers the airbag for use alone. Its stated aim is to increase the effectiveness of the primary restraint system, the lap/shoulder belt (Sources: Kahane, 1996 for airbag-only estimate, others from Evans, 1991).

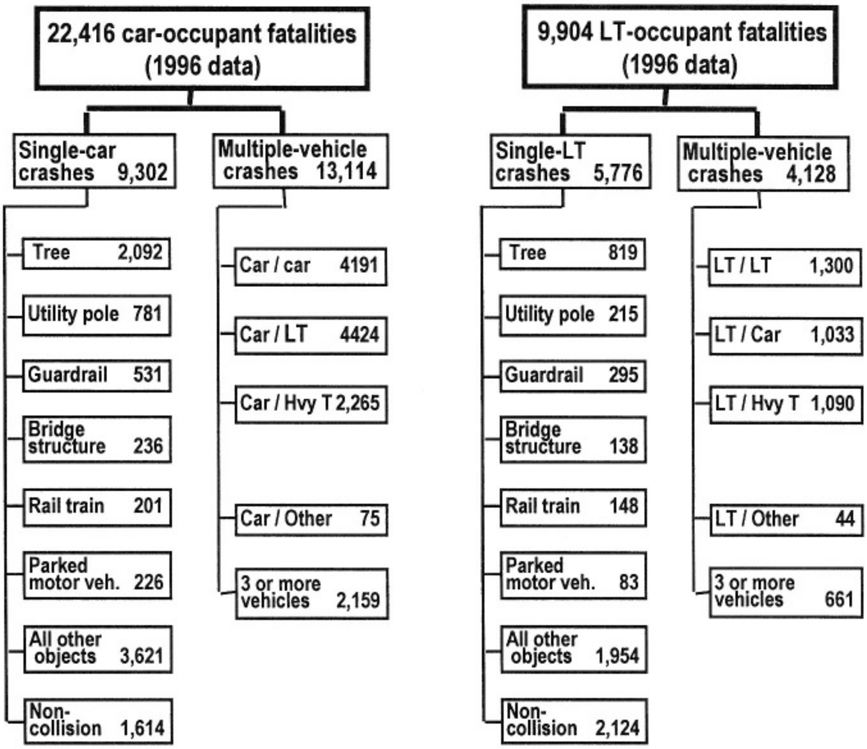
By far the most effective occupant protection device is the familiar lap/shoulder safety belt. This not only reduces the likelihood and severity of impact with the interior of the vehicle, but is highly effective at preventing ejection from the vehicle. Ejection quadruples the risk of death in a crash -- about one in four unbelted drivers killed are ejected from their vehicles. The effectiveness of the lap/shoulder belt is, on average, enhanced by airbags.

In this chapter airbag refers to frontal airbags, which are designed to inflate rapidly in order to place a cushioning barrier between occupant and vehicle structure when sensors detect a frontal crash with severity exceeding some pre-set limit, typically equivalent to striking a barrier at a speed in the range 10 to 20 km/h. The driver rides down the crash in contact with the airbag, which spreads the impact forces over a larger area and reduces forces due to the belt. Side airbags are being introduced without any estimate of their overall effectiveness, which is expected to be, at most, far lower than for frontal airbags.

Increased size and weight of a vehicle increase protection. Doubling the weight reduces occupant risk by about half. All vehicles being heavier does not eliminate the safety benefits of increased vehicle weight (Evans, 1994; 1995), because, in the US, 41% of car drivers and 58% of light-truck drivers who are killed are killed in single-vehicle crashes (Table 4-10). The drivers of two large cars crashing into each other are at lower risk than the drivers of two small cars crashing into each other (Evans, 2001a, Wood and Simms, 2002).

The influence of weight on crash risk is so great that even adding the weight of a passenger generates clearly measurable effects. If a car with a passenger crashes into a car with a lone driver, the accompanied driver is 14.5% less likely to be killed than the unaccompanied driver (Evans, 2001).

Table 4-10. Distributions of 32,320 occupants killed in cars and light trucks. (Data from FARS 1996).



Riders of two-wheeled vehicles are at dramatically higher risks than occupants of even the smallest four-wheel vehicle. Helmets reduce motorcycle driver or passenger fatality risk by 28%. An unhelmeted motorcyclist is about 22 times as likely to be killed as is an average car driver. Wearing a helmet reduces this risk to 18 times that of the car driver. Riders of two-wheeled vehicles (whether engine or human powered) and pedestrians are particularly vulnerable road users. Such road users account for a large fraction of all traffic deaths in the early stages of motorization.

While much attention has been devoted to possible modifications to vehicle design to better protect pedestrians in crashes, the opportunities are intrinsically much less than for vehicle occupants. The main opportunity to prevent such harm is by changing how pedestrians and drivers behave.

4.5 Human Factors Of Road-Users

Discussion of the influence of human factors of drivers (and of road users in general) on road safety must make the clearest distinction between two deceptively similar but fundamentally different concepts:

- Driver performance -- what the driver can do, or is capable of doing
- Driver behavior -- what the driver in fact does

Driver Performance

Studies have concluded that driver error is a contributory factor in over 95% of traffic crashes. Such findings have generated suggestions that the first priority for better safety is to teach higher levels of skill and knowledge about driving. That is, to improve levels of driver performance. While driver training, especially of motorcycle riders, has reduced crash rates in some cases, it has not generally been found to do so. A number of considerations show why crash risk is not determined mainly by driver performance.

Everywhere young male drivers have the highest crash rates (see also section 4-6, Older and younger drivers). Yet this is the very age group with the best visual acuity, swiftest reaction times, and fastest cognitive processing skills. Males tend to be more knowledgeable about and interested in driving and automobiles. Racing-car drivers have higher on-the-road crash rates than average drivers. Much more important than what the driver can do is what the driver chooses to do.

Driver Behavior

The average driver has a crash about once per decade (usually a minor property damage crash -- for fatal crashes it is about one per 4,000 years). Drivers tend to dismiss their crashes as unpredictable and unpreventable bad luck, or the other involved driver's fault. A more appropriate interpretation is that average driving produces one crash per ten years. Feedback once per decade is unlikely to affect behavior. Every crash-free trip reinforces the driver's incorrect conclusion that average driving is safe driving. Individual experience is a false teacher. I wonder how many of us would fly on commercial aircraft if a pilot's method of learning how to avoid crashes was by experiencing them?

A crucial factor that contributes to the high level of commercial airline safety (Table 4-5) is that pilots follow procedures based on expert analyses of the experience of many. For road vehicles, traffic law attempts to fulfill a parallel role. However, ground vehicle drivers routinely violate such laws. Table 4-11 compares various safety characteristics of road and air traffic.

Two of the factors most affecting road-traffic fatality risk are travel speed and alcohol consumption. Research indicates that the risk of crashing increases approximately in proportion to travel speed, injury risk in proportion to travel speed squared, and fatality risk in proportion to travel speed to the fourth power. When speed limits on the US rural intrastate system were reduced in 1974 from 70 mph to 55 mph following the October 1973 Arab oil embargo, average travel speed dropped

Table 4-11. Comparison of safety characteristics of US commercial air carriers and road transportation.

	Commercial Airline	Road Traffic
Deaths per billion km of occupant travel	1.0	6.6
Countermeasures with most success and potential	Crash prevention	Crash prevention
Main US policy emphasis	Crash prevention	Crashworthiness
Impact of vehicle design or manufacturing flaws	Very important	Minimally important
Driver selection	Strict	Essentially everyone
Importance of driver skill and knowledge	High	May increase or decrease crash risk
Main influence on driver behavior	Following increasingly effective procedures	Experience and personal judgment
Violations of pertinent laws	Rare	Typically, many times per trip
Use of alcohol/drugs	Rare	In about 40% of fatalities
Value of high technology driver training simulators	Enormously high value	Zero or minimal value
Time to react to crash-threatening situations	Often more than many seconds or minutes	Usually less than a second
Value of crash-avoidance advanced technology	Enormously high value	Minimal value
Key to making largest improvements in safety	Safer aircraft flown by better trained pilots adhering to better procedures	Behavior changes resulting from changes in social norms, legislation and enforcement

from 63.4 mph to 57.6 mph. This change leads to a predicted fatality risk decrease of 32%, remarkably close to the 34% decline observed. Case-control studies found casualty crash to double with each 5 km/h increase in speed (Kloeden et al., 1997).

Drunk driving is a major traffic safety problem in all countries in which alcohol is used widely, often accounting for about half of all fatalities. Reducing the availability of alcohol has in many cases led to reduced traffic deaths. When all US states increased the minimum age to purchase or consume alcohol to 21 years, from earlier ages of 18 to 20 years in various states, a 13% reduction in fatal-crash involvements of affected drivers followed. Police use of random breath testing to enforce drunk driving laws more effectively has reduced casualties. The Australian state of New South Wales tests about a third of all drivers each year, many of them more than once. This intervention decreased overall fatalities by about 19%.

Driver behavior is a crucial factor in occupant protection because the most effective occupant protection device, the safety belt, works only when fastened.

Mandatory wearing laws have been introduced in most countries, though wearing rates and level and type of enforcement vary greatly. The best evaluated wearing law was that for the United Kingdom, where fatality rates for drivers and left front passengers declined by about 20%.

Vehicles are used for purposes that go beyond transportation, including competitiveness, sense of power and control, or more generally, hedonistic objectives -- the pursuit of sensual pleasure for its own sake. Speed and acceleration appear to produce pleasurable excitement even when no specific destination lies ahead and there is no point in haste. While most drivers are motivated by non-transportation motives at some times, as they mature the mix of motives evolves in a more utilitarian direction. This is likely one reason why crash risk is so much lower for 40-year-olds than for 20-year-olds. It seems plausible that as a nation's motorization matures, a similar evolution occurs and contributes to a lowering of crash rates. Drivers in newly motorized countries are likely to be the first generation to drive, and to approach the activity with a sense of novelty, excitement and adventure. In motorized countries, children grow up with the motor vehicle playing an essential role in even the most routine and mundane aspects of daily life.

Crash risk relates to the deepest human characteristics. Factors at the very core of human personality influence behavior in traffic. A comparison of the gender and age dependence of involvement rates in severe single-vehicle crashes and in crimes unrelated to traffic offenses (say, burglary, as a typical example) show remarkable similarities (Figure 4-7). No one would suggest seriously that 40-year-olds commit fewer burglaries than 20-year-olds solely because the 40-year-olds have learned how not to commit burglaries! This should invite a parallel caution against interpreting lower crash rates for 40-year-old drivers compared to those for 20-year-old drivers to mean that the 40-year-olds have simply learned how to not crash. The most compelling interpretation of the similarity between the two curves in Figure 4-7 is that there are fundamental human characteristics related both to involvement in severe crashes and arrests for offenses unrelated to driving; neither conduct is likely to be changed dramatically by increasing knowledge or skill.

Figure 4-8 compares male and female pedestrian deaths. If male and female rates were similar, the data would lie randomly above and below the dotted line indicating equality. An entirely different, and remarkably consistent, picture emerges. At all ages, plotted in one-year intervals, the male rate exceeds the female rate, including the first year of life. For this first year, with average age close to six months, there were 93 male deaths compared to 59 female deaths, or 93/59. The corresponding numbers of fatalities for ages 1.5, 2.5, and 3.5 years are 590/418, 1131/730, and 1353/741, respectively. Such large robust differences suggest an intrinsic gender difference at the most basic level, likely linked to testosterone.

In driving behavior, as in most human activities, social norms play a central role. People drive in a way that they think will win the approval of those whose approval they desire. A change in social norms regarding drunk driving has taken place in the US. The drunk driver is no longer the amiable comic character of the past, a change that has contributed to reductions in drunk driving. While the fictional portrayal of drunk driving as a harmless activity has become uncommon, the same cannot be said for the portrayal of illegal and life-threatening driving in general, which is often presented as humorous or heroic in television programs and movies specifically aimed at young people. The possibility that such behavior may lead to tragic consequences is rarely addressed. Claims that fictional portrayals do not influence behavior ring hollow in the light of the billions of dollars spent for television advertising. These expenditures are predicated on the firm belief that they do influence behavior. Surely the programs must have a dramatically greater influence than the advertisements. Shaming the entertainment industry into desisting from some current practices would, in my view, save the lives of many young people.

The dominant role of driver behavior. As discussed above, reducing the speed limit from 70 to 55 miles per hour reduced fatality rates on US rural interstate roads by 34%, mandatory safety-belt wearing in the United Kingdom reduced front-seat occupant fatalities by 20%, and random breath testing for alcohol in the Australian state of New South Wales reduced driver fatalities by 19%. Hingson et al. (1996) report similarly large changes in risk in response to programs aimed at changing behavior.

In the 1970s, major independent studies in the US and in Britain identified factors associated with large samples of crashes. The US study found the road user to be the sole factor in 57% of crashes, the roadway in 3%, and the vehicle in 2%; the corresponding values from the British study were 65%, 2% and 2% respectively. In nearly all cases the vehicular factor was in fact a vehicle maintenance problem, such as bald tires or worn brake linings. The road user was identified as a sole or contributing factor in 94% of crashes in the US study and in 95% of crashes in the British study.

4.6 Older And Younger Drivers

Much information is available on how various rates depend on age and gender because these variables are nearly always coded in large data sets. Little additional information on the personal characteristics of people involved in road crashes is available, in large part because of privacy concerns.

Demographic projections of increasingly large numbers of increasingly older drivers have generated concerns captured in the phrase “the older-driver problem”. Examining how rates depend on age and gender addresses the older driver problem and the younger-driver problem. Behavior is already identified above as the

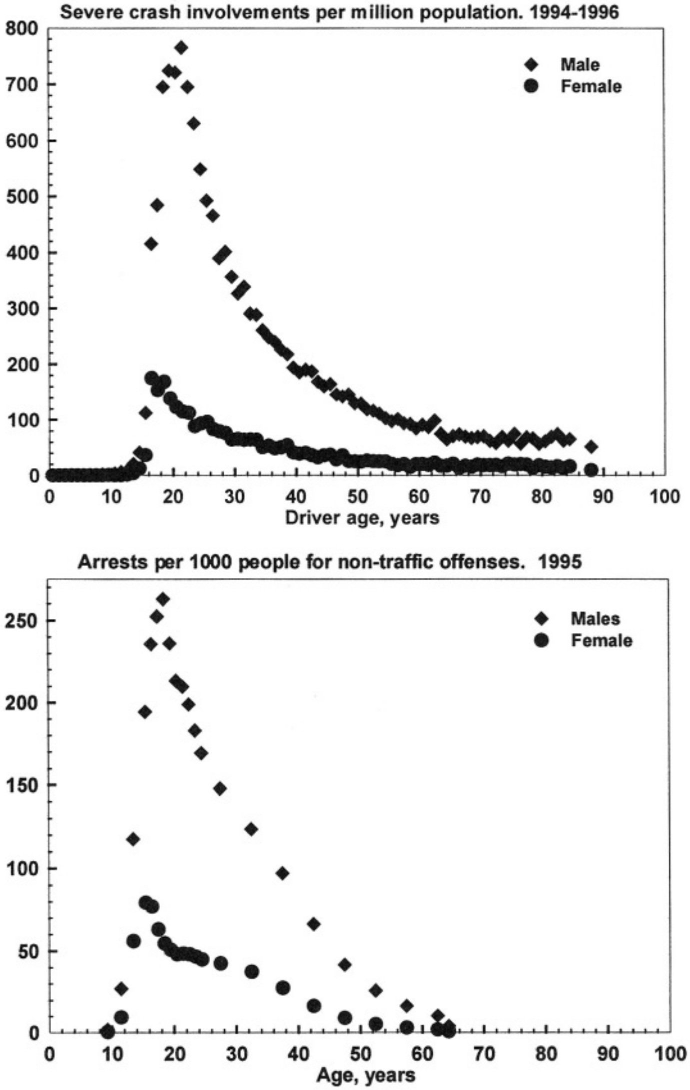


Figure 4-7. *Top:* Estimated driver involvements per capita in severe single-vehicle crashes. *Bottom:* Number of arrests per capita for non-traffic-related offenses.

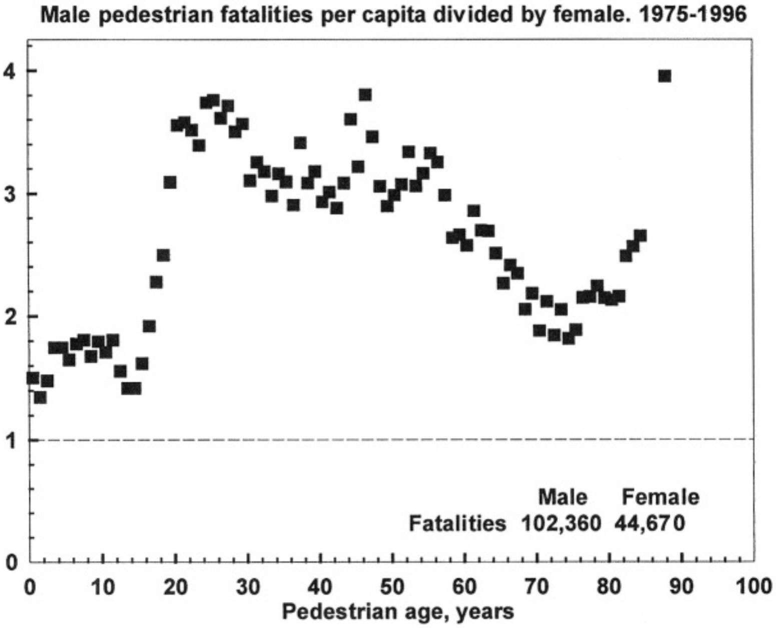


Figure 4-8. Male pedestrian deaths per capita divided by female pedestrians per capita.

as the dominant factor in the elevated rates of younger drivers, but performance becomes more critical with increasing age. Because age effects include *performance* and *behavior* factors, the topic is treated in this separate section. The material presented is based on Evans (1991) and Evans (2000).

Changes in driving risk with increasing age are best separated into two distinct components:

- Changing risks to the drivers themselves, and
- Changing risks they impose on other road users.

These risks are of a different nature. There is near universal agreement that society should take stronger measures to prevent its members from doing things that endanger others than to prevent them from doing things that endanger only themselves. Public safety makes a stronger claim on public resources than does personal safety, which can be supported often using personal resources. Differences between the risks we assume ourselves and those we impose on others impact legislation, licensing policy, police enforcement, and so on.

Changing Risks Drivers Face As They Age

Figures 4-9 to 4-12 show fatality data normalized for the same length of time, the same number of people, the same number of licensed drivers, and for the same distance of travel. Three of the relationships exhibit a characteristic "U-shape," exhibiting particularly strong increases at the oldest ages.

Involvement rates in severe crashes. Increases with age like those in the above figures have often been interpreted in terms solely of the older drivers' risk of involvement in a crash. Such an interpretation misses the crucial point that the number of drivers of given age and gender killed is the product of two factors:

1 The number of involvements in very serious crashes, and

2. The probability that involvement proves fatal.

The first factor reflects influences due to all use and behavioral factors, such as amount and type of driving, driver capabilities, type of vehicle driven, time of day, degree of intoxication, and driving risks. The second factor can be influenced also by such behavioral factors as safety belt wearing and alcohol consumption. Apart from such considerations, the probability that a given crash results in death is essentially physiological rather than behavioral in nature. The graphs that follow are based on the relationships given on page 26 of Evans, 1991. These are not materially different from more recent and more precise estimates (Evans 2001b, 2001c):

$$R_{\text{males}}(A) = \text{Exp } 0.0252\ (A - 20), \text{ and}$$

$$R_{\text{females}}(A) = 1.311 \text{ Exp } 0.0216\ (A - 20)$$

where R(A) is the fatality risk to an individual of age A compared to the risk to an individual of age 20 when both are subject to the same physical insult, or impact. When driver age is 16 to 20, we assume R = 1 for males and R = 1.311 for females; that is, the fatality risk from the same severity crash is the same as for a 20-year-old driver of the same gender. These relationships are applicable from age 20 to age 80. Fatality rates focus on the outcome, not the severity of the crash that led to the death. Figures 4-13 and 4-14 show involvement rates in crashes of similar severity by considering crashes in a severity range greater than or equal to that sufficient to likely kill 80-year-old male drivers, for which case R has a value of 4.0. Comparing Figures 4-13 and 4-11 shows that most of the increase in the fatality rate per licensed driver results from the same severity crash being more likely to lead to death. When this is factored out, an increase at older age remains, but of smaller magnitude. The rate of involvement for the same distance of travel increases with

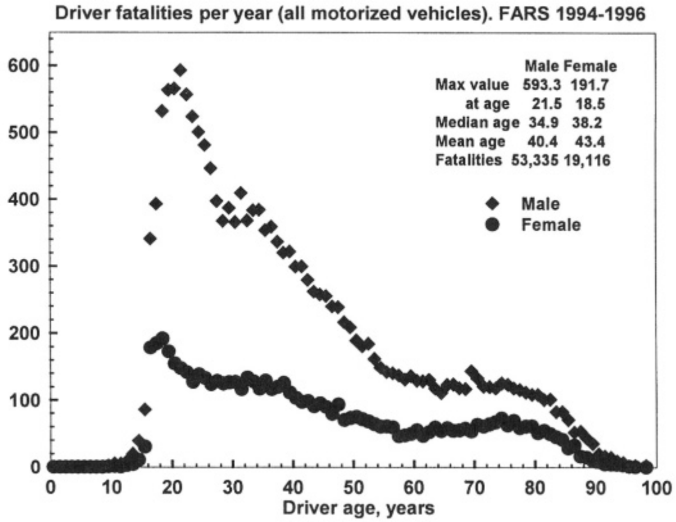


Figure 4-9. Average number of driver fatalities per year (all motorized vehicles) versus gender and age. (Based on FARS, 1994-1996).

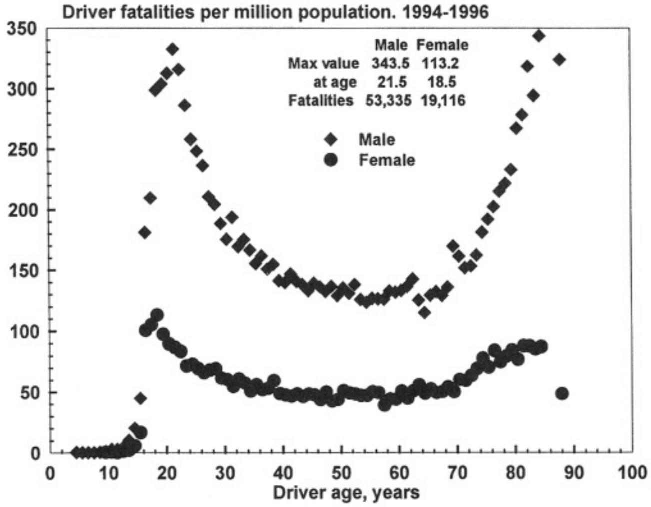


Figure 4-10 Driver fatalities (all motorized vehicles) per million population versus gender and age. (Based on FARS and US Bureau of the Census, 1994-1996).

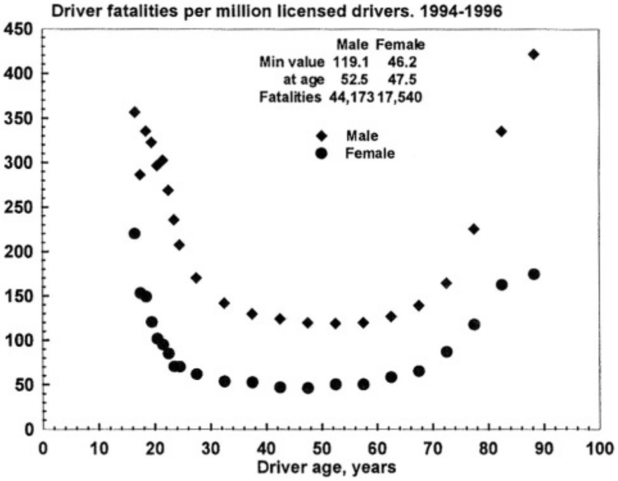


Figure. 4-11. Fatally injured drivers (all motorized vehicles) with valid driver licenses per million licensed drivers versus gender and age. (Based on FARS and Federal Highway Administration data, 1994-1996).

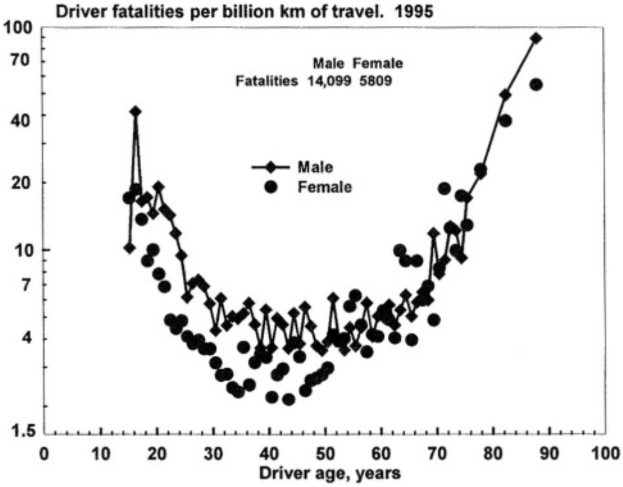


Figure 4-12 Driver fatalities (all motorized vehicles except large commercial trucks) per billion km of travel versus gender and age. (Based on FARS 1995, National Personal Transportation Study 1995).

increasing driver age for ages above about 60. However, the increase is smaller than in Figure 4-12; even at the oldest age plotted, the rates for males and females are still less than those for male drivers under 30.

Threat To Other Road Users

All the above focused on how the age and gender of a driver influence the threat to the driver's own life. Here we investigate the threat to other road users by examining, in Figures 4-15 and 4-16, the number of crashes in which pedestrians are killed as a function of the age and gender of drivers (of any type of motorized vehicle) involved in the crashes. No assumption is made regarding responsibility in pedestrian fatality crashes; the FARS data show about one third of fatally-injured pedestrians have blood alcohol concentrations in excess of 0.1 percent by volume, the legal limit for intoxicated driving in most US states (in Sweden the legal limit is 0.02 percent).

Figures 4-15 and 4-16 may be compared to Figures 4-13 and 4-14. Figure 4-16 indicates that very old drivers may pose an increased risk to other road users for the same distance of driving. However, the risk posed per licensed driver shows no such trend. The difference arises because as drivers age, they drive much less. The similarity of Figures 4-13 and 4-15 supports the interpretation that each is measuring, approximately, the risk of involvement in crashes in general (likewise Figures 4-14 and 4-16).

Table 4-12 addresses the risks that drivers impose on other road users by comparing the rates of 80-year-olds to drivers of age 40 and 20. For male drivers, licensing an 80-year-old poses 26% less risk than licensing a 40-year-old. Licensing a 20-year-old poses 140% more risk than licensing an 80-year-old. In

Table 4-12. **Risks* 80-year-old drivers pose to other road users compared to the risks posed by 40-year-old drivers contrasted to the risks posed by 20-year-old drivers compared to the risks posed by 80-year old drivers. The first two values for male drivers indicate that licensing an 80-year-old driver poses 26% *lower* risk to society than licensing to a 40-year-old driver, whereas licensing a 20-year old male driver poses a 140% *higher* risk to society than licensing an 80-year old driver.**

	Male		Female	
	$\frac{\text{Age 80}}{\text{Age 40}}$	$\frac{\text{Age 20}}{\text{Age 80}}$	$\frac{\text{Age 80}}{\text{Age 40}}$	$\frac{\text{Age 20}}{\text{Age 80}}$
Per licensed driver (Fig. 4-15)	**0.74**	**2.40**	**0.70**	**3.67**
For same distance of driving (Fig. 4-16)	**3.71**	**0.91**	**1.88**	**2.01**

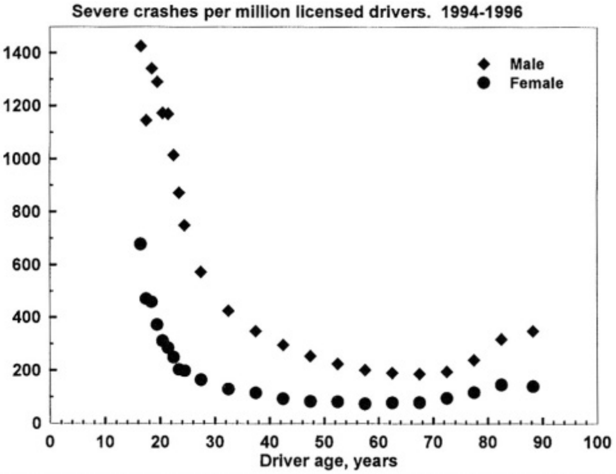


Figure 4-13. Estimated licensed driver involvements (all motorized vehicles) per million licensed drivers in crashes of sufficient severity to likely kill 80-year-old-male drivers versus gender and age of the driver. (Based on FARS and Federal Highway Administration data for 1994-1996).

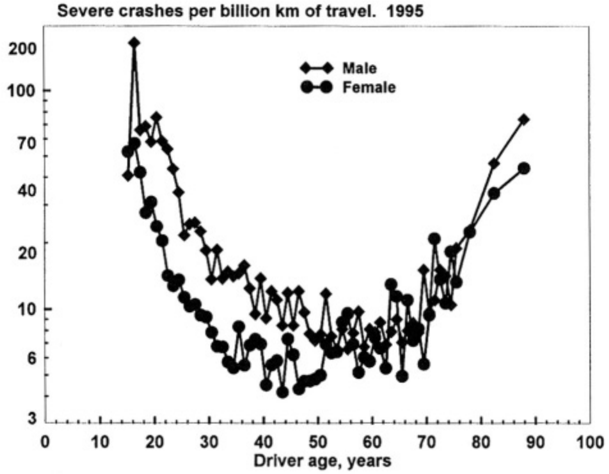


Figure 4-14. Estimated driver involvements (all motorized vehicles) per billion km of travel in single-vehicle crashes of sufficient severity to likely kill 80-year-old-male drivers versus gender and age.

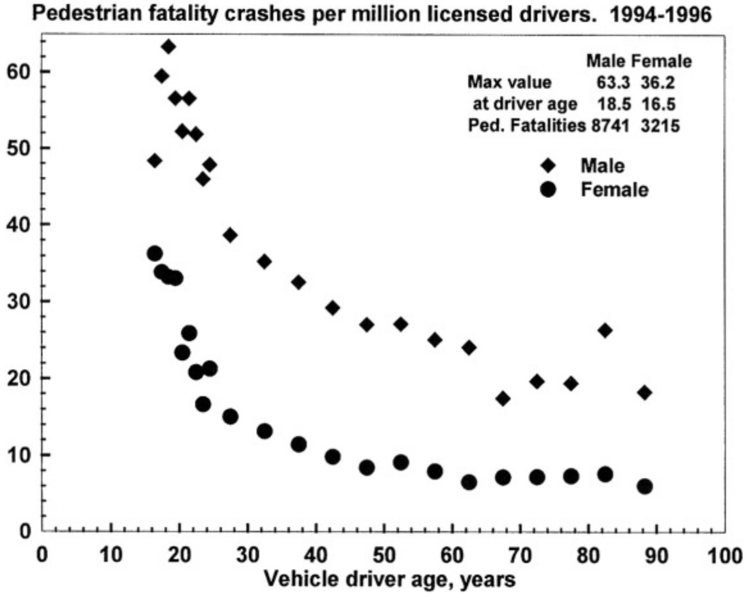


Figure 4-15 Number of single vehicle crashes per million licensed drivers in which one or more pedestrians was killed versus the age and gender of driver.

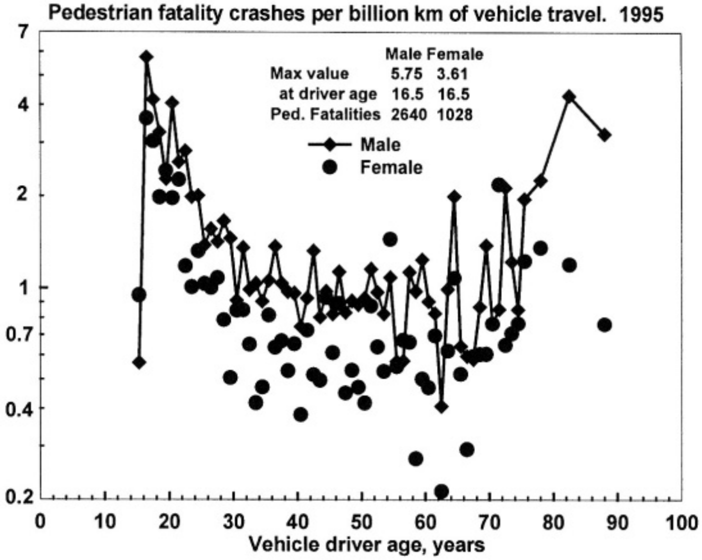


Figure 4-16. Number of single vehicle crashes per billion km of travel in which one or more pedestrians was killed versus the age and gender of the driver. (Based on FARS, and Nationwide Personal Transportation Study, 1995).

terms of the threat posed for the same distance of travel, the 80-year-old is 271% more likely to harm someone else than is a 40-year-old. The 80-year-old male driver's risk is, nominally, larger than the 20-year-old male driver's risk (by 9%). For females drivers, the 80-year-old risk is approximately double that of the 40-year-old, but about twice that for the 20-year-olds per unit distance.

Traffic Deaths Relative To All Deaths

A noticeable feature of the ratio of traffic deaths to all deaths (Fig. 4-17) is the lack of a clear difference between the genders. Indeed, from the 20s through the 70s the fraction of all deaths that are traffic deaths declines at an approximately constant rate of 8% per additional year of life for both genders.

Conclusions Regarding Age Effects

The relationships presented here suggest:

1. Licensing an older driver (data goes up to age 80) does not pose a greater threat to other road users than licensing younger drivers -- indeed it poses substantially less risk than licensing a 20-year-old.
2. As drivers age, most measures indicate that they face an increased risk of becoming a traffic fatality, with the increase accelerating at very old ages.
3. Given that a death occurs, the probability that it is a traffic fatality declines steeply with age, from well over 20% for late teens through mid twenties, to under one percent at age 65, and under half a percent at age 80.

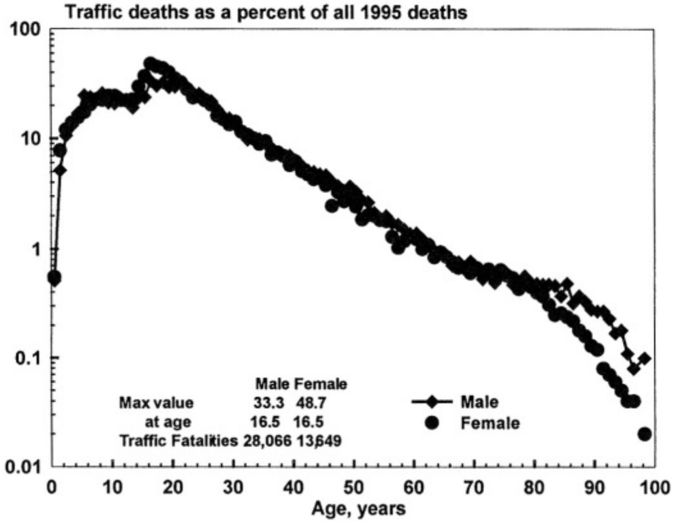


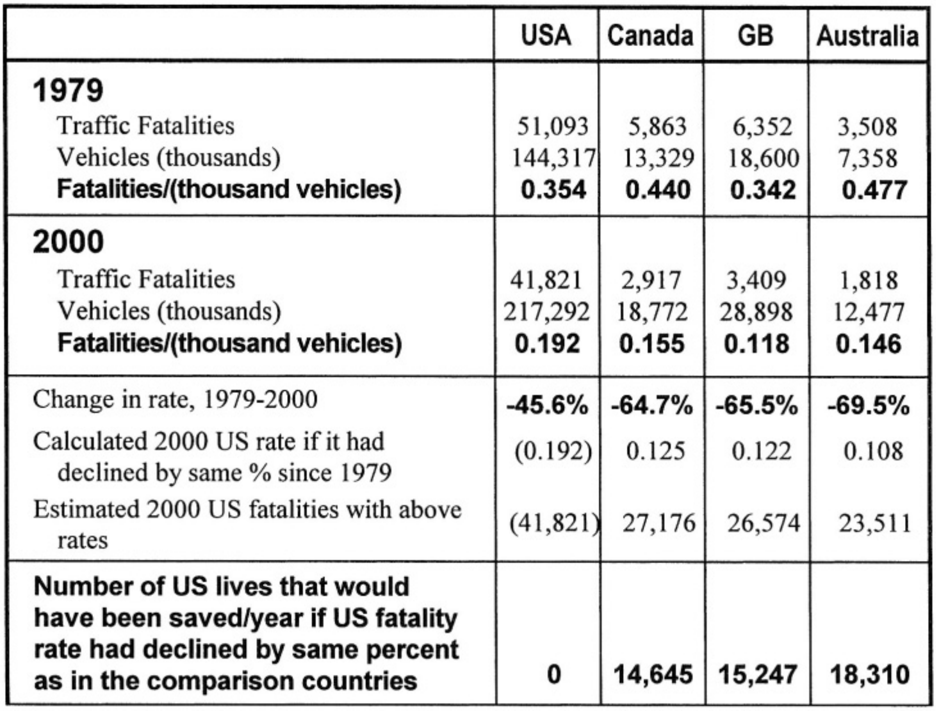
Figure 4-17. Traffic deaths expressed as a percentage (on a logarithmic scale) of total deaths from all causes (including traffic). All data are for 1995.

4.7 US Safety Compared To Safety In Other Countries

In the 1960s the US had, by far, the lowest fatality rates in the world, whether measured by deaths per same distance of travel or per registered vehicle. A series of tabulated rates for the US and 11 other major industrialized countries for the years up to 1978 justified the headline "U.S. the Safest Place for Driving" (Motor Vehicle Manufacturers Association, 1981, p. 52). For every year, the US rate was substantially lower than for any of the other countries listed. The remainder of this section focuses mainly on one rate, the number of deaths per thousand registered vehicles, which will be called the fatality rate. The US began to fall from its leadership position in the late 1970s. Now the International Road Traffic and Accident Database (2001) lists 12 countries (Australia, Canada, Finland, Germany, Iceland, Japan, Luxembourg, Netherlands, Norway, Sweden, Switzerland and the United Kingdom) with rates lower than the US.

Table 4-13 compares US safety in 2000 to safety in 1979, and contrasts the US changes to those occurring in Canada, Great Britain, and Australia. These three comparison countries were chosen because they have much in common with the US.

Table 4-13: Comparison of traffic safety changes, 1979-2000, in the US and other "similar" countries.



	USA	Canada	GB	Australia
1979				
Traffic Fatalities	51,093	5,863	6,352	3,508
Vehicles (thousands)	144,317	13,329	18,600	7,358
Fatalities/(thousand vehicles)	**0.354**	**0.440**	**0.342**	**0.477**
2000				
Traffic Fatalities	41,821	2,917	3,409	1,818
Vehicles (thousands)	217,292	18,772	28,898	12,477
Fatalities/(thousand vehicles)	**0.192**	**0.155**	**0.118**	**0.146**
Change in rate, 1979-2000	**-45.6%**	**-64.7%**	**-65.5%**	**-69.5%**
Calculated 2000 US rate if it had declined by same % since 1979	(0.192)	0.125	0.122	0.108
Estimated 2000 US fatalities with above rates	(41,821)	27,176	26,574	23,511
Number of US lives that would have been saved/year if US fatality rate had declined by same percent as in the comparison countries	**0**	**14,645**	**15,247**	**18,310**

The 45.6% decline in the US rate in this 21-year period might seem substantial. It corresponds to an average reduction of 2.9% per year. However, this is less than the average reduction of 3.2% over the entire prior period, 1900 to 1978.

Canada, Great Britain and Australia all had fatality rate reductions of more than 64%. If the US rate had declined by the same 64.7% experienced by Canada, then 2000 fatalities would have been 27,176, rather than the 41,821 that occurred. That is, 14,645 fewer Americans would have been killed in 2000. Matching the British and Australian performance would have reduced 2000 US road deaths by 15,247 and 18,310 respectively.

The calculation is reasonably robust with regard to choosing other approaches and reference years different by a few years from 1979. A calculation based on the changes in total fatalities from 1979 to 2000 (data in Table 4-13), rather than on the rates, gives similar estimates. For the US, the 2000 fatality count is 23.4% below the peak value of 54,589 attained in 1972. For Canada, Britain and Australia, the corresponding reductions are 56.5%, 57.3% and 52.1% below their respective peaks. All three comparison countries more than halved their peak fatalities. If US fatalities had declined by half of the peak value, the 2000 total would be 27,300. The observed number exceeds this by more than 14,000. The overall conclusion is that if US safety performance had matched that in any one of the three comparison countries, substantially more than ten thousand Americans who were killed in 2000 road traffic would now be alive.

The calculation in Table 4-13 was repeated to compare every year from 1979 through 2000, with the results shown in Table 4-14. Summing over the period gives estimates of the total numbers of American lives that would have been saved over the period 1979-2000 if US safety performance had matched that in the comparison, countries as follows:

If US matched Canada,	196,604 fewer US fatalities
If US matched Great Britain	146,733 fewer US fatalities
If US matched Australia	226,796 fewer US fatalities.

In Britain, the rate for the same distance of travel (Road traffic statistics, 2000) declined by 70.5% from 1979 to 2000, compared to a 54.54% drop in the US (Fig. 4-18). If each year from 1979 through 2000 the number of traffic deaths for the same distance of travel had declined in the US by the same percent as occurred for Britain, then 185,913 fewer Americans would have been killed in the 21-year interval. This is larger than the estimates based on fatalities per vehicle because the average distance traveled per vehicle per year increased more in Britain than in the US from 1979 to 1997. While estimates of distance traveled per vehicle per year are unavailable for Canada and Australia, it is expected that estimates based on fatalities for the same distance of travel would likely also generate corresponding larger estimates of additional US fatalities.

Table 4-14: *Calculated reductions in US traffic fatalities if US fatality rate (in the* comparison countries. The calculation is based on standardizing all rates to the value 1 for 1979.

Year	Observed US fatalities	Calculated US fatalities if US rate change had matched that in:			Calculated reduction in US Fatalities if US rate change had matched that in:		
		Canada	GB	Australia	Canada	GB	Australia
1979	51,093	51,093	51,093	51,093	0	0	0
1980	51,091	47,247	47,652	47,108	3,844	3,439	3,983
1981	49,301	46,761	46,650	46,510	2,540	2,651	2,791
1982	43,945	35,442	46,960	43,734	8,503	-3,015	211
1983	42,589	35,699	42,987	36,636	6,890	-398	5,953
1984	44,257	36,574	44,342	37,698	7,683	-85	6,559
1985	43,825	39,359	41,938	39,771	4,466	1,887	4,054
1986	46,087	36,241	43,336	38,904	9,846	2,751	7,183
1987	46,390	37,562	41,343	37,934	8,828	5,047	8,456
1988	47,087	36,319	39,888	39,861	10,768	7,199	7,226
1989	45,582	36,098	41,699	38,427	9,484	3,883	7,155
1990	44,599	34,631	40,394	31,641	9,968	4,205	12,958
1991	41,508	33,669	36,007	29,437	7,839	5,501	12,071
1992	39,250	31,420	32,659	26,456	7,830	6,591	12,794
1993	40,150	32,773	29,967	26,186	7,377	10,183	13,964
1994	40,716	29,762	28,869	25,759	10,954	11,847	14,957
1995	41,817	31,179	29,160	26,991	10,638	12,657	14,826
1996	42,065	29,193	28,594	25,871	12,872	13,471	16,194
1997	42,013	28,648	28,157	22,913	13,365	13,856	19,100
1998	41,501	27,316	26,797	22,472	14,185	14,704	19,029
1999	41,717	27,638	26,605	22,695	14,079	15,112	19,022
2000	41,821	27,176	26,574	23,511	14,645	15,247	18,310
Total US lives saved if changes in US fatality rate (fatalities per thousand vehicles) had matched changes in the indicated country.					**196,604**	**146,733**	**226,796**

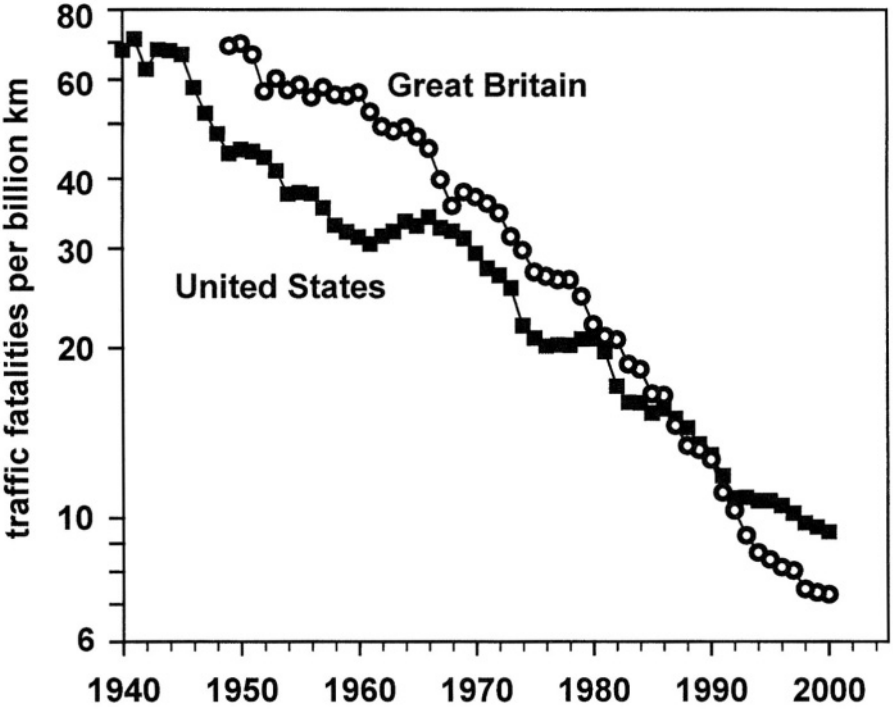


Figure 4-18. Fatalities for the same distance of travel in the US and Great Britain

While there is uncertainty in all the above estimates, they do justify the conclusion that, if US safety performance had been similar to that in any of the three comparison countries, well over one hundred thousand Americans who are now dead would be alive. An additional 100,000+ Americans being killed so overshadows any other transportation safety matter that it is treated in some detail below in an attempt to reach for explanations.

Airbag Mandate And Vehicle Factors At Core Of US Policy

No safety issue has consumed so much time and effort as the requirement that all new cars (plus some other vehicles) sold in the US must come equipped with frontal airbags (hereafter called airbags). This mandate makes the US the only nation in the world whose inhabitants are prohibited from purchasing a new vehicle without an airbag.

Claims

The mandate was enacted because advocates claimed that airbags:-

1. Are passive (require no user knowledge or action)

2. Replace belts (permit vehicles to not have belts)
3. Reduce driver fatality and injury risk by 40%
4. Reduce risk regardless of gender, age, etc.
5. Hurt nobody

Reality

All 5 claims are false.

1. Are passive. Drivers, passengers, and parents must know an ever-increasing list of rules on how to avoid death and injury from deploying airbags. Arguably, airbags are the least passive safety device ever installed on vehicles; they might better be called *belligerent restraints.* The manual belt is far more passive, requiring only one simple rule, "buckle up."

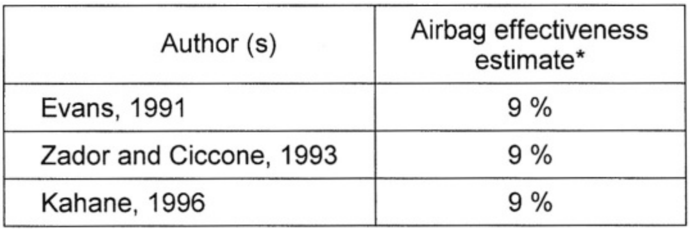
2. Replace belts. The government's estimate of airbag cost included $ 18 saved by not installing belts (FR Doc.77-19137, 1977). Yet today no manufacturer in the world offers airbags as complete occupant protection devices. In all cases they are offered as supplemental devices to increase the effectiveness of the primary occupant protection device, the lap-shoulder belt.

3. Reduce driver fatality risk by 40%. This claim is not merely incorrect – it is absurd, and was know to be absurd when the claims were made. Airbags deploy only in frontal-impact crashes, which are responsible for just over half of fatalities. For an airbag to be 40% effective, its effectiveness in frontal crashes would have to be nearly 80%, a performance level that 1970's knowledge readily dismissed as impossible. The government disparaged a well-executed study (Wilson and Savage, 1973) reporting an overall effectiveness of 18% (the latest government estimate (Kahane, 1996) is 13%). Using data from a fleet of 10,000 airbag-equipped cars sold in the mid 1970s, Pursel et al. (1978) estimated that the airbag alone reduces severe injury risk by 9%.

Table 4-15 summarizes estimates of the effectiveness of airbags in reducing risk to belted drivers (driving unbelted is illegal in all US states except New Hampshire). Barry et al. (1999) claim that these estimates are too high.

The 9% fatality reduction for belted drivers is consistent with the finding (Table 4-9) that adding an airbag increases the effectiveness of safety belts from 42% to 47%, a difference of 5 percentage points. Figure 4-19 clarifies the difference, and shows that as belt use rates increase from 0% to 100%, deaths prevented by airbags decline from 13 to 5 per original 100 fatalities.

Table 4-15. Effectiveness of airbags in reducing fatality risk to belted drivers.



Author (s)	Airbag effectiveness estimate*
Evans, 1991	9 %
Zador and Ciccone, 1993	9 %
Kahane, 1996	9 %

*All estimates based on "naïve assumption" (discussed later) that occupant protection does not influence driver behavior.

4. Airbags reduce risk regardless of gender, age, etc. All the above estimates are averages for all drivers. There is now considerable evidence that airbags increase risk to many large portions of the population, including possibly older drivers (Kahane, 1996). Dalmotas et al. (1996) find that while airbags reduce net harm to males by 12%, they increase net harm to females by 9%. The evidence that airbags increase risks to children is clear (Kahane, 1996, Graham et al. 1998.) Graham et al (2000) find that airbags increase fatality risk to unbelted children by 84% and to belted children by 31%.

5. Airbags hurt nobody. More than 200 people in the US have been killed by the forces of deploying airbags in crashes they otherwise would have survived, in many cases uninjured. The victims have been mainly children and babies in the front passenger seat, and short female drivers. Vastly larger numbers have sustained many other types of injuries, including eye injuries, hearing loss and respiratory disease.

Well prior to the airbag mandate technical information raised questions regarding risks airbags posed to children. Papers were published with titles including *Possible effects of air bag inflation on a standing child* (Aldman, 1974) and *Airbag effects on the out-of-position child.* (Patrick and Nyquist 1972). Yet the agency responsible for mandating airbags writes "air bags will provide substantial crash protection to otherwise unrestrained small children in crashes" (National Highway Traffic Administration, 1980, p. 71). On page 70 of the same document the agency cites, and dismisses, statements by General Motors, based on their own animal testing and other technical considerations, that a "child might be injured by an inflating bag". Ralph Nader, while engaged in promoting airbags, is photographed in July 1977 demonstrating an airbag "safely" deploying into the face of an unbelted three-year old girl (Photograph reproduced in Evans, 2002; see also http://www.scienceservingsociety.com/nader.htm). Airbags in fact increase fatality risk to unbelted children by 84% (Glass et al., 2000). Even for belted children, airbags increase fatality risk by 31% (Glass et al., 2000).

Airbags cause additional harm, including eye injuries (Duma et al., 1996), hearing loss (Yaremchuk and Dobie, 1999; Buckley, 1999) and asthmatic attacks (Gross et al., 1994; 1995)

All the above relates to frontal airbags. Many manufacturers now offer side airbags. It is difficult to see how they could be more than about 10% as effective as frontal airbags, given how much less deployment space is available. This means that the theoretical maximum reduction in overall occupant fatality risk can be no more than a percent or so. It seems almost inevitable that a child asleep against the deploying unit will be killed.

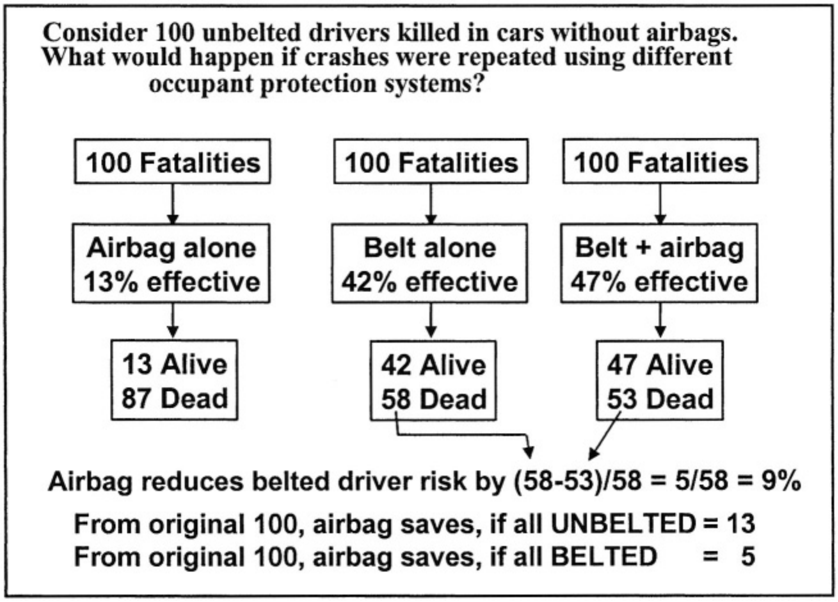

Consider 100 unbelted drivers killed in cars without airbags.
What would happen if crashes were repeated using different occupant protection systems?
100 Fatalities
100 Fatalities
100 Fatalities
Airbag alone 13% effective
Belt alone 42% effective
Belt + airbag 47% effective
13 Alive 87 Dead
42 Alive 58 Dead
47 Alive 53 Dead
Airbag reduces belted driver risk by (58-53)/58 = 5/58 = 9%
From original 100, airbag saves, if all UNBELTED = 13
From original 100, airbag saves, if all BELTED = 5

Figure 4-19. Number of belted and unbelted driver lives saved by airbags, (based on the "naïve assumption" that occupant protection does not influence driver behavior).

Airbag Mandate and the Technology/Human Interface

The conclusion that airbags reduce belted driver risk by 9% refers strictly to the change in risk, given identical numbers of identical crashes. From this, one cannot infer the change in fatalities due to a policy requiring universal airbag installation. Such an inference requires a crucial, and false, assumption (one which is implicitly included, without comment, in all estimates of lives saved by airbags). The false assumption is that beliefs about airbag effectiveness have zero effect on driver behavior.

Table 4-16 compares two models of how technological changes affect safety. An analysis of 24 studies (Evans, 1996) shows that the naive model can be grossly in error, even to the point of estimating increases in safety when reductions actually occur, and vice versa.

Driver behavior changes have been reliably observed in response to technologies that provide clear feedback. Anti-lock brakes provide a clear example (Evans, 1999, Farmer et al., 1997). For technologies that affect only injury risk, behavior effects are expected to be smaller and therefore difficult to measure. While behavior responses to injury risk are generally difficult to determine empirically, the following construct establishes that they occur.

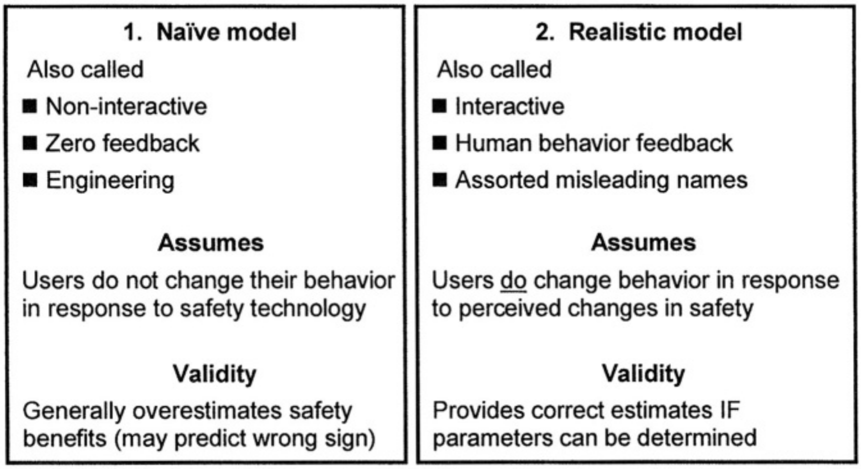
Consider two hypothetical cars, identical in all respects except that one has the magical property that its occupants cannot be hurt in any crash, while the other is wired with dynamite to explode on the slightest impact. No one would claim that the two cars would be driven in identical ways. The same conclusion applies even if the cars were in fact identical, but falsely believed to possess the hypothesized properties. Changes in perceived protection can be viewed as lying along a continuum bounded by hypothetical extremes.

While there are no empirical estimates of changes in driver behavior in response to US airbag policy, the considerations below suggest that the airbag mandate not only increased driver risk-taking, but by more than the meager actual benefit of the device in a crash. For over 30 years the public was inundated with messages grossly overestimating the benefits and ignoring the negatives of airbags. It was widely believed that airbags were so effective that belt wearing was unnecessary. Slow-motion movies convinced many that in a crash they would glide forward into the gentle caress of a soft cushion. Such massive inputs must lead to outputs, most likely including less belt wearing and faster speeds.

When it became clear that airbags were killing short ladies, a number of short ladies told me "When I discovered the airbag could kill me, I started to drive more cautiously." If one accepts this statement, it is hard to dispute the corresponding conclusion that a belief that the airbag dramatically reduces risk must lead to less cautious driving.

If beliefs about airbags led to an undetectable 3% increase in average speed, a 13% increase in fatality risk would result. Instead of reducing fatalities by 9%, the intervention would increase fatalities by 4%. Government calculations, based on assuming the naïve model, that airbags have saved over 2,000 lives (mainly of unbelted male drivers) in the more than ten years since 1986, should not be accepted even as gross approximations. They are, however, closer to the truth than the claims made to support the airbag mandate (Federal Register 1977) that driver and passenger airbags would prevent 12,100 deaths per year (or 120,000 per decade, the unit apparently adopted today in order to associate larger numbers of lives saved with airbags). Even assuming the naïve model, the calculated benefits are relatively small, and, as discussed above (Figure 4-19), will decline sharply as belt-wearing rates increase.

Table 4-16. Contrast between the naïve model, which ignores the technology/human interface, and the realistic model, which attempts to take into account the technology/human interface.



1. Naïve model	2. Realistic model
Also called ■ Non-interactive ■ Zero feedback ■ Engineering	Also called ■ Interactive ■ Human behavior feedback ■ Assorted misleading names
Assumes	**Assumes**
Users do not change their behavior in response to safety technology	Users <u>do</u> change behavior in response to perceived changes in safety
Validity	**Validity**
Generally overestimates safety benefits (may predict wrong sign)	Provides correct estimates IF parameters can be determined

The question "Did the US airbag mandate increase or decrease traffic fatalities?" cannot be answered without knowing the magnitude of its effect on driver behavior. However, the airbag mandate offers insight into why 100,000 more Americans died in traffic than would have been killed if US safety performance had matched that of Canada, Britain, or Australia.

Priorities In US Safety Policy

The relative contributions of different factors to traffic safety discussed earlier are synthesized in the non-quantitative sketch in Figure 4-20. Not reflected in this sketch is another large and fundamental distinction between engineering and human-factors interventions. Even if a regulated vehicle design change actually reduces risk, it takes a number of years to incorporate it into a vehicle, and another decade before essentially all vehicles on the road have it. Belt wearing and drunk-driving legislation start reducing harm from the time the laws take effect (perhaps even from the time it is discussed).

While other countries formulated effective policies consistent with Figure 4-20, US priorities were ordered almost perfectly opposite to where benefits are greatest. An obsessive focus on the airbag mandate and on minimally important vehicular factors misled the public into making more dangerous choices than would otherwise have occurred, and largely precluded the adoption of effective countermeasures.

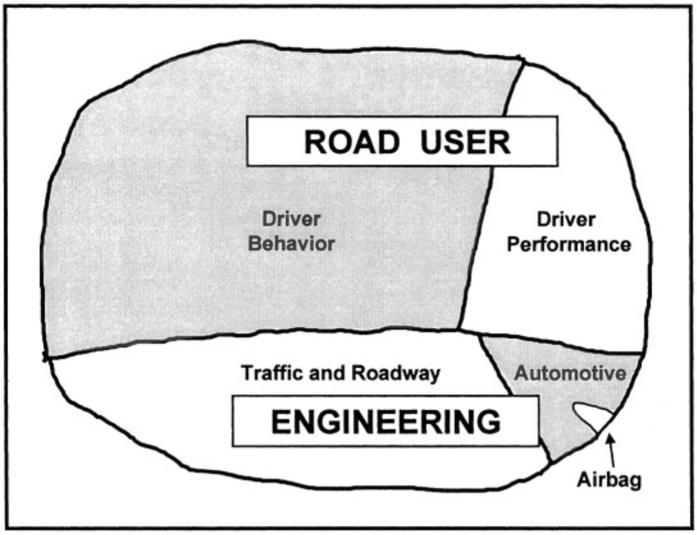


Figure 4-20. Sketch of non-quantitative judgmental estimate of relative importance of different factors.

The decision makers, and those whose council they welcomed, neither understood nor respected technical information. By the early 1970s there was already more than sufficient information in the technical literature to dismiss the claims of airbag performance offered to justify the mandate.

Why does technical knowledge influence safety policy in the US so much less than in other motorized countries? It seems to me that the explanation is, in part, because no other nation bears a burden resembling the US legal system. In other democracies, elected legislators with varied backgrounds are influenced by inputs from diverse sources, including the technical community. In the US, lawyer legislators get nearly all their inputs from other lawyers. It is therefore not too surprising that measures that open deep pockets for legal assault are more appealing than measures, which reduce harm. Only the most gullible can imagine that any net good emerges from the resulting system which lavishly supports an enormous community of the richest people in America, "expert" witnesses, consultants skilled at identifying jurors lowest in knowledge and reasoning skills, and a vast court superstructure, all of which are, in their scope, unknown anywhere else on earth. Even advocates of the US system rarely conclude that US cars must be much safer than Swedish cars because the US spends astronomically more per capita on litigation than does Sweden.

It is only in the US that traffic crashes serve as catalysts to transfer vast wealth from the public to the legal system. Even if all alleged defects in the engineering or manufacture of the vehicle, road, or traffic control system could be miraculously fixed, it is hard to see how this could reduce fatalities by as much as a percent or so. However, I am convinced that indirect effects of US litigation are enormously greater. The broad message from so much litigation is that crashes flow from the failings of asset-rich institutions, a factor over which drivers have no direct control. Even believing a little in this may tend to make drivers less responsible and careful, factors that have an enormous influence on crash risks.

It is only in the US that citizens asked to identify anyone important in traffic safety would produce a list comprised almost exclusively of lawyers. It was a lawyer, Joan Claybrook, who, when head of the National Highway Traffic Safety Administration (NHTSA) from 1977 to 1980, spearheaded the airbag mandate with the five false claims discussed above. A NHTSA official is quoted as saying "Joan came to NHTSA with a mission and that mission was air bags" (Graham, 1989, p. 109).

In a November 1983 television interview Joan Claybrook says of airbags:

> **"They're much better than seat belts, according to the government's most recent data"**

and continues to dismiss safety belts as

> **"the most rejected technology we have. So I believe that airbags would add a great dimension to cars and car safety, would protect all front seat occupants in those types of crashes where 55% of the public is now killed"**

Claybrook continues

> **"Airbags are really the best solution -- they fit all different sizes and types of people, from little children up to 95th percentile males, very large males. ... So they really work beautifully and they work automatically and I think that that gives you more freedom and liberty than being either forced to wear a seat belt or having a car that's not designed with the safety engineering we know today."**

The main pressure to retain the airbag mandate and keep it the focus of national safety attention, rather than let consumers choose whether or not they wish to purchase the device, still comes from the non-technical lawyers responsible for the mandate. They now have allies in the massive airbag industry, which has been likened to the military/industrial complex of an earlier era. What industry would not enthuse over a government requirement that everyone must purchase their expensive product, regardless of whether they want it or are even prepared to pay to have it disconnected? The purchasers of the vehicles that comprise the current US fleet paid about 25 billion dollars for airbags. A microscopic fraction of such a sum properly applied could generate far larger safety benefits than those claimed (falsely) for airbags.

The uniquely US fixation on vehicle factors can be traced to the mid-1960s efforts of lawyer Ralph Nader and his proteges, including Joan Claybrook. Legislation focussing overwhelmingly on the vehicle followed, starting with the 1996 National Traffic and Motor Vehicle Safety Act and the Highway Safety Act. As discussed above, it takes about a decade or so before any reductions claimed for vehicle safety improvements could begin to show. Let us refer to the period before about the mid 1970s as the *Pre-Naderite* period, and the period after about the mid 1970s as the *Post-Naderite* period. During the *Pre-Naderite* period, US traffic was, by a large margin, the safest in the world. In the *Post-Naderite* period the US has dropped from the number one ranking to number 13, and is still sinking. As a result of the US following the lead of lawyers rather than adopting policies illuminated by technical understanding, well over ten thousand additional Americans are being killed in traffic each year. The number, equivalent to an additional 30 deaths per day, will increase if current trends continue.

One of the great ironies is that the very same lawyers responsible for this disaster continue to exercise decisive influence to keep US safety policy on the same wrong track. What is even more ironic is that the media continue to respectfully refer to these same non-technical architects of policies that have killed over a hundred thousand Americans as *safety advocates.*

4.8 Bibliography

Many of the topics in this chapter are treated in greater detail in *Traffic Safety and the Driver* by Leonard Evans (Van Nostrand Reinhold, NY, 1991). In order to reduce repetition, the absence of a citation in the text implies that additional information and references are available in this book. *Traffic Safety and the Driver* is available from *amazon.com, bn.com* and directly from the author at *scienceservingsociety.com.* Many of the themes treated here will be expanded in the author's forthcoming book *Traffic Safety,* expected in 2003.

4.9 References

Aldman, B. (1974). Possible effects of air bag inflation on a standing child. Proceedings of the 18th Annual Conference of the American Association for Automotive Medicine.

American Automobile Manufacturers Association (1997) Motor vehicle facts and figures. Detroit, Michigan.

Barry, S., Ginpil, P and O'Neill, T.J. (1999) The effectiveness of airbags. *Accident Analysis and Prevention,* **31**, 781-787.

Bollen, K.A.; Philipps, D.P. (1981) Suicidal motor vehicle fatalities in Detroit: a replication. *Am. J Sociology* **87**, 404-412.

Buckley, G., Setchfield, N. and Framption, R. (1999) Two case reports of possible noise trauma after inflation of air bags in low speed car crashes. *British Medical Journal* **318**, 499-500

Bureau of Transportation Statistics, National Transportation Statistics 2000, http://www.bts.gov/programs/btsprod/nts/ (consulted 8 Dec 2001).

Company captain (2001) http://www.execpc.com/~reva/ejsmith.htm (consulted 8 Dec 2001)

Pless, B., Davis R.M. (2001) BMJ bans "accidents": Accidents are not unpredictable, *British Medical Journal* **322**,1321 -2

Dalmotas, D.J., Hurley, J., German, A and Digges, K (1966) Air bag deployment crashes in Canada. Paper 96-S1O-05, 15th Enhanced Safety of Vehicles Conference, Melbourne, Australia, 13-17 May 1996.

Department of Environment, Transport and the regions (1998) Road accidents Great Britain: The 1997 Report; The Statistical Office.

Doege, T.C. (1978) Sounding board -- an injury is no accident. *New England Journal of Medicine* **298**, 509-510.

Duma, S.M., Kress, T.A., Porta, D.J., Woods, C.D., Snider, J.N., Fuller, P.M., Simmons, R.J. (1996) Air Bag Induced Eye Injuries: A Report of 25 Cases. *The Journal of Trauma,* **41**, 114-119.

Evans, L. (1991) *Traffic Safety and the Driver.* New York, Van Nostrand Reinhold. (Copies may be obtained from *amazon.com* (etc.) or directly from the author).

Evans, L. (1994) Small cars, big cars: what is the safety difference? *Chance -- New Directions for Statistics and Computing,* a journal of the American Statistical Association. Vol. 7, No. 3, p 9-16.

Evans, L. (1995) Car size and safety -- a review focused on identifying causative factors. Proceedings of the 14th Enhanced Safety of Vehicles Conference, Munich, Germany, 23-26 May 1994, US Government Printing Office: 1995-381-067, Vol. 1, p 721-733.

Evans, L. (1996) Traffic safety measures, driver behavior responses, and surprising outcomes. 5th Westminster Lecture, Parliamentary Advisory Council for Transport Safety (PACTS), Queen Elizabeth II Conference Center, London SW1, 6 December 1994; *Journal of Traffic Medicine* **24**,5-15.

Evans, L. (1997) A Crash Course in Traffic Safety. *Encyclopœdia Britannica Medical and Health Annual,* p. 126-139.

Evans, L. (1999) Antilock Brake Systems and risk of different types of crashes in traffic, *Crash Prevention and Injury Control,* **1**, 5-23

Evans, L. (2000). Risks older drivers face themselves and threats they pose to other road users. *International Journal of Epidemiology,* **29**: 315-322. (available at http://www.scienceservingsociety.com/pl32.pdf)

Evans, L. (200la). Causal influence of car mass and size on driver fatality risk, *Am J. Public Health,* **91**, 1076-81. (available at http://www.scienceservingsociety.com/pl38.pdf)

Evans, L. (2001 b). Age and Fatality Risk from Similar Severity Impacts. *J Traffic Medicine,* **29**: 10-19. (available at http://www.scienceservingsociety.com/pl35.pdf)

Evans, L. (2001c). Female Compared to Male Fatality Risk from Similar Physical Insults, *J Trauma,* **50**: 281-288. (available at http://www.scienceservingsociety.com/pl36.pdf)

Evans, L. (2002). Traffic Crashes, *American Scientist,* **90**,:244-253. (see also http://www.scienceservingsociety.com/nader.htm).

Farmer, C.M.; Lund, A.K.; Trempel, R.E; Braver, E.R. (1997) Fatal crashes of passenger vehicles before and after adding antilock braking systems. *Accident Analysis and Prevention,* **29**, 745-757.

Federal Register (1977) Federal Register, Vol. 42, No 128, Part 571 – Federal Motor Vehicle Standards: Occupant protection systems, Docket No. 75-14, Notice 10, pages 34289-34305, 5 July 1977.

FR Doc.77-19137 (1977) Federal Register, Vol. 42, No 128, July 5, 1997, Table III, page 34299.

Glass, R.J., Segui-Gomez, M., and Graham, J.D. (2000) Child passenger safety: Decisions about seating location, airbag exposure, and restraint use. *Risk Analysis,* **20**, 521-527.

Graham, J.D. (1989) *Auto safety: Assessing America's performance,* Auburn House Publishing Company, Dover, Massachusetts.

Graham, J.D., Goldie, S.J., Segui-Gomez, M, Thompson, K.M., Nelson, T., Glass, R., Simpson, A and Woerner, L.G. (1998) Reducing risks to children in vehicles with passenger airbags, *Pediatrics* **102**, 1 July 1998. (see also http://www.pediatrics.org/cgi/content/full/102/1/e3).

International Road Traffic and Accident Database (OECD) (2001). http://www.bast.de/htdocs/fachthemen/irtad/english/englisch.html

Hernetkoski, K.; Keskinen, E. (1998) Self-destruction in Finnish Motor Traffic accidents in 1974-1992, *Accident Analysis and Prevention,* **30**, 697-704.

Hingson, R., McGovern, T., Howland, J., Heeren, T, Winter, M and Zakocs (1996). Reducing alcohol-impaired driving in Massachusetts: the Saving Lives Program. *Am J. Public Health,* **86**, 791-797.

Kahane, C.J. (1996) Fatality reduction by air bags: Analysis of accident data through early 1966, NHTSA technical report HS 808 470, US Department of Transportation, Washington, DC.

Kloeden, C.N., McLean, A.J., Moore, V.M., and Ponte, G. (1997) Travel speed and the risk of crash Involvement. Report VR 172, Federal Office of Road Safety, GPO Box 594, Canberra ACT 2601, Australia, Volume 1: Findings, November 1997.] (see also http://www.raru.adelaide.edu.au/ruralspeed/RURALSPEED.PDF for a related study).

Langley, J.D. (1988) The need to discontinue the use of the term "accident" when referring to unintentional injury events. *Accident Analysis and Prevention* **20**, 1-8.

Motor Vehicle Manufacturers Association of the United States (1981) MVMA *Facts and Figures 1981,* Detroit, Michigan.

National Highway Traffic Safety Administration (1980) Automobile occupant crash protection, Progress report no. 3, US Department of Transportation, DOT-HS-805-474

National Safety Council (1997) *Accident Facts,* Chicago, IL

Ohberg, A., Penttila, A. and Lonnqvist, J. (1997) Driver suicides. *British Journal of Psychiatry,* **171**, 468-472

Patrick, L.M. and Nyquist, G.W. 1972. Airbag effects on the out-of-position child. SAE paper 720442. Society of Automotive Engineers; Warrendale, PA:

Philipps, D.P. (1979) Suicide, motor vehicle fatalities, and the mass media: evidence towards a theory of suggestion. *American Journal of Sociology* **84**, 1150-1174.

Pursel H.D.; Bryant R.W.; Scheel J.W.; Yanik A.J. (1978) Matched case methodology for measuring restraint effectiveness. SAE paper 780415. Warrendale, PA: Society of Automotive Engineers.

Road Traffic Statistics: 2000, Transport Statistics Bulletin, Statistics report SB (01) 19 http://www.transtat.dtlr.gov.uk/tables/2001/rts/pdf/rts.pdf (consulted 23 Jan 2002)

U.S. Department of Transportation (2001), National Transportation Statistics 2000, Bureau of Transportation Statistics, BTS01-01 (Washington, DC, U.S. Government Printing Office) http://www.bts.gov/programs/btsprod/nts/

Wilson, R.A.and Savage, C.M. (1973) Restraint system effectiveness -- a study of fatal accidents. Proceedings of Automotive Safety Engineering Seminar, sponsored by Automotive Safety Engineering, Environmental Activities Staff, General Motors Corporation; 20-21 June 1973.

Wood D. P. (1997) Safety and the car size effect: A fundamental explanation. *Accident Analysis and Prevention,* **29**, 139-151.

Wood, D.P. and Simms C.K. (2002) Car size and injury risk: a model for injury risk in frontal collisions. *Accident Analysis and Prevention* **34**, 93-99

WHO (2001). A 5-Year WHO strategy for road traffic injury prevention. Unpublished document WHO/NMH/VIP/01.03; available from Department of Injuries and Violence Prevention, World Health Organization, 1211 Geneva 27, Switzerland.

Yaremchuk, K and Dobie, R. (1999) The otologic effects of airbag deployment. *Journal of Occupational Hearing Loss,* **2**, 67-73

Zador, P.L. and Ciccone, M.A.. (1993) Automobile driver fatalities in frontal impacts: air bags compared with manual belts. *Am J Public Health.* **83**, 661-666.

5 TRANSPORTATION QUEUEING

Randolph W. Hall

5.1 Introduction

Since the time that humans first gathered into societies, there have been queues. They have existed whenever people have demanded more of a service than that service could provide. Though queueing is by no means new, the study of queues is relatively recent, dating only to the beginning of the twentieth century and the work of A.K. Erlang (Brockmeyer *et al,* 1948). Erlang's investigations centered on determining capacity requirements for telephone systems, a then very new technology. Even to this day, much of the research in queueing has been directed at applications in communication. The first textbook on the subject, *Queues, Inventories and Maintenance,* was written in 1958 by Morse. The first textbook focusing on queueing applications in transportation (*Applications of Queueing Theory*) was written by Newell in 1971.

Research on queueing in transportation has evolved in its own distinct direction, in part due to the influence of Newell's work, and in part due to the unique aspects of transportation systems. Unlike applications of queueing in communication or production, queues in transportation tend to be much more predictable and, as a consequence, much of the research on queues in transportation has been directed at non-stationary (time varying) systems. Non-stationarities arise in transportation because:

- People prefer to travel at set times of the day and week, largely corresponding to their work schedules. These demand surges create much of the queueing in transportation, and

- In many transportation systems (e.g., mass transit, trucking, railroads and intersections), customers are served in bulk.

From the standpoint of capacity provision, transportation often relies on major investments in infrastructure, such as roadways, runways or railroad lines. Infrastructure intensive systems have only limited latitude for adjusting capacity to responsd to fluctuating demand. Thus, queues recur at known times when customers arrive at a faster rate than the infrastructure can accommodate.

Another unique aspect of transportation is that the customer service mechanism is often defined by the spacing between vehicles along a guideway, and not by how quickly a person or piece of equipment can process customers. Thus, the time to serve a customer is determined by the customer's behavior. A queueing system also behaves as a continuum of serial servers, with extremely short service times, interacting with each other. Therefore, the system model depends not only on the number of customers that queue at a particular location, but the physical length of that queue, and whether that queue spills back into other servers. These phenomena are the core subject matter of *traffic flow theory,* covered in Chapter 6.

Finally, transportation is different from most other queueing applications because the service mechanism is frequently government owned. As a consequence, pricing normally is not used to level out demand patterns, and there tends to be much less flexibility in varying capacity to match fluctuating demand.

Most textbooks in queueing theory emphasize modeling stochastic characteristics of queues that occur in steady-state (i.e., the probability distribution for the state of the system is not time dependent). Unfortunately, for the reasons mentioned above, this theory is not always relevant to transportation. Instead, queueing models in transportation are more likely to concentrate on the non-stationary characteristics of queueing, as well as on the optimization of system design and system control. Examples include:

- Determining the best cycle length and phase lengths for traffic signals.

- Evaluating the consequences of adding lanes or changing the geometric configuration of a highway on "recurrent" (peak period) and "non-recurrent" (incident produced) delay.

- Optimizing the frequency at which buses or trucks should be dispatched along a route, taking cost of operation and service quality into consideration.

- Determining how many service vehicles are needed to respond to randomly occuring demand that is spread over a service region.

5.2 Elements of a Queueing System

Queueing systems are defined by three elements: customers, servers and queues. Customers are the persons or things that await service. They can be travelers, or the vehicles that they travel in. Customers can also be the good, piece of freight or container that is being shipped. The server is the resource that provides the service to the customer. It could be a piece of roadway, a bus, or gate in an airport, to name a few examples. The queue is the group of customers waiting to be served, along with the place they are waiting. Queues can occur as orderly lines, but they also can be groups of customers spread out in a terminal waiting area or perhaps a warehouse. All queueing systems have customers and servers, though occasionally they don't have queues. This occurs when the system refuses to accept customers when they cannot be served immediately.

The performance of the queueing system is defined by the arrival process, service process and queue discipline. The arrival process represents the time pattern by which customers enter the queueing system. Arrival processes in transportation are usually non-stationary, meaning the average arrival rate varies in some predictable way. Arrival processes also exhibit some level of stochastic variation, which is usually represented by the probability distribution for the inter-arrival time (the time separation between two successive arrivals). The service process represents the time and resources required to serve a customer. Service process, like arrival processes, exhibit stochastic variations and often non-stationary patterns (when capacity varies by time). The service time can also depend on the type of customer. The queue discipline is the rule for sequencing customers. Typically, this is a first-come-first-serve pattern. However, other disciplines are used to account for priorities, or to group customers for efficiency (such as a traffic signal, which groups by turn pattern).

Queueing systems are important in transportation because of their effects on customers, and because of the cost of providing the service. The dominant effect is delay, which might be measured in such ways as "time in system", "average speed," or "waiting time." Fundamentally, queueing analysis is used to determine the difference between how long it takes to complete a trip, and how long it would have taken if there were no queueing or congestion. The following are examples of the performance measures that can be predicted with queueing models or measured in the field:

Throughput: Rate at which customers are processed by the system

Crowding or Congestion: Separation between customers, or density of customers (e.g., vehicles per lane-mile of roadway).

Lost Customers: Number of customers that do not travel because of queueing.

Queue Percentage: Percentage of customers that encounter a queue prior to receiving service (instead of being served immediately).

Service Cost: The annual or per customer expense of providing the service.

Productivity: In some cases, the productivity of the server depends on the amount of queueing and whether the system is saturated.

In some instances, queues are stochastic, reflecting a momentary surge in demand or drop in capacity. In others, queues are predictable, following a regular daily pattern. And in some cases queues are perpetual, being present whenever a facility is open for business. One of the objectives in designing a queueing system is remove perpetual and predictable queues, and then to minimize the occurrence of stochastic queues.

5.3 History of Research on Transportation Queueing

Nearly all of the published research on queueing in transportation is motivated by a modal application, such as vehicles on roadways or mass transit. Nevertheless, there is considerable cross-over in concepts and methods between modes. This section provides a few examples.

Traffic: Vehicular Flow on Highways

Controlled access highways were first constructed in the 1930s and 1940s, and only became widely available in the United States in the 1950s and 1960s. Even in the 1990s, they are uncommon in many parts of the world. Research on queueing on highways paralleled this pattern, with the 1950s and 60s seeing a surge of activity, with more or less steady activity ever since. To this day, problems in highway traffic flow have influenced our understanding of queueing phenomena more than any other mode of transportation (for instance, see Hankin and Wright, 1958; Lovas, 1994; and Older, 1968; as examples of how vehicular traffic research has influenced modeling of pedestrian traffic). Its three greatest contributions have been: (1) modeling speed and capacity as functions of vehicle concentrations, (2) modeling the formation and size of queues with shock waves, and (3) application of cumulative diagrams to represent non-stationary phenomena. Secondarily, it has stimulated thinking on congestion pricing, though this research has yet to be applied in a significant way.

Queues on highways are typically manifest in slowed, rather than completely stopped, traffic, making queues difficult both to count and model. It was observed as early as 1935 (Greenshields), that traffic has a natural tendency to slow as the vehicle concentration (vehicles per unit length of roadway) increases, because vehicles naturally reduce speed to provide safe spacing. Extremely large concentrations only occur under jammed conditions, when both vehicle speeds and vehicle flows (product of concentration and speed) are small. Vehicle flows are maximized at moderate concentrations, when vehicle speeds are only slightly impeded by congestion. The maximum flow value is referred to as the highway capacity.

Lighthill and Whitham (1955) and Richards (1956) used speed/concentration curves in their kinematic wave theory to model the formation of queues behind roadway bottlenecks – that is, places where capacity is lower than upstream or downstream sections. The end of a queue is modeled as a shock-wave, representing an abrupt change in traffic density and speed. So long as traffic arrives at the bottleneck at a faster rate than its capacity, the shock-wave grows upstream.

The 1950s is notable for introducing concepts from physics into the study of traffic queues, as in the kinematic wave theory of Lighthill and Whitman, and and also the thermodynamic theories of Newell (1955). It also was a period that established traffic science, and more generally transportation science, as a field of research that blends empirical and theoretical investigation. This is especially evident in the work of Wardrop (1952), Edie (1956), Edie and Foote (1958) and Edie (1961), and Herman *et al* (1959). Edie and Foote's investigations are especially famous, and are based on extensive data collection on traffic flows and speeds in the Holland and Lincoln tunnels in New York.

Non-stationary phenomena are critical to analysis of queueing on highways, due to peaking of traffic during commute periods. This type of queueing is sometimes called "recurrent congestion", as it occurs on a daily basis. Recurrent congestion is distinguished from "non-recurrent congestion", representing delay caused by random occurrences, such as accidents. Considerable research has been devoted to analyzing the effects of random incidents on highway, often by the same basic methods as non-stationary phenomena. However, research on vehicular queueing usually does not account for random variations in inter-arrival or service times, as is common in mainstream queueing literature. Queueing caused by this type of randomness is viewed as secondary relative to queues caused by accidents or queues caused by non-stationary traffic patterns.

Cumulative diagrams have been a part of the traffic science literature for some time as a representation of non-stationary phenomena. They are used to show the cumulative count of vehicles passing a point along a roadway, but they are applied more generally in queueing to represent cumulative counts of arriving and departing customers. They can be used to measure vehicle concentrations, queue sizes, travel times and delays. They are used to model empirically observed processed (i.e., based on actual counts) and also to deterministically model average system performance. Finally, they are used to represent non-recurrent incidents by randomizing event times, durations and magnitudes. The methodology is documented in the texts by Newell (1971, 1982) and Hall (1991), and later in this chapter.

According to Newell (1993), empirically based cumulative curves first appeared in published literature in 1960 (Edie and Foote), and were first used as a predictive tool in 1965 (Gazis and Potts), though they had already been used for some time within state transportation departments. May and Keller (1967) represented traffic as a continuously flowing fluid within a cumulative diagram to model the formation and dissipation of a queue caused by peaking in traffic flows. More recently, in 1993, Newell merged the concepts of cumulative diagrams with wave theory, relying on a three-dimensional version of the cumulative diagram (traffic is a function of both time and space; Makigami *et al,* 1971).

Roads in most countries have been financed through the imposition of taxes, most commonly paid when purchasing fuel. As a consequence, road users do not ordinarily pay additional charges on costly roads. And it is very rare for roadway charges to be related to how heavily the roadway is utilized or the amount of congestion on the roadway. As a consequence, economists have argued that roadways are overutilized during peak periods. (This is because drivers impose more delay on other vehicles during congested periods than they personally incur.)

Vickrey (1963, 1969) proposed that queues can be eliminated through the application of a continuously variable toll, and that all road users would benefit (despite that added toll). Beckmann *et al* (1956), Beckmann (1965) and Dafermos and Sparrow (1971) proposed route based tolls to influence traveler routes, and to optimize use of roadway capacity on primary and parallel routes. Numerous papers have been written since, but the basic approach has remained constant. Prices are set such that travelers optimally equilibrate across travel times and travel routes, greatly reducing or eliminating queueing. The equilibration is based on a combination of direct cost and indirect cost (representing the inconvenience of traveling on a secondary time or at a non-preferred time). In general, however, the models are highly speculative, as realistic data are not available to verify their underlying behavioral assumptions, and because pricing policies are dictated by politics, technology and practicality more than idealized toll structures.

Traffic: Signalized Intersections

Signalized intersections operate as bulk service systems, in which the server alternates between different customer types. A customer type represents a vehicular path through the intersection, defined by a “from direction”, a “to direction” and possibly by a lane. Unlike bulk service systems in production, intersections allow different customer types to be served simultaneously, provided that their trajectories do not intersect, allowing for many ways to combine trajectories into flow patterns.

The queueing delay for any trajectory through an intersection depends on the signal’s cycle length, green phase length (portion of cycle that signal is green for the trajectory), and the synchronization of the green phase with the pattern of vehicle arrivals. It also depends on intersection parameters, such as vehicle service rates during the green phase and the average arrival rate. The usual pattern is that queues accumulate during a red phase, dissipate at a rate matching signal capacity at the start of the green phase and, after the queue vanishes and until the signal turns red again, vehicles are served as they arrive.

Research on intersections has centered on optimizing cycle length, phase lengths, phase patterns and signal offsets (representing time lags between adjacent intersections). Cycle lengths are typically extended when it is necessary to increase an intersection's capacity. This is because capacity losses occur at each phase change; hence, enlarging the cycle length reduces the capacity lost per unit time. If arrival rates are small, cycle lengths are set shorter, so as to minimize cycle delays. [If rates are very small, traffic may be better served by a stop sign or uncontrolled intersection, further reducing cycle delays at the expense of lower capacity (Tanner, 1962; Cheng and Allam, 1992).] Phase lengths are apportioned according to arrival rates and service rates.

As general practice, phase lengths must be at least large enough to serve all vehicles that arrive in a cycle, and should sometimes be even longer if the arrival rate is much larger for a traffic stream than others. Offsets are set to provide synchronization between intersections. Ideally, a signal should enter its green phase as the vehicles begin arriving from an upstream signal. These vehicles arrive in "platoons" (i.e., clusters of vehicles), which have a tendency to disperse as they travel away from an intersection (Pacey, 1956; Grace and Potts, 1964). When intersections are spaced far apart, platoon dispersion (as well as turning traffic) makes it impossible and perhaps unnecessary to synchronize traffic signals. Closely spaced intersections, on the other hand, can be synchronized to minimize cyclic delays and provide for a smoother progression of traffic (e.g., Allsop, 1970; Robertson, 1969; Little *et al,* 1981).

Synchronization is easily accommodated on isolated one-way streets. However, perfect synchronization is usually impossible on two-way streets (in which case opposing directions may arrive at different times) or in signal grids (in which case crossing streets may require different synchronizations). In any case, synchronization demands identical, or integer-multiple, cycle lengths, to ensure that settings do not drift apart. This in itself forces a compromise, as traffic levels at some intersections may demand longer cycle lengths than others.

Grids of signals can also experience blocking effects. This can occur when signals are closely spaced and poorly synchronized, and is exacerbated by poor driver behavior. When a signal operates close to saturation, vehicles may queue back to the preceding intersection. When the preceding intersection turns green, they are blocked from passing through the intersection because the downstream segment is already occupied. The situation worsens when the signals are out of phase with each other, and can be especially problematic in a tight grid of intersections. Intersection blocking in Manhattan is the source of the term "gridlock", which has lately become synonymous with any form of queueing.

Essential trade-offs between cycles length, phase length and queue time were captured as early as 1941 in the work of Clayton, but has since been enhanced through consideration of stochastic effects and different signal configurations and control policies. Most of the literature treats arriving and departing vehicles as fluids, flowing at constant rates within time intervals. In some cases, these rates are stochastic, and in others arrival rates may vary within a cycle (accounting for effects of upstream signals). In Webster's classic work (1958), arrival patterns were simulated, and empirical relationships were statistically estimated for waiting time as a function of signal parameters. Newell (1965) examines signal through analytical expressions in which queue parameters are random variables, but once these parameters are determined the intersection behaves as a deterministic/fluid system. He, along with Miller (1963), examined the effects of spillover from one traffic cycle to the next, which can significantly exacerbate queueing when operating close to capacity.

Transit and Trucking

Mass transit and truck systems have similar characteristics in that they serve "customers" (people in the case of transit, and shipments in the case of trucking) in groups (called bulk service). Bulk service also occurs in production systems, such as batch chemical processes, printing, and metal stamping, and therefore research on queueing systems is somewhat intertwined among these applications. In all cases, the basic issues are to determine when bulk services should occur, how many customers should be served in each bulk service, and which customers should be served. The decisions are optimized against cost objectives (e.g., cost of providing the service), customer service objectives (e.g., average time waiting or average time in inventory), and throughput objectives (e.g., ensuring that customers can be served as fast as they arrive). Unlike traffic signals, bulk service in trucking and transit occurs virtually instantaneously, as the vehicle departs. Furthermore, bulk service models for transit and trucking usually do not consider what happens to the resource (vehicle) when it completes its service.

Perhaps the most famous and widely used model is the Wilson economic-order-quantity model, which was developed in the early 20^{th} century. The basic premise is that a total cost function (sum of inventory and set-up cost) is minimized by optimizing the number of customers served in each bulk service. The resulting equation provides a square-root relationship between order quantity and the arrival rate of customers. Similar ideas can be found in the transportation literature, most notably in the work of Newell (1971), Blumenfeld *et al* (1985), Burns *et al* (1985) and Hall (1996). Newell demonstrated how to optimize the interval between dispatches for non-stationary/deterministic systems. The other three papers determined how the Wilson model can be applied in transportation contexts, accounting for inventory at both the source and destination of a trip, synchronization with arrival and departure processes at the trip ends, and multiple-stop vehicle routes. (These topics are covered in depth in Section 5.5)

One of the most interesting application papers in queueing is Oliver and Samuel's (1967) study of mail processing. This is one of a few papers that examines sortation in terminals and transportation to and from the terminal as a linked process. But the paper is most significant for determining how capacity should be determined within a serial queueing system under non-stationary demand. Their fundamental conclusion was that staffing should be allocated in a way that evens out capacities, thus providing minimal queueing once the customer has passed through the initial server.

A second area of interest is real-time control of routes, governing the release of vehicles from stops in response to random arrival rates. Again, the earliest work in this area falls outside of the transportation literature (Bailey, 1954; Neuts, 1967). More recent work includes Powell (1985), Powell and Humblet (1984), and Powell (1986), who investigated a variety of policies for dispatching or canceling services based on the elapsed time from the previous service and the number of customers waiting. Similar policies have been investigated for transfer terminals by Hall *et al* (2001) and cyclic truck routes (Hall, 2002). These contributions fall in the tradition of dispatching policies form the production literature.

A final area concerns schedule control of vehicles traveling on routes with multiple stops. Here, the application is almost exclusively transit. In this context, it is usually impossible to hold vehicles at stops if there are insufficient customers. First, it would be unwise to base a dispatch policy on just one stop when the bus will later serve many downstream locations. Second, most transit systems advertise a schedule that is relied on by customers. Finally, the majority of the service cost is incurred whether the vehicle is in motion or stopped, so there is little cost advantage in holding a vehicle or canceling a trip.

In routes providing frequent service (headways of 10 minutes or less), the objective in schedule control is largely to ensure consistency in headways (time separation between vehicle arrivals or departures). Customers on short-headway lines typically do not consult schedules before arriving at their stops, and therefore arrival patterns are reasonably stationary relative to the schedule. Second, as demonstrated in Osuna and Newell (1971), average waiting time increases with the square of the coefficient of variation in the headway (ratio of standard deviation to the mean). Completely random Poisson vehicle arrivals generate twice the average wait of deterministic arrivals. In fact, waiting time can be worse then the Poisson case, as vehicles on frequent lines have a tendency to bunch. Headways on very frequent lines are inherently unstable: when a bus falls slightly behind schedule, it tends to pick up more passengers, causing it to slow further, until it eventually bunches with the trailing bus (Newell, 1975, Barnett, 1974). This can be controlled, to some degree, by slowing down a trailing bus when it is catching up with the preceding bus. However, the added delay for passengers already on the trailing bus limits the applicability of this (and other) control strategies, except at the very start of lines.

The behavior of infrequent lines differs substantially from frequent lines. Customers generally do consult schedules, making arrival patterns non-stationary. Therefore, waiting time is not defined by the headway, but instead by the random deviations in the bus arrivals at the stop, along with the customer's selected arrival time relative to the schedule. Finally, because late bus generally do not pick up additional passengers, schedules tend to be much more stable.

Aircraft and Airports

As in road transportation, a fundamental issue in air transportation is accommodating peak traffic loads. And though techniques such as fluid models have been applied in air transportation (e.g., Newell, 1979), a separate branch of research has evolved in which stochastic phenomena are explicitly modeled. Unlike highway traffic, the number of customers (represented by aircraft) that may reside in a queue is relatively small, making it relatively easy to measure the system state as a discrete entity, and also making round-off errors introduced in fluid models somewhat more significant. Consequently, this line of research is linked more directly to mainstream queueing research.

Air transport research is dominated by the phenomena of runway queues. Runways are traditionally a weak link in the air transport system, likely due to the high cost and environmental constraints in their construction, and safety requirements in operation. A complication in modeling runway queues is that the service time for an aircraft depends on the type of preceding aircraft, which is defined by speed and size (creating wake effects that can impose safety risks to trailing aircraft), and whether it is taking off or landing. Therefore, as in many production systems, it can be advantageous to sequence customers in a way that optimizes throughput (Newell, 1979).

Stochastic modeling of runway queues is represented in the work of Gallagher and Wheeler (1958), Koopman (1972), Peterson *et al* (1995a,b) and Odoni and Roth (1983). Odoni and Roth, for instance, developed an approximation for the time constant within an exponential decay function, representing the difference between the expected state of the system at a time t and the limiting state as t goes toward ∞. Newell (1982) is also notable for development of relaxation times, representing the approximate time for a system to reach steady derive. Newell demonstrated that the relaxation time goes toward infinity as the arrival rate approaches capacity, and that steady-state equations are inherently inaccurate for systems that operate close to capacity, even if arrival rates fluctuate only slightly. These were derived from diffusion models, and were not intended for a specific modal application.

Runway queues have also been evaluated within the context of "ground holding", which is a form of network flow control in which the release of aircraft from a departure airport is based on congestion and weather at the destination airport. Ground holding is advantageous because queues are shifted from the airspace to the airports, saving operating costs and enhancing safety. Stochastic programming methods have been used to study the problem (Odoni, 1987; Andreatta and Romanin-Jacur, 1987; Richetta and Odoni, 1993).

A second queueing application is the baggage claim process (Horonjeff, 1967; Ghobrial *et al.,* 1982; Robuste and Daganzo, 1992). Here, a service is not completed until two events occur: the arrival of the passenger (or passenger group) and the arrival of the bag (or bag group). Hence, the service is defined by the maximum of a set of random variables. Horonjeff's analysis is based of actual bag and passenger arrival patterns, which are expressed relative to the time that an aircraft begins disembarking passengers. Ghobrial *et al* offer an extension, in which the time required to retrieve a bag is a function of the passenger density surrounding the baggage carousel, and Robuste and Daganzo examine baggage sortation and containerization strategies (similar issues arise in rail and ocean terminals).

Railways

Railways are somewhat unique as transportation modes in two ways: shipments are grouped into long serial units during transportation, and vehicles have no steering capability. Each situation has led to research on queueing.

Train transportation exhibits strong scale economies, meaning that the cost per unit declines substantially when trains operate in longer lengths. However, it is unusual for a single origin/destination pair to generate sufficient traffic to create a long train. Therefore, different origins and destinations must somehow be grouped together. This is accomplished in classification terminals. Each arriving train brings cars from a common set of origins. The train is then broken apart and sorted according to groups of destinations. The sorted cars are finally formed into outbound trains. The process is sometimes repeated multiple times, and sometimes pre-sorting at one terminal to reduce work at a downstream terminal.

The classic work on train sortation can be found in the book by Beckmann *et al* (1956), who modeled the expected number of train breaks (and associated service time) as a function of the number of sortation categories and their probabilities. More detailed models were not developed until the 1970s, and is represented in the work of Petersen (1977a,b), Turnquist and Daskin (1982), and Daganzo *et al* (1983). These authors developed models representing the time required to process a train based on how cars are grouped into sortation blocks on outbound trains.

Because trains cannot be steered, and because the guideway is restricted to narrow track, vehicles can only pass each other at prescribed locations (sidings), and with the assistance of switching. This contrasts with most other forms of transportation, where vehicles can pass by steering into another lane or otherwise outside the trajectory of the other vehicle. Most railroads are designed to have either one track (shared by opposing directions) or two tracks (one for each direction). In the first case, trains must be switched into sidings to allow faster trains to overtake slower trains (e.g., a passenger train passing a slower freight train), or whenever trains meet from opposing directions, no matter how fast they are traveling. With two tracks, sidings are only needed to allow faster trains to pass slower trains.

Queueing research has centered on design, including: (1) provision of one or two tracks, (2) separation between sidings, (3) operating policies, with respect to speed, passing priority and train scheduling. Railroads must consider whether the benefits of operational flexibility and reduced delay justify the added expense of constructing additional track or sidings. This investment is typically only justified when traffic levels are sufficient. Research on the subject is represented by Frank (1966), Petersen (1974) and Welch and Gussow (1986). A common technique is to utilize time-space diagrams (the vehicle trajectory, showing position as a function of time) to identify train "interference" (i.e., the intersection of vehicle trajectories). Petersen determines the interference frequency as a function of the train separations and speeds, and associates these with interference delays siding locations. This research is closely related to the traffic flow literature, both in its use of time-space diagrams and in its modeling of interference.

Spatial Queueing

A final application area spans transportation and location science, and concerns queueing for spatially separated resources, such as police or fire service. The general question is to allocate resources in a way that minimizes a measure of response time, while staying within an available budget. In some cases, the resources reside at fixed bases (e.g., fire), and in other cases the resources are mobile (e.g., police). Versions also exist where the customer travels to the server, rather than the server traveling to the customer.

The response time typically includes a combination of travel time (from where the resource is located to where it is needed), call processing time, and queueing time. One of the interesting phenomena is that when the system gets busy, it is the travel time that suffers rather than queueing time. This is because when nearby resources are busy, a more distant resource is dispatched instead – creating a longer response time. Simultaneously, the throughput degrades, as it takes longer to serve calls when travel distance increases.

Much of the work on the topic can be attributed to a series of projects conducted by the RAND Corporation in New York City in the early 1970s. Examples of research in the area include Chaiken and Larson (1972), Green and Kolesar (1989), Ignall *et al* (1978), Kolesar (1975), Kolesar and Blum (1973), Kolesar *et al* (1975), Larson (1972) Rider, (1976). The work is most notable for how it has blended empiricism, theory, and application. This includes modeling response distance as a square-root function of the average territory served by each resource, explicitly representing resource allocation and call rates as non-stationary functions, precisely modeling service time distributions and verifying results against actual performance.

5.4 Representation of Queueing Processes

Cumulative diagrams and fluid models are the most important contributions of transportation to the queueing literature, and this is our emphasis here. They have been applied to all modes of transportation, and are useful in displaying and modeling queueing phenomena, and in system optimization.

Basic Concepts

A cumulative diagram indicates how many customers (often vehicles) have passed a point in the transportation system as a function of time (measured from an initialization time). A cumulative arrival diagram indicates how many customers have entered the system, and a cumulative departure diagram indicates how many customers have left the system. Figure 5.1 provides an example empirical cumulative diagram. In an empirical diagram, individual customers are represented by steps in the curve, corresponding to the time instants when events occurred (either an arrival or a departure). Additional curves can be created, is desired, for intermediate points, as when customers pass through serial servers.

Cumulative diagrams are important because they provide many performance measures in one simple picture. Let:

$$A(t) = \text{cumulative arrivals from time 0 to time t}$$
$$D_s(t) = \text{cumulative departures from the system from time 0 to time t}$$

The number of customers in the system at any time t is simply:

$$L_s(t) = \text{number of customers in the system at time t}$$
$$A(t) - D_s(t) \qquad (5.1)$$

And the total time spent by customers in the system up to time t is:

$$W(t) = \text{total time spent by customers up to time t}$$
$$= \int_0^t L_s(\tau)d\tau \; + \int_0^t [A_s(\tau)\text{-}D_s(\tau)]d\tau \qquad (5.2)$$

Two critical performance measures are the average number of customers in the system and the average time in system per customer. The average number of customers is easily derived from W(t):

L(t) = average customers in system, time 0 to time t

$$L(t) = W(t)/t \tag{5.3}$$

In cases where the system begins and ends in an empty state (i.e., $L_s(0) = L_s(t)=0$), the average waiting time is also easily defined:

W(t) = average time in system from time 0 to time t

$$\mathbf{W}(t) = W(t)/A(t) \tag{5.4}$$

Combining these expressions, it can be seen that

$$tL(t) = A(t)\mathbf{W}(t) \tag{5.5a}$$

$$L(t) = [A(t)/t]\mathbf{W}(t) \tag{5.5b}$$

Equation 5.5b is a special case of Little's formula (1961), which states that the average number of customers in the system asymptotically approaches the average time in system multiplied by the customer arrival rate for a wide class of systems.

All of these results are clearly seen in a cumulative diagram, as Figure 5.1 illustrates. The number of customers in system (queue size) is the vertical separation between the cumulative curves, and the total waiting time is the area between the curves. The average time in system is the average horizontal separation and the average customers in system is the average vertical separation. If customers are processed in a FCFS order, the diagram also shows the time in system for individual customers, also measured by the horizontal separation. If the sequence is not FCFS, then another graphical device, such as a GANTT chart, is needed to show the time in system for individual customers.

Fluid Models

In a fluid model, individual customers are represented as a continuously flowing fluid rather than discrete entities. This has the effect of smoothing out the steps in the arrival and departure curves. Fluid models are often used to predict the future performance of queueing systems, or just to simplify the representation of observed phenomena.

In a fluid model, arrival rate, $\lambda(t)$, and departure rate, $\mu(t)$, are defined by the derivatives of their corresponding cumulative curves:

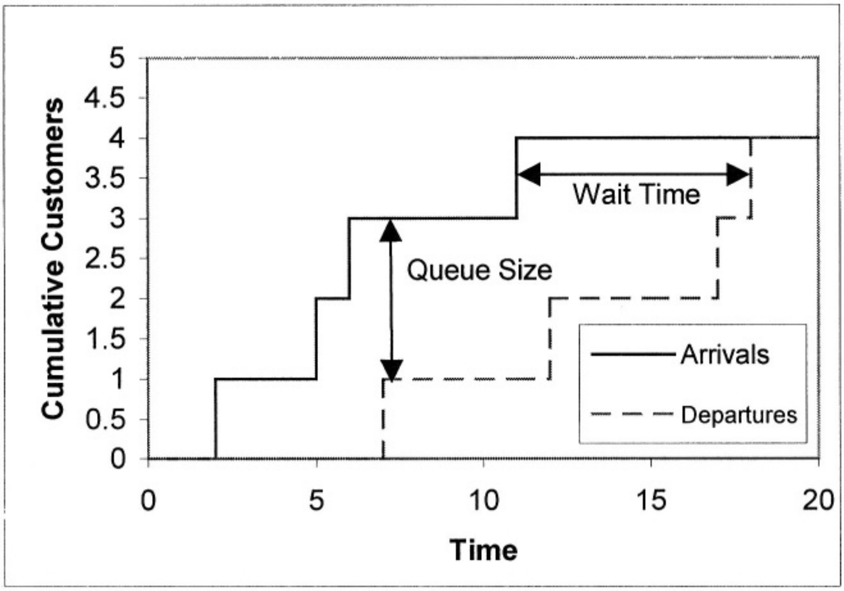


Figure 5.1 Cumulative Diagram

$$\lambda(t) = dA(t)/dt \qquad (5.6a)$$
$$\mu(t) = dD(t)/dt \qquad (5.6b)$$

In bulk service systems, the service rate can be undefined; otherwise it reflects three factors: (1) the speed at which customers can be processed by the server, (2) the size of the queue, and (3) the rate at which customers arrive. Servers ordinarily operate at their fastest rate when queues are present, and operate at the same rate at which customers arrive when queues are not present. Exceptions exist, as service capacity can be variable, depending on demand, and service times can sometimes change as queue lengths change.

To illustrate fluid models, we consider a simple system in which the service rate is limited to a capacity c, but service times are very short. This might represent queueing at a highway toll plaza, for instance. The arrival process and departure process are both non-stationary, but are assumed to be deterministic, for purposes of illustration. Under these conditions, Figure 5.2 illustrates how the queues would evolve over a period of peak arrivals. The system is shown to evolve through a series of four phases:

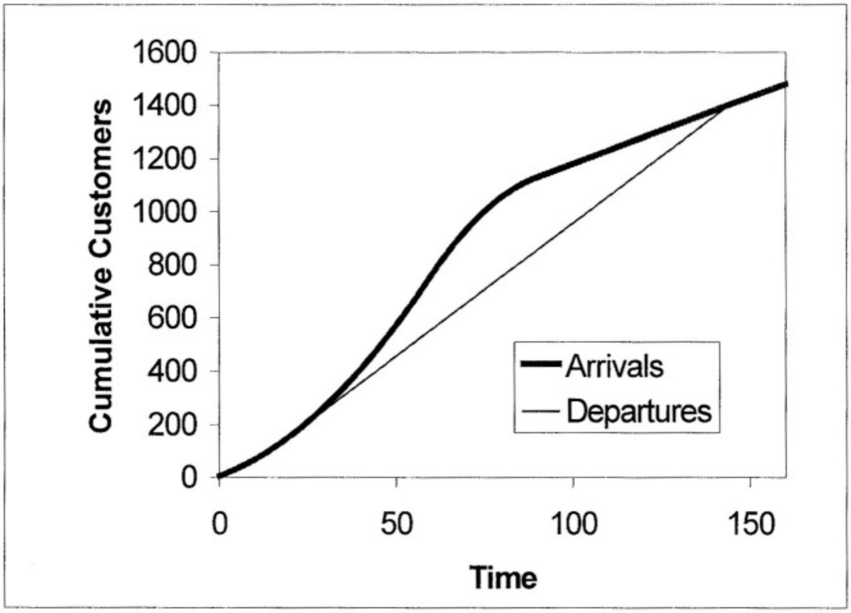


Figure 5.2 Cumulative Fluid Model

Phase 1: Stagnant

$$A(t) = D_s(t) \qquad \lambda(t) \leq c \qquad \mu(t) = \lambda(t) \qquad dL(t)/dt = 0 \qquad (5.7)$$

Phase 1 represents the initial period when customers can be processed as fast as they arrive (time 0 to time 20 in the figure).

Phase 2: Growth

$$A(t) > D_s(t) \qquad \lambda(t) > c \qquad \mu(t) = c \qquad dL(t)/dt = \lambda(t) - c > 0 \qquad (5.8)$$

Phase 2 represents the period in which the queue grows because customers cannot be served as fast as they arrive (time 20 to time 80 in the figure).

Phase 3: Decline

$$A(t) > D_s(t) \qquad \lambda(t) \leq c \qquad \mu(t) = c \qquad dL(t)/dt = \lambda(t) - c \leq 0 \qquad (5.9)$$

Phase 3 begins when the queue reaches its maximum length, which occurs when the arrival rate drops down to capacity. It ends when the queue vanishes. (Time 80 to time 140 in the figure.)

Phase 4: Stagnant

$$A(t) = D_s(t) \qquad \lambda(t) < c \qquad \mu(t) = \lambda(t) \qquad dL(t)/dt = 0 \tag{5.10}$$

Phase 4 is when the queue is again stagnant at 0, with customers arriving slower than they can be served. An interesting phenomenon is that $\mu(t)$ exhibits a discontinuity at the time the queue vanishes (5.3), dropping suddenly from c to the current arrival rate. Thus, the departure rate pattern is highly asymmetrical in queueing systems.

Analysis Through Cumulative Diagrams

Through perturbation analysis, it is possible to optimize the design of the queueing system. It is relatively straight forward, for instance, to model the effects of changing system capacity. Increasing capacity has a non-linear effect on time in system, as it causes both the duration of the queue (length of Phase 2 and 3), and the magnitude of the queue to decline. And when capacity exceeds the maximum arrival rate, the queue vanishes. Comparison of departure curves can be used to select a capacity from a set of discrete options.

Cost trade-offs can be evaluated through use of marginal analysis, in which capacity is continuously varied. We define a total cost function as:

$$C = \text{capacity cost} + \text{waiting cost}$$
$$C = \alpha c + \beta w(c) \tag{5.11}$$

where:

α = capacity cost per unit capacity
β = waiting cost per unit customer time
W(c) = total waiting time when capacity equals c

A necessary condition for optimality is that cost must not decrease if the capcity is changed by a small amount Δc. The change in cost, ΔC, if c is increased by Δc can be written as the sum of the change in capacity cost and the change in waiting cost:

$$\Delta C = a\Delta c + \beta[W(c+\Delta c) - W(c)] \tag{5.12}$$

The change in waiting time (the term within the brackets) can be calculated from the cumulative diagrams. Assume, as in Figure 5.3, that one predictable queue occurs per time period. Then, for small values of Δc, the change in waiting time can be approximated from the area of the triangle shown in the figure. That is:

$$W(c+\Delta c) - W(c) \approx ½ T(c) [T(c)\Delta c] \qquad (5.13)$$

ΔC represents the marginal change in cost, which must equal zero at the optimum (provided that W(c) is continuously differentiable). Substitution of Eq. 5.13 in Eq. 5.12 provides the following optimality criterion:

$$T^*(c) = \sqrt{2\alpha/\beta} \qquad (5.14)$$

Eq. 5.14 states that the optimal capacity is represented by the duration of the queueing period – time from when the queue first forms until it vanishes – and not by the arrival rates during the queueing period. The optimal duration increases with the square root of the capacity cost (when capacity is expensive, longer duration queues can be tolerated) and decreases with the square-root of the waiting cost (when waiting is expensive, queues should be shorter in duration).

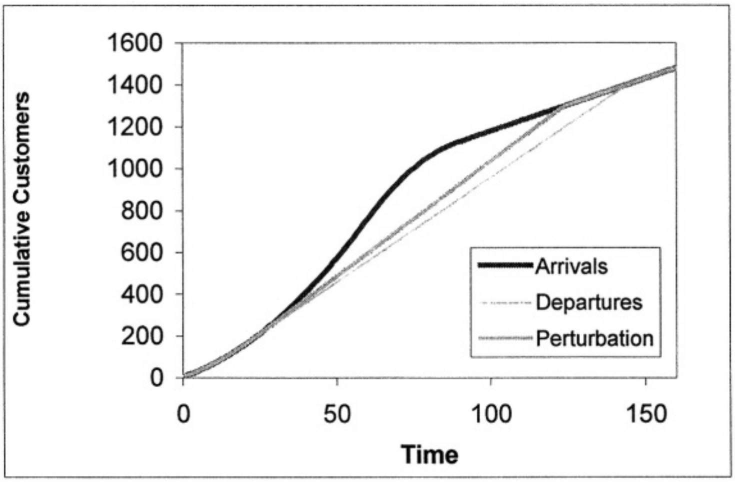


Figure 5.3 Marginal Analysis

Extensions

The cumulative modeling approach has been extended in a variety of ways.

- Investigation of the combined effects of stochastic variability and non-stationarity, principally through application of diffusion models.

- Optimization of other system attributes, such as staffing plans and time-off scheduling.

- Measuring the effects of incidents that cause capacity to decline over short intervals

- Estimating effects of behavioral responses, causing arrival rates to be a function of queue lengths, waiting times or tolls.

- Evaluating queueing in bulk-service systems, such as signalized intersections, transit and trucking.

Bulk service will be examined in some depth in the following section. But first, we note that incidents often have a pronounced effect on system performance. This is especially true when incidents occur around the time that a queue begins to form, as it affects everyone who arrives over the queue's entire duration. The effect is not nearly so great when an incident occurs later, as it only affects those customers that arrive later. As a consequence, queue management demands special care during Phase 2, both to prevent harmful incidents, and to persuade customers, if possible, to arrive at other times.

5.5 Bulk Service Models

Economic Order Quantity (EOQ) and Economic Production Quantity (EPQ) models have been used for many years in transportation and manufacturing to optimize cycle lengths, load sizes and batch quantities for bulk service. While research in this area today is focused on complex scheduling systems, many of the underlying assumptions of the EOQ/EPQ models have been retained, especially in transportation applications.

This section describes how the EOQ/EPQ methodology is applied, taking both input processes and output process into account. To this end, a set of "characteristic cumulative diagrams" is developed to represent a range of scenarios. The principal assumptions are: (1) input and output processes occur at constant and deterministic rates (in some scenarios, rates are allowed to alternate between "on" and "off" phases through batch processing). (2) Set-up and order costs are independent of batch size. (3) Batches can be initiated instantaneously when the queue size drops to zero. (4) Queueing costs are linear functions of the average queue size.

The systems considered will have three components: an input process, a bulk transportation system, and customers. The models explicitly represent bulk transportation of goods, but they are easily adapted to represent other transportation systems, such as traffic signals and buses. Hence, the input represents a production process. The section is organized to demonstrate the effects of: (1) Synchronization of input batch sizes with output batch sizes; and (2) Coordination of input and output when there are multiple customer or product types.

Basic Methodology

The general approach is to represent total cost per unit time as the sum of a queue cost and a "set-up" cost. The queue cost equals the average queue level multiplied by a queue cost parameter. The set-up cost equals the number of set-ups or orders per unit time (the demand rate divided by the batch size) multiplied by the cost per set-up.

The following parameters are used to represent the system. In some cases, these parameters are subscripted to denote an individual customer or product.

d = output rate (items/time)
p = input rate (items/time)
S = input set-up cost (money/set-up)
A = transportation "set-up" cost (money/order)
h = queue cost (money/customer per unit time).

To simplify expressions, transportation lead time (i.e., the transportation time from origin to destination) is assumed to be zero. With respect to optimizing batch sizes, this assumption results in no loss in generality, provided that lead times are independent of the other parameters. While it is not difficult to incorporate a "pipeline" cost to represent lead-time, the cumulative diagrams lose clarity. Queueing cost is also assumed to be identical at source and destination, again with the intention of highlighting principles. For similar reasons, the time to perform the set-up is assumed to be negligible relative to the run time. Finally, batch sizes are assumed to be unconstrained.

The decision variables are the order and production batch sizes, which in turn define the order and production cycles:

Q_p = production batch size
Q_t = transportation order quantity
T_p = production cycle time = Q_p/d
T_t = transportation cycle time = Q_t/d.

While in most cases the production and order quantities are held constant, it will, in some instances, be less costly to allow for varying quantities.

Queue holding costs are defined by the cumulative production at the source and cumulative demand at the customer (or customers).

P(t) = cumulative production from time 0 to time t
D(t) = cumulative demand from time 0 to time t.
I(t) = customers in the system at time t = P(t)-D(t).

The order and set-up costs depend on P(t) and D(t), as achieving a small queue requires more frequent set-ups and orders.

Dispatching Rule We now define a general characteristic of batch transportation systems under optimal control. The characteristic is a necessary condition for optimality when the following four conditions apply, but as a matter of practice applies more broadly:

(1) transportation set-up cost is fixed with respect to shipment size,
(2) queue cost is a linear function of the total queue in the system (i.e., P(t)-D(t)),
(3) vehicle size is unlimited,
(4) P(t) and D(t) are non-decreasing and represent a single product.

Let:

T(t) = cumulative items dispatched from the manufacturer, from time 0 to time t.

Then at the time of any dispatch:

T(t) = D(t) immediately before dispatch
T(t) = P(t) immediately after dispatch.

In words, the dispatching rule states that a shipment should be sent as soon as the queue is exhausted at the customer, and that the order quantity (i.e., shipment size) should be identical to the queue on-hand at the manufacturer: P(t)-D(t). Visually, this rule is manifest in the cumulative graphs presented later through the staircase pattern for T(t), which alternately "bounces" between D(t) and P(t).

The optimality of the dispatching rule can be proved by contradiction. From any solution that violates the rule, it is possible to construct a solution which obeys the rule, with equal or lower cost. Specifically, if T(t) does not equal D(t) immediately before dispatch, then the shipment can be delayed until T(t) = D(t), with no increase in queue cost, and a possible decrease in transportation cost (if two shipments can be consolidated). If T(t) does not equal P(t) immediately after dispatch, then the shipment size could be increased, with no change in queue cost, and a possible decrease in transportation cost (if a subsequent shipment can be eliminated).

Queue Models

This section creates a set of seven characteristic cumulative diagrams, each representing a different cyclic queueing pattern. In a subsequent section, these curves are used as building blocks for developing EOQ/EPQ models. While the diagrams represent production/distribution, they are easily adapted to represent other situations in transportation.

The average queue level equals the average separation between the cumulative production and cumulative demand curves, which is determined by calculating the area of separation and dividing by the elapsed time. The separation depends on the batch sizing policies, both in production and transportation. In its simplest form, production and demand are characterized by Figure 5.4 or 5.5. Figure 5.4 is the textbook version of the EOQ model, as it assumes instantaneous production and transportation. Figure 5.5 is the textbook version of the EPQ model, as it assumes production occurs at some set rate, and transportation occurs continuously, and not in batch.

In a more general sense, average queue level may be defined by any of the following types of cumulative production and demand diagrams, which will be called the "characteristic curves." (Recall that, in all cases, constant demand is assumed.) The set of cases is not completely exhaustive, but does encompass most reasonable patterns that apply to direct transportation routes in production/distribution.

1. Instantaneous Production/Batch Distribution (Synchronized) This is the textbook EOQ model (Figure 5.4).

$$\text{Average Queue Level} = \quad Q_t/2$$

2. Instantaneous (or Constant) Distribution/Batch Production As in Figure 5.5, production is immediately available for consumption, eliminating batch size inventories in distribution. Figure 5.5 is equivalent to the textbook EPQ model.

$$\text{Average Queue Level} = (Q_p/2)(1\text{-}d/p)$$

3. Constant Production/Batch Distribution As in Figure 5.6, production and demand occur at a constant rate. Inventories exist at both point of production and

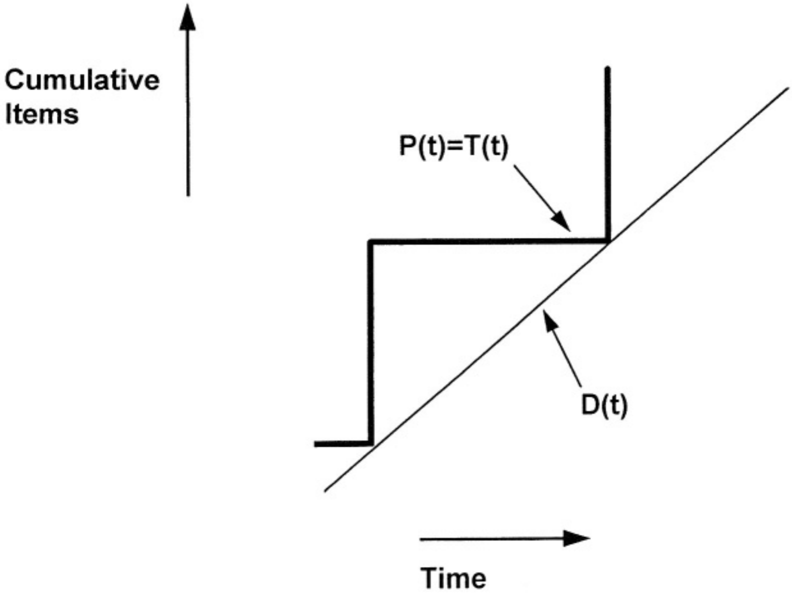


Figure 5.4 Instantaneous Production/Batch Distribution

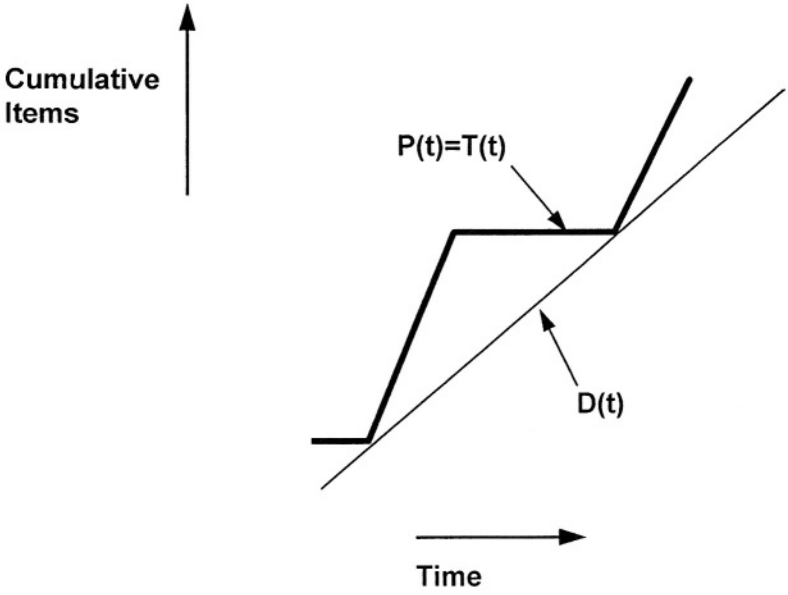


Figure 5.5 Instantaneous Distribution/Batch Production

point of demand as a result of distribution batch sizes with constant sizes and constant separation.

$$\text{Average Queue Level} = Q_t$$

4. Batch Production/Batch Distribution In all of these cases, the product is both manufactured in batches and transported in batches.

4a. Synchronized/lot-for-lot As in Figure 5.7, transportation is synchronized with production, so that a dispatch occurs as soon as a batch is manufactured. The average queue level at the customer is the transportation batch size divided by two. The average queue level at the manufacturer is one-half the production batch size, multiplied by the proportion of time that the machine is running (d/p).

$$\text{Average Queue Level} = (Q_p/2)(d/p) + Q_t/2 = (Q/2)(1 + d/p),$$

where $Q = Q_t = Q_p$ (due to lot-for-lot production). If d=p, the machine runs continuously and the average queue level is the same as for case 3. If p>>d, production is effectively instantaneous, and the average queue level is the same as case 1. Finally the ratio of average queue level relative to case 2 (instantaneous distribution) is (p+d)/(p-d). Hence, if d<<p, average queue levels are approximately the same. As d approaches p, the ratio approaches infinity, indicating that the EPQ model greatly underestimates queue level in batch distribution when production and demand rates are similar.

4b. Synchronized/multiple transportation lots Due to this case's complexity, the queue model will be presented later within the context of a specific system scenario (Scenario F).

4c. Non-synchronized/lot-for-lot As in Figure 5.8, the transportation and production cycle lengths are identical. However, transportation is not scheduled to coincide with the end of a production run. (This may occur if multiple products, each with a different start/end time, are transported in each cycle.)

$$\text{Average Queue Level} = (Q_p/2)(d/p) + Q_t/2 + \tau d = (Q/2)(1 + d/p) + \tau d$$

where τ is the time lag between the end of the production run and the time of dispatch ($Q=Q_t=Q_p$).

4d. Non-synchronized As in Figure 5.9, production and transportation both occur in batches, but are not synchronized. As a result, the average queue is the sum of case 2 and case 3 (Blumenfeld *et al,* 1985).

$$\text{Average Queue Level: } (Q_p/2)(1\text{-}d/p) + Q_t$$

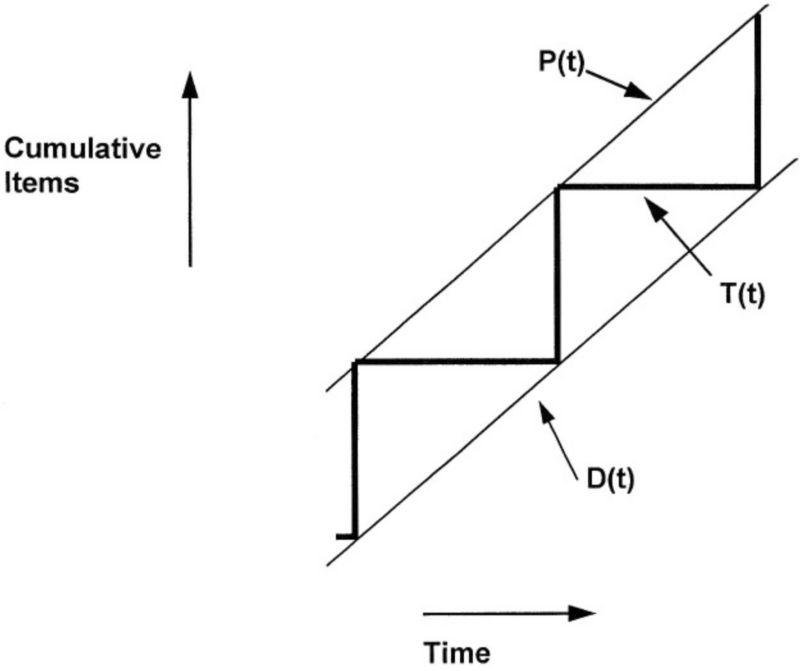


Figure 5.6 Constant Production/Batch Distribution

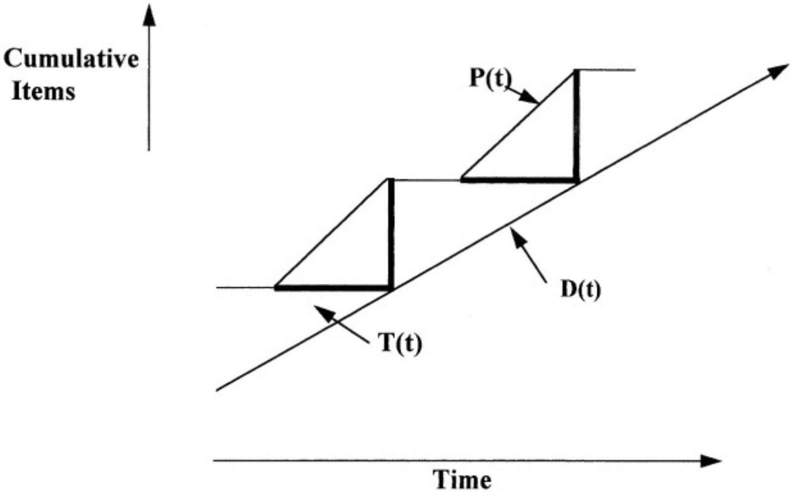


Figure 5.7 Synchronized Lot-for-Lot

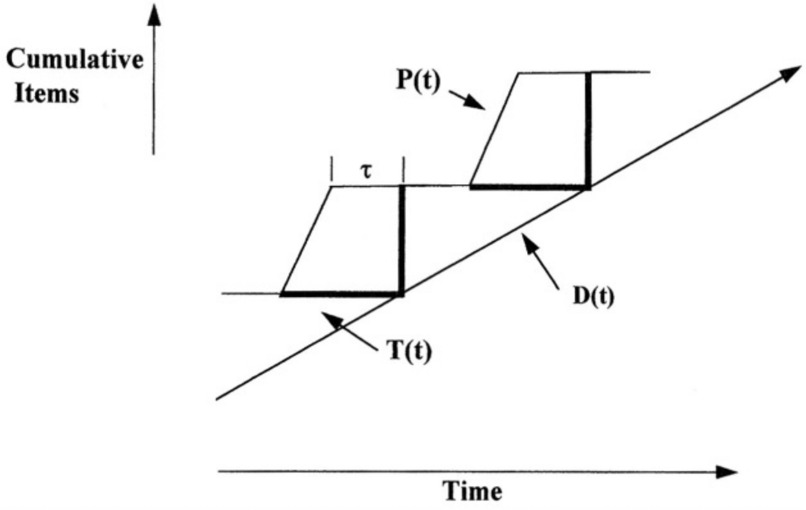


Figure 5.8 Non-Synchronized Lot-for-Lot

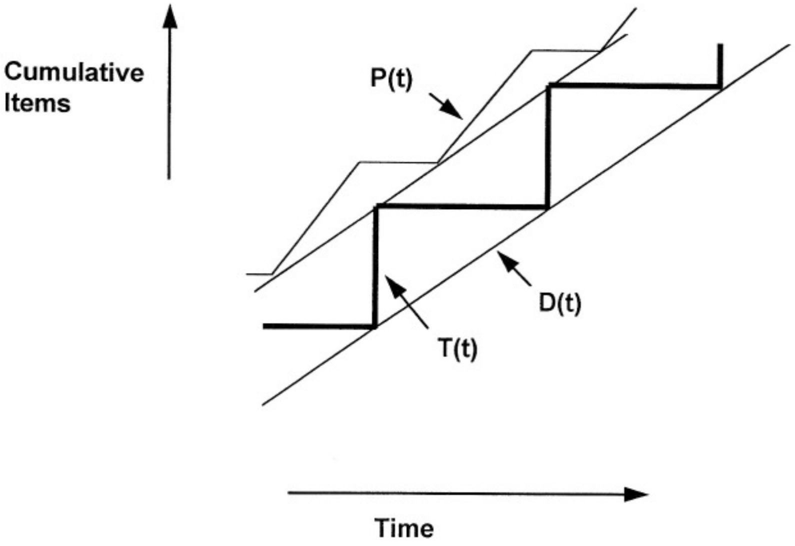


Figure 5.9 Non-Synchronized

Set-Up Cost Models

The set-up cost per unit time is the cost per set-up (or order), multiplied by the number of set-ups (or orders) per unit time. For consistency with the queue models, it is necessary to derive the number of set-ups per unit time as a function of the batch size (or sizes). In the classic EOQ and EPQ models, this function is simply the following:

$$f_p = \text{production set-up frequency}$$

$$f_p = d/Q_p = 1/T_p \tag{5.15a}$$

$$f_t = \text{transportation "set-up" frequency}$$

$$f_t = d/Q_t = 1/T_t \ . \tag{5.15b}$$

These models are adequate when a single product is manufactured/distributed, but more precision is needed for multiple products. In the transportation process, in particular, it is customary to serve multiple products within the same batch. Hence, for any origin/destination pair, there is a single transportation cycle length, which is identical for all products:

$$f_t = d_1/Q_{1t} = d_2/Q_{2t} = \ldots = d_i/Q_{it} = \ldots \tag{5.16}$$

where the first subscript on d and Q denotes product number, and where Q is interpreted as the shipment size per dispatch. Equivalently, a "composite product" can be defined, where demand is the sum across all products, expressed in a common unit (such as weight or dollar value). Then Eq. 5.1 would apply, provided that Q and d are interpreted in this common unit. In Case 4b, where batch sizes vary within a production cycle, a further modification is needed. The transportation set-up frequency will be the number of orders per production cycle (n) multiplied by the production set-up frequency (d/Q_p).

Problem Dimensions

The number of potential variations to the EOQ and EPQ model is quite enormous. Our purpose is to present a range of scenarios, and later discuss the implications of the more significant variations on cost. This will be accomplished by identifying the "characteristic cumulative diagram" that applies to the scenario, computing total cost, and optimizing the production batch size and transportation order quantity.

The scenarios are defined at two levels. At the top level, the defining attributes are the number of customers and the number of plants. At the lower level, scenarios are defined by the number of machines within each plant and the number of products:

Top-level Attributes

1) Single Customer/Single Plant
2) Multiple Customer/Single Plant
3) Single Customer/Multiple Plants
4) Multiple Customer/Multiple Plants

Lower-level Attributes

a) Single Machine/Single Product
b) Multiple Machines/One Product per Machine
c) Single Machine/Multiples Products per Machine
d) Multiple Machines/Multiple Products per Machine .

Attributes (a) and (b) do not require production changeovers; hence, production is continuous, and the transportation order quantity is the only decision variable. Attributes (c) and (d) demand changeovers between products; hence, both production batch size and order quantity must be optimized. Table 1 summarizes the scenarios covered in the section, which are constructed by combining attributes. The first three are fairly straight-forward, and do not entail schedule interactions among products. The second three are more complex.

Cost Analysis: Simple Scenarios

This section develops cost models for three simple scenarios, which illustrate the effects of accounting for: (1) queue costs at both the manufacturer and customer; (2) consolidation of multiple products from multiple machines; and (3) costs for unsynchronized systems. These are classified as simple cases because all treat one product at a time.

A. Single Customer/ Single Plant (Queue at Manufacturer and Customer) In this scenario, one machine operates at a constant rate (equaling the demand rate), producing a single product, without interruption, for a single customer. Set-ups do not occur because product change-overs are not needed. Hence, the only decision variable is the transportation order quantity.

The cumulative diagram in Figure 5.6 (constant production/batch distribution) characterizes the situation. The objective function, and its optimal solution, are then:

$$C(Q_t) = A(d/Q_t) + Q_t h \tag{5.17a}$$

$$Q_t^* = \sqrt{Ad/h} \tag{5.17b}$$

$$C^* = 2\sqrt{Ahd}. \tag{5.17c}$$

B. Single Plant/Multiple Machines/Single Customer (Consolidation Effect) In this scenario, each machine produces a single product at a constant rate, for which demand also occurs at a constant rate. The products are manufactured at a single plant, and distributed to a single customer. Unlike the prior scenario, different products are consolidated in the transportation process. This situation illustrates a major difference between EPQ and EOQ models. Whereas batch production does not allow different products to be processed simultaneously (rather, alternating phases are needed), batch transportation virtually mandates simultaneous service. That is, from the standpoint of cost minimization, it is cheaper to consolidate products in the same vehicle than to transport each product separately.

Blumenfeld *et al* (1985) examined this situation, and introduced the concept of a composite product to represent the portfolio of product characteristics contained in the load. Hence, Figure 5.6 is interpreted as the demand among all products sent between the manufacturer and customer. The cost model, and optimized results, are shown below. Cycle length is used as the decision variable, rather than batch size, because batch size varies among products:

$$C(T_t) = A/T_t + T_tHD \qquad (5.18a)$$

$$T_t^* = \sqrt{A/HD} \qquad (5.18b)$$

$$C^* = 2\sqrt{AHD}, \qquad (5.18c)$$

where:

$$HD = \Sigma\, h_jd_j\,. \qquad (5.19)$$

C. Multiple Plants and Customers (Unsynchronized) In a system with multiple plants and customers, it may be impossible to synchronize transportation and production cycles due to scheduling conflicts. As a result, larger queues must be held at the manufacturer to buffer against cyclic fluctuations. In this scenario, batch production and batch distribution are assumed. The system is decomposed to individual plant/customer/product combinations, assuming the absence of synchronization, as in Figure 5.9. The cost model, and optimized results, are shown below.

$$C(Q_p,Q_t) = S(d/Q_p) + A(d/Q_t) + [(Q_p/2)(1\text{-}d/p) + Q_t]h \qquad (5.20a)$$

$$Q_t^* = \sqrt{Ad/h} \qquad (5.20b)$$

$$Q_p^* = \sqrt{2Sd/h(1\text{-}d/p)} \qquad (5.20c)$$

$$C^* = 2\sqrt{Ahd} + \sqrt{2Shd(1\text{-}d/p)}\,. \qquad (5.20d)$$

In this case, the production and distribution results are decoupled. Further, the production batch size is identical to the textbook EPQ model. The transportation order quantity, on the other hand, is identical to Eq. 5.17b. Hence, the base comparisons for the transportation order quantity are the same as those presented in the single customer/single plant scenario.

More Complicated Scenarios

Within this section, cost analysis is shown for three more complicated scenarios, to illustrate issues involving multiple products and multiple customers. In the first example, a single machine produces a single product to serve multiple customers. In the second, a single machine produces multiple products for a single customer. In the last, a single machine produces multiple products for multiple customers, with one product per customer.

Within the framework of EOQ/EPQ modeling, it is impossible to fully account for complex scheduling systems. In the examples, schedule conflicts are avoided by assuming either (or both) of the following: (1) products are manufactured sequentially in a common rotation cycle, or (2) production rate greatly exceeds demand. Within a rotation cycle, the production rate for a machine is assumed to be the same as the total demand for the machine. The large production rate case will only be used for multiple customer scenarios.

D. Single Machine/Multiple Customers In this scenario, a single machine produces a single product at a constant rate for multiple customers, without interruption. Though set-ups do not occur, production must still be divided into time segments, corresponding to customers. Consequently, the average queue depends both on the time to produce and the time to consume a quantity. These values are different because the production rate, by the necessity to serve multiple customers, must exceed the demand rate of any one customer. This effectively results in production batches without the need for production set-ups. Hence, the characteristic queue curve for any one customer is a batch production/batch distribution case (Figure 5.7), but the production set-up has a cost of zero.

For an individual customer, the cost can be expressed as:

$$C(Q_t) \;= A(d/Q_t) + h(Q_t/2)(1 + d/p)\,. \tag{5.21}$$

Assume that customers are served in a common rotation cycle (length T_t), and that the production rate matches the sum of the demand rates. Because there is a common product, further assume that the queue holding cost is the same for all customers. Using T_t as the decision variable, the total cost for the rotation can be expressed as:

$$\begin{aligned} C(T_t) &= nA'/T_t + \Sigma\,[h(T_t d_i/2)(1 + d_i/\Sigma\, d_j)] \qquad (5.22)\\ C(T_t) &= nA'/T_t + h(T_t/2)(nd' + E(d_i{}^2)/d') \\ C(T_t) &= nA'/T_t + h(T_t/2)d'(n + 1 + C^2), \end{aligned}$$

where:

d' = average demand rate
A' = average value of A_i, among customers i=1,...,n
C = coefficient of variation of the demand rate
$E(d_i^2)$ = average of enclosed quantity.
The optimized values of T_t and $C(T_t)$ are then:

$$T_t^* = \sqrt{2nA'/hd'(n+1+C^2)} \tag{5.23a}$$

$$C(T_t^*) = \sqrt{2nA'hd'(n+1+C^2)}\,. \tag{5.23b}$$

Note that if n=1, C must equal zero, and the model reduces to the same form as Eq. 5.17, or the simple single plant/single customer case. As n approaches infinity, the model converges toward something like the classic EOQ model, with $T^* = \sqrt{2A'/hd'}$, and $C(T^*)/n = \sqrt{2A'hd'}$. However, they are based on averages among all customers, not individual customer values. The scenario demonstrates that when the demand for an individual customer falls well below the production capacity, the queue model is much like the classic EOQ

If the production capacity greatly exceeds the demand rate, it might be reasonable to optimize order quantities on an individual customer basis. Eq. 5.21 could then serve as the objective function, resulting in the following solution:

$$Q^* = \sqrt{2Ad/h(1+d/p)} \tag{5.24a}$$

$$C(Q^*) = \sqrt{2Ahd(1+d/p)}\,. \tag{5.24b}$$

If p>>d, these results reduce to the exact same form as the classic EOQ.

E. Single Machine/Multiple Products: Single Customer In this scenario, demand occurs at a constant rate for each product, but production is cycled among products on a single machine, with set-ups and changeovers. First, products are assumed to be produced at the same rate, with the same queue holding cost. Later, this assumption is relaxed. As stated at the beginning of the chapter, set-up times are assumed to be negligible.

Figure 5.9 is the characteristic cumulative diagram for individual products. Given that each product must be produced at a different time (recall, a single machine is used), it is impossible to synchronize all products with distribution. The aggregate queue diagram for the rotation cycle (Figure 5.10) is more revealing. The similarity to Figure 5.6 is a striking feature of Figure 5.10, for it suggests that a rotation cycle can bear the same queue cost as simple single product cycles. That is, queues are built up at a rate p-d during a production phase, and depleted to zero at a

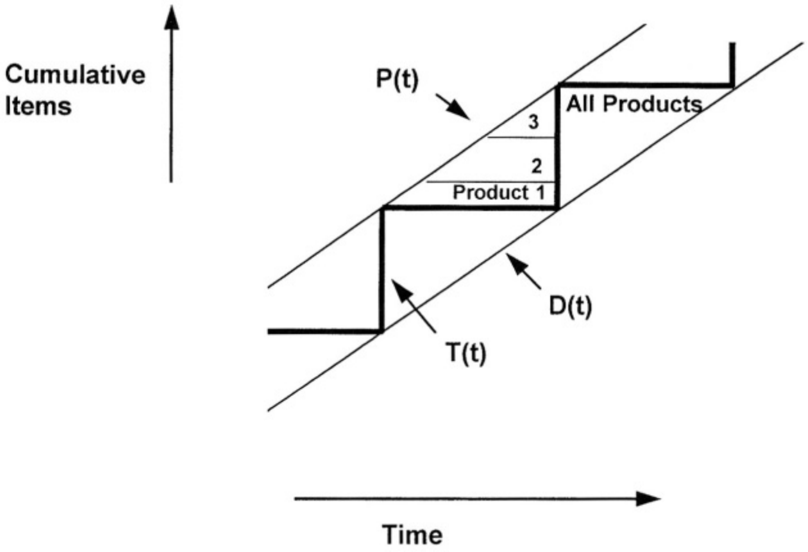


Figure 5.10 Single Machine/Multiple Products/Single Customer

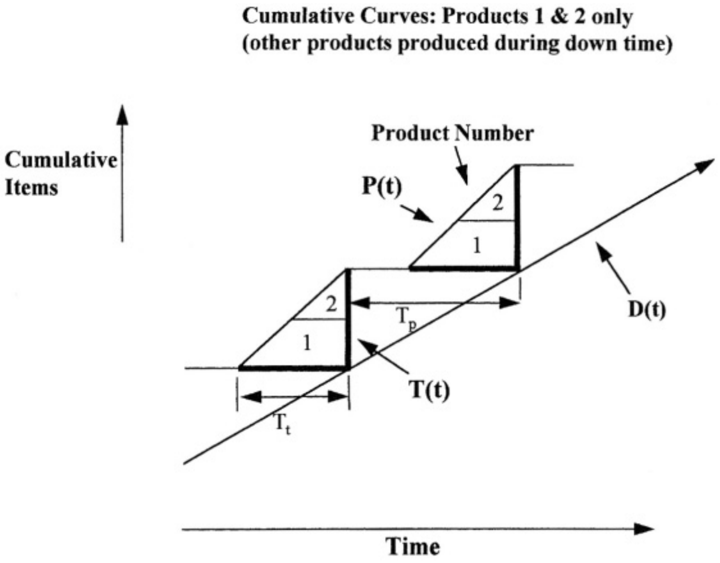

Cumulative Curves: Products 1 & 2 only
(other products produced during down time)
Cumulative Items
Product Number
P(t)
2
1
2
1
T_p
D(t)
T(t)
T_t
Time

Figure 5.11 Single Machine/Multiple Products/Single Customer: Decoupled Production and Distribution Cycles

rate d when production is cycled off. The batch transfer process acts to consolidate products into the same load independently of their position within the rotation. Hence, the first product in the rotation, which must wait nearly a full cycle before dispatch, is transported at the same time as the last product in the rotation.

The cost formulation can now be represented as follows:

$$C(T) \quad = (A+nS')/T + Tnhd' , \tag{5.25a}$$

Where

$$S' = \text{average of } S_j\text{, among customers } i=1,\ldots,n \tag{5.25b}$$

The optimized result is then:

$$T^* = \sqrt{(A+nS')/nhd'} \tag{5.26a}$$
$$C(T^*) = 2\sqrt{(A+nS')nhd'}. \tag{5.26b}$$

These results are the same as Scenario A (single plant/single customer), with the exceptions that the "set-up cost" includes both the order cost and the combined set-up cost across all products, and that the demand is the total demand across all products.

In some instances, it is preferable to decouple production and transportation cycles, with the latter occurring more frequently than the former. These decisions can be totally decoupled when one ignores the round-off errors that result when a dispatch occurs in the middle of a product's production run. The average queue at the manufacturer is then one-half the distribution batch size. The average aggregate queue at the customer is one-half the production batch size (Figure 5.11). Again assuming a rotation cycle, the total cost is the following:

$$C(T_p,T_t) = A/T_t + nS'/T_p + (T_t/2)nhd' + (T_p/2)nhd' . \tag{5.27}$$

The optimized results are then:

$$T_t^* = \sqrt{2A/nhd'} \tag{5.28a}$$
$$T_p^* = \sqrt{2S'/hd'} \tag{5.28b}$$
$$C(T_t^*,T_p^*) = \sqrt{2Anhd'} + n\sqrt{2S'hd'} . \tag{5.28c}$$

To be implemented, the cycle lengths must be adjusted so that the manufacturing cycle is an integer multiple of the transportation cycle.

F. Single Machine/Multiple Products: One Product per Customer In this final scenario, each production batch serves a single customer, and is fully synchronized with distribution. As soon as a production run is completed, all queue for the given product is dispatched to the customer. Production can either occur on a lot-for-lot basis, or with multiple distribution lots per production cycle.

Simple Rotation Cycle In a simple rotation cycle, production and transportation are synchronized with the same cycle length for all products/customers. The queue pattern for this scenario is batch production/batch distribution, synchronized lot-for-lot (Figure 5.7). The total cost for a cycle is then:

$$C(T) = n(A' + S')/T + \Sigma (T/2)h_id_i(1 + d_i/p_i) \tag{5.29}$$

The optimized cycle length and cost are:

$$T^* = \sqrt{2(A'+S')/[HD+E(h_id_i^2/p_i)]} \tag{5.30a}$$
$$C(T^*) = n\sqrt{2(A'+S')[HD+ E(h_id_i^2/p_i)]} \tag{5.30b}$$

As a point of contrast, Scenario D (single machine/single product/multiple customers) did not include set-up costs, and the production rate was simply the sum of the demand rates. This leads to a relatively higher set-up cost in Eq. 5.29, and a slightly modified queue holding cost. Hence, the optimal cycle length is longer for the multiple product scenario (F) than the single product scenario (D).

As a second point of contrast, Scenario E (single machine/multiple products/single customer) uses only one transportation set-up per cycle, and queue cost is larger. Hence, the optimal cycle length is longer for the multiple customer scenario (F) case than the single customer scenarios (E).

Large Production Capacity If the production capacity greatly exceeds the demand rate, it might be reasonable to optimize order and production quantities on an individual customer basis. The following could then serve as the objective function for an individual product:

$$C(T) = (A + S)/T + (T/2)hd(1 + d/p). \tag{5.31}$$

The optimized results are then:

$$T^* = \sqrt{2(A+S)/hd(1+d/p)} \tag{5.32a}$$
$$C(T^*) = \sqrt{2(A+S)hd(1+d/p)} \tag{5.32b}$$

Allowing for Multiple Dispatches If queue holding costs are sufficiently high, it might be reasonable to provide multiple dispatches per production cycle. Let:

I_0 = queue in the system at the start of the production run
Q_i = size of transportation batch i (i = 1,2,...), within a production run

The initial queue, I_0, is exhausted at the moment that batch 1 is transported. Hence, Q_1 equals the production during the time required to consume I_0 units of queue:

$$Q_1 = p(I_0/d)\,. \tag{5.33}$$

Similarly, all subsequent batch sizes are dictated by the prior batch sizes, in the following fashion:

$$Q_i = p(Q_{i-1}/d) = I_0(p/d)^i\,. \tag{5.34}$$

I_0 can now be derived, by recognizing that the sum of the transportation batch sizes within a cycle must equal the production batch size:

$$Q_p = \Sigma\, Q_i = I_0\ \Sigma\ (p/d)^i\,, \tag{5.35a}$$

or

$$I_0 = Q_p/[\ \Sigma\ (p/d)^i]\,. \tag{5.35b}$$

Referring to Figure 5.12, the average queue level can now be characterized as the sum of a base level, I_0, and an EPQ type quantity:

$$\text{Average Queue Size} = I_0 + Q_p[(p-d)/p]/2\,. \tag{5.36}$$

With I_0 given in Eq. 5-35b, the average queue size becomes:

$$C(T_p) = (mA+S)/T_p + hT_p[1/\sum_{i=1}^{m}(p/d)^i + (1-d/p)/2]\,, \tag{5.37}$$

where m is the number of transportation cycles per production cycle. Through a combination of search techniques and calculus, it is not difficult to optimize m and T_p within the above expression.

Summary of More Complicated Scenarios Introduction of scheduling considerations complicates EOQ and EPQ calculations in several ways. First, to avoid schedule conflicts, either a rotation cycle must be optimized, or simplifying assumptions must be made with respect to production capacity. Second, the combination of batch production and batch distribution results in somewhat non-standard forms for the queue equations. Third, both production and transportation set-up costs must be considered when optimizing cycle length.

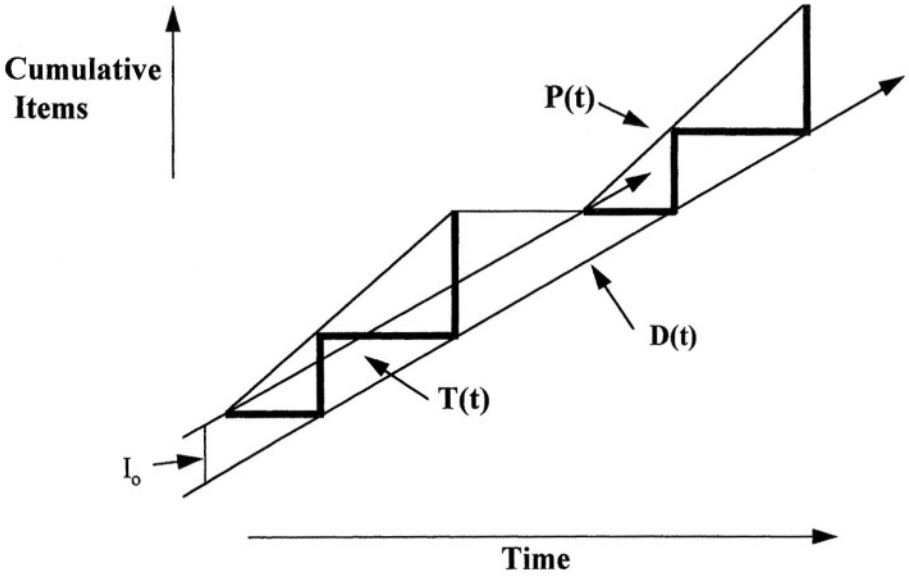


Figure 5.12 Synchronized/Multiple Transportation Lots

Extensions

The scenarios presented in this section served to illustrate a methodology, and to illustrate the complexity of accurately accounting for queue sizes when input and output processes are discontinuous. Many extensions have been covered in the literature, including the following:

Random Cycle Length Batch processes seldom occur precisely on schedule. Consequently, the headways between batches vary randomly, causing average and maximum queues sizes to increase. This occurs because customers are more likely to arrive during longer headways, and because the average wait for a long headway is greater than the average wait for a short headway. In the special case where customers arrive at random relative to batch times, the average wait is given by:

$$E(W) = [E(h)/2][1 + C^2(H)] \tag{5.38}$$

Where E(h) is the mean headway and C(H) is the headway coefficient of variation. If batches occur with the randomness of a Poisson process, E(W) = E(h), which reflects the memoryless property of the exponential distribution (the headway distribution for a Poisson process).

Non-Stationary Demand Headways between batch services should vary in relationship to the demand rate. Larger demand invites shorter headways, according to an inverse square-root relationship. In some systems, however, the total waiting time per dispatch should stay constant for all demand rates. For example, if demand increases by a factor or 2, then headway should decrease by a factor of $\sqrt{2}$, batch size should increase by a factor of $\sqrt{2}$, and the product stays constant. Another common characteristic of optimal batching is that the arrival time at the time of service equals the ratio of the number of customers served to the time until the subsequent dispatch.

Multiple Stop Transportation Routes Scenarios can be further delineated by transportation characteristics, principally, whether or not transportation equipment is shared among customers and plants. Sharing, in the form of multiple-stop pick-up and delivery routes, can provide substantial savings in transportation and queue cost in low demand systems (Burns *et al,* 1985; Daganzo, 1985; Hall, 1985). This naturally adds complexity, as it may be desirable to serve some customers less frequently than others, yet also ensure that their service intervals are synchronized so that all shipments within a territory occur on a common schedule.

Capacity Considerations Capacity is important in two ways. First, batch sizes may be limited by the size of available vehicles and, second, the batch service system may be limited in the total rate at which customers are processed. Either factor leads to solutions that violate the "Dispatching Rule" presented in this section. In the former case, the optimal *feasible* batch size is generally the minimum of two values: the vehicle size or the cost minimizing batch size, as determined in this section. In the latter case, the batch size may need to be enlarged, to reduce the batch frequency, and reduce loss times when initiating batches. This is especially relevant to traffic signals, where cycle lengths are typically defined by capacity considerations rather than set-up costs.

Real-time Control Random variations in demand can make it desirable to alter headways and batch sizes in real-time. When the number of customers is insufficient, a headway can be extended or a batch can be cancelled. Dispatch times might also be altered to provide greater consistency in headways, thus minimizing its coefficient of variation and reducing waiting time.

5.6 Future Directions

As it has in the past, future research on queueing in transportation is likely to respond to innovations in the methods of transportation. Technologies for automating and controlling vehicle movements on highways has already stimulated queueing research, addressing delays and capacities associated with lane-following strategies, lane-assignment and entrance/exit processes. Changes in aircraft routing and control, possibly allowing aircraft to travel in free-space rather than on prescribed paths, is also likely to stimulate original research.

Future research will also be directed at gaps in the literature. A notable example is the paucity of research on queueing within terminals, and on the interactions between sorting processes and transportation processes. Relatively little is known on how terminal queues interact with vehicular queues. Yet the problem grows in importance, as more shipments are transported through parcel transportation companies, in which sortation is a critical cost driver.

Finally, despite the considerable accomplishments in understanding the behavior of queues on roadways, researchers have been largely unsuccessful in actually eliminating vehicular queues. It appears inevitable, as observed long ago, that in the absence of road pricing queues will exist. Developing and testing pricing methods for roadways, and then creating a mechanism by which they can be implemented, is perhaps the most important challenge to the field. But success in this area demands far more than an understanding of the mathematics of queues; it demands accurate representations of human behavior, along with knowledge of the institutional and technical aspects of toll collection.

One clear aspect of research on queueing in transportation is that the most significant papers have offered a blend of empiricism and theory, and have been innovative in exploring new applications. It is simply insufficient to develop the mathematical theorems. The papers that best explain important “real-world” phenomena, or provide generalizable methods for system design and operation, have been the most significant, and will likely continue to be in the future.

5.7 Acknowldgement

Portions of the chapter appeared earlier in Transportation Research (Hall, 1996). These excerpts are reprinted with the permission of the publisher, Pergammon Press. Research was supported in part by National Science Foundation grant DMI-9732878.

5.8 References

Allsop, R.E. (1970). Optimisation techniques for reducing delay to traffic in signalised road networks. Ph.D. Thesis, University of London.

Andreatta, G. and Romanin-Jacur, G. (1987). Aircraft flow management under congestion. *Transportation Science,* **21**, 249-253.

Bailey, N.T.J. (1954). On queueing processes with bulk service, Journal of the Royal Stastical Society B, **16**, 80-87.

Barnett, A. (1974). On controlling randomness in transit operations. *Transportation Science,* **8**, 102-116.

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

Beckmann, M.J. (1965). On optimal tolls for highways,tunnels and bridges, In: *Vehicular Traffic Science* (Edie, Herman and Rothery, eds.), 331-341. Elsevier, New York.

Blumenfeld, D.E., Burns, L.D., Diltz, J.D. and Daganzo, C.F. (1985). Analyzing trade-offs between transportation, inventory and production costs on freight networks. *Transportation Research,* **19B**, 361-380.

Blumenfeld, D.E., Burns, L.D. and Daganzo, C.F. (1991). Synchronizing production and transportation schedules. *Transportation Research,* **25B**: 23-27.

Burns, L.D., Hall, R.W., Blumenfeld, D.E. and Daganzo, C.F. (1985). Distribution strategies that minimize transportation and inventory cost. *Operations Research,* **33**, 469-490.

Chaiken, J. and Larson, R. (1972). Methods for allocating urban emergency units: a survey. *Management Science,* **19**, 110-130.

Cheng, T.E.C. and Allam, S. (1992). A review of stochastic modelling of delay and capacity at unsignalized interjections. *European Journal of Operations Research,* **60**, 247-259.

Clayton, A.J.H. (1941). Road traffic calculations, *Journal of Institute of Civil Engineers,* **16**, 247-284, 558-594.

Dafermos, S. C. and Sparrow, F.T. (1971). Optimal resource allocation and toll patterns in a user-optimized transportation network. *Journal of Transportation Economic Policy,* **5**, 198-200.

Daganzo, C.F. (1982). Supplying a single location from heterogeneous sources. *Transportation Research,* **19B**: 409-420.

Daganzo (1989). On the coordination of inbound and outbound schedules at transportation terminals, Institute of Transportation Studies Research Report, Berkeley, California.

Daganzo, C.F., Dowling, R.G. and Hall, R.W. (1983). Railroad classification yard throughput: the case of multistage triangular sorting. *Transportation Research,* **17A**, 95-106.

Edie, L.C. (1956). Traffic delays at toll booths. *Journal of Operations Research,* **4**, 107-138.

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

Edie, L.C., and Foote, R.S. (1958). Traffic flow in tunnels, *Proceedings of the Highway Research Board,* **37**, 334-44.

Edie, L.C., and Foote, R.S. (1960). Effect of shock waves on tunnel traffic flow. *Proceedings of the Highway Research Board,* **39**, 492-505.

Frank, O. (1966). Two-way traffic in a single line of railway. *Operations Research.* **14**, 801-811.

Gallagher, H.P. and Wheeler, R.C. (1958). Nonstationary queueing probabilities for landing congested aircraft. *Operations Research,* **6**, 264-275.

Ghobrial, A., Daganzo, C.F. and Kazimi, T. (1982). Baggage claim area congestion at airports: an empirical model of mechanized claim device performance. *Transportation Science,* **16**, 246-260.

Grace, M.M. and Potts, R.B. (1964). A theory of diffusion of traffic platoons. *Operations Research,* **12**, 255-275.

Green, L. and Kolesar, P. (1989). Testing the validity of a queueing model of police patrol. *Management Science,* 35, 127-148.

Greenshields, B.D. (1935). A study of traffic capacity. Poceedings of the Highway Research Board, **14**.

Hall, R.W. (1985). Determining vehicle dispatch frequency when shipping frequency differs among suppliers. *Transportation Research,* **19B**: 421-431.

Hall, R.W. (1996). On the integration of production and distribution: economic order and production quantity implications. *Transportation Research,* **30B**, 387-403.

Hall, R.W. (1996). On the integration of production and distribution: economic order and production quantity implications. *Transportation Research,* **30**, 387-403.

Hall, R.W. (1991). *Queueing Methods for Services and Manufacturing,* Prentice Hall, Engelwood Cliffs, New Jersey.

Hall, R.W. (2002). Control of vehicle dispatching on a cyclic route serving trucking terminals. *Transportation Research,* **36A**, 257-276.

Hall, R.W., M. Dessouky and Q. Lu (2001). Optimal holding times at transfer stations, *Computers and Industrial Engineering,* **40**, 379-397.

Hankin, J.D. and Wright, R.A. (1958). Passenger flow in subways. *Operations Research Quarterly,* **9**, 81-88.

Herman, R., Montroll, E.W., Potts, R.B. and Rothery, R.W. (1959). Traffic dynamics: analysis of stability in car following, *Operations Research,* **7**, 86-106.

Horonjeff, R. (1969). Analysis of passenger and baggage flows in airport terminal buildings. *Transportation Research,* **6**, 446-451.

Ignall, E.J., Kolesar, P. and Walker, W.E. (1978). Using simulation to develop and validate analytic models: some case studies. *Operations Research,* **26**, 237-253.

Kolesar, P. (1975). A model for predicting average fire engine travel times. *Operations Research,* **23**, 603-613.

Kolesar, P. and Blum, E.H. (1973). Square root laws for fire engine response distances. *Management Science,* **19**, 1368-1378.

Kolesar, P.J., K.L. Rider, T.B. Crabill and Walker, W.E. (1975). A Queueing-linear programming approach to scheduling police patrol cars. *Operations Research.* 23, 1045-1062.

Koopman, B.O. (1972). Air-terminal queues under time-dependent conditions. *Operations Research,* **20**, 1089-1114.

Larson, R.C. (1972). *Urban Police Patrol Analysis.* The MIT Press, Cambridge, Massachusetts.

Lighthill, M.J. and Whitham, G.B. (1955). On kinematic waves, II, a theory of traffic flow on long straight crowded roads. *Proceedings of the Royal Society,* London, Series A **229**, 317-345.

Little, J.D.C. (1961). A proof for the queueing formula $L = \lambda W$. *Operations Research,* **9**, 383-387.

Little, J.D.C., Kelman, M.D. and Gartner, N.H. (1981). MAXBAND: a program for setting signals on arterials and triangular networks. *Transportation Research Record,* **795**, 40-46.

Lovas, G.G. (1994). Modeling and simulation of pedestrian traffic flow. *Transportation Research,* **28B**, 429-443.

May, A.D. and Keller, H.E. (1967). A deterministic queueing model. *Transportation Research,* **1**, 117-128.

Makigami, Y., Newell, G.F. and Rother, R. (1971). Three-dimensional representations of traffic flow. *Transportation Science,* **5**, 302-313.

Miller, **A.J.** (1963). Settings for fixed-cycle traffic signals. *Operational Research Quarterly,* **14**, 373-386.

Morse, P. (1958). *Queues, Inventories and Maintenance.* New York: John-Wiley.

Neuts, M.F. (1967). A general class of Poisson queues with Poisson input. *Annals of Mathematical Statistics,* **38**, 759-770.

Newell, G.F. (1965). Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Review,* **2**, 223-240.

Newell, G.F. (1971, 1982). *Applications of Queueing Theory,* Chapman and Hall, London.

Newell, G.F. (1975). Control of pairing of vehicles on a public transportation route, two vehicles, one control point. *Transportation Science,* **9**, 248-264.

Newell, G.F. (1979). Airport capacity and delays, *Transportation Science,* **13**, 201-241.

Newell, G.F. (1993). A simplified theory of kinematic waves in highway traffic, Parts I-III. *Transportation Research,* **27B**, 281-313.

Newell, G.F. (1955). Mathematical models for freely-flowing highway traffic, *Operations Research,* **3**, 176-186.

Odoni, A.R. (1987). The flow management problem in air traffic control. In A.R. Odoni, L. Bianco and G. Szego (Eds.), *Flow Control of Congested Networks,* 269-288. Springer Verlag, New York.

Older, S.J. (1968). Movement of pedestrians on footways in shopping streets. *Traffic Engineering and Control,* **13**, 434-438.

Oliver, R.M. and Samuel, A.H. (1967). Reducing letter delays in post offices. *Operations Research,* **10**, 839-892.

Osuna, E.E. and Newell, G.F. (1972). Control strategies for an idealized public transportation system. *Transportation Science,* **6**, 52-72.

Pacey, G.M. (1956). Progress of a bunch of vehicles released from a traffic signal. Repart RN/2665, Dept. of Scientific and Industrial Research, Road Research Laboratory, England.

Petersen, E.R. (1974). Over-the-road transit time for a single track railway. *Transportation Science,* **8**, 65-74.

Petersen, E.R. (1977a). Railyard modeling, part I, prediction of put-through time. *Transportation Science,* **11**, 37-49.

Petersen, E.R. (1977b). Railyard modeling, part II, the effect of yard facilities on congestion. *Transportation Science,* **11**, 50-59.

Peterson, M.D., Bertsimas, D.J. and Odoni, A.R. (1995). Decomposition algorithms for analyzing transient phenomena in multiclass queueing networks in air transportation. *Operations Research,* **43**, 995-1011.

Peterson, M.D., Bertsimas, D.J. and Odoni, A.R. (1995). Models and algorithms for transient queueing congestion at airports, *Management* Science, **41**, 1279-1295.

Powell, W.B. (1985) Analysis of vehicle holding and cancellation strategies in bulk arrival, bulk service queues, *Transportation Science,* **19**, 352-377.

Powell, W.B. (1986) Approximate closed form moment formulas for bulk arrival, bulk service queues. *Transportation Science,* **20**, 13-23.

Powell, W.B. and Humblet, P. (1984) The bulk service queue with a general control strategy, *Operations Research,* **19**, 352-377.

Richards, P.I. (1956). Shock waves on the highway. *Operations Research,* **4**, 42-51.

Richetta, O. and A.R. Odoni (1994). Dynamic solution to the ground-holding problem in air traffic control. *Transportation Research,* **28A**, 167-185.

Rider, K.L. (1976). A parametric model for the allocation of fire companies in New York City, *Management Science,* **28**, 146-158.

Robertson, D.I. (1969). TRANSYT: a traffic network study tool. Road Research Laboratory, LR 253, Crowthorne, England.

Robuste, F. and Daganzo, C.F. (1992). Analysis of baggage sorting schemes for containerized aircraft. *Transportation Research A,* **26A**, 75-92.

Tanner, J.C. (1962). A theoretical analysis of delays at an uncontrolled intersection. *Biometrica,* **49**, 163-170.

Turnquist, M.A. and Daskin, M.S. (1982). Queueing models of classification and connection delay in railyards. *Transportation Science,* **16**, 207-230.

Vickrey, W. (1963). Pricing in urban and suburban transportation. *American Economic Review,* **53**, 452-465.

Vickrey, W. (1969). Congestion theory and transport investment. *American Economic Review,* **59**, 251-261.

Wardrop, J.G. (1952). Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers II.* **1**, 325-364.

Webster, F.V. (1958). Traffic signal settings, Road Research Laboratory Technical Paper, Her Majesty's Stationery Office, London.

Welch, N. and Gussow, J. (1986). Expansion of Canadian Nation Railway's line capacity. *Interfaces,* **16:1**, 51-64.

This page intentionally left blank

6 TRAFFIC FLOW AND CAPACITY

Michael J. Cassidy

6.1 Introduction

The design of highways, runways, ports or any transportation facility is guided by knowledge and theory of the traffic streams they serve. A facility's scale, its geometry and its control measures are selected to affect certain properties of its traffic, such as the travel delay, the separation between vehicles, etc. In the case of highway traffic, the emphasis of this chapter, these are usually properties that are collected from, or averaged over, some number of vehicles. This is because the behavior of one driver differs from that of another, sometimes in complicated or even unexpected ways, and the traffic engineer typically seeks properties that are reproducible or predictable; i.e., properties that are not sensitive to driver variations.

Chapter 6 is devoted to methods of measuring traffic stream properties and of predicting how these properties evolve over time and space. Certain emphasis is given to flow restrictions, or bottlenecks, and to the estimation of their capacities since traffic streams are often impacted by these restrictions.

Section 6.1 provides some important definitions along with descriptions of some graphical tools for analyzing the motion of objects on transport systems. Most of what is presented here is applicable to any mode of transportation. Moreover, this information is necessary background for the treatment of highway traffic offered in the remaining sections of the chapter. Section 6.2 describes methods of processing traffic data measured, for example, by loop detectors to identify bottleneck locations along highway facilities without traffic signals or other exogenous controls. The use of these methods to estimate bottleneck capacities is likewise shown here. Section 6.3 presents methods of estimating capacities and vehicle delays at highway intersections controlled by traffic signals or stop signs. Theories for predicting the evolution of highway traffic are the subject of section 6.4. A simple theory is described here in some detail and other theories are briefly noted.

The chapter provides references for all of the topics covered. Notes on the historical developments and future research directions are likewise included for many of the subjects.

Basic Concepts

This first section includes definitions for some of the properties commonly used to characterize traffic streams. So-called generalized definitions, which preserve useful relations between the properties, are part of this discussion. Also described in section 6.1 is a three-dimensional representation of traffic streams. This representation makes clear the conservation concepts that are fundamental to theories of traffic evolution. In particular, it illustrates the relation between two important graphical tools for presenting and interpreting traffic data: 1) curves of cumulative vehicle count and 2) trajectories plotted on time-space diagrams. A description of the latter tool is the starting point for this section.

Before embarking on this discussion, however, there are two points that deserve mention. First, the subjects covered in section 6.1 do not involve theory or conjecture. Rather the concepts are true by definition. Secondly, the discussion in this section owes much to notes composed by Newell (unpublished) for a graduate course in transportation engineering and to a book written by Daganzo (1997).

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 6-1 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_ℓ (*t*).

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.”

The literature includes numerous illustrations of how these diagrams, even when drawn approximately, can be used in solving problems that frequently arise in transport. As examples, Daganzo (1997) shows how trajectory plots can help to select desirable scheduling policies in rail and in sea transportation; Newell (1979) used them in deriving expressions of airport runway capacity; and they are a widely used tool for synchronizing traffic signals along an arterial (Newell, 1989).

This chapter will frequently rely upon time-space diagrams to illustrate fundamental concepts. They are used immediately below to convey the precise meanings of some important properties of the traffic stream.

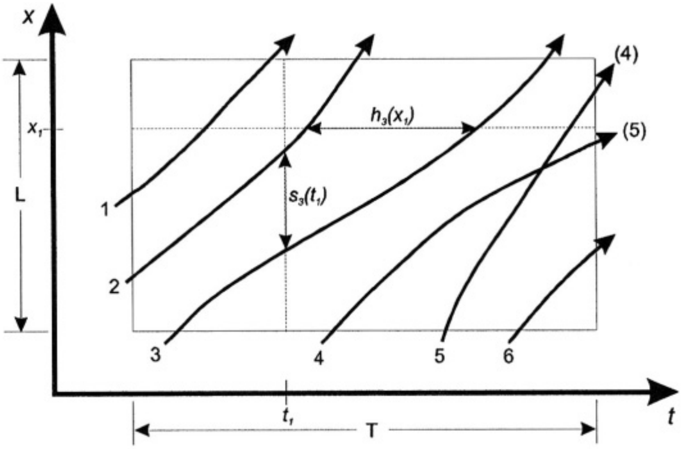


Figure 6-1. Time-space diagram.

Definitions of Some Traffic Stream Properties. It is evident from Figure 6-1 that the slope of the ℓth trajectory is object ℓ's instantaneous velocity, $v_\ell(t)$, i.e.,

$$v_\ell(t) \equiv dx_\ell(t)/dt\,, \tag{6.1}$$

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 6-1, 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.,

$$h_i(x_1) \equiv t_i(x_1) - t_{i-1}(x_1) \tag{6.2}$$

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

$$q(T, x_1) \;\equiv\; m/T. \tag{6.3}$$

For observation intervals containing large *m*,

$$\sum_{i=1}^{m} h_i(x_1) \approx T \tag{6.4}$$

and thus,

$$q(T,x_1) \approx \frac{1}{\frac{1}{m}\sum_{i=1}^{m} h_i(x_1)} = \frac{1}{\bar{h}(x_1)}, \tag{6.5}$$

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_l, $s_j(t_l)$, is the distance separating *j* from the next downstream object; i.e.,

$$s_j(t_l) \equiv x_{j-1}(t_l) - x_j(t_l). \tag{6.6}$$

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

$$k(L, t_l) \equiv n/L. \tag{6.7}$$

If the *L* contains large *n*,

$$\sum_{j=1}^{n} s_j(t_1) \approx L \tag{6.8}$$

and

$$k(L,t_1) \approx \frac{1}{\frac{1}{n}\sum_{j=1}^{n} s_j(t_1)} = \frac{1}{\bar{s}(t_1)}, \tag{6.9}$$

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

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_l over observation interval *T*,

$$\alpha(T, x_1) = \frac{1}{m}\sum_{i=1}^{m}\alpha_i(x_1), \tag{6.10}$$

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.,

$$\alpha(L, t_1) = \frac{1}{n}\sum_{j=1}^{n}\alpha_j(t_1). \tag{6.11}$$

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.

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 6-2, 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 6-2 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 6-2, 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.

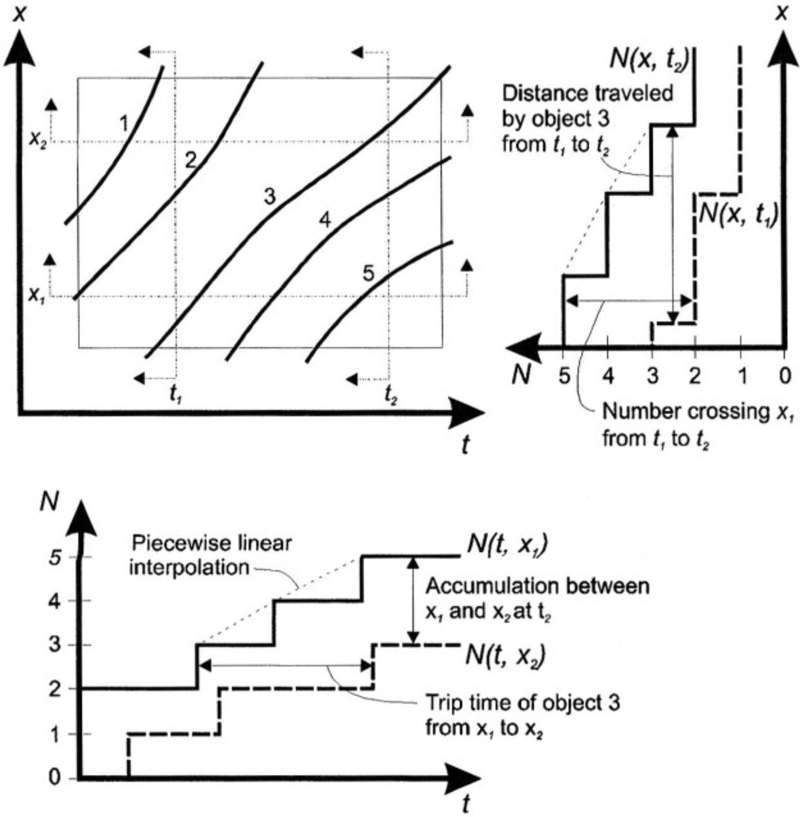

x
x
1
2
3
4
5
x_2
x_1
t_1
t_2
t
$N(x, t_2)$
Distance traveled by object 3 from t_1 to t_2
$N(x, t_1)$
N 5 4 3 2 1 0
Number crossing x_1 from t_1 to t_2
N
5
4
3
2
1
0
Piecewise linear interpolation
$N(t, x_1)$
Accumulation between x_1 and x_2 at t_2
$N(t, x_2)$
Trip time of object 3 from x_1 to x_2
t

Figure 6-2. Three-dimensional representation.

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). 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.

By examining trends on the cumulative curves, one can observe how flows and densities change with time and space, respectively. This can be a powerful diagnostic and examples are provided in later sections. Suffice to say that by defining flows and densities as they are displayed on the cumulative curves, their values may be taken over intervals that exhibit fixed trends (i.e., 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).

The Conservation Law. The existence of the surface $N(t, x)$ implies that objects did not enter or exit within the region of interest. If the N can be replaced by a smooth surface N' so that at all points within the region the *instantaneous* flows and densities can be defined as $\partial N'(t,x)/\partial t$ and $\partial N'(t,x)/\partial x$, respectively, and if N' has second derivatives (i.e., flow and density are smooth), then $\partial^2 N'(t,x)/\partial x \partial t$ must be equal to $\partial^2 N'(t,x)/\partial t \partial x$ and thus

$$\frac{\partial q(t,x)}{\partial x} = -\frac{\partial k(t,x)}{\partial t}. \tag{6.12a}$$

The more common form of this conservation equation is

$$\frac{\partial q(t,x)}{\partial x} + \frac{\partial k(t,x)}{\partial t} = 0. \tag{6.12b}$$

It is by direct consequence of the conservation equation that the speed of an interface separating two (different) stationary traffic conditions, u, is the change in flow across the interface over the change in density across the interface; i.e.,

$$u = \Delta q / \Delta k\,. \tag{6.13}$$

The reader may refer to Daganzo (1997, pp. 97-103) for the simple derivation of (6.13) and for further discussion of this. The conservation equation also gives rise to the well-known expression for computing the relative flow measured by a moving observer in (stationary) traffic; see again Daganzo (1997).

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 6-3 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 × time. The numerator 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 6-3, 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.

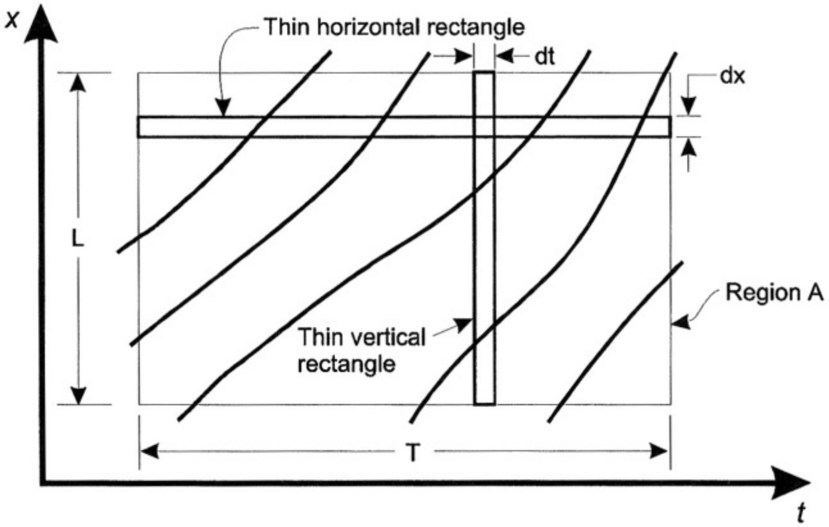


Figure 6-3. Trajectories in time-space region.

As the analogue to this, the thin, vertical rectangle in Figure 6-3 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, $d(A) = m \cdot dx$, the generalized mean velocity becomes

$$v(A) = \frac{d(A)}{t(A)} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{v_i}}, \tag{6.14}$$

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

$$v(A) = \left[\frac{1}{m}\sum_{i=1}^{m} p_i\right]^{-1}. \tag{6.15}$$

Eq. 6.15 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).

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 6-4. The (generalized) occupancy in the region A, $\rho(A)$, 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 $\rho(A)$ 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 $\rho(A)$ to the $k(A)$,

$$\textit{average vehicle length} = \frac{\rho(A)}{k(A)} = \frac{\textit{area of the shaded strips}}{|A|} \cdot \frac{|A|}{t(A)}. \tag{6.16}$$

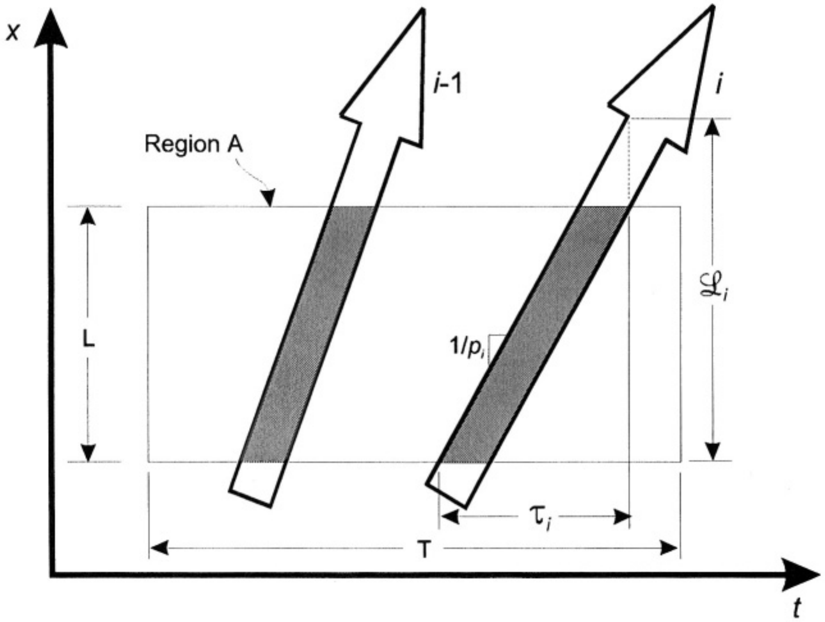


Figure 6-4. Trajectories of vehicle fronts and rears.

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, $\mathfrak{L}_i$ is the summed length of the detection zone and the length of vehicle i. Therefore,

$$\sum_{i=1}^{m} \tau_i = \sum_{i=1}^{m} \mathcal{L}_i \cdot \frac{1}{v_i} = \sum_{i=1}^{m} \mathcal{L}_i p_i \tag{6.17}$$

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

$$\frac{\sum_{i=1}^{m} \tau_i}{T} = \frac{\frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i}{\frac{1}{m} \cdot T} = q(A) \cdot \frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i, \tag{6.18}$$

it follows that

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

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

$$\rho = k(A) \cdot \left[\frac{\sum_{i=1}^{m} \mathcal{L}_i \cdot p_i}{\sum_{i=1}^{m} p_i} \right], \tag{6.19}$$

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 (6.19) 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 $\mathcal{L}_i$ 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.

6.2 Freeway Bottlenecks and their Capacities

This section provides description of some simple diagnostics for locating bottlenecks on freeways and on highways without control devices, such as traffic signals. Also described here are techniques for estimating the capacities of these bottlenecks. Cumulative curves constructed from counts and occupancies measured at neighboring locations along the roadway serve as the diagnostics. By transforming and visually inspecting these curves as described below, one can verify the occurrence of an active bottleneck, where the word active is used to denote 1) a queue's presence immediately upstream, which ensures that vehicles are discharging through the bottleneck at a maximum rate, and 2) the absence of any downstream effects that would impede this discharge. Once having identified these two essential conditions, the bottleneck's capacity may be estimated. Plotting the cumulative counts that have been measured immediately downstream of the bottleneck can aid in this endeavor.

To illustrate these diagnostics, they are applied to a bottleneck that formed downstream of a merge; some details of this site are described below. Of note, the diagnostics may be used for examining bottlenecks caused by other types of geometric inhomogeneities, including curves, lane reductions, and diverges.[1] They can also be applied to bottlenecks formed by incidents, such as vehicle stalls or collisions.

An Example Application

The diagnostics will be described with data taken from the freeway section shown in Figure 6-5, a segment of the Queen Elizabeth Way in Ontario, Canada. All the data presented here came from a single morning rush and were measured with loop detectors installed at four locations along this freeway segment. These four detectors are labeled in the figure as per the numbering strategy that had been adopted by the region's transportation authority. The vehicle counts and occupancies were measured in 30-second intervals and the resulting step-wise cumulative curves were smoothed using piece-wise linear interpolations.

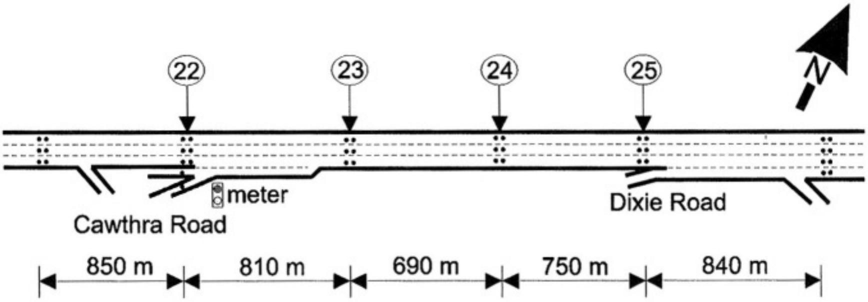


Figure 6-5. Segment of Queen Elizabeth Way.

[1] At a diverge bottleneck, the queue(s) formed by vehicles wishing to exit the freeway may entrap some through-moving vehicles and thereby affect the bottleneck's capacity. The reader may note for now these special circumstances that surround diverge operation. More is said about this in section 6.4.

As a useful aside, cumulative count curves may be obtained from vehicle arrival times measured by human observers stationed roadside when loop detectors are not deployed near a bottleneck of interest. A detailed description of one such experiment can be found in Smilowitz, et al. (1998).

Locating the Active Bottleneck

Figure 6-6 shows cumulative count, or N-curves. These were constructed from counts measured across all lanes at the four detector stations during the onset of queueing. The counts at each detector were started ($N = 0$) with the passage of an imaginary reference vehicle. These passage times were based upon estimated free-flow trip times between each detector because vehicle N . 0 did not encounter queueing at the site. The curves in Figure 6-6 have been transformed in the following two ways.

First, each curve, along with its respective time axis, was shifted to the right by the average free flow trip time between the respective detector and downstream detector 25. Having done this, the vertical displacements between curves are the excess vehicular accumulations due to traffic delays. Such a shift is advantageous because two superimposed curves indicate that traffic in the intervening segment is flowing freely; every feature of an upstream N-curve is passed to its downstream neighbor a free-flow trip time later. Secondly, Figure 6-6 shows only the differences between each curve of cumulative count to time t and the line $N = q_o t'$, where t' is the elapsed time from the curve's starting point (N=0) and q_o is the rate used for re-scaling the cumulative curve. This is important because reducing the N (displayed by the curves) by a background flow q_o magnifies details without changing the excess accumulations (Cassidy and Windover, 1995).

The superimposed curve portions in this figure indicate that traffic was initially in free flow and remained in free flow between detectors 24 and 25. The marked separation of curves 24/25 from curve 23 from about 6:27 onward (as shown by the darkened arrow) indicates that a bottleneck was activated a little earlier between detectors 24 and 23. The subsequent separation of curve 23 from curve 22 indicates when the queue arrived to detector 23.

This illustrates how transformed cumulative count curves expose active bottlenecks by revealing the excess accumulations upstream and the free-flow conditions downstream. Further verification of a bottleneck's activation can be obtained by using re-scaled curves of cumulative occupancy, as described below.

Additional evidence of the bottleneck. A bottleneck's location may be confirmed using curves of cumulative occupancy versus time (T-curves) where cumulative occupancy is the total vehicular trip time over the loop detector by time t (Lin and Daganzo, 1997). To illustrate this, Figure 6-7 presents T-curves for the four detector stations. As before, these curves were constructed for times near the onset of

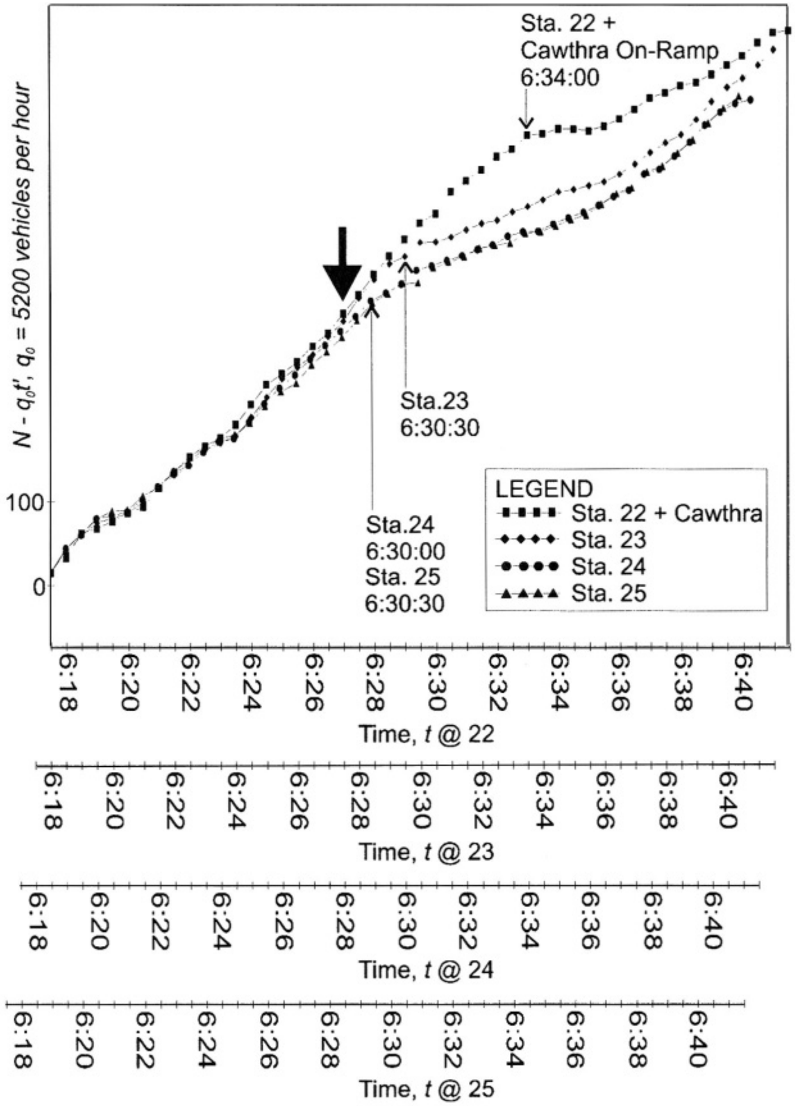


Figure 6-6. Transformed N-curves.

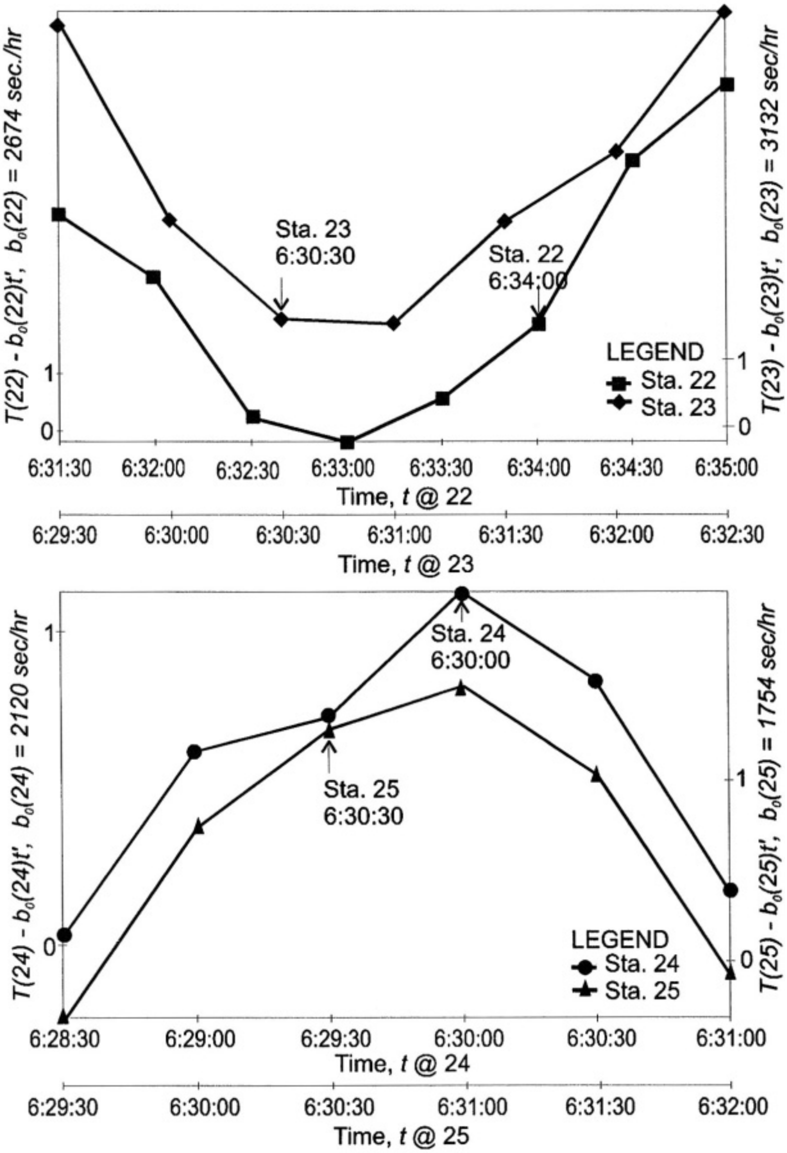


Figure 6-7. Re-scaled T-curves.

queueing and the occupancies were those measured in all travel lanes. Again for the purpose of magnifying details, the figure presents the differences between each curve of cumulative occupancy to time t and the line $T = b_o(x)tN$, where $b_o(x)$ is the background occupancy rate used at detector x and tN is the elapsed time from the curve's starting point ($T = 0$).

The T-curves in the lower half of Figure 6-7 display concave shapes, indicating sudden reductions in the occupancy rates at detectors 24 and 25. These lower occupancies prevailed during times that coincided (approximately) with the flow reductions previously revealed by the N-curves; the times marking the onset of these flow reductions are labeled on the lower portion of Figure 6-7. Conversely, the upper half of Figure 6-7 indicates that upstream detectors 23 and 22 each measured a rather abrupt increase in the occupancy rates. These occupancy changes coincided closely with the flow reductions previously identified at these detectors. The times marking the onsets of these upstream flow reductions are labeled in the top half of the figure.

The interpretation. The traffic patterns described above reveal that a forward-moving interface signaling lower flow and occupancy,[2] along with a backward-moving queue, emanated from between detectors 23 and 24. This confirms that a bottleneck was activated somewhere in the intervening segment.

Repeated observations. As part of a study on freeway capacity (Cassidy and Bertini, 1999), traffic conditions on the freeway segment in Figure 6-5 were examined using data collected on several weekday mornings. During each of these mornings, the bottleneck formed between detectors 23 and 24. The reason(s) that the bottleneck occurred about a kilometer or more downstream of the merge is a subject of ongoing research. For now, it suffices to note that a bottleneck's location is not always obvious. This, in turn, underscores the value of the diagnostics illustrated above.

The bottleneck's persistence. Knowing the duration that a bottleneck remains active is important for certain tasks such as estimating the bottleneck's capacity or predicting the evolution of its queue. (These topics are covered later in this section and in section 6.4). Toward determining a bottleneck's persistence, one can construct transformed N-curves that are similar to those in Figure 6-6, but that have

[2] One would expect to view this type of forward-moving interface following a sudden restriction in the flow upstream. This characteristic has been consistently observed to accompany the onset of upstream queueing, despite the notable absences of any traffic accidents or other exogenous causes of flow reduction (Cassidy and Bertini, 1999). Why queue formations give rise to this traffic characteristic is currently under investigation.

been constructed using counts taken over an extended time period; i.e., one that spans the entire rush. By constructing them over a prolonged period, the displacements in the curves reveal the persistence of upstream queuing. Some examples of this are provided in Cassidy and Bertini (1999).[3]

It is likewise advantageous to construct re-scaled curves of *N* and of *T,* again for extended durations spanning the rush, using measurements taken downstream of the bottleneck. (In the context of the present example, curves could be constructed from measurements at detectors 24 and/or 25). One can examine these *N* and *T* collectively to check for the arrival of any queue that may have spilled-over from further downstream and restricted flow through the bottleneck of interest. Such an arrival is marked by a reduction in the *N* accompanied by an increase in the *T,* as shown previously. Long-run *N*- and *T*-curves constructed at detector 25 are shown as part of the discussion presented next.

Estimating Bottleneck 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

$$q_{max} \equiv \lim_{T \to \infty} \left(\frac{N_{max}}{T} \right), \qquad (6.20)$$

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 $\hat{q}_{max}$ (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, $E(\hat{q}_{max}).q_{max}$. For this reason, one would collect samples (counts) immediately downstream of an active bottleneck so as to measure vehicles discharging at a maximum rate. The amount that $\hat{q}_{max}$ 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

$$\hat{q}_{max} = \sum_{m=1}^{M} n_m /(M \cdot \tau), \qquad (6.21)$$

where n_m is the count collected in the *m*th interval and each of these *M* intervals has a duration of τ. If the $\{n_m\}$ can be taken as independent, identically distributed random

[3] Comparing cumulative curves that have been constructed over long time periods requires accurate measurements since errors can accumulate over time.

variables (e.g. the counts were collected from consecutive intervals with τ sufficiently large), then the variance of $\hat{q}_{max}$ can be expressed as

$$variance(\hat{q}_{max}) = \frac{1}{\tau^2}\left[\frac{variance(n)}{M}\right] = \frac{1}{T}\left[\frac{variance(n)}{\tau}\right] \tag{6.22}$$

since q_{max} is a linear function of the independent n_m and the (finite) observation period T is the denominator in (6.21).

The bracketed term $variance(n) / \tau$ 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 $E(n) / \tau = q_{max}$, one obtains

$$variance(\hat{q}_{max}) = \frac{\gamma}{T}, \tag{6.23}$$

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

The $variance(\hat{q}_{max})$ is the square of the standard error. Thus, by isolating T in (6.23) and then multiplying both sides of the resulting expression by q_{max}, one arrives at

$$q_{\max} \cdot T = \frac{\gamma}{\varepsilon^2}, \tag{6.24}$$

where $q_{max}AT$. N, the number of observations (i.e., the count) needed to estimate capacity to a specified percent error ε. Note, for example, that $\varepsilon = 0.05$ 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.

A different definition of capacity. The *Highway Capacity Manual* (TRB, 1994), a widely circulated guidebook, offers a definition of capacity different from the one above. The *Manual* recommends taking as an estimate of q_{max} the highest flow measured over some interval, usually 15 minutes, during periods when "*sufficient demand exists.*" The *Manual* is not specific about what constitutes the existence of "sufficient demand." This omission may be intentional since there appears to be a lack of consensus as to whether capacity is a bottleneck's long-run queue discharge rate or the higher flow sometimes reported to occur prior to upstream queueing (Agyemang-Duah and Hall, 1991; Banks, 1990, 1991; Cassidy and Bertini, 1999). A rationale for treating capacity as the former, and an illustration of a deficiency in the *Highway Capacity Manual's* recommendation, are both offered later in this section. For now, it suffices to note that (6.24) yields the sample size for estimating capacity to within a specified error given the traffic features (i.e., the γ) at the bottleneck of interest. This has obvious appeal as compared with taking samples over some interval of arbitrary duration, such as 15 minutes.

Also of note, the *Highway Capacity Manual* provides default values for estimating a bottleneck's capacity without collecting samples in the field. These default values may be useful for certain long-range planning applications when coarse estimates suffice. They should only be used, however, when suitable sample counts cannot be collected.

An illustration of freeway capacity. Figure 6-8 presents re-scaled *N*- and *T*-curves that were constructed for a period spanning the morning rush. These measurements were taken downstream of the bottleneck (at detector 25) which was found on this day to be active from 6:30:30 a.m. to 7:54:00 a.m., as labeled on the figure. During this period of nearly 90 minutes, the *N* and the *T* display similar features, indicating that the measurements were not influenced by traffic conditions (i.e., queues) from further downstream.

The *N*-curve reveals that the observed pattern of flow can be described as sequences of sustained surges followed by reductions. This pattern is highlighted in Figure 6-8 by means of linear approximations superimposed on the *N* and by labels designating the flows (in units of vehicles per hour) for each period marked by quasi-linear arrivals. Despite these variations, the queue discharge flows exhibit a constant long-run trend, shown by the dashed line in the figure. While the bottleneck was active, the *N* never deviated from this trend by more than about 50 vehicles. By constructing the *N*-curve with a smaller background flow reduction and/or by plotting it over a longer time, curve portions measured during the period of queue discharge would appear to have a constant slope. Thus, the bottleneck's queue discharge rate may be described as being nearly constant over the rush.

One could determine with (6.24) a suitable sample size for estimating the average discharge rate. In this case, it would be advantageous to estimate the γ with a presample that captures both the surges and the reductions in the discharge flow; e.g. the $\{n_m\}$ could be periodically sampled over the entire period marked by the discharging queue. In general, such a sampling scheme would be feasible. If, for example, measurements were taken (automatically) using loop detectors, counts would usually be available for the entire day, including the rush. If instead, one incurred the expense of collecting counts by deploying human observers roadside, it would make sense to extend the data collection period so as to include the entire rush.

In light of the above, it would make even more sense to treat the queue discharge flow observed over an entire rush as an estimate of the long-run average. Eq. 6.24 could be rearranged to obtain ε, the precision of this estimate.

Treating the long-run discharge rate as capacity. Findings indicate that a bottleneck's average discharge rate, when measured over the rush, is reproducible from day to day (Cassidy and Bertini, 1999). Given this reproducibility, and in light of its near-constancy over the rush, it seems reasonable to take the estimated long-run average discharge rate as the bottleneck capacity. In the present example, the estimated capacity is therefore 6,420 vehicles per hour (vph), as annotated on Figure 6-8.

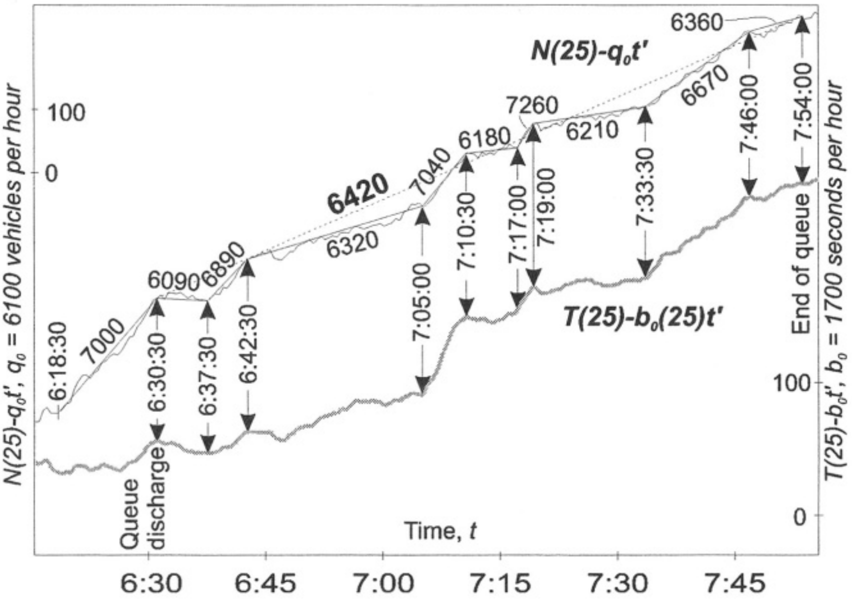


Figure 6-8. Re-scaled *N*- and *T*-curves downstream of the bottleneck.

Revisiting the Highway Capacity Manual's recommendation. If one takes as capacity the maximum flow observed over an interval of about 15 minutes (TRB, 1994), then, for the present example, the estimate is higher than 6,420 vph. This is because a flow of 7,000 vph prevailed for 12 minutes before the bottleneck's activation; this very high flow is labeled in Figure 6-8. Although some studies have reported that bottlenecks can support very high flows prior to their activation, these high flows have typically been observed only for time periods that are short relative to the rush. Figure 6-8 also shows that queue discharge rates comparable to 7,000 vph occasionally arose, but again, only for brief durations. Notably, these periods of very high flow are not only short-lived, their (short) durations vary from one day to the next (Cassidy and Bertini, 1999). Thus, these very high flows appear to be unstable and whether they can be prolonged through control measures such as on-ramp metering remains a question of active research. In light of this, the higher estimate of capacity (i.e., the one that follows from the *Highway Capacity Manual's* recommendation) may be unduly optimistic, even misleading.

6.3 Intersection Capacity and Vehicle Delay

This section describes techniques for estimating the capacity of highway intersections controlled by either traffic signals or stop signs, as these are often bottlenecks. Methods of estimating the vehicle delays at these facilities are likewise discussed, since delay minimization is a commonly used objective when developing intersection control schemes. This section does not specifically address such control schemes, however, as these are covered in Chapter 7 of the handbook.

Signalized Intersections

At a busy intersection, a traffic signal periodically interrupts vehicle movements to serve traffic in conflicting directions. Green times are extended so that consecutive vehicles in a queue may discharge through the intersection at a high rate, termed the saturation flow. It is by serving vehicle movements in this batched manner that a signal can increase the rates by which (conflicting) traffic streams traverse the intersection.

Much like the queue discharge rates through freeway bottlenecks, an intersection's saturation flows are affected by its geometric features and by certain attributes of its traffic streams such as the percentage of trucks. Moreover, saturation flows often vary with the type of traffic movement (i.e., turning or through-moving) served. Intuitively, a traffic movement's capacity is the product of its saturation flow and the proportion of "green time" available for this discharge. Methods of capacity estimation are described below.

Saturation flow and capacity. If by the end of the signal's red time, a traffic movement exhibits a sufficiently long queue (perhaps 6 or more vehicles in each lane), one can estimate its saturation flow by sampling the times consecutive vehicles pass a fixed point near the intersection, such as the stop bar, and plotting these departures cumulatively. Figure 6-9 shows a hypothetical cumulative count curve for the vehicles observed (e.g. in a single lane) during a time period greater than one signal cycle length, *C*. The step-wise curve has a strictly horizontal portion beginning some time near the end of the green indication. This period of "zero flow" extends until some time after the initiation of the green because the queued vehicles do not begin discharging at the instant of the green's initiation (Webster, 1958).

By collecting the departure times over K cycles and setting time t equal to zero at the initiation of each green, the arithmetic average of the cumulative number of vehicles to enter the intersection by $t, \overline{N}(t)$, is $\frac{1}{K}\sum_{k=1}^{K} N_k(t)$, where $N_k(t)$ is the cumulative number entering the intersection by t during the kth cycle. For a sufficiently large K, the cumulative curve is smooth, as exemplified in Figure 6-10. The slope of this smooth curve rises gradually from zero at $t = 0$ to a maximum of s, the saturation flow. This maximum slope eventually transitions to a smaller value (equal to the average arrival rate, $\overline{q}$) because the queue vanished during the green.

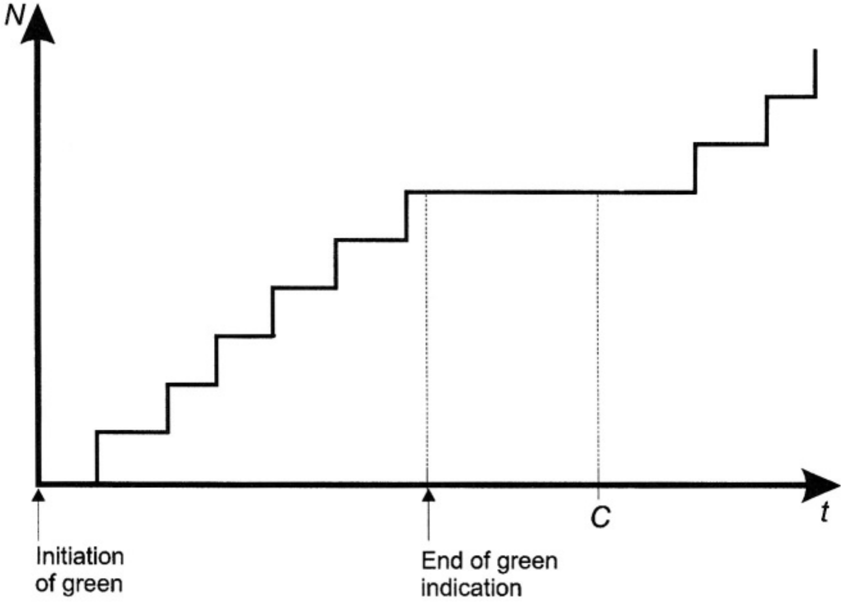


Figure 6-9. Typical N-curve at a traffic signal (Newell, 1989).

The "effective" start of the green time can be identified by extrapolating the curve portion with slope s backwards in time until it intersects with the line $\overline{N} = 0$. In similar fashion, the end of this green is found (approximately) by extrapolating the linear curve portion with slope $\overline{q}$ forward in time until it intersects with an extrapolation of the horizontal curve portion, as shown in Figure 6-10. The effective green period, G, is the time available in each cycle for serving vehicles at rate s. Thus, the intersection's capacity to serve the traffic movement is $sA(G/C)$.

One can usually assume that an $\bar{N}$-curve like the one described above is reproducible from day to day. Moreover, the precision of an estimate of s may be obtained using Eq. 6.24.

If drawn on "standard sized" paper, the $\bar{N}$-curve would not require any re-scaling, such as a background flow reduction. This is because the $\bar{N}$ and the t used to diagnose the intersection flows are small as compared with those used in diagnosing a freeway bottleneck.

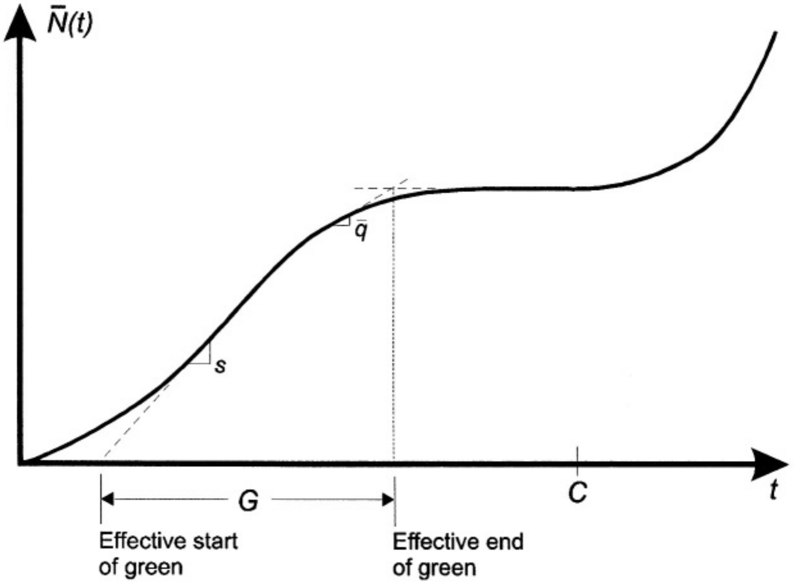


Figure 6-10. Average counts at a traffic signal, interpretation of effective green time (Newell, 1989).

Vehicle delay. If the traffic movement of interest is under-saturated (i.e., if the $\bar{q}$ does not exceed its capacity), an average departure curve like the one in Figure 6-10 can be used for measuring the vehicle delays. To this end, the vehicle arrival times may be measured upstream of any queueing caused by the traffic signal and a cumulative curve of the average arrivals per cycle can be constructed in the manner just described. One obtains a queueing diagram, and with it delay information, by horizontally translating this average arrival curve so that it is superimposed on the average departure curve for the period when the queue was not present.

Notably, the average arrival curve is merely a straight line (with slope equal to $\bar{q}$) if the vehicle arrivals are not influenced by any other traffic control device located further upstream. In these instances, the data needed to construct a queueing diagram can be collected by a single observer. A complete set of instructions for conducting such an experiment is provided by Pitstick (1990).

In many cases, one can predict intersection delays at some future time merely by altering the average arrival curve and/or the average departure curve as appropriate for the conditions (e.g. arrival rates, signal timing, etc.) that have been projected. The reader may again refer to Pitstick for discussion of this.

Stop-controlled Intersections

At an approach controlled by a stop-sign, the capacity to serve a certain traffic movement may be estimated by sampling the times that its queued vehicles enter the intersection. The issues here are much like those described in section 6.2, with the notable difference that, for a traffic movement controlled by a stop sign, capacity is influenced by the conflicts created by vehicles on other approaches.

The effects of vehicular conflicts are especially complex when stop signs are used to control only the traffic approaching from the minor street. The capacities for these minor street movements depend upon the propensity of its drivers to utilize headways exhibited by vehicles entering from the major street; more precisely, these drivers are forced to use "gaps" in the major street's traffic streams(s). Complexities arise because this gap-acceptance behavior is influenced by a host of factors, including the geometry of the major street and the velocity of the vehicles traveling on it, the sight distance(s) available to drivers on the minor street, their intended maneuvers and their personalities (e.g. aggressive or timid).

Gap acceptance models and delay prediction. Models for predicting gap-acceptance behavior have been developed to serve as tools for planning purposes. Some of these assume that drivers are both homogeneous and consistent; i.e., presumably a gap larger than some critical value is invariably utilized by the motorist waiting at the stop bar and motorists always decline to use a gap that is smaller than this critical value (TRB, 1994). Other models assume that the critical value used for accepting or rejecting a gap varies across drivers (Cohen, et al., 1955; Solberg and Oppenlander, 1966; Miller, 1972; Daganzo, 1981). Some gap acceptance functions assume that a driver's critical value may change, for example, as she grows impatient while waiting at the stop bar (Mahmassani and Sheffi, 1981; Madanat, et al., 1994; Cassidy, et al., 1995).

To predict vehicle delays at stop-controlled intersections, gap acceptance functions have been incorporated into analytical queueing models or they have been used in computer simulations. As a practical matter, it is worth noting that intersection delays become substantial only when the traffic intensity (i.e., the ratio of arrival rate to capacity) reaches a value of about 0.8 (Webster, 1958). Such high ratios are seldom observed at stop-controlled intersections because of often-used warrants (MUTCD, 1988) that provide for the installation of a traffic signal even when traffic intensities are rather small. Moreover, all gap acceptance models are estimated through statistical means and the applicability of any one model is limited by the intersection characteristics used for its estimation.

6.4 Traffic Flow Theory

It was noted in the previous two sections that bottlenecks exhibit predictable features. In section 6.2, for example, the discharge flow from an active freeway bottleneck was shown to be nearly constant over the rush. It was further noted that freeway bottlenecks consistently arise at the same locations and that a bottleneck's average discharge rate is reproducible from day to day. The queue discharge rates at signal-controlled and stop-controlled intersections can also be characterized as reproducible, as noted in section 6.3.

In light of their predictable features, it seems reasonable to theorize about how traffic evolves upstream and downstream of bottlenecks. Such theories have been developed and they may be used to predict important attributes of the traffic stream, such as a queue's growth due to accidents or geometric restrictions, the flow patterns generated by traffic control measures, etc. These theories rely upon given sets of boundary conditions and these might entail, for example, the rate of trip generation at all origins along the highway, the routes of all trips and the vehicle trip times, where the latter are functions of the time-dependent flows along the highway.

This section begins with a description of a remarkably simple theory of highway traffic flow. Proposed by Newell (1993), this theory is a version of one originally developed by Lighthill and Whitham (1955) and by Richards (1956) whereby traffic is treated as a continuum; i.e., the models describe the collective or average motion of vehicles in a traffic stream. Newell's version is described as being a "simplification," partly because of the relation it assumes, in effect, between flow and vehicle trip time and an explanation of this is provided later in the section. The simple theory predicts the shapes of standard (i.e., untransformed) cumulative count curves at locations of interest along the roadway. With this as its framework, the theory exploits the advantages of cumulative curves already described in the previous three sections. Accordingly, the theory is presented here in a purely graphical way. To highlight the theory's intuitive attributes, it is described in the context of a simple scenario involving a single highway segment with a bottleneck of fixed capacity located somewhere further downstream. References are made to some of the empirical findings that support the theory. Extensions needed to apply the theory to more complex scenarios are likewise noted. Prior to concluding this section, some of the limitations of the simple theory (and of its predecessors) are mentioned and attention is briefly given to some other traffic theories that have been proposed in light of these limitations.

A Simple Theory of Traffic Evolution

The upper portion of Figure 6-11 presents some hypothetical trajectories passing measurement (e.g. detector) location x_1 and proceeding past downstream measurement location x_3. It is assumed here that the intermediate location x_2 has no device(s) for traffic measurement. Thus, the theory will be used to describe the time-dependent traffic conditions, or more specifically, to construct the cumulative count curve at this location. The reader will note that the theory could be used for constructing the cumulative curves at any intermediate locations of interest. To satisfy an essential boundary condition, the cumulative curve at x_1, $N(t, x_1)$, used for the present example was not affected by any queueing from downstream.

The trajectories drawn in Figure 6-11 describe one freely flowing traffic state, labeled *a,* and three different states in queued traffic, labeled *b* through *d.* All vehicles are assumed to exhibit identical headways, spacings and velocities within a given state. Thus, no vehicle overtaking occurs. Furthermore, all vehicles are assumed to travel at a free-flow velocity v_f whenever traffic is freely flowing. Thus, for the interval $t_2 - t_1$, the cumulative count curve at x_2, is obtained by constructing the *N*-curve at x_1 and shifting it horizontally to the right by a vehicle's free-flow trip time from x_1 to x_2, $1/v_f\text{A}(x_2 - x_1)$. The curve labeled *1* was constructed by translating $N(t, x_1)$ in this manner. (Step-wise curves are shown in Figure 6-11 to make more obvious the relation between the trajectories and the cumulative curves).

In queued state *b,* the lead vehicle, labeled 0, decelerated from its previous (free-flow) velocity. It was stopped in state *c* and it accelerated upon entering state *d.* Of note, the theory assumes that vehicle accelerations (and decelerations) occur instantaneously and thus the theory can only hold over dimensions of time and space that are large relative to the separations between vehicles.

All vehicles of higher arrival number behave precisely as vehicle *0.* In queued traffic, an *i*th vehicle's trajectory is assumed to adopt the features of the *i-1*th trajectory following some fixed time lag; the time lag is the same for all queued states. This, along with the assumption that vehicles exhibit uniform spacings within a given state, means that the interface between any two queued states propagates at a fixed speed *u,* as shown by the dashed lines in the upper portion of Figure 6-11. Moreover, in queued traffic, the *i*th trajectory can be constructed by shifting the *i-1*th trajectory forward by the fixed time lag and downward by a fixed spacing. Study of the time-space diagram in Figure 6-11 reveals that this spacing is the one adopted by vehicles that have come to a complete stop, the so-called jam-density spacing. This is true even in the absence of a "jammed" traffic state like state *c.* In fact, state *c* was included in Figure 6-11 merely to aid the reader in verifying that the appropriate downward translation is always the jam density spacing.

It follows that in queued traffic, the curve at x_2 is obtained by shifting the *N*-curve at x_3, $N(t, x_3)$, to the right by a distance equal to the trip time of an interface between these two locations, $1/u\text{A}(x_3 - x_2)$, and upward by the number of vehicles that pass through the interface during this time. It should be clear that the latter is the jam-

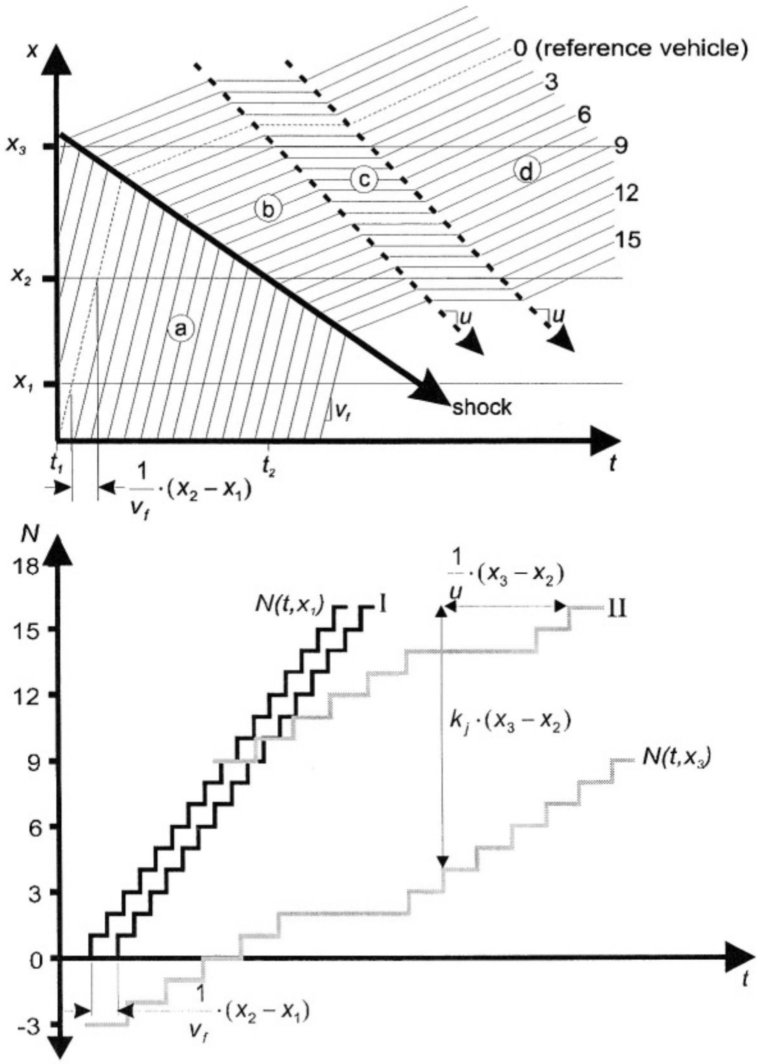


Figure 6-11. Simple example of traffic evolution.

density storage, k_j A$(x_3$ - $x_2)$, where k_j , the jam density, is the maximum density the road segment can accommodate.

These horizontal and vertical curve translations produced the curve labeled *II* in Figure 6-11. By referring to the time-space diagram in this figure, the reader may verify that shifting $N(t, x_3)$ as described above produces the *N*-curve that would have been measured for queued traffic conditions at x_2, had measurement devices existed there.

The curve translated forward from x_1, labeled *I*, intersects the curve translated from downstream x_3, labeled *II*. Notably, this intersection occurs at time t_2, the time when the back of the queue (i.e., state *b*) arrived to location x_2. Time t_2 is thus said to mark the arrival of a *shock* at location x_2, where, in this simple theory, a shock is an interface between queued and freely flowing traffic.[4] The lower envelope of curve *I* and curve *II* in Figure 6-11 is the resulting *N*-curve at x_2. It is intuitive that flow is constrained at x_2 following the shock's arrival and the two translated curves intersect at this arrival time because vehicles are conserved across the shock's path.

A shock may exhibit a multitude of possible speeds, positive or negative, as dictated by the traffic conditions on its upstream and downstream sides. Moreover, a shock's speed changes when it intersects interfaces or other shocks. Notably, the use of *N*-curves as described above does not require one to trace the paths of shocks and interfaces over time and space. Such (tedious) analyses are required in the Lighthill and Whitham and Richards (LWR) versions.

An assumed bivariate relation. Implicit in the simple theory and LWR is a key postulate that there exists some relation between traffic properties, such as flow and vehicle trip time, that may vary with location along the highway, but not with time. These relations are purely empirical; i.e., they are obtained through measurement. Not only would they depend upon the highway geometry, these relations would also be affected by environmental factors, such as weather conditions, as well as by attributes of the traffic stream, such as its proportion of large trucks, the tendencies of its drivers, etc. Although the bivariate relation between any number of traffic properties may be used as a boundary condition for continuum models, it is customary to use a relation between flow, *q,* and density, *k.* The assumptions of the simple theory just described imply a *q-k* relation that is triangular in form, like the one shown in Figure 6-12, and the reasons for this are explained below.

To begin, the right branch of the relation describes conditions in queued traffic, whereby *k* increases with decreasing *q.* From Eq. 6.13, the speed of an interface between stationary traffic states, *u,* is $\Delta q/\Delta k$. That the relation's right branch is linear means that interfaces between queued states presumably propagate at a single speed, a previously noted assumption of the simple theory.

The left branch of the triangular relation describes freely flowing traffic. As noted in section 1, the average vehicle velocity is the ratio of flow to density. Thus, the simple theory assigns to the left branch a slope of v_f, the presumed velocity of all

[4] In the theory developed by Lighthill and Whitham and by Richards, shocks arise when backward-moving interfaces collide. These collisions do not occur in the simplified theory since all backward interfaces are presumed to travel at the same speed *u.*

freely flowing vehicles. That this slope is constant also implies that changing (q, k) states move forward with the vehicles in freely flowing traffic.

It was previously noted that, in the simple theory, a shock separates freely flowing and queued traffic. The shock's speed is therefore the slope of the chord connecting the (q, k) states as they lie on each side of the triangular relation.

The existence of well-defined bivariate relations is an assumption widely adopted in traffic and transportation engineering. In addition to their role in continuum theories of traffic flow, these relations are commonly used in highway design and in transportation planning. In fact, additional discussion on these presumed relations may be found in virtually any traffic engineering text or handbook, including (TRB, 1994). In particular, chapter 4 of Daganzo (1997) contains a thorough introduction to the subject.

Some empirical evidence. Despite their role in continuum theories like the simple one described above, the assumption that bivariate relations are independent of time is known to be incorrect. Interfaces in the traffic stream exhibit characteristic widths where traffic is not stationary because vehicles are adjusting from one state to another. When collected in the presence of these nonstationary regions, bivariate data do not give rise to well-defined relations. To the contrary, plots of these data are widely scattered (Hall, et al., 1992).

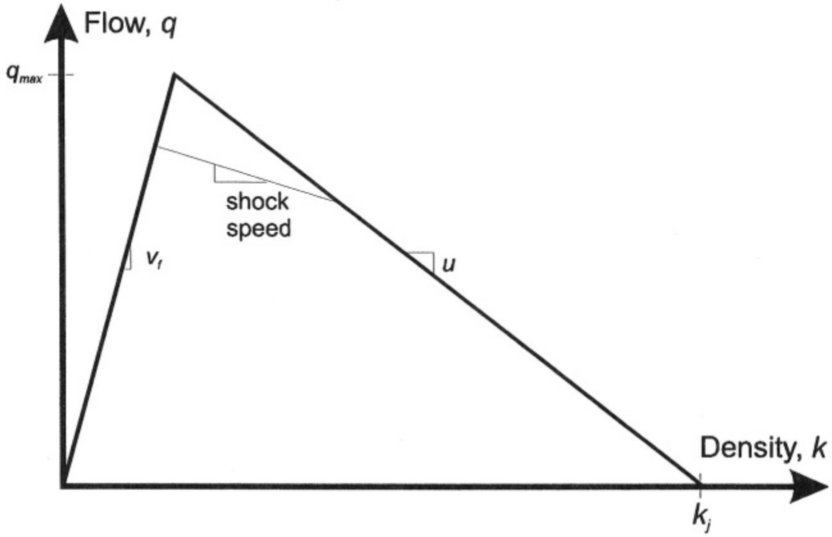


Figure 6-12. Triangular flow-density relation.

On the other hand, one might "*reasonably expect drivers to do the same on average under the same average conditions*" (Daganzo, 1997, p 80). It therefore seems reasonable to postulate that well-defined bivariate relations exist for stationary traffic. Indeed, there is the empirical evidence supporting this postulate. One study found that the average values of flows and occupancies taken only from nearly stationary traffic produced well-defined relations that appeared to be reproducible from day to day (Cassidy, 1998). The research showed that (freeway) traffic is stationary in the absence of any interfaces and that one may describe these conditions on a highway segment using some bivariate relation. It follows that an interface must propagate as specified by the highway segment's bivariate relation, even though the relation does not describe traffic behavior in the immediate vicinity of the interface.

In another study, re-scaled cumulative count curves were used to examine numerous backward-moving interfaces in individual travel lanes (Windover, 1998). These curves were constructed in series using counts taken (by loop detectors) on a 2-km freeway segment upstream of an active bottleneck. They were plotted for time periods long enough so as to display numerous interfaces and these interfaces separated a variety of queued (q, k) traffic states. The curves were observed to have similar forms; i.e., slope changes on the downstream curve later appeared on the upstream curves, indicating that drivers responded in similar ways to changes in traffic conditions as per LWR and Newell's simple theory. Furthermore, each (entire) curve could be nearly superimposed upon its neighbor following a vertical and a horizontal translation. This latter observation indicates that between any two of the measurement locations, all interfaces passed through nearly the same number of vehicles and had nearly the same trip times. The finding indicates that the adoption of triangular q-k relations like the one in Figure 6-12 may be a reasonable approximation for describing traffic.

Applying the simple theory to complex scenarios. The simple theory was described above using a highway segment of fixed geometry and a single bottleneck downstream. Newell describes how the theory can be applied to more complicated roadways by using suitable boundary conditions. In addition to those required for the previous analysis, these boundry conditions include the differences between the cumulative number of vehicles that enter and exit the roadway by time t at each junction; i.e., the net cumulative count.[5] One would also specify (possibly different) triangular q-k relations for each section of highway.

As regards the net cumulative counts, a complication arises in estimating the exit counts at each junction. Even if one knows the routes taken by vehicles and the times they enter the roadway, the times they actually exit will be dictated by the prevailing traffic conditions whenever queueing occurs upstream. This complication was originally recognized by Vaughn, Hurdle and Hauer (1984) and also by Vaughn and

[5] If entrance and exit ramps are closely spaced, their separations may be ignored.

Hurdle (1992). Newell (1993, part *III*) provided a solution by specifying that a vehicle's trip time between successive junctions is independent of its origin and destination and that at all junctions, one must evaluate the component cumulative curves for each destination. The reader may refer to the source for a complete description of this.

On a related note, a diverge formed by an off-ramp can cause queueing in at least two ways: 1) a queue from the off-ramp spills over and blocks traffic; and 2) the off-ramp is not queued but an increase in the flow of vehicles wishing to exit creates a queue (on the freeway or highway) when the off-ramp reaches capacity (Daganzo et al., 1997). A mathematical theory of the diverge can be found in sections 3.3 and 4 of Daganzo (1995) and a preliminary theory for wide freeways where traffic may sort itself by lane depending upon destination can be found in Daganzo (1997*a*).

An additional complication arises in determining the net cumulative count at some on-ramp after a queue has propagated beyond it. This is because the rate by which vehicles merge onto a congested roadway requires some observation or some intuition. Studies by the California Department of Transportation (Newman, 1986), for example, found that when queues are present both on an entrance ramp and on the freeway, merging vehicles share available roadway space with vehicles in the adjacent freeway lane on a one-to-one basis, creating the so-called "zipper effect." Daganzo (1997a) offers a simple theory to predict both the ramp and the freeway (output) flows when two merging traffic streams compete for the capacity available downstream.

Finally, a computer program has been developed to aid in applying Newell's simple theory to complex highway sections. As of the date of this publication, the program is available on the World Wide Web through the civil engineering department at the Georgia Institute of Technology.

Some deficiencies and some alternatives. The simple theory, as well as the LWR theory, are known to have deficiencies. The most noteworthy of these, along with some of the other theories developed to address these deficiencies, are described below. The coverage here is admittedly light. The intent is merely to bring toward closure the discussion of traffic flow theory and to provide the reader with references to some of the other models common in the literature.

Traffic instabilities. As apparent from the previous discussion, the simple theory does not predict the "stop and go" instabilities often observed in queues. Nor are instabilities described by the LWR theory. Models that qualitatively match many (although not all) of the features observed in unstable traffic do exist. One of the earliest of these (Newell, 1962) belongs to a class of models that are, in effect, extensions to the LWR theory and its simplified version. This class of models share a number of similarities with the latter continuum models. In all of these models, for example, a platoon is described by state (q, k) on the flow-density plane. Likewise, each of these models assume that all vehicles in a platoon always respond in the same way to a sustained change in the lead car's velocity. However, the class of models which we describe as being extensions are specified by defining two families of *q-k* curves; one family describes all possible evolutions of decelerating platoons and the other corresponds to accelerating traffic (Daganzo, et al., 1997). Models of this type are thus compatible with a well-known finding of Edie (1965) that decelerating platoons adopt a set of states on the *q-k* plane that are consistently different from accelerating ones.

Efforts to explain traffic instabilities also led to the development of models to describe a driver's response to the changing trajectory of a lead vehicle. These so-called car-following or microscopic traffic models predict how a driver adjusts her vehicle's velocity because of a stimulus, such as the (time-dependent) spacing between the subject vehicle and its leader (Chandler, et al., 1958; Herman, et al., 1959; Gazis, et al. ,1959; Herman and Potts, 1961). To calibrate and test these models, considerable data collection took place using instrumented cars on test tracks. This line of research is notable in that the objective was to describe the behavior of individual drivers and the interested reader can refer to May (1990) for discussion on some of the early developments in this area.

To their credit, the car-following models do predict the occurrences of instabilities. However, there has been little success in matching the oscillatory behavior observed in real traffic with the predictions from these car-following models. Perhaps this is because the theories assume that drivers apply control continuously and that a driver responds only to the motion of the car immediately downstream and not, for example, to cues collectively displayed by the (possibly many) downstream vehicles visible to the driver. Driver behavior is probably more complex than this.

Indeed, the complexities of driver behavior likely explain why instabilities exhibit periods of oscillation and growth that are site specific. For example, a bottleneck studied in New York's Holland Tunnel displayed regular oscillation periods of about 2 minutes (Edie and Foote, 1961), while observations upstream of other bottlenecks have revealed more sporadic characteristics (Kerner and Rehborn, 1997, 1996, 1996*a,* Smilowitz et al., 1998). This is not surprising given that the vehicular interactions are different for different types of bottlenecks (e.g. merges, diverges, lane reductions, etc.) and a thorough understanding of each type of bottleneck will likely come only by studying them individually.

What may be most important here, however, are the details of traffic instabilities that are understood at the present time. Namely, that instabilities occur well upstream of bottlenecks and that they do not appear to affect a bottleneck's discharge flow.[6]

[6] Contrary to claims made in some of the literature (Kerner, et al., 1995; Kerner and Rehborn, 1997), this author has seen to date no conclusive evidence suggesting that instability phenomena may cause freely flowing traffic to break-down and form queues spontaneously. The interested reader may refer to Daganzo et al. (1999) for more discussion on these issues.

The detailed behavior of queued traffic therefore has little effect on the delay caused by a bottleneck. Furthermore, to predict time-dependent queue lengths, (e.g. due to control actions) it might suffice to predict the general, coarse shapes of cumulative count curves without attempting to predict the occurrences of wiggles that characterize instabilities. This is precisely the objective behind Newell's simple theory.

Theories from other scientific fields. In efforts to improve the LWR models, researchers have borrowed theories from other fields of scientific inquiry. Discussion of methods adopted from other fields are not the emphasis of this handbook. Nonetheless, a few of these adaptations deserve mention, albeit brief, because they are prominent in the literature, they are used in practice and they are not without shortcomings of their own.

For example, Navier-Stokes-like equations commonly used in fluid mechanics have been proposed for describing traffic instabilities (Kerner, et al., 1995). Researchers have likewise advocated other second-order partial differential equations similar to those used for fluid approximations to explain the motion of vehicles passing through shocks (Payne, 1971; Kühne and Beckschulite, 1993); the reader will recall that the simple theory and its predecessors assume that vehicles change velocities instantaneously and this is a very coarse approximation. These second-order models can generate unrealistic predictions of traffic evolution because these models borrow features of materials flow that are unreasonable for describing highway traffic. For example, del Castillo et al. (1993) have noted that these models predict that interfaces may overtake particles (e.g. fluid molecules) and that when applied to highway traffic, this implies that drivers collectively respond to stimuli from behind. This would not seem to be a reasonable depiction of the driving process. Daganzo (1995a) offers extended discussion on the pitfalls in applying second-order fluid approximations to highway traffic.

Lastly, models based upon the kinetic theory of gases have been proposed for describing vehicular interactions in light traffic (Phillips, 1977; Prigogine, 1961; Prigogine and Herman, 1971). Continuum theories are deficient for these conditions because they do not describe the variations in velocities across vehicles and thus, they do not predict the overtaking and the natural spreading of platoons that occur in low densities. The kinetic models of traffic flow were intended to improve the LWR theory by considering the distribution of vehicle velocities at each point in time and space. However, these models were derived from the integration of molecular properties, such as positions, collisions and velocities, that do not accurately describe highway traffic. Daganzo (1995a), for example, has noted that these models assume that a distribution of desired vehicle velocities "*can be defined exogenously at every point in time and space independent of the drivers who happen to be there.*" In so doing, these models fail to recognize that individual drivers have personalities (e.g. aggressive and timid) which they retain with their motion (Cassidy and Windover, 1998). The reader may refer to Newell (1995) for further discussion on these issues.

As an aside, a theory of very light traffic with weak overtaking interactions is fairly complete and its description is likewise found in Newell (1995).

Some Final Comments

Although section 6.4 has made reference to a number of theories for describing highway traffic, it has emphasized a simple continuum model proposed by Newell. In the interest of completeness, we note that Newell described his recipe using a coordinate system whereby the time at each location along the roadway is measured from the passage of a freely-flowing reference vehicle (e.g. labeled $N = 0$). In effect, this so-called "moving time coordinate system" horizontally shifts the cumulative count curves so that neighboring curves display vehicle delays and excess accumulations. This is analogous to the horizontal shifts that were applied to count curves in section 6.2 and the reader may refer to the original source (Newell, 1993) for more details on the use of moving time.

Far more important than these details, however, is the question of reliability; i.e., the adequacy of Newell's simple theory for predicting traffic evolution remains an active research question. The need for changing or refining the model may become apparent, for example, as ongoing empirical studies reveal more about roadway bottlenecks and the features of the queues they create. What is noteworthy about Newell's recipe, however, is its exploitation of cumulative count curves. Such a framework can be used to predict virtually any traffic feature likely to be of interest, including vehicle delays, the spatial extent of queueing, etc. Thus, it would seem that any models of highway traffic flow developed in the future, or any future refinements to existing models, should make use of these cumulative curves.

6.5 References

Agyemang-Duah K. and Hall F.L. (1991) Some issues regarding the numerical value of capacity. *Proc. Int. Symp. on Highway Capacity,* pp. 1-15, A.A. Balkema, Germany.

Banks J.H. (1990) Flow processes at a freeway bottleneck. *Transpn Res. Rec.* ***1287***, 20-28.

Banks J.H. (1991) Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering? *Transpn Res. Rec.* ***1320***, 83-90.

del Castillo J.M., Pintado P. and Benitez F.G. (1993) A formulation for the reaction time of traffic flow models. *Proc. Int. Symp. on Transportation and Traffic Theory,* (C.F. Daganzo, ed.), pp. 387-405, Elsevier, New York.

Cassidy M.J. (1998) Bivariate relations in nearly stationary highway traffic. *Transpn Res.* ***32B***, 49-59.

Cassidy M.J. and Bertini R.L. (1999) Some traffic features at freeway bottlenecks. *Transpn Res.* ***33B***, 25-42.

Cassidy M.J. and Coifman B. (1997) Relation among average speed, flow and density and the analogous relation between density and occupancy. *Transpn Res. Rec.* ***1591***, 1-6.

Cassidy M.J., Madanat S.M. and Wang M.H. (1995) Unsignalized intersection capacity and level of service: revisiting critical gap. *Transpn Res. Rec.* ***1484***, 16-23.

Cassidy M.J. and Windover J.R. (1995) Methodology for assessing dynamics of freeway traffic flow. *Transpn Res. Rec.* ***1484***, 73-79.

Cassidy M.J. and Windover J.R. (1998) Driver memory: motorist selection and retention of individualized headways in highway traffic. *Transpn Res.* ***32A***, 129-137.

Chandler R.E., Herman R. and Montroll E.W. (1958) Traffic dynamics: studies in car-following. *Opns. Res.* ***6***, 165-184.

Cohen, E., Dearnaley J. and Hansel C. (1955) The risk taken in crossing a road. *Opns. Res. Qrtly,* ***6***, 120-128.

Daganzo C.F. (1981) Estimation of gap acceptance parameters within and across the population from direct roadside observation. *Transpn Res.* ***15B***, 1-15.

Daganzo C.F. (1995) The cell transmission model. *II*: Network traffic. *Transpn Res.* ***29B***, 79-93.

Daganzo C.F. (1995a) Requiem for second-order fluid approximations of traffic flow. *Transpn Res.* ***29B,*** 277-286.

Daganzo C.F. (1997) *Fundamentals of transportation and traffic operations.* Elsevier, New York.

Daganzo C.F. (1997*a*) The nature of freeway gridlock and how to prevent it. *Proc. Int. Symp. on Transportation and Traffic Theory,* (J.B. Lesort, ed.), pp. 629-646, Pergamon, Tarrytown.

Daganzo C.F., Cassidy M.J. and Bertini R.L. (1999) Possible explanations of phase transitions in highway traffic. *Transpn Res.* ***33A***, 365-379.

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

Edie L.C. (1974) *Traffic Science* (D.C. Gazis, ed.), pp. 8-20, Wiley, New York.

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

Edie L.C. and Foote R.S. (1961) Experiments on single-lane flow in tunnels. *Proc. Int. Symp. on the Theory of Traffic Flow* (R. Herman, ed.), pp. 175-192, Elsevier, New York.

Gazis D.C., Herman R. and Potts R.B. (1959) Car-following theory of steady state flow. *Opns. Res.* ***7***, 499-505.

Hall F.L., Hurdle, V.F. and Banks, J.H. (1992) A synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transpn Res. Rec.* ***1365***, 12-18.

Herman R., Montroll E.W., Potts R.B. and Rothery R. (1959) Traffic dynamics: analysis of stability in car-following. *Opns. Res.* ***7***, 86-106.

Herman R. and Potts R.B. (1961) Single-lane traffic theory and experiment. *Proc. Int. Symp on the Theory of Traffic Flow* (R. Herman, ed.), pp 120-146, Elsevier, New York.

Kerner B.S., Konhauser P. and Schilke M. (1995) Deterministic spontaneous appearance of traffic jams in slightly inhomogeneous traffic flow. *Phys. Rev. E* ***51***, 6243-6246.

Kerner B.S. and Rehborn H. (1996) Experimental properties of complexity in traffic flow. *Phys. Rev. E* ***53***, R4275-R4278.

Kerner B.S. and Rehborn H. (1996*a*) Experimental features and characteristics of traffic jams. *Phys. Rev. E* ***53***, R1297-R1300.

Kerner B.S. and Rehborn H. (1997) Experimental properties of phase transitions in traffic flow. *Phys. Rev. Let.* ***79***, 4030-4033.

Kühne R.D. and Beckschulte R. (1993) Non-linearity stochastics in unstable traffic flow. *Proc. Int. Symp. on Transportation and Traffic Theory* (C.F. Daganzo, ed.), pp. 367-386, Elsevier, New York.

Lighthill M.J. and Whitham G.B. (1955) On kinematic waves. *I*: Flood movement in long rivers. *II*: A theory of traffic flow on long crowded roads. *Proc. Royal Soc.* ***A229***, 281-345.

Lin W.F. and Daganzo C.F. (1997) A simple detection scheme for delay-inducing freeway incidents. *Transpn Res.* ***31A***, 141-155.

Madanat S.M. Cassidy, M.J. and Wang M.H. (1994) A probabilistic model of queueing delay at stop-controlled intersection approaches. *J. Transpn Eng.* ***120***, 21-36.

Mahmassani H. and Sheffi Y. (1981) Using gap sequences to estimate gap acceptance functions. *Transpn Res.* ***15B***, 243-248.

Makagami Y., Newell G.F. and Rothery R. (1971) Three-dimensional representation of traffic flow. *Transpn Sci.* ***5***, 302-313.

May A.D. (1990) *Traffic flow fundamentals*. Prentice Hall, New Jersey.

Miller A. (1972) Nine estimators of gap acceptance parameters. *Proc. Int. Symp. on the Theory of Traffic Flow and Transportation*, (G.F. Newell, ed.), pp. 215-235, Elsevier, New York.

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

MUTCD (1988) *Manual on Uniform Traffic Control Devices*. U.S. DOT, Government Printing Office, Washington, D.C.

Newell G.F. (unpublished) Notes on transportation operations. Univ. of California, Berkeley, U.S.A.

Newell G.F. (1962) Theories of instability in dense highway traffic. *J. Opn. Res. Soc. Japan* ***5***, 9-54.

Newell G.F. (1971) *Applications of Queueing Theory*. Chapman Hall, London.

Newell G.F. (1979) Airport capacity and delays. *Transpn Sci.* ***13***, 201-241.

Newell G.F. (1989) Theory of highway traffic signals. Institute of Transportation Studies UCB-ITS-CN-89-l, Univ. of California, Berkeley, U.S.A.

Newell G.F. (1993) A simplified theory of kinematic waves in highway traffic *I:* General theory. *II*: Queueing at freeway bottlenecks. *III*: Multi-destination flows. *Transpn Res.* ***27B***, 281-313.

Newell G.F. (1995) Theory of highway traffic flow 1945 to 1965. Institute of Transportation Studies, Special Report, Univ. of California, Berkeley, U.S.A.

Newman L. (1986) Freeway Operations Analysis Course Notes. Institute of Transportation Studies, University Extension, Univ. of California, Berkeley, U.S.A.

Payne H.J. (1971) Models of freeway traffic control. *Proc. Math. Models Publ. Sys. Simul. Council* ***28***, 51-61.

Phillips W.F. (1977) Kinetic theory for traffic flow. Research Report for U.S. DOT. Logan, UT: Dept. of Mech. Eng., Utah State Univ., U.S.A.

Pitstick M. (1990) Measuring delay and simulating performance at isolated signalized intersections using cumulative curves. *Transpn Res. Rec.* ***1287***, 34-91.

Prigogine I (1961) A Boltzman-like approach to the statistical theory of traffic flow. *Proc. Int. Symp. on the Theory of Traffic Flow* (R. Herman, ed.), pp. 158-164, Elsevier, New York.

Prigogine I. and Herman R. (1971) *Kinetic theory of vehicular traffic*. Elsevier, New York.

Richards P.I. (1956) Shock waves on the highway. *Opns. Res.*, ***4***, 42-51.

Smilowitz K.R., Daganzo C.F., Cassidy M.J. and Bertini R.L. (1998) Some observations of traffic in long queues. Institute of Transportation Studies UCB-ITS-RR-98-6, Univ. of California, Berkeley, U.S.A. *to be published in Transpn Res. Rec.*

Solberg P. and Oppenlander J. (1966) Lag and gap acceptance at stop-controlled intersections. *Highway Res. Rec.* ***118***, 48-67.

TRB (1994) Special Report 209, *Highway Capacity Manual.* Transportation Research Board, Washington, D.C.

Vaughn R. and Hurdle, V.F. (1992) A theory of traffic flow for congested conditions on urban arterial streets. *I*: Theoretical development. *Transpn Res.* ***26B****,* 381-396.

Vaughn R., Hurdle, V.F. and Hauer E. (1984) A traffic flow model with time-dependent o-d patterns. *Proc. Symp. on Transportation and Traffic Theory,* (J. Volmuller and R. Hamerslag, eds.), pp. 155-178, VNU Science Press, Utrecht, The Netherlands.

Webster F.V. (1958) Traffic signal settings. Road Research Technical Paper, No. 39, Road Research Laboratory, Ministry of Transport, HMSO, London.

Windover J.R. (1998) Empirical studies of the dynamic features of freeway traffic. *PhD thesis.* Univ. of California, Berkeley, U.S.A.

This page intentionally left blank

7 AUTOMATED VEHICLE CONTROL*

Petros Ioannou and Arnab Bose

7.1 Introduction

Mankind's love for automobiles dates back more than one hundred years to when they were first introduced. The major functions performed while driving an automobile are lateral and longitudinal control of the vehicle. While the former is necessary to ensure that the vehicle does not lose track of the desired path, the latter is used to keep the vehicle at a safe speed dependent on the surrounding conditions and at a safe distance from the preceding vehicle (if any). This function is carried out by humans whose senses detect changes in the environment and act as stimuli. The human driver then reacts to these stimuli by applying either the brakes or the gas pedal. This reaction defines the driving behavior of an individual driver.

It is a well-known fact that the driving habits differ from person to person. The behavior of the human driver also has randomness associated with it that adversely affects safety and traffic flow characteristics. In the last decade considerable research has been done to automate the vehicle-highway system in an effort to improve safety, capacity and traffic flow characteristics. It has been envisioned that removing the human from the vehicle-driver control loop will eliminate the randomness associated with today's manual traffic and satisfy the above requirements (Stevens, 1997). This motivates the concept of automated vehicle control.

The degree of automation in a vehicle determines the involvement of the human driver in the driving loop. While the use of partial automation as driver aids guarantees improved traffic flow characteristics, the same cannot be said of safety and throughput (Bose and Ioannou, 1998). However, full automation in the longitudinal direction is expected to benefit safety, capacity and traffic flow

* This work is supported by the California Department of Transportation through PATH of the University of California. The contents of this chapter reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This chapter does not constitute a standard, specification or regulation.

characteristics. Likewise, completely replacing the human driver in the driving loop with longitudinal and lateral automation, as in a fully automated vehicle, is expected to further improve safety. The use of actuators and sensors is deemed to improve safety, capacity and traffic flow characteristics for the following reasons: firstly, electronic sensors do not get fatigued, tired or distracted and are therefore more reliable (Ward, 1997). Secondly, the actuators react much faster than the average alert human driver, who has a time delay of about 1.0s to 1.5s (Milestone 2 Report, Appendix J, 1996). This implies that vehicles can travel closer, which translates into higher capacity/throughput (Bose and Ioannou, 1998). Thirdly, the deterministic response of the controllers in comparison to the random behavior of human drivers smoothes traffic flow (Bose and Ioannou, 1999). Different system configurations for automated vehicle deployment have been outlined in Hall (1997).

The lateral and longitudinal functions of an automated vehicle are performed with the use of lateral and longitudinal controllers that work in conjunction with on-board sensors (Walker and Harris, 1993). The longitudinal controller comprises two subsystems, namely throttle and brake controllers that do not work simultaneously (Ioannou and Xu, 1994). A switching logic dictates the switching from one controller to the other. The lateral controller uses a lateral control system (Peng and Tomizuka, 1990) that uses a road referencing/sensing system that measures the position and orientation of the vehicle relative to the road (Hessburg et al., 1991).

Chapter 7 deals with vehicle automation, how it is achieved and the subsequent benefits. We begin with vehicle longitudinal and lateral dynamics models in section 7.2 that are used to design automated controllers. Different manual vehicle models that provide a better understanding of the human-driver interface are overviewed in section 7.3. This is useful when designing an automated controller to mimic the behavior of the human driver and partially/completely replacing him/her in the driving loop. A design of an automated longitudinal controller is briefly outlined in section 7.4. This is followed by discussions on longitudinal controllers for heavy-duty vehicles, vehicle-to-vehicle communication designs and sensor requirements that are needed to ensure safe and proper operation of automated vehicles. Next the intervehicle spacing of an automated vehicle that is dictated by the longitudinal controller and the expected safety level due to its use are highlighted.

The lateral control of light and heavy-duty vehicles using automated controllers are briefly outlined in the next section, 7.5, followed by necessary lateral/side sensor requirements. Different sensor technologies available today are evaluated for their applicability as both longitudinal and lateral sensors. Increased level of safety due to the use of an automated lateral controller in addition to an automated longitudinal controller is discussed next.

A beneficial effect of automation, namely automated lane changing, is investigated in section 7.6 using different acceleration profiles of the lane changing vehicle. String stability in vehicle following is evaluated for manual and automated traffic in section 7.7 using the models presented in the previous sections. Lastly, mixed manual/automated driving is discussed and benefits are evaluated as a function of the market penetration of automated vehicles in section 7.8. The chapter ends with a summary in section 7.9.

7.2 Vehicle Dynamics

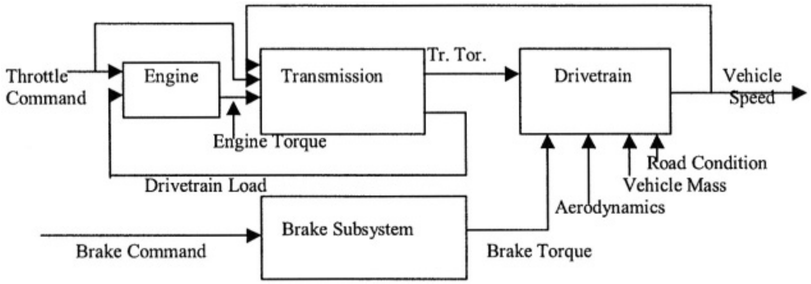


Figure 7-1. Longitudinal vehicle model.

Fig. 7-1 shows a block diagram of the longitudinal vehicle model (Ioannou and Xu, 1994). Each block can be considered as a subsystem with various inputs and outputs. The output of the engine subsystem is the engine torque that is a nonlinear function of the air/fuel ratio, the exhaust gas recirculation (EGR), the cylinder total mass charge, the spark advance, the engine speed and the drivetrain as well as the throttle angle. The spark advance, EGR and air-to-fuel ratio are the outputs of an internal controller (inside the engine block of Fig. 7-1) whose inputs are the throttle position, engine speed and drivetrain load. The transmission subsystem considered is an automatic transmission with hydraulic coupling and four transmission gears. This subsystem transfers transmission torque (Tr. Tor. in Fig. 7-1) to the drivetrain as a function of vehicle speed and engine condition. The latter then outputs the vehicle speed and acceleration/deceleration that are affected by the road condition, aerodynamic drag and vehicle mass.

The brake subsystem has a brake actuator that receives braking commands and outputs braking torque. It acts like a low pass first order filter with some delays which have been verified to be noticeable only at the beginning and negligible later on. The dynamics of the brake subsystem are faster than those of the drivetrain.

Longitudinal

The longitudinal vehicle speed V is a nonlinear function of the throttle angle θ and can be expressed as

$$V = F(\theta, t, \tau) \tag{7.1}$$

where $0 \leq \tau \leq t$ indicates the presence of dynamics.

Swaroop et al. (1994) developed a three state variable lumped parameter longitudinal model of a vehicle which is outlined as follows. Assuming that ideal gas law holds in the intake manifold and that its temperature is constant, the intake manifold dynamics are given by

$$\dot{m}_a = \dot{m}_{ai} - \dot{m}_{ao} \tag{7.2}$$

$$P_m V = m_a R T_m \tag{7.3}$$

where m_a is the mass of air in the intake manifold, $\dot{m}_{ai}$, $\dot{m}_{ao}$ are the mass flow rates through the throttle and into the cylinders (entering the combustion chamber), respectively; P_m, V, T_m are the intake manifold pressure, volume and temperatures, respectively while R is the universal gas constant for air. The empirical relationship for $\dot{m}_{ai}$ is defined as

$$\dot{m}_{ai} = MAX.TC(\alpha).PRI\left(\frac{P_m}{P_a}\right) \tag{7.4}$$

where *MAX* is a constant dependent on the size of the throttle body, $TC(\alpha)$ is the throttle characteristic which is the projected area the flow sees as a function of α, *PRI* is the "pressure ratio influence" which describes the choked flow relationship which often occurs through the throttle valve and P_a is the atmospheric pressure. The empirical relationship for $\dot{m}_{ao}$ is defined as (Hedrick at al., 1991)

$$\dot{m}_{ao} = c_1 \eta_v (P_m, \omega_e) m_a \omega_e \tag{7.5}$$

where η_v is the volumetric efficiency which is a measure of the effectiveness of an engine's induction process and defined as the volume flow rate of air into an engine divided by the rate at which the volume is displaced by the piston (Cho and Hedrick, 1989), ω_e is the engine speed and c_1 is a physical constant defined by (Cho and Hedrick, 1989)

$$c_1 = \frac{V_e}{4\pi V_m} \tag{7.6}$$

where V_e is the engine displacement equal to $0.0038m^3$ and V_m is the intake manifold volume equal to $0.0027m^3$.

Assuming that the drive axle is rigid and the torque converter is locked, the rotational dynamics of the engine is described by

$$\dot{\omega}_e = \frac{T_{net} - r(hF_{tr} + T_{br})}{I_e} \tag{7.7}$$

where T_{net} is the net combustion torque (indicated torque-friction torque) and is a nonlinear function of ω_e and P_m, r is the gear ratio, h is the tire radius, I_e is the effective rotational inertia of the engine when the inertia wheel is also referred to the engine side, T_{br} is the brake torque at the wheels and F_{tr} is the tractive force defined as

$$F_{tr} = K_r sat\left(\frac{i}{i_{\max}}\right) \tag{7.8}$$

where K_r is the longitudinal tire stiffness and i is the slip between the tires and the ground, defined as

$$i = 1 - \frac{v}{rh\omega_e} \tag{7.9}$$

where v is the longitudinal velocity of the vehicle.

Assuming that the brakes obey first order linear dynamics, we have

$$\dot{T}_{br} = \frac{T_{bc} - T_{br}}{\tau_b} \tag{7.10}$$

where T_{bc} is the commanded brake torque and τ_b is the brake actuator time constant. The final equation for longitudinal vehicle velocity is given by

$$\dot{v} = \frac{F_{tr} - c_a v^2 - F_f}{M} \tag{7.11}$$

where c_a is the drag coefficient, F_f is the force due to rolling resistance and M is the effective mass of the vehicle.

The above vehicle model consists of nonlinearities that have been dealt with, when designing automated controllers, by using techniques such as input/output linearization (Swaroop et al., 1994) and modified sliding control (Hedrick et al., 1991). Another approach is to linearize the nonlinear vehicle model (7.1) by considering different operating points for vehicle speed and throttle angle (Ioannou and Xu, 1994). The linearization procedure is described as follows.

Let V_0 be the steady state speed for throttle input θ_0. Define $\overline{V} = V - V_0$ and $\overline{\theta} = \theta - \theta_0$ as the deviations from vehicle speed V_0 and throttle angle θ_0, respectively. Using the validated nonlinear longitudinal vehicle model, the linearized model relating $\overline{V}, \overline{\theta}$ over a wide range of speed V_0 (i.e. from 0 to 36ms^{-1}), for any fixed gear state, is of the form

$$\frac{\overline{V}}{\overline{\theta}} = \frac{b_0}{s^3 + a_2 s^2 + a_1 s + a_0} = \frac{b_0}{(s+p_1)(s+p_2)(s+p_3)} \tag{7.12}$$

where the coefficients b_0, a_2, a_1, a_0 are functions of the operating point (θ_0, V_0).

For all operating points considered, b_0>0, p_1>0 and p_2 and p_3 (which may be real or complex conjugates) have positive real parts. Furthermore, Re (p_2), Re (p_3)>>p_1, where "Re" denotes the real part, and 0<p_1 ≤0.2. A variable μ gives a measure of how far Re (p_2) and Re (p_3) are from p_1 and is defined as

$$\mu = \sup_{\theta_0 \in \Theta} \max \left[\frac{p_1}{\mathrm{Re}(p_i)}, i = 2,3 \right] \tag{7.13}$$

where Θ is the full domain of θ. Simulations show that μ<0.05, which indicates that -p_1 is the dominant pole and the fast dynamics associated with p_2 and p_3 can be neglected, leading to the simpler model

$$\frac{\overline{V}}{\overline{\theta}} = \frac{b}{s+a} \tag{7.14}$$

where a and b vary with V_0. The effects of the neglected fast modes and uncertainties are modeled as a disturbance d, giving the final model

$$\dot{\overline{V}} = -a\overline{V} + b\overline{\theta} + d$$

or equivalently

$$\dot{V} = -a(V - V_0) + b\overline{\theta} + d \tag{7.15}$$

In vehicle following the vehicle position is also taken into account with the vehicle speed. So the complete dynamic equation of the throttle angle to vehicle speed and position subsystem is

$$\begin{aligned}\dot{X} &= V\\ \dot{V} &= -a(V-V_0)+b\bar{\theta}+d\end{aligned} \tag{7.16}$$

Sheikholeslam and Desoer (1991) developed another model for the dynamics of the i-th vehicle assuming a horizontal road surface with no wind gust. Following Newton's second law of motion, the vehicle dynamics are described as

$$m_i\ddot{x}_i = m_i\xi_i - K_{di}\dot{x}_i^2 - d_{mi} \tag{7.17}$$

$$\dot{\xi}_i = -\frac{\xi_i}{\tau_i(\dot{x}_i)} + \frac{u_i}{m_i\tau_i(\dot{x}_i)} \tag{7.18}$$

where $K_{di}\dot{x}_i^2$ is the air resistance, $K_{di} = \frac{\rho A_i C_{di}}{2}$, ρ is the specific mass of air, A_i is the cross-sectional area of the i-th vehicle and C_{di} is the drag coefficient of the i-th vehicle; $m_i\xi_i$ is the engine force applied to the i-th vehicle, the constant d_{mi} is the mechanical drag of the i-th vehicle and u_i is the i-th vehicle throttle input. The input-output behavior of each vehicle (in a platoon) is linearized by substituting ξ_i in (7.18) from (7.17) to get

$$\dot{\xi}_i = -\frac{1}{\tau_i(\dot{x}_i)}\left[\ddot{x}_i + \frac{K_{di}}{m_i}\dot{x}_i^2 + \frac{d_{mi}}{m_i}\right] + \frac{u_i}{m_i\tau_i(\dot{x}_i)} \tag{7.19}$$

and differentiating (7.17) and substituting $\dot{\xi}_i$ from (7.19) to obtain

$$\dddot{x}_i = -2\frac{K_{di}}{m_i}\dot{x}_i\ddot{x}_i - \frac{1}{\tau_i(\dot{x}_i)}\left(\ddot{x}_i + \frac{K_{di}}{m_i}\dot{x}_i^2 + \frac{d_{mi}}{m_i}\right) + \frac{u_i}{m_i\tau_i(\dot{x}_i)} \tag{7.20}$$

Eqn. (7.20) can be written as

$$\dddot{x}_i = b(\dot{x}_i, \ddot{x}_i) + a(\dot{x}_i)u_i \tag{7.21}$$

where

$$b(\dot{x}_i,\ddot{x}_i):=-2\frac{K_{di}}{m_i}\dot{x}_i\ddot{x}_i-\frac{1}{\tau_i(\dot{x}_i)}\left(\ddot{x}_i+\frac{K_{di}}{m_i}\dot{x}_i^2+\frac{d_{mi}}{m_i}\right)$$

$$a(\dot{x}_i):=\frac{1}{m_i\tau_i(\dot{x}_i)}$$

An exogeneous input c_i, that is related to the vehicle throttle input u_i, is created to linearize the i-th vehicle's nonlinear dynamics, by

$$u_i=\frac{1}{a(\dot{x}_i)}\left[c_i-b(\dot{x}_i,\ddot{x}_i)\right] \tag{7.22}$$

Substituting (7.22) into (7.21) gives a system of linear equations representing the dynamics of the i-th vehicle after linearization by state feedback, namely for i=1 ,2...

$$\begin{aligned}\frac{d}{dt}x_i&=\dot{x}_i\\ \frac{d}{dt}\dot{x}_i&=\ddot{x}_i\\ \frac{d}{dt}\ddot{x}_i&=c_i\end{aligned} \tag{7.23}$$

Lateral

Lateral vehicle dynamics modeling is used for designing automated lateral controllers. Nathoo and Healey (1978) showed that a coupled vertical and lateral vehicle model consists of sprung mass, two independent front suspension and wheel unsprung masses and a solid rear axle comprising the rear wheel unsprung mass. Peng and Tomizuka (1990) proposed a complex and a simple model that describe the lateral vehicle dynamics. The complex model has six degrees of freedom: three translational and three rotational. The simple model includes only the lateral and yaw motions of the vehicle and are expressed in state space as follows

$$\frac{d}{dt}\begin{bmatrix}y_r\\ \dot{y}_r\\ \varepsilon-\varepsilon_d\\ \dot{\varepsilon}-\dot{\varepsilon}_d\end{bmatrix}=\begin{bmatrix}0&1&0&0\\ 0&\frac{A_1}{V}&-A_1&\frac{A_2}{V}\\ 0&0&0&1\\ 0&\frac{A_3}{V}&-A_3&\frac{A_4}{V}\end{bmatrix}\begin{bmatrix}y_r\\ \dot{y}_r\\ \varepsilon-\varepsilon_d\\ \dot{\varepsilon}-\dot{\varepsilon}_d\end{bmatrix}+\begin{bmatrix}0\\ B_1\\ 0\\ B_2\end{bmatrix}\delta+\begin{bmatrix}0\\ d_1\\ 0\\ d_2\end{bmatrix} \tag{7.24}$$

where y_r is the lateral deviation of the mass center from the reference, ε is the yaw angle of the vehicle, ε_d is the desired yaw angle from the road curve and δ is the front wheel steering angle (Fig. 7-2). The A_i's and B_i's are defined as

$$A_1 = \frac{-4C_s}{m} \qquad B_1 = \frac{2}{m}(C_s + F_x)$$

$$A_2 = \frac{2C_s}{m}(l_2 - l_1) \qquad B_2 = \frac{2l_1}{I_z}(C_s + F_x)$$

$$A_3 = \frac{2C_s}{I_z}(l_2 - l_1) \qquad d_1 = \frac{F_{disturb}}{m} - \frac{V^2}{\rho} + \frac{A_2}{V}\dot{\varepsilon}_d$$

$$A_4 = \frac{-2C_s}{I_z}(l_1^2 + l_2^2) \qquad d_2 = \frac{T_{disturb}}{I_z} + \frac{A_4}{V}\dot{\varepsilon}_d - \ddot{\varepsilon}_d$$

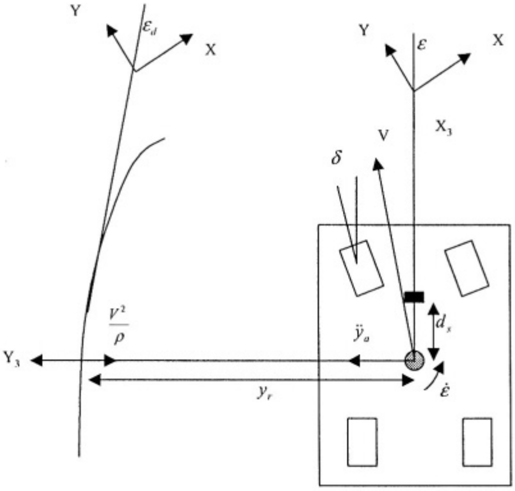


Figure 7-2. Schematic diagram of the simplified vehicle model (Peng and Tomizuka, 1990).

where m and I_z are the mass and moment of inertia of the vehicle, V is the longitudinal velocity of the vehicle, ρ is the radius of curvature of the road, l_1 and l_2 are the distance from the mass center to the front and rear axle respectively, F_x is the traction force from the tire, $F_{disturb}$ and $T_{disturb}$ are the disturbance force and torque acting on the vehicle respectively, and $C_{s.}$ is the cornering stiffness. As is evident from (7.24), the cornering stiffness C_s and the longitudinal velocity V of the vehicle are important parameters in lateral vehicle dynamics. While V is measured directly, the same cannot be done for C_s, which is affected by factors such as tire pressure, load, velocity, among others, and is estimated using a parameter identification scheme.

7.3 Manual Control

In order to study the interaction of human control actions with vehicle dynamics, investigators studied mathematical models such as Fig. 7-3 that mimic the behavior of human drivers (Ioannou and Chien, 1993). Using the structure of Fig. 7-3, investigators came up with models based on vehicle following in a single lane with no passing.

We present an overview of the following different human driver models: (1) Pipes model, (2) Optimal Control Model and (3) Look-Ahead model. The reader is directed to the references for a more detailed discussion of these models.

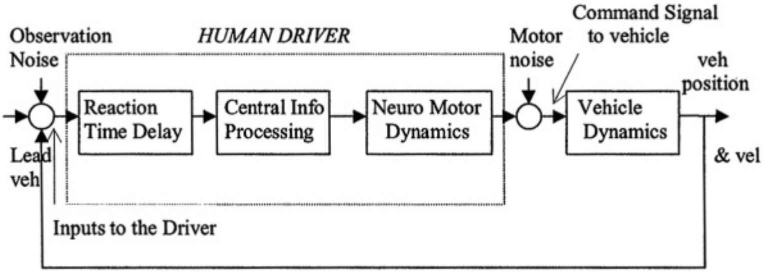


Figure 7-3. Structure of the human driver model.

Pipes Model

This model was first proposed by Pipes (1953) and pertains to a single lane dense traffic. The vehicle following theory assumes that each driver reacts to a stimulus that is the velocity difference and responds with an acceleration command, i.e.

$$Response(t) = Sensitivity \times Stimulus(t - \tau)$$

where τ is the reaction time of the driver-vehicle system.

The vehicle dynamics is modeled by an integrator and the driver's central processing and neuromuscular dynamics are modeled by a constant. The block diagram of the model is shown in Fig. 7-4. Chandler et al. (1958) used vehicle-following data to validate this model at the General Motors Technical Center and showed experimentally that the reaction time is 1.5s while the constant gain is $0.37s^{-1}$. It is mathematically represented as

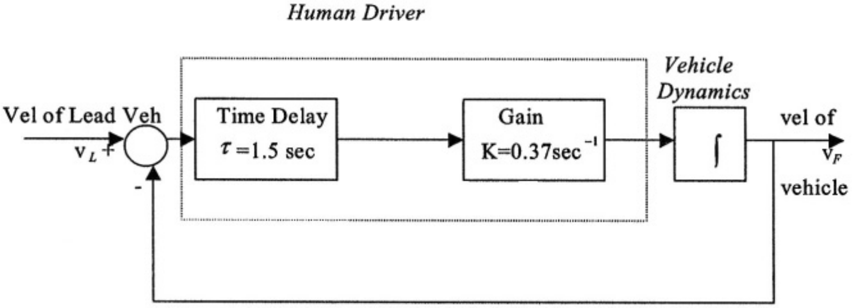


Figure 7-4. Linear follow-the-leader model – Pipes model.

$$a_F(t) = \frac{\lambda}{M}[v_L(t-\tau) - v_F(t-\tau)] \tag{7.25}$$

where v_L is the velocity of the leading vehicle, v_F is the velocity of the following vehicle, a_F is the acceleration of the following vehicle, M is the mass of the following vehicle and λ is a sensitivity factor.

Optimal Control Model

The optimal controller is based on a quadratic cost function that penalizes the weighted sum of the square of the intervehicle spacing and the square of the relative velocity (Tyler, 1964). The quadratic performance criterion function is chosen such that the following cost function is minimized

$$J = \frac{1}{2}\int_0^\infty \{[s_L(t) - s_F(t) - \lambda v_F(t)]^2 q_1 + [v_L(t) - v_F(t)]^2 q_2 + ru^2(t)\}dt \tag{7.26}$$

where λ is the chosen headway of the vehicle and q_1, q_2, r are the associated weights.

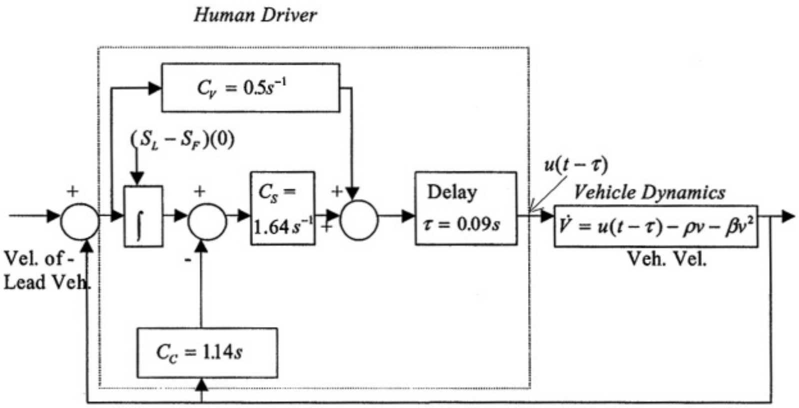


Figure 7-5. Linear optimal control model.

With the supplementary assumption that the leading and following vehicle dynamics are identical, the solution of this optimization problem is the control *u(t)* given by

$$u(t) = C_S[s_L(t) - s_F(t) - C_C v_F(t)] + C_V[v_L(t) - v_F(t)] \qquad (7.27)$$

where subscripts L and F represent the leading and following vehicles respectively and s, v stand for the distance and velocity of the vehicles respectively, while C_C, C_S and C_V are constant gains. The values of these parameters were obtained from actual traffic data.

As the weights differ from driver to driver, the best the optimal controller can perform is as a controller that simulates the behavior of a particular human driver. Another drawback of this model is that it neglects the driver's reaction time, the neuromuscular dynamics and the nonlinearities of vehicle dynamics. These reasons prompted Burnham et al. (1974) and Bekey et al. (1977) to first modify the optimal controller structure by introducing the effects of driver reaction time and vehicle nonlinearities and then estimate the unknown parameters by fitting real traffic data. The model is shown in Fig. 7-5 and the vehicle dynamics are modeled as

$$\dot{v}_F = u(t-\tau) - \rho v_F - \beta v_F^2 \qquad (7.28)$$

where ρ is the mechanical drag coefficient (about 10^{-5} to 10^{-4} $\sec^{-1}$) and β is the aerodynamic drag coefficient (about 10^{-3} to 10^{-2} m^{-1}).

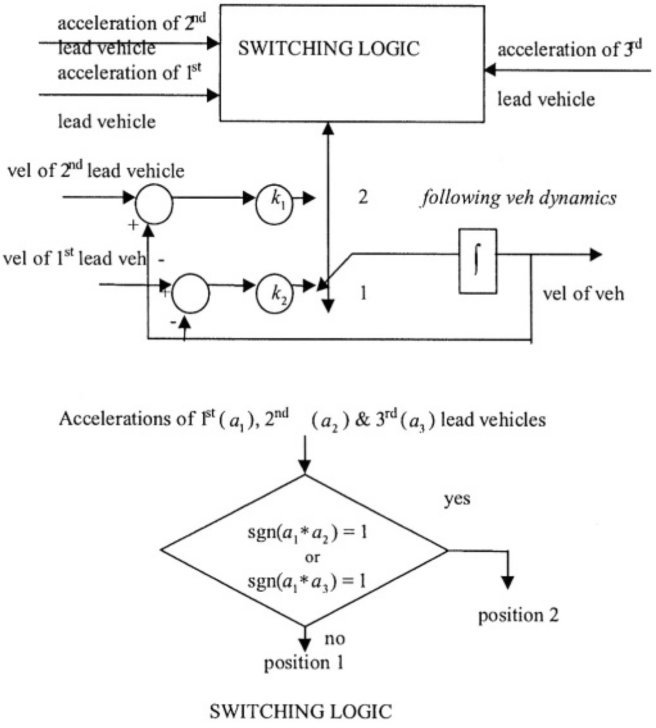


Figure 7-6. Look Ahead model.

Look Ahead Model

This model (Burnham et al., 1974; Bekey et al., 1977) is based on the hypotheses that the human driver observes the behavior of three vehicles directly ahead. It incorporates a switching logic that determines whether to follow the velocity of the first or the second lead vehicle. The switching logic determines the majority direction of acceleration and then actuates the mode switch accordingly. Parameter values obtained by fitting actual traffic data are k_1=0.2s^{-1} and k_2=0.65s^{-1} (Fig. 7-6).

7.4 Automated Longitudinal Control

The control phase of automobile driving is concerned with actuation of steering wheel, accelerator and brakes in such a way that the vehicle follows its preceding vehicle or the desired path with a desired velocity and with acceptable precision. In this section we investigate a design for automatic longitudinal vehicle control taken from Ioannou and Xu (1994) followed by discussions on longitudinal controller design for heavy-duty vehicles, vehicle-to-vehicle communication systems, sensor requirements, spacing and safety improvements. The longitudinal vehicle model used for the controller design is given by (7.16). The time delay and the dynamics of the brake system are ignored in the modeling. For longitudinal control, the system may be considered as having two input variables: throttle angle command and brake command, and one output variable: vehicle speed. The other inputs such as aerodynamic drag, road conditions and vehicle mass changes are treated as disturbances. We deal with each subsystem separately as the throttle and the brake subsystems are not allowed to act simultaneously.

Longitudinal Controller Design

Throttle angle to vehicle speed and position model. Using the complete dynamic equation of the throttle angle to vehicle speed and position subsystem (from (7.16))

$$\begin{aligned}\dot{X} &= V\\ \dot{V} &= -a(V-V_0)+b\bar{\theta}+d\end{aligned}$$

the closed loop dynamic equations of the throttle subsystem are

$$\begin{aligned}\dot{X}_r &= V_l - V_f\\ \dot{V}_f &= -a(V_f-V_l)+b\bar{\theta}_f+d\\ \delta &= X_r - hV_f - S_0\\ V_r &= V_l - V_f\end{aligned} \tag{7.29}$$

where V_f, V_l is the velocity of the following and the leading vehicles, respectively, $\bar{\theta}_f = \theta_f - \theta_0$ is the throttle angle deviation from the desired position, a,b are coefficients determined by V_l, δ is the spacing error relative to desired spacing $S_d = hV_f + S_0$, measured from the rear of the leading vehicle to the front of the following vehicle, h is the desired time headway of the automated vehicle, defined as the time taken to cover the distance $S_d - S_0$; S_0 is a nonnegative constant and d is the disturbance term to represent the fast dynamics of the vehicle model.

For a Proportional-Integral-Differential (PID) with fixed gain the following controller is used

$$\theta_f = \theta_0 + k_1V_r + k_2\delta + \int_0^t (k_3V_r + k_4\delta)d\tau \tag{7.30}$$

where k_1, k_2, k_3, k_4 are gains chosen to meet the control objective $V_r, \delta \rightarrow 0$ as $t \rightarrow \infty$.

The term θ_0 is obtained from a look-up table that describes the relation of θ_0 and V_l, i.e.

$$\theta_0 = f^{-1}(V_l) \tag{7.31}$$

The final closed loop control is given by combining (7.29) and (7.30):

$$\dot{V}_f = -(a + bk_1)(V_f - V_l) + bk_2\delta + b\int_0^t (k_3 V_r + k_4 \delta) d\tau + d \tag{7.32}$$

Due to passenger ride comfort, the control objective is to be achieved under the following human factors design constraints:

C1 $a_{\min} \leq \dot{V}_f \leq a_{\max}$ where $a_{\min}$ and $a_{\max}$ are specified

C2 the absolute value of the jerk $\ddot{V}_f$ should be as small as possible

Taking into consideration constraints C1 and C2, it is seen that rapid changes in V_l create a large error, which violates the constraints. This might happen when the following vehicle switches following from one vehicle to another due to lane changes. Such a change might introduce large initial position error $|\delta(0)|$ and speed error $|V_r(0)|$ leading to high accelerations/decelerations that might violate constraints C1, C2. In order to prevent that, V_l is passed through an acceleration limiter shown in Fig. 7-7, where p is a positive constant. Instead of following V_l, $\hat{V}_l$ is followed. The acceleration limiter eliminates any erratic or sudden changes in V_l during transients and presents a smoother trajectory to the follower. It also serves as a low pass filter for V_l and the $\hat{V}_l$ trajectory is defined as

$$\dot{\hat{V}}_l = \begin{cases} a_{\max} & if \quad p(V_l - \hat{V}_l) \geq a_{\max} \\ p(V_l - \hat{V}_l) & if \quad a_{\min} < p(V_l - \hat{V}_l) < a_{\max} \\ a_{\min} & if \quad p(V_l - \hat{V}_l) \leq a_{\min} \end{cases} \tag{7.33}$$

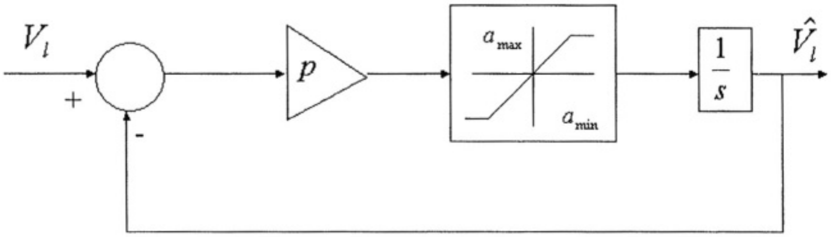


Figure 7-7. Acceleration limiter.

The effect of the large initial error $|\delta(0)|$ is taken care of by introducing a saturation element $sat(\delta)$ that is defined as

$$sat(\delta) = \begin{cases} e_{\max} & if \quad \delta > e_{\max} \\ e_{\min} & if \quad \delta < e_{\min} \\ \delta & otherwise \end{cases} \tag{7.34}$$

and limits the measurements seen by the controller to be within $e_{\min}$ and $e_{\max}$. Note that, since negative position error means short intervehicle spacing, $|e_{\min}|$ is much larger than $e_{\max}$.

The low pass character of the throttle actuator together with the acceleration limiter and the saturation function give a smooth throttle angle response and help reduce the amount of jerk. To further reduce jerk and attenuate the effect of sensor noise in measurements, the following low pass filter

$$\frac{c_0}{s + c_0}$$

is used to filter X_l, V_l and V_f, where c_0 is some constant chosen based on the sampling rate and noise level. For simplicity, these filters were not introduced in the equations. The throttle controller was validated using the actual nonlinear model (7.1).

Brake torque to vehicle speed and position model. It is assumed that when the automated vehicle is braking, the throttle is at a minimum value that corresponds to idle engine speed. In this case the transmission torque is very small compared to the braking torque and is therefore neglected. The dynamic equations of the braking torque to vehicle speed and position subsystem based on Newton's second laws of motion and on the assumption that the wheels are not locked, are given as:

$$\begin{aligned} &\dot{X} = V \\ &\dot{V} = \frac{1}{M}(-c_1 T_b - f_0 - c_2 V - c_3 V^2) \end{aligned} \tag{7.35}$$

where T_b is the braking torque, M is the mass of the vehicle, $c_1 T_b$ is the braking force, f_0 is the static friction force, $c_2 V$ is the rolling friction, $c_3 V^2$ is the air resistant force and c_1, c_2, c_3, f_0 are known constants obtained from experiments.

The brake controller is designed to control the brake line pressure, which is approximately proportional to the brake torque. The dynamics of the brake torque are given by (from (7.35))

$$\dot{V}_f = \frac{1}{M}(-c_4 P_r - f_0 - c_2 V_f - c_3 V_f^2) \tag{7.36}$$

where P_r is the control input and c_4 is a known constant obtained from experiments. Using feedback linearization, (7.36) is converted into the linear system

$$\dot{V}_f = u \tag{7.37}$$

where $u = \frac{1}{M}(-c_4 P_r - f_0 - c_2 V_f - c_3 V_f^2)$.

Using the feedback control

$$u = k_5 V_r + k_6 \delta \tag{7.38}$$

gives the following equation

$$\dot{V}_f = k_5 V_r + k_6 \delta \tag{7.39}$$

where k_5, k_6 are gains chosen to meet the control objective $V_r, \delta \to 0$ as $t \to \infty$.

Thus the desired line pressure is obtained as

$$P_r = \frac{-1}{c_4}[M(k_5 V_r + k_6 \delta) + f_0 + c_2 V_f + c_3 V_f^2] \tag{7.40}$$

To take into account constraint C1, a saturation is put on $k_5 V_r + k_6 \delta$, leading to the modified expression for the line pressure:

$$P_r = \begin{cases} 0 & \text{if} \quad -M(k_5V_r + k_6\delta) < f_0 + c_2V_f + c_3V_f^2 \\ -(Ma_{\min} + f_0 + c_2V_f + c_3V_f^2)/c_4 & \text{if} \quad k_5V_r + k_6\delta < a_{\min} \\ -[M(k_5V_r + k_6\delta) + f_0 + c_2V_f + c_3V_f^2]/c_4 & \text{otherwise} \end{cases} \tag{7.41}$$

The desired P_r might have discontinuities at the time the brake controller is turned on. Since the brake actuator acts as a low pass filter, the output of the actuator is smooth, leading to a smooth braking force and therefore any discontinuities in P_r will not lead to rough braking.

The throttle and brake controller follows a logic design to switch from one mode to another whose details can be found in Ioannou and Xu (1994). The throttle controller is turned off with the throttle angle set to a minimum when the brake controller is on and the brake controller is turned off when the throttle controller is on.

Other longitudinal controller designs include the one by Hedrick et al. (1991) where a combined throttle/brake controller is outlined using a modified sliding control method. In Ioannou et al. (1992) an autonomous controller is designed for a constant time headway policy. In Raza et al. (1997) a model and a computer controller is developed for the brake subsystem for implementation in automatic vehicle following.

Heavy-duty Vehicles

The control algorithm discussed above is applicable to light duty passenger vehicles. For heavy-duty vehicles such as commercial trucks and buses, however, different vehicle dynamics modeling is necessary along with different control algorithms that have been developed by Kanellakopoulos and Tomizuka (1997). The reasons put forth for using different controllers include: (1) because of increased weight, heavy-duty vehicles have low actuation-to-weight ratio; (2) heavy-duty vehicles have roll and yaw instability modes that are insignificant in light duty vehicles; (3) strong interaction between the longitudinal and lateral dynamics in heavy-duty vehicles; (4) pronounced actuator delays and nonlinearities and (5) increased sensitivity to disturbances such as wind gusts. The authors developed a model for truck-semitrailer vehicles and linearized the longitudinal model around different operating points determined by different fuel command/vehicle mass combinations. The resultant sixth order model is finally reduced to a first order by neglecting the fast mode dynamics associated with angular velocity of the wheel, fuel systems, intake manifold pressure, engine speed and the turbocharged diesel engine rotor speed.

Different control algorithms using a proportional and an integral controller are implemented and simulated with the full nonlinear longitudinal model for a group (platoon) of ten tractor-semitrailer combination vehicles. The error term for the control algorithm is defined as

$$v_r + k\delta \tag{7.42}$$

where v_r is the relative speed, i.e. the difference between the leading and following vehicle speeds, δ is the spacing error and k is a positive design constant. Variations of the Proportional-Integral (PI) controller are simulated such as adding a nonlinear signed quadratic (Q) term to the PI controller, which thus becomes a Proportional-Integral-Quadratic (PIQ) controller, using adaptive gains in the PIQ controller, using intervehicle communication, variable time headway and variable separation error weighting k. Results show that the different procedures offer different advantages and disadvantages. For example, using intervehicle communication improves the throughput significantly but at the expense of increasing complexity required to establish and maintain the intervehicle communication.

Vehicle-to-Vehicle Communication

Automated vehicles may be equipped with communication systems. Such vehicles can communicate with each other and exchange information about vehicle status and traffic flow conditions that help in the longitudinal control of the vehicles. For example, when an automated vehicle detects a stopped vehicle on the freeway it communicates to other similarly equipped automated vehicles about the obstacle. After receiving this information, the automated vehicles start slowing down, change lanes and propagate the message to other automated vehicles behind. As a consequence, the automated vehicles perform soft braking (i.e. braking at a comfortable rate), which slows down the whole traffic stream including any manually driven vehicles between the automated vehicles. Thus the disturbance caused by stopped vehicles is attenuated and the traffic flow is smoother.

The vehicle-to-vehicle communication system design may be based on two-way communication, with each vehicle simultaneously transmitting and receiving information. The transmitted signal is acknowledged by each receiving vehicle, thus allowing the automated vehicles to detect the surrounding vehicles.

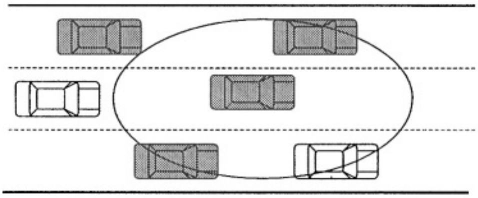


Figure 7-8. 'Zone of relevance' of an automated (shaded) vehicle.

The frequency of operation is an open issue. Frequencies as high as 64GHz have been proposed (Kaltwasser and Kassubek, 1994). Each automated vehicle has a 'zone of relevance' around it (Kaltwasser and Kassubek, 1994) to which communication and data exchange are restricted (Fig. 7-8). It is obvious that this zone may include automated vehicles with communication capability as well as manually driven vehicles. An appropriate strategy for dealing with this is the following: when a vehicle in the 'zone of relevance' does not acknowledge the transmission, it is automatically classified as a manually driven vehicle. This improves traffic coordination, as the automated vehicles know where other automated vehicles are in the immediate surrounding. Furthermore, it circumvents the potential danger due to failures of the communication system on an automated vehicle. An automated vehicle with a non-functional communication system is treated as a manually driven vehicle.

For each pair of automated vehicles, both the leader and the follower exchange information like the 'Double Boomerang Transmission System' (Mitzui et al., 1994). Exchange of vehicle information like braking capability and tire pressure in addition to traffic conditions reduces the minimum safe intervehicle spacing. The required information data transfer rate is over 1Mbps, and the processing rate is between 1000MIPS and 9000MIPS (Milestone 2 Report, Appendix B, 1996). Contingencies exist for emergency measures (like hard braking) which will override any ongoing message and are given top priority.

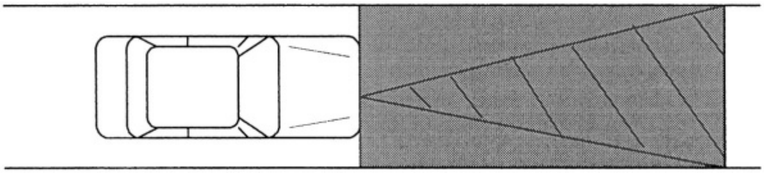


Figure 7-9. The shaded area denotes ideal sensor coverage, while crossed area denotes actual coverage for single beam sensor.

Longitudinal Sensors

Automated vehicles are equipped with sensors that measure the relative distance and relative speed to all vehicles in the immediate neighborhood. Naturally, vehicles ahead must be detected with the highest accuracy and precision. Relative speed readings need to be accurate and sensitive to small speed changes of less than 2mph.

The forward looking longitudinal sensors must have a sufficient range to allow the vehicle to come to a stop even under the assumption of a "brick wall scenario". For example, a simple calculation shows that a vehicle traveling at 80mph which has a maximum deceleration ability of 0.65g needs 100 meters to come to a complete stop. However, while deciding on the range of the frontal longitudinal sensors, the degradation in braking capabilities due to wet/icy road conditions must also be taken into consideration. Furthermore the front sensors must be able to distinguish and resolve the position of all the target vehicles in two dimensions, i.e. relative distance and relative angle. The sensors must be able to track the target vehicle regardless of the presence of other vehicles in the adjacent lanes, in straight roadway segments and also along curves. It is quite a task and may require the combined powers of sophisticated radar systems and real-time image processing.

The longitudinal ranging sensor may be chosen to be a single beam sensor as depicted in Fig. 7-9 (Ioannou et al., 1994). The primary reason is that a single beam sensor has lower interference and hence less chance of false alarms. However, the shaded region shows the ideal sensor coverage needed to detect obstacles in the lane other than vehicles such as motorcycles, animals, etc. and vehicles in cut-in situations, while the crossed region in Fig. 7-9 shows actual sensor coverage. The single beam sensor may not be able to detect sudden cut-in vehicles from adjacent lanes. Moreover, with increase in the market penetration of automated vehicles, most vehicles may have radar-type ranging sensors at similar frequencies, which means greater probability of interference and shorter radar ranges. In this case a combination of a narrow beam radar with a video camera may provide the desired properties of a ranging and obstacle detection sensor. Other possibilities include multiple beam sensors like the one used by Bastian et al. (1998) during field tests on the German autobahn.

The backward looking sensors have to measure the relative position and relative speed of the following vehicle and must be able to detect potential rear-end collision threats. It is also needed to evaluate the available spacing during lane changing and merging. They are similar to the front sensors and are subject to same difficulties discussed above. However, they are not as essential as the frontal longitudinal sensors and automated vehicles may or may not be equipped with them.

Intervehicle Spacing

The automated longitudinal controller of an automated vehicle chooses the intervehicle spacing. Different policies such as constant intervehicle spacing (Shladover, 1977) or constant time headway spacing (Ioannou et al., 1992; Ioannou and Xu, 1994) can be used. The intervehicle spacing or time headway used (depending on the controller) is chosen such that it guarantees stopping of the following vehicle in any braking scenario. The controller uses algorithms such as the one outlined in Kanaris, Ioannou and Ho (1997) to determine the Minimum Safe Spacing (MSS), which is defined as the minimum intervehicle spacing that gives no-collision for all possible braking scenarios. The algorithm uses a “worst case” analysis assuming conservative braking capabilities of the following and leading vehicles.

Bose and Ioannou (1998, 1999) contend that the initial stages of vehicle automation include partially or semi-automated vehicles that have automation only in the longitudinal direction using ICC systems. Some semi-automated vehicles have a frontal collision warning system (FCWS) and depend on the human driver during emergencies. Such a semi-automated vehicle is equipped with ICC that enables it to automatically follow a vehicle in a lane, but requires the human driver to take over the control of the vehicle in case of an emergency. These semi-automated vehicles are exposed to the problem that the driver may be incapable of taking over the control of the vehicle. Different systems for monitoring driver vigilance such as tracking eye movement of the driver (Kogure et al., 1993) and others (Cointot et al., 1993; Dingus et al., 1987; Scolfield et al., 1993; Skipper and Wierwille, 1986) have been proposed to tackle this problem. The other type of semi-automated vehicles has frontal collision avoidance system (FCAS) and can handle emergencies without any driver intervention.

Bose and Ioannou (1998) showed that the MSS for semi-automated vehicles with FCWS is greater than the average used in today's manual traffic, the reason being that due to the automatic longitudinal control of the semi-automated vehicle, the human driver has only lateral control and may tend to relax. Thus he/she may have a larger reaction time than usual, which demands greater intervehicle spacing. On the other hand, semi-automated vehicles with full longitudinal automation are equipped with ICC and frontal collision avoidance system (FCAS) and rely on their own sensors, throttle and brake controllers and intelligence to operate in the highway environment. In these vehicles the human driver has no functionality in the longitudinal direction and is replaced by sensors and controllers operated by an on-board computer. This makes the time difference between the onset of braking of the leading and the following vehicle smaller than the average reaction time of a human driver. Therefore, the following automated vehicle requires shorter MSS than a manually driven vehicle.

Safety

The degree of automation in an automated vehicle determines the level of safety it provides. In semi-automated vehicles the human driver is not completely out of the driving loop. Depending on the degree of automation in the longitudinal direction in a semi-automated vehicle, the functionality of a human driver varies. In semi-automated vehicles with partial longitudinal automation, the human driver has almost the same functions as in today's manual driving except that he/she is aided by devices such as ICC and FCWS. It is a different story, however, for semi-automated vehicles with full longitudinal automation (equipped with FCAS) where the driver is given an opportunity to take control of the vehicle and perform necessary collision avoidance. On detection of a potentially hazardous situation, the semi-automated vehicle performs automatic soft braking and informs the driver. This allows greater reaction time for the driver who otherwise was only responsible for lane keeping (and lane changing). This is done to include cases where lateral collision avoidance like lane changing may be necessary. However, if he/she does not respond, then the semi-automated vehicle initiates hard braking (i.e. braking at maximum possible rate). Thus, the level of safety in a semi-automated vehicle is affected by the possibility of human intervention during collision avoidance maneuvers (as shown in the flowchart in Fig. 7-10).

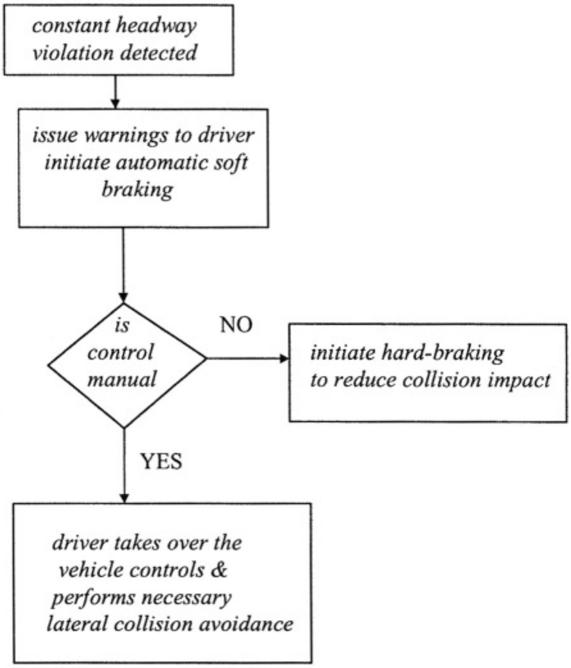

constant headway violation detected
issue warnings to driver initiate automatic soft braking
is control manual
NO
initiate hard-braking to reduce collision impact
YES
driver takes over the vehicle controls & performs necessary lateral collision avoidance

Figure 7-10. Flowchart showing the emergency procedure in a semi-automated vehicle equipped with FCAS.

7.5 Automated Lateral Control

Lateral control of vehicles has been achieved using different types of systems. Vision-based control system with on-board camera is discussed in Jurie et al. (1993) where the position (state) of the vehicle is estimated using extended Kalman filtering. Other systems include navigation along a known road network using a priori information and low-level road detection (Zhang and Thomas, 1993). Peng and Tomizuka (1990) use a magnetic road marker-based system for lateral control of vehicles, which integrates a feedback controller with a feedfoward loop as shown in Fig. 7-11. The roadway reference/sensing system is based on a series of magnetic markers placed in the center of the roadway to be followed by the automated vehicle (Hessburg et al., 1991). Hall effect magnetometers mounted on the front center of the vehicle sense the magnetic field from the markers. A complex nonlinear and a simple linear model have been used to represent the lateral dynamics of a front-wheel-steering rubber-tire vehicle. The controller is designed using the simple model and evaluated on the complex model.

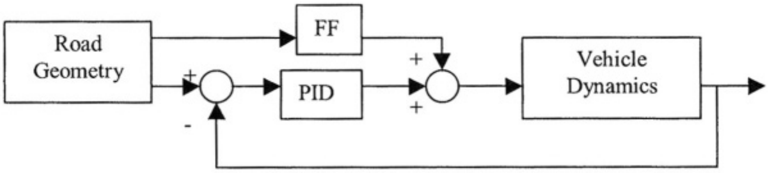


Figure 7-11. Block diagram of lateral controller.

Lateral Controller Design

The overall lateral controller consists of a feedback and a feedforward controller that are elaborated in the following subsections.

Feedback controller. The frequency-shaped linear quadratic (FSLQ) control theory is used for the design of the feedback controller. The feedback controller is designed to minimize a performance index. The ride quality is included in the performance index and the high-frequency robustness of the controller is improved by properly choosing the weighting factors.

The performance index is given by:

$$\begin{aligned} J = \frac{1}{2\pi}\int_{-\infty}^{\infty} & [a^*(j\omega)Q_a(j\omega)a(j\omega) + y_r^*(j\omega)Q_y(j\omega)y_r(j\omega) + (\varepsilon(j\omega) - \varepsilon_d(j\omega))^* \\ & Q_\varepsilon(j\omega)(\varepsilon(j\omega) - \varepsilon_d(j\omega)) + y_s^*(j\omega)Q_i(j\omega)y_s(j\omega) + \delta^*(j\omega)\delta(j\omega)]d\omega \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty} & [a^*(j\omega)\frac{q_a^2}{1+\lambda_a^2\omega^2}a(j\omega) + y_r^*(j\omega)\frac{q_y^2}{1+\lambda_y^2\omega^2}y_r(j\omega) + (\varepsilon(j\omega) - \varepsilon_d(j\omega))^* \\ & \frac{q_\varepsilon^2}{1+\lambda_\varepsilon^2\omega^2}(\varepsilon(j\omega) - \varepsilon_d(j\omega)) + y_s^*(j\omega)\frac{q_i^2}{(j\omega)^2}y_s(j\omega) + \delta^*(j\omega)\delta(j\omega)]d\omega \end{aligned} \tag{7.43}$$

where a is the difference between the lateral acceleration $\ddot{y}_a$ and its desired value, expressed as

$$a = \ddot{y}_a - \frac{V^2}{\rho} \approx \ddot{y}_r \tag{7.44}$$

and ρ is the radius of curvature of the road, y_r is the lateral deviation of the mass center from the reference, ε is the yaw angle of the vehicle, ε_d is the desired yaw angle from the road curve and δ is the front wheel steering angle. The weighting factors on the tracking error terms in the performance index (7.43) is shaped in the frequency domain to enhance the controller robustness when the plant experiences high frequency measurement noises or unmodeled dynamics. At the same time, the tracking performance is not deteriorated in the low frequency region. The weighting factor on the lateral acceleration in (7.43) is chosen to ensure good ride quality in the frequency range where passengers are sensitive.

The coefficient λ_a, which is crucial for ride quality, determines the weighting on the lateral acceleration and is set to $\frac{1}{60\pi}$. The coefficients λ_y and λ_ε are tuned to enhance the high-frequency robustness of the controller while maintaining good tracking capability. From the analysis presented in Peng and Tomizuka (1990), the coefficients are chosen to be

$$\lambda_y, \lambda_\varepsilon = \left(\frac{D_f}{2\pi a_{\max}}\right)^{\frac{1}{2}} \tag{7.45}$$

where D_f is spacing between two adjacent markers in the discrete marker scheme and $a_{\max}$ is the maximum allowable lateral acceleration. Thereafter choosing, λ_a, λ_y and λ_ε, parameters q_a, q_y, q_ε and q_i are tuned to compromise ride quality and tracking error.

Feedforward controller. If the radius of curvature of the road, ρ, is known, then the corresponding steady state steering angle δ is computed from

$$\delta = \frac{mV^2(l_2 - l_1) + 2C_s(l_1 + l_2)^2}{2\rho(C_s + F_x)(l_1 + l_2)} \tag{7.46}$$

where m is the mass of the vehicle, V is the longitudinal velocity of the vehicle, l_1 and l_2 are the distance from the mass center to the front and rear axle respectively, F_x is the traction force from the tire and C_s is the cornering stiffness. A parameter identification scheme using a least squares algorithm with a forgetting factor is proposed in Peng and Tomizuka (1990) to calculate the cornering stiffness from the measurements of lateral acceleration and yaw rate signals.

Heavy-duty Vehicles

Kanellakopoulos and Tomizuka (1997) state that automated lateral control of single-unit heavy-duty vehicles with constant longitudinal velocity can be achieved using techniques such as frequency-shaped linear quadratic (FSLQ) and gain scheduling that have been used for light duty passenger vehicles. However, in cases of varying longitudinal velocity, a different controller such as sliding mode control (SMC) has to be used. The equilibrium point in the SMC is zero lateral error and lateral velocity, and the control laws move the state to the sliding surface from where it slides to the equilibrium point. The method introduces chattering due to high gain, which is rectified using a saturation function in place of sign function in SMC. Simulations demonstrate low tracking error is obtained by using this method.

For articulated vehicles, two linear control algorithms are outlined: linear quadratic optimal control and FSLQ optimal control. The latter has a better performance at low speeds. In comparison, an input/output feedback linearization scheme combined with backstepping is shown to give better lateral tracking.

Lateral Sensors

The lateral/side sensors are needed mostly to assist an automated vehicle during lane changing (Fig. 7-12). They detect if there is any vehicle in the target/destination lane, if any vehicle is merging from the other side or if any vehicle is approaching at a threatening speed in the target lane. They should be able to reliably detect all kinds of vehicles, even motorcycles.

The demands on the lateral sensor systems for a fully automated vehicle are quite complex. Candidate technologies include ultrasound, radar and video systems (Hovanessian, 1988). Ultrasound sensors detect target position and range by bouncing acoustic energy pulses off a target and estimating time-of-flight. Radar sensors measure range and relative speed using the echo from radio frequency pulses and measuring time-of-flight as well as the Doppler effect. Video based sensors rely on efficient real-time image processing for target recognition (Graefe and Kuhnert, 1988; Smith and Brady, 1993; Trassoudaine et al., 1993). They all have individual advantages and disadvantages and combinations of sensor types may offer the only reliable way of meeting all the complex requirements on them. Different sensor technologies available today are presented and evaluated for their applicability in Table 7-1.

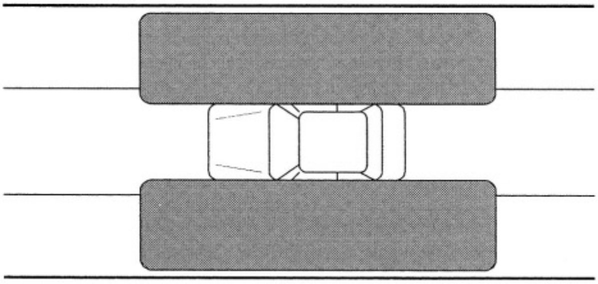
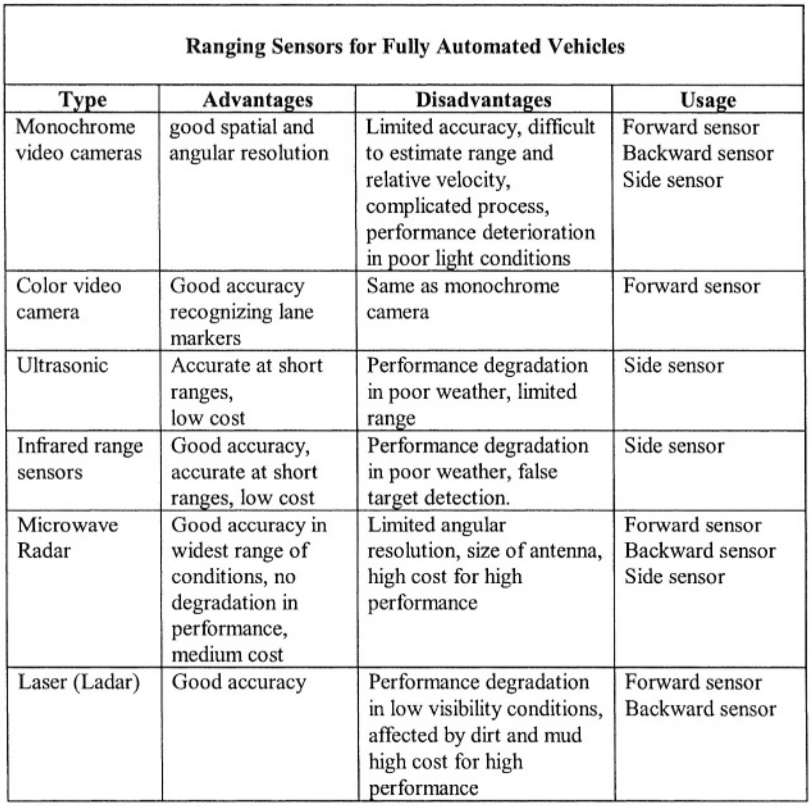


Figure 7-12. Coverage of lateral/side sensors on both sides of an automated vehicle.



Ranging Sensors for Fully Automated Vehicles

Type	Advantages	Disadvantages	Usage
Monochrome video cameras	good spatial and angular resolution	Limited accuracy, difficult to estimate range and relative velocity, complicated process, performance deterioration in poor light conditions	Forward sensor Backward sensor Side sensor
Color video camera	Good accuracy recognizing lane markers	Same as monochrome camera	Forward sensor
Ultrasonic	Accurate at short ranges, low cost	Performance degradation in poor weather, limited range	Side sensor
Infrared range sensors	Good accuracy, accurate at short ranges, low cost	Performance degradation in poor weather, false target detection.	Side sensor
Microwave Radar	Good accuracy in widest range of conditions, no degradation in performance, medium cost	Limited angular resolution, size of antenna, high cost for high performance	Forward sensor Backward sensor Side sensor
Laser (Ladar)	Good accuracy	Performance degradation in low visibility conditions, affected by dirt and mud high cost for high performance	Forward sensor Backward sensor

Table 7-1. Summary of different available sensor technologies for automated vehicles and their applicability.

Safety

A fully automated vehicle with automatic longitudinal and lateral control systems is expected to provide a high level of safety due to automated lane changing and automated longitudinal/lateral collision avoidance. Before beginning a lane change, an automated vehicle uses its lateral sensors to verify if the necessary spacing in the target/destination lane is available. The lateral sensors check for the presence and position of other vehicles in the target lane. They detect if there are vehicles changing lanes simultaneously from other lanes and if any vehicle in the target lane is approaching at a threatening speed. The backward looking sensors (if present) also assist in this matter. If any of the above conditions exist, the lane-change is aborted.

If not, then the automated vehicle executes the lane change. The flowchart in Fig. 7-13 describes the automated lane-changing procedure.

The automated vehicle guarantees a safer collision avoidance maneuver by completely removing the human driver from the diving loop. The flowchart in Fig. 7-14 shows the collision avoidance maneuver in an automated vehicle. The driver can, however, override the automatic control system and take over the control of the automated vehicle after a smooth transition procedure that guarantees the driver is not put in a situation that he/she cannot handle.

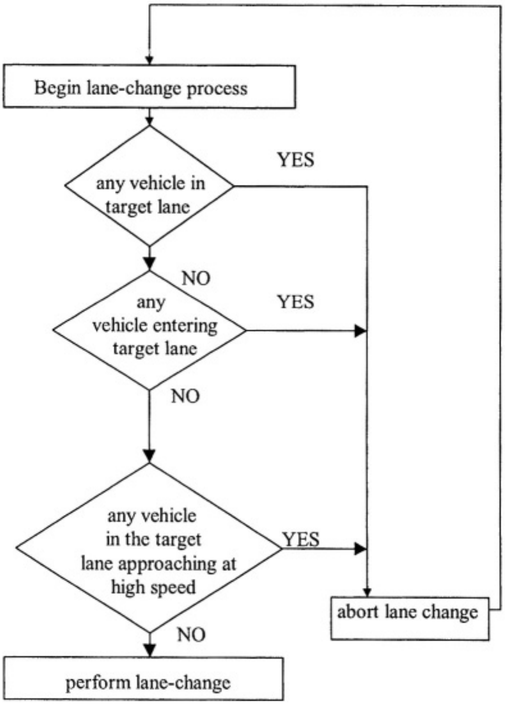

Begin lane-change process
any vehicle in target lane
YES
NO
any vehicle entering target lane
YES
NO
any vehicle in the target lane approaching at high speed
YES
NO
abort lane change
perform lane-change

Figure 7-13. Flowchart for automated lane changing.

7.6 Lane Changing

In 1994, 50.4% of accidents on interstate highways were caused by driver actions, out of which 27.6% occurred during lane changing (Interstate Hazard Analysis, 1994). Rear end, angle and sideswipe are the possible collisions that may occur during lane changing/merging (Wang and Knipling, 1993). Fully automated vehicles

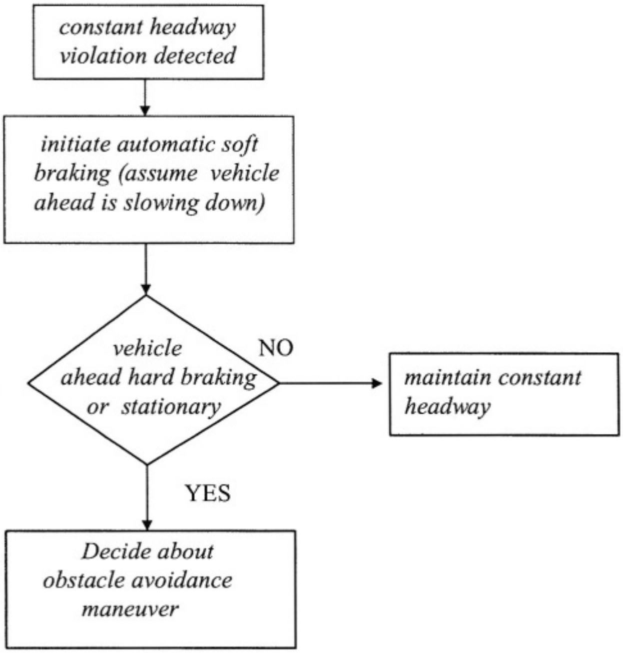

constant headway violation detected
initiate automatic soft braking (assume vehicle ahead is slowing down)
vehicle ahead hard braking or stationary
NO
maintain constant headway
YES
Decide about obstacle avoidance maneuver

Figure 7-14. Flowchart showing the emergency procedure in a fully automated vehicle.

have automated lane changing that is deemed to improve safety and reduce lane change accidents. Jula et al. (1998) conducted an in-depth analysis into the kinematics of lane changing and collisions that may occur during the maneuvers.

The study considered a general configuration of a merging/lane changing vehicle between a leading and a following vehicle in an originating lane, wanting to enter between another pair of leading/following vehicles in a destination lane. Analyses show calculations of the MSS between the merging and the four different vehicles. Simulations demonstrate lane changing scenarios considering different longitudinal acceleration profiles for the merging vehicle.

Assuming constant longitudinal velocity for all vehicles, Fig, 7-15 shows the “safe” and “unsafe” regions during lane changing. The relative spacing and relative speed of the vehicle pair are obtained using on-board sensors. This analysis can be used by the on-board controller before a lane change/merge maneuver to determine whether the vehicle is in the safe region. However, it has been argued that in real life scenarios, the merging vehicle has a changing/switching longitudinal acceleration profile that can be assumed to be as shown in Fig. 7-16. Under such circumstances,

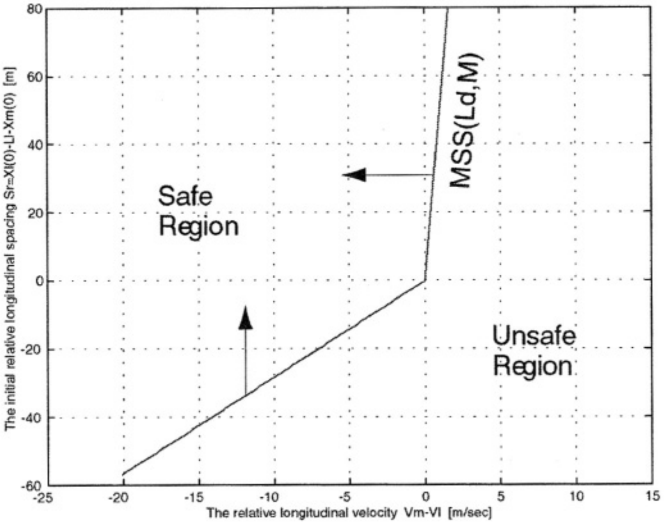


Figure 7-15. The collision region between the merging and the leading vehicle in the destination lane (constant speed).

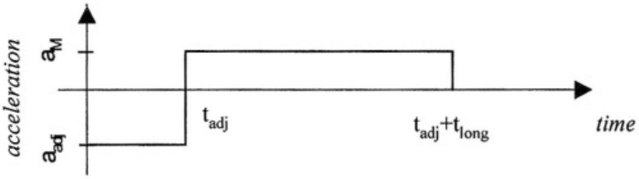


Figure 7-16. Switched longitudinal acceleration profile of the merging vehicle.

two cases are considered: first, $t_{adj} = 0$ implying that the merging vehicle only accelerates during the maneuver. The safety region expands for this scenario as shown in Fig. 7-17. Second, for $t_{adj} > 0$, simulation results in Fig. 7-18 indicate that if the merging vehicle is initially in the unsafe region, it can move into the safe region by applying adequate deceleration.

The demerit of the previous scheme is that the merging vehicle may not always be in a position to perform such decelerations because of the following vehicle in the originating lane. In such cases, the authors considered another acceleration profile as shown in Fig. 7-19, which shows that a similar transition from the unsafe into the safe region is possible with a lower deceleration (Fig. 7-20). These analyses enable

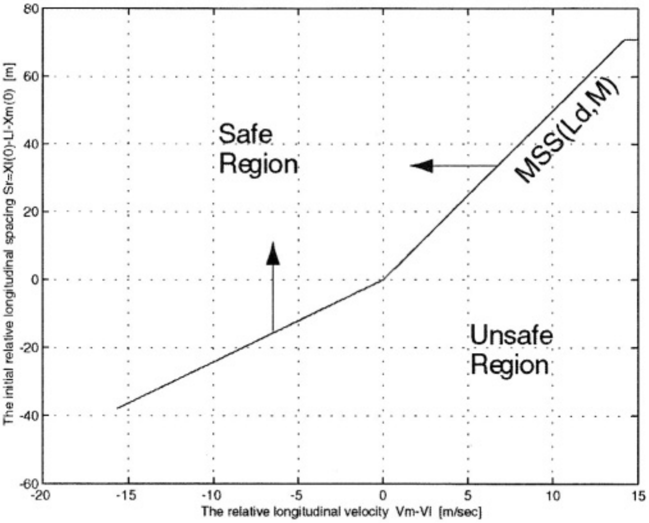


Figure 7-17. The collision region between the merging and the leading vehicle in the destination lane (switching longitudinal acceleration).

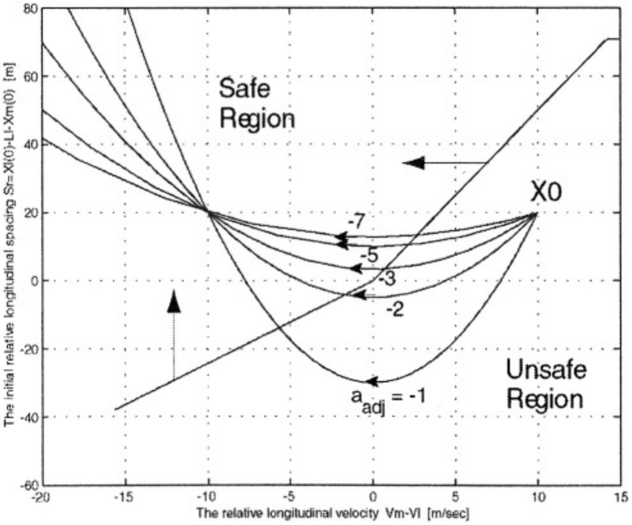


Figure 7-18. Merging vehicle applying deceleration to move from the unsafe to the safe region.

the merging automated vehicle to perform safe lane change/merge with minimum risk of collision.

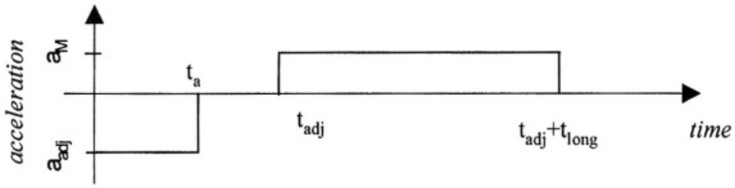


Figure 7-19. Modified switching longitudinal acceleration profile of the merging vehicle.

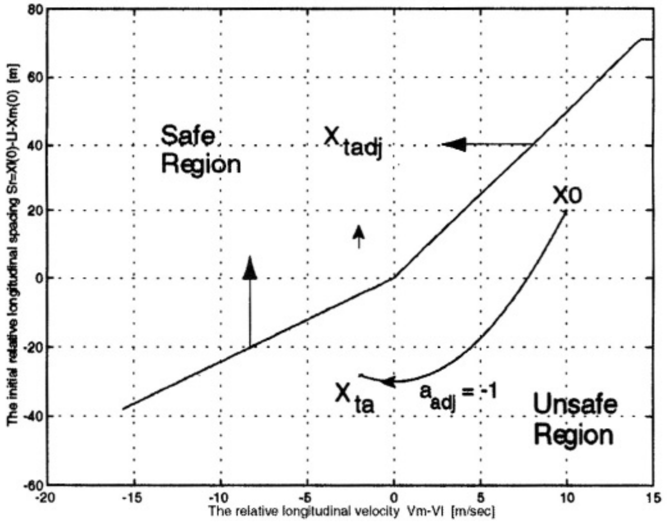


Figure 7-20. Merging vehicle applying modified deceleration to move from the unsafe to the safe region.

7.7 String Stability

In vehicle following the dynamics of each vehicle are coupled with other vehicles leading to a larger dynamical system. Even though each vehicle may have stable behavior and good performance, the behavior of the overall coupled system may not be desirable. For example, transients caused by a single vehicle changing its speed may be amplified upstream leading to what is known as "slinky-type effect"

(Sheikholeslam and Desoer, 1991) or string instability. String stability (Swaroop and Hendrick, 1996) in vehicle following implies that any nonzero position, velocity and acceleration error of an individual vehicle in a string of vehicles does not get amplified as it propagates upstream. Bose and Ioannou (1999) carried out an analysis on string stability of manual and automated traffic using the models presented in sections 7.3 & 7.4.

A system of vehicles in a single lane under moderately dense traffic conditions can be considered as a countable infinite interconnected system. Such a system shown in Fig. 7-21 is modeled as

$$v_i = G_i(s)v_{i-1} \tag{7.47}$$

where $i \in N$, the total number of vehicles considered, v_i is the velocity of the i-th vehicle and $G_i(s)$ is a proper stable transfer function that represents the input-output behavior of the i-th vehicle. The system represents traffic in a single lane without passing in which every vehicle tries to match the speed of the preceding vehicle with some precision and intervehicle spacing.

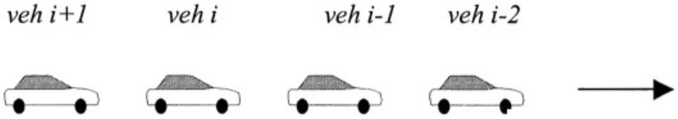


Figure 7-21. Interconnected system of vehicles following each other in a single lane.

The following errors are defined for the i-th vehicle:

$$\begin{aligned} \delta_i &= x_{i-1} - x_i - L_i - S_i && \text{(position error)} \\ v_{ri} &= v_{i-1} - v_i && \text{(velocity error)} \\ a_{ri} &= a_{i-1} - a_i && \text{(acceleration error)} \end{aligned}$$

where x_i denotes the abscissa of the rear bumper of the i-th vehicle; L_i is the length of i-th vehicle, S_i is the constant intervehicle spacing followed by the i-th vehicle, measured from the rear of the (i-1)-th vehicle to the front of the i-th vehicle; v_i denotes the velocity of the i-th vehicle and a_i denotes the acceleration of the i-th vehicle.

It can be shown that for vehicles with constant intervehicle spacing policy, the error propagation is given by

$$\frac{\delta_i}{\delta_{i-1}}=\frac{v_{ri}}{v_{ri-1}}=\frac{a_{ri}}{a_{ri-1}}=\frac{(1-G_i)}{(1-G_{i-1})}G_{i-1}=\overline{G}_i(s) \qquad (7.48)$$

and for vehicles with constant time headway policy, it is given by

$$\frac{\delta_i}{\delta_{i-1}}=\frac{1-G_i-sh_iG_i}{1-G_{i-1}-sh_{i-1}G_{i-1}}G_{i-1}=\hat{G}_i(s) \qquad (7.49)$$

$$\frac{v_{ri}}{v_{ri-1}}=\frac{a_{ri}}{a_{ri-1}}=\frac{(1-G_i)}{(1-G_{i-1})}G_{i-1}=\overline{G}_i(s) \qquad (7.50)$$

To have string stability for a class of interconnected system of vehicles, Bose and Ioannou (1999) showed that the impulse response $g_i(t)$ of the error propagation transfer function $G_i(s)$, $\overline{G}_i(s)$ or $\hat{G}_i(s)$, as the case may be, for each individual vehicle in this class must satisfy

$$\|g_i\|_1 \leq 1 \qquad \forall \quad i \in N \qquad (7.51)$$

Assuming identical input/output characteristics in a fleet of vehicles, the string stability of the manual vehicle and the automated longitudinal models presented in the earlier sections is investigated.

Pipes Model

The transfer function of the Pipes model is

$$G_p(s)=\frac{v_i}{v_{i-1}}=\frac{0.37e^{-1.5s}}{s+0.37e^{-1.5s}} \qquad (7.52)$$

The impulse response is such that $\|g_p\|_1 = 1.1$ which shows that the Pipes model does not belong to the class of systems that guarantee string stability. The frequency response shown in Fig. 7-22 has magnitude greater than unity for very small frequencies. In Fig. 7-23 it is demonstrated that a string of vehicles represented by Pipes model lacks string stability.

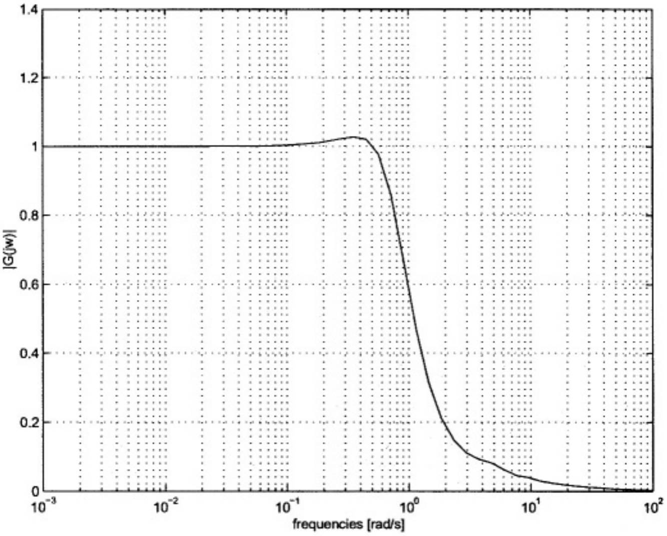


Figure 7-22. Pipes linear car following model: $|G_p(j\omega)|$ vs ω.

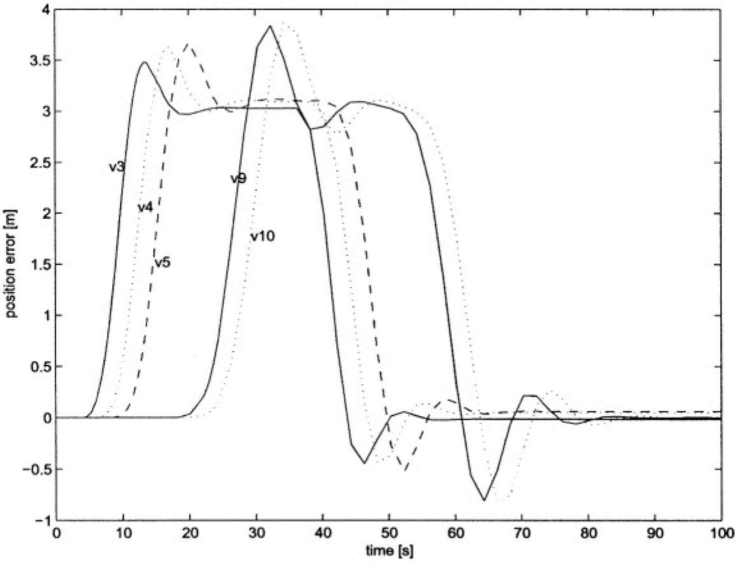


Figure 7-23. 10 vehicles in manual traffic (Pipes model) following a lead vehicle. Position error of vehicles 3 to 5 (v3-v5) and 9,10 (v9, v10) demonstrate lack of string stability.

Optimal Control Model

The optimal control model (7.28) is linearized around three different operating speeds. After linearization, a parameter A is obtained that depends on the speed at which linearization is performed and is defined as

$$A = \begin{cases} 0.005s^{-1} & 0m/s \\ 0.0125s^{-1} & 15m/s \\ 0.0175s^{-1} & 25m/s \end{cases} \tag{7.53}$$

The transfer function of the linearized model is given by

$$G_o(s) = \frac{v_i}{v_{i-1}} = \frac{(0.5s+1.64)e^{-0.09s}}{s^2 + As + (2.3696s + 1.64)e^{-0.09s}} \tag{7.54}$$

It is seen that $\|g_o\|_1 = 1$ for all values of A which shows that the optimal control model belongs to the class of systems that guarantee string stability.

Look Ahead Model

The transfer functions for the two positions of the model are

$$G_{l1}(s) = \frac{v_i}{v_{i-1}} = \frac{0.2}{s+0.2} \quad \text{(position 1)} \tag{7.55}$$

$$G_{l2}(s) = \frac{v_i}{v_{i-1}} = \frac{0.65}{s+0.65} \quad \text{(position 2)} \tag{7.56}$$

It is clearly seen that (7.55) and (7.56) satisfy (7.51) and the model belongs to the class of systems that guarantee string stability.

Automated Longitudinal Controller

The automated vehicle longitudinal controller presented in section 7.4 is designed using constant time headway policy. For automated vehicle longitudinal controllers using a constant spacing policy, it has been shown that vehicle-to-vehicle communication is necessary to guarantee string stability (Shladover, 1977; Sheikholeslam and Desoer, 1991; Swaroop et al., 1994).

Throttle controller. The transfer function of the throttle subsystem (Ioannou and Xu, 1994) is given by

$$G_{th}(s)=\frac{\delta_i}{\delta_{i-1}}=\frac{(a+bk_1)s^2+b(k_2+k_3)s+bk_4}{s^3+(a+bk_1+bk_2h)s^2+b(k_2+k_3+k_4h)s+bk_4} \qquad (7.57)$$

where k_1,k_2,k_3,k_4 are designed controller parameters; h is the time headway desired and a, b are coefficients determined by the operating point which is the speed of the vehicle ahead. Using parameter values in (7.57) that have been used on a validated nonlinear vehicle model and tested using simulations to satisfy the performance criteria given in Ioannou and Xu (1994), the following transfer function is obtained

$$G_{th}(s)=\frac{1.2s^2+0.24s+0.012}{s^3+1.4s^2+0.25s+0.012} \qquad (7.58)$$

that has the property $\|g_{th}\|_1=1$ which shows that the throttle controller belongs to the class of systems that guarantee string stability.

Brake controller. For the closed loop brake subsystem we have the following transfer function (Ioannou and Xu, 1994)

$$G_{br}(s)=\frac{v_i}{v_{i-1}}=\frac{k_5s+k_6}{s^2+(k_5+k_6h)s+k_6} \qquad (7.59)$$

where k_5,k_6 are the brake controller gains and h is the desired time headway. Using parameter values in (7.59) that satisfy the performance criteria given in Ioannou and Xu (1994), we get $\|g_{br}\|_1=1$, which shows that the brake controller belongs to the class of systems that guarantee string stability.

Therefore the longitudinal control design satisfies string stability. The automated longitudinal controller developed by Kanellakopoulos and Tomizuka (1997) for heavy-duty vehicles was designed to satisfy string stability such that the errors decrease in magnitude as they propagate upstream.

7.8 Mixed Traffic

The introduction of automated vehicles into the current manual traffic system where they will coexist with manually driven vehicles will usher the stage of mixed manual/automated traffic. In this section we investigate the possible effects of the introduction of automated vehicles on throughput, transients in vehicle following, fuel consumption and the environment.

Throughput

The use of automation guarantees that an automated vehicle with full longitudinal automation always follows a time headway that is lower than the average time headway followed by manual traffic. As a consequence, mixed automated/manual traffic promise improvement in traffic throughput. Concepts such as platooning (Ioannou, 1997; Kanaris, Ioannou and Ho, 1997) in fully automated traffic promise greater improvement in traffic throughput as shown in Shladover (1997) and Carbaugh et al. (1998). These studies show that larger number of vehicle platoons under certain assumptions about the braking capabilities of vehicles is expected to significantly improve throughput.

Throughput model (without communication). Random sequencing of automated and manual vehicles in mixed traffic operations produce different combinations of pair of vehicles. We assume that the automated vehicles do not have communication capability. In other words, an automated vehicle following a vehicle is not able to determine the type of vehicle ahead, i.e. whether it is following a manual or an automated vehicle. Therefore, it always maintains the same time headway in traffic at a given speed. Likewise, we assume that a manual vehicle follows the same time headway at a given speed regardless of the type of the vehicle ahead.

We adopt the model used in Milestone 2 Report (Appendix I, 1996). We assume an identical speed for the automated and the manual vehicles and that all manual vehicles follow an identical average time headway at a given speed. Let the market penetration of the automated vehicles be *a*. The probability that a vehicle is automated or manual is given by

P(automated vehicle) = *a*
P(manual vehicle) = *1-a*

This means that when *a* =0.1, 10% of the vehicles in mixed traffic are automated.

Let the following represent the time headways for manual and automated vehicles:

H_a: time headway of automated vehicles
H_m: time headway of manual vehicles

The average time headway of mixed traffic, assuming the total number of vehicles on the highway is very large compared to the number under analysis, is given by

$$\textit{av\_head}=[\,H_a\,\mathrm{P(automated)} + H_m\,\mathrm{P(manual)}] \tag{7.60}$$

and the throughput is calculated as

$$\textit{throughput} = 3600/\textit{av\_head} \tag{7.61}$$

Throughput model (with communication). An automated vehicle with communication capability determines if the leader is automated or manual by attempting to communicate to the vehicle ahead. If the leader does not acknowledge, then it is assumed to be a manually driven vehicle. Accordingly, the following vehicle selects and maintains an appropriate time headway.

We define the following probabilities

P(A,M) = probability that an automated vehicle is followed by a manual vehicle
P(A,A) = probability that an automated vehicle is followed by an automated vehicle
P(M,A) = probability that a manual vehicle is followed by an automated vehicle
P(M,M) = probability that a manual vehicle is followed by a manual vehicle

So we have from previous notation

$$P(A,M) = a \times (1-a)$$
$$P(A,A) = a \times a$$
$$P(M,A) = (1-a) \times a$$
$$P(M,M) = (1-a) \times (1-a)$$

The throughput in Milestone 2 Report (Appendix I, 1996) is formulated based on intervehicle data for the four possible outcomes. We carry out the analysis based on time headway data and follow the notation given below

H(A,M): time headway of a manual vehicle following an automated vehicle
H(A,A): time headway of an automated vehicle following an automated vehicle
H(M,A): time headway of an automated vehicle following a manual vehicle
H(M,M): time headway of a manual vehicle following a manual vehicle

The average time headway of mixed traffic is given by

$$av\_head=[P(A,A)\,H(A,A)+P(A,M)\,H(A,M)+P(M,A)\,H(M,A)+P(M,M)\,H(M,M)] \tag{7.62}$$

and the throughput is calculated using (7.61).

Time headway of a manual vehicle H_m. The time headway of manual vehicles follows a shifted lognormal distribution (Fig. 7-24 (Cohen, 1991)). It can be considered to have a mean value of 1.8s, which is about 2000veh/hr/lane that has been termed as 'national average' in the Highway Capacity Manual (TRB, 1985). There is considerable influence of age, gender and experience on the manual traffic time headway (Fancher et al., 1995). A part of the highway might have time headway other than 1.8s due to any of the above factors. However, for simplicity we use the average value of 1.8s for our analysis. The same value is used for H(A,M) and H(M,M) in (7.62).

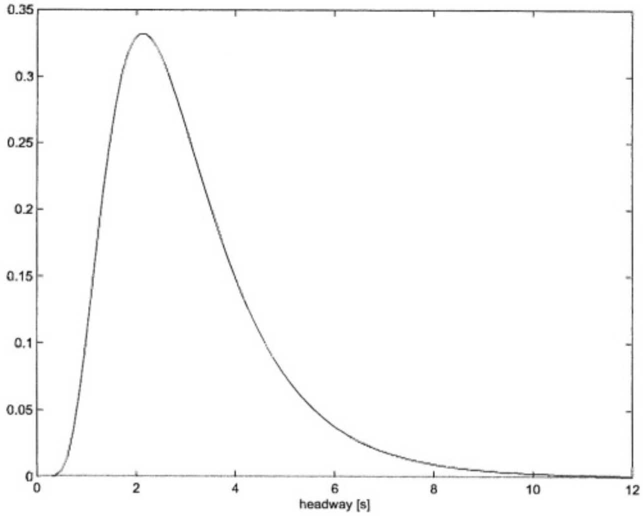


Figure 7-24. Empirical time headway distribution for manual traffic.

Time headway ofan automated vehicle $H_{a.}$ The MSS is calculated based on a worst case stopping scenario with vehicle characteristics obtained from actual vehicles.

We use the SPACING software tool (Kanaris, Grammagnat and Ioannou, 1997) to calculate the time headway of an automated vehicle. Assuming the worst case scenario that the vehicle ahead is a manual one, we consider that it performs hard braking with a deceleration of 0.90g. On the other hand, automated vehicles are assumed to meet a minimum performance criterion and have a lower bound for maximum deceleration of0.75g. The sensor detection delay ofan automated vehicle time headway system and the brake actuation delay are taken from a study by Ioannou et al. (1994). The parameter values are listed in Table 7-2.

Parameters	Leading vehicle	Following vehicle
Speed	55mph	55mph
Initial Accl.	-	0.21g
Normal jerk	-	40ms^{-3}
Emergency jerk	70ms^{-3}	60ms^{-3}
Detection delay	-	0.3s
Emergency delay	-	0.4s
Normal Decl.	-	0.098g
Friction Coeff.	1	1
Emergency Decl.	0.90g	0.75g

Table 7-2. Parameter values used for calculating headway.

The intervehicle spacing that measures the distance from the rear of the leading vehicle to the front of the following vehicle obtained using the SPACING software is 13.44m for speed of 55mph. The manual traffic time headway (Fig. 7-24) measures the time it takes a manual vehicle to cover the distance between the same points of successive vehicles, i.e. it includes the vehicle length. Assuming an average vehicle length of 5m, we get an intervehicle spacing of 18.44m that translates into an automated vehicle time headway of 0.75s at 55mph for the automated vehicles. This value is also used for H(M,A) in (7.60). However, when an automated vehicle is following another automated vehicle and both are exchanging information, then the time headway H(A,A) is smaller than H(M,A). When the lead vehicle performs hard braking, it immediately notifies the following vehicle about its rate. After a propagation and actuation delay, the following vehicle performs the exact maneuver at the same rate. Considering this braking scenario, we obtain an intervehicle spacing of 9.84m using SPACING. Adding the average vehicle length of 5m this translates into a time headway of 0.61s. Using (7.60), (7.61) and (7.62), Fig. 7-25 shows how increased penetration of automated vehicles improves traffic throughput. We also observe that increased level of automation in an automated vehicle such as vehicle-to-vehicle communication further improves throughput.

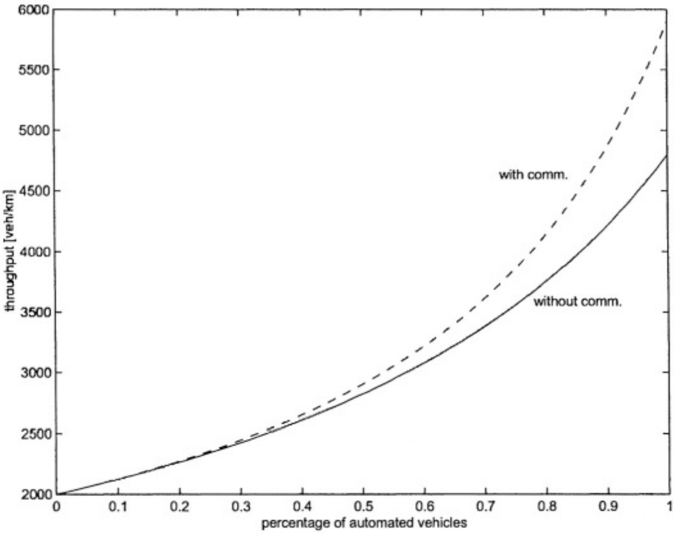


Figure 7-25. Increase in throughput with increase in market penetration of automated vehicles with and without communication capability.

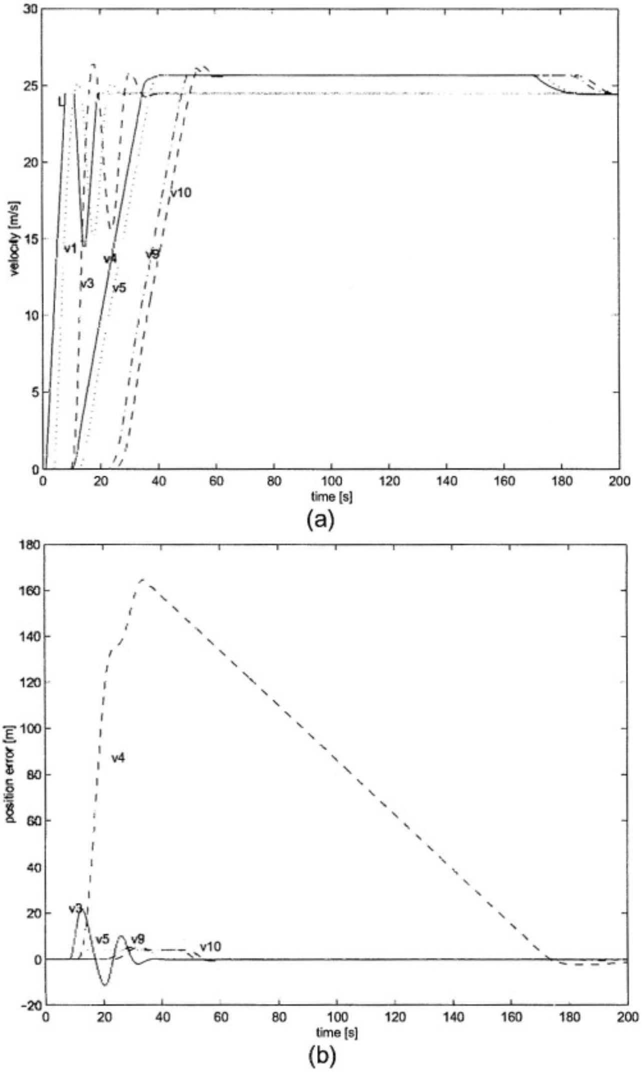
Traffic Flow Characteristics

The presence of automated vehicles improves traffic flow characteristics in manual traffic. Bose and Ioannou (1999) show that automated vehicles do not contribute to the “slinky-type effect” (Sheikholeslam and Desoer, 1991) observed in today’s manual traffic. Simulations demonstrate that automated vehicles smooth traffic flow by filtering the transients caused by rapidly accelerating manual vehicles. Consider a string of 10 vehicles in mixed manual/automated traffic. We use the Pipes model for manual vehicles as it models the slinky-type effects we observe in today’s manual traffic. The ICC model in Ioannou and Xu (1994) is used to simulate automated vehicles. Assume the 4th vehicle to be automated that corresponds to 10% mixing of automated with manual vehicles. The lead vehicle accelerates at 0.35g from 0m/s to 24.5m/s, maintains a constant speed at 24.5m/s, thereafter decelerates to 14.5m/s at 0.3g and finally accelerates to 24.5m/s at 0.25g. The acceleration and deceleration values used are typical for many passenger cars (Consumer Reports, 1998). The velocity responses in Fig. 7-26(a) show that the automated vehicle v4 filters the response of the rapidly accelerating vehicle v3 in an effort to maintain smooth driving. As a result the responses of vehicles v5, v9 and v10 are less oscillatory than that of v1 and v3. However, this is done at the expense of large position error in v4 (Fig. 7-26(b)).

Travel time? If the presence of an automated vehicle in mixed traffic improves traffic flow stability, the question remains if it changes the total travel time. The automated vehicle has limited acceleration and is not able to keep up with fast accelerating manual vehicles. We perform the same rapid acceleration maneuver by the lead vehicle and compare the difference in the distance as a percentage of the total distance covered by the 10th vehicle in manual traffic and mixed traffic. If there is no difference in the distance covered by the 10th vehicle in manual and mixed traffic, then there is no difference in the distances covered by the other vehicles ahead and therefore in the total travel time. This is because each vehicle follows the vehicle ahead using a constant time headway policy. As shown in Fig. 7-27, though there is a transient percentage difference in the distance covered by the 10th vehicle when the automated vehicle cannot keep up with the accelerating leader, it finally goes to zero. This happens when the lead vehicle attains a constant speed after rapid acceleration and theautomated vehicle catches up with it. Therefore, the presence of an automated vehicle is not expected to change the total travel time for the following vehicles.

Fuel Consumption and Pollution

Barth (1997) lists vehicle parameters such as second-by-second velocity, acceleration and grade that determine the emission levels and fuel consumption. Bose and Ioannou (2001) used the above stated parameters in simulations and experiments using actual vehicles to examine the environmental effect of automated



(a)

(b)

Figure 7-26. 10 vehicles in mixed manual (Pipes model)/automated traffic following a rapidly accelerating lead vehicle. The 4th vehicle (v4) is automated. (a) Velocity response of leader (L), 1st vehicle (v1) and vehicles 3 to 5 (v3-v5) and 9,10 (v9, v10); (b) position error of vehicles 3 to 5 (v3-v5) and 9,10 (v9, v10).

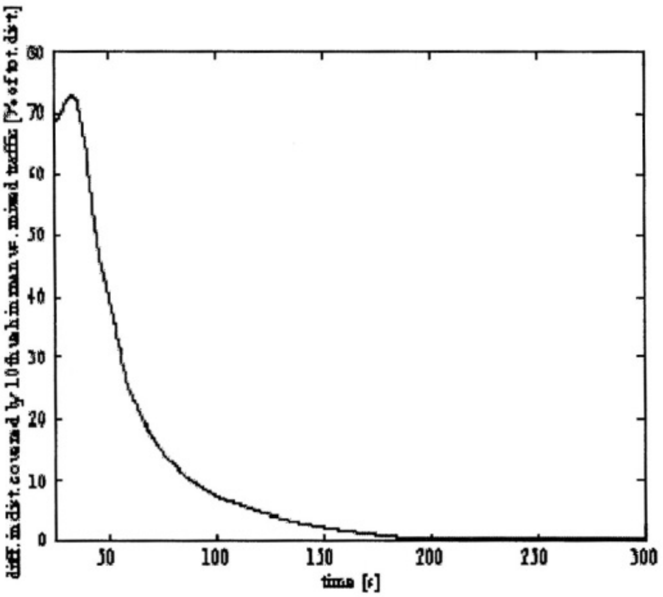


Figure 7-27. 10 vehicles follow a rapidly accelerating lead vehicle. Difference in the distance as a percentage of the total distance covered by the 10th vehicle in manual and mixed traffic. Mixed traffic is when the 4th vehicle is a semi-automated vehicle and the rest are manual.

vehicles among manual ones. The quantities measured are the tailpipe emissions of unburnt hydrocarbons (HC), carbon monoxide (CO), CO_2, oxides of nitrogen (NO, NO_2, denoted by NO_x in this Chapter) and fuel consumption. The Comprehensive Modal Emissions Model (CMEM) version 1.00 developed at UC Riverside is used to analyze the vehicle data and calculate the air pollution and fuel consumption [3]. It is a high fidelity, recently developed model that is more sensitive to transients than previous TRAF models (Barth, 1998). The model calculates vehicle emissions and fuel consumption as a function of the vehicle operating mode, i.e. idle, steady state cruise, various levels of acceleration/deceleration, among others.

In simulations, smooth and rapid acceleration scenarios were evaluated for a string of 10 vehicles following a lead vehicle in a single lane without passing in manual and mixed traffic. It is seen that the accurate speed and position tracking and the smoothing of traffic flow by the automated vehicle translates into lower air pollution and fuel savings that are significant during rapid acceleration transients, as shown in Table 7-3.

Furthermore, three actual vehicles were used in experiments since it was not possible to use 10 vehicles. Therefore to see how the simulation results compare with the experimental results, we reran the simulations using only 2 vehicles following a lead vehicle in manual traffic and mixed traffic. The lead vehicle speed profiles obtained during the experiments were used. The speed responses of the models were collected and analyzed using CMEM. The environmental benefits measured due to the presence of the automated vehicle during experiments and simulations are presented in Table 7-4. The simulation results are conservative compared to environmental benefits in actual driving, a consequence of the fact that the Pipes model gives a smooth approximation of actual manual driving.

	Smooth Acceleration	Rapid Acceleration
CO emission	18.4%	60.6%
CO_2 emission	8.1%	19.8%
NO_x emission	13.1%	1.5%
HC emission	15.5%	55.4%
Fuel consumption	8.5%	28.5%

Table 7-3. Percentage savings in pollution emission and fuel consumption for mixed traffic over manual traffic (simulation results)..

	Smooth Acceleration		Rapid Acceleration	
	Experiment	Simulation	Experiment	Simulation
CO emission	1.2%	0.8%	19.2%	12.3%
CO_2 emission	0.4%	0.2%	3.4%	3.3%
NO_x emission	1.6%	1.3%	25.7%	19.2%
HC emission	0.8%	0.4%	9.8%	6.6%
Fuel consumption	0.4%	0.2%	3.6%	3.4%

Table 7-4. Percentage savings in pollution emission and fuel consumption for mixed traffic over manual traffic.

7.9 Conclusion

In this chapter we outlined a design for a longitudinal and a lateral controller and discussed communication technologies and sensor requirements for automated vehicles. We highlighted benefits in safety that may be expected due to use of automation. We presented how the presence of lateral sensors and automated lane changing in fully automated vehicles has the potential to improve safety on freeways. We observed that automated vehicles among manual ones in mixed traffic behave like a filter and smooth the traffic flow characteristics when following fast accelerating manual vehicles, without changing the total travel time. This smoothing out of the traffic flow is expected to have a beneficial environmental impact and decrease fuel consumption.

References

Barth, M.J. (1997) Integrating a modal emissions model into various transportation modeling frameworks. *ASCE Conference Proceedings.*

Barth, M.J. (1998) *CMEM User's Manual,* UC Riverside.

Bastian, A. et al. (1998) Autonomous cruise control: A first step towards automated driving. *SAE Technical Paper Series* 981942.

Bekey, G.A., Burnham, G.O. and Seo, J. (1977) Control theoretic models of human drivers in car following. *Human Factors* **19***,* 399-413.

Bose, A. and Ioannou, P.A. (1998) Issues and analysis of mixed semi-automated/manual traffic. *SAE Technical Paper Series* 981943.

Bose, A. and Ioannou, P.A. (1999) Analysis of traffic flow with mixed manual and semi-automated vehicles. *Proc. American Control Conference,* 2173-2177.

Bose, A. and Ioannou, P.A. (2001) Evaluation of the environmental effects of Intelligent Cruise Control vehicles. *Journal of the Transportation Research Board* **1774**, 90-97.

Burnham, G.O., Seo, J. and Bekey, G.A. (1974) Identification of human driver models in car following. *IEEE Transactions on Automatic Control* **19**, 911-915.

Carbaugh, J., Godbole, D.N. and Sengupta, R. (1998) Safety and capacity analysis of automated and manual highway systems. *Transportation Research Part C (Emerging Technologies)* **6C**:1-2, 69-99.

Chandler, P.E., Herman, R. and Montroll, E.W. (1958) Traffic dynamics: studies in car following. *Operations Research* **6***,* 165-184.

Cho, D. and Hedrick, J.K. (1989) Automotive powertrain modeling for control. *Trans ASME, J. of Dyn. Sys. Meas. & Contr.* **111**:4, 568-576.

Cohen, S. (1991) Flow variables. *Concise Encyclopedia of Traffic & Transportation Systems,* (M. Papageorgiou, ed.) Pergamon Press.

Cointot, B. et al. (1993) Detection of driver's low vigilance periods on motorway. *26th ISATA Intern. Symp. Automotive Technology and Automation,* 347.

Consumer reports online, (1998).

Dingus, T., Hardee, L. and Wierwille, W. (1987) Development of models for on-board detection of driver impairment. *Accid. Anal. And Prev.* **19**:4.

Fancher, P. et al (1995) Fostering development, evaluation and deployment of Forward Crash Avoidance System (FOCAS). *NHTSA,* Report No. DOT HS 808 437.

Graefe, V. and Kuhnert, K.-D. (1988) Towards a vision-based robot with a driver's license. *Proceedings, IEEE International Workshop on Intelligent Robots and Systems,* 627-632.

Hall, R.W. (1997) System configurations: Evolutionary deployment considerations. *Automated Highway Systems* (P. A. Ioannou, ed.), pp. 49-71, Plenum Press, New York.

Hedrick, J.K., McMahon, D., Narendran, V. and Swaroop, D. (1991) Longitudinal vehicle controller design for IVHS systems. *Proceedings of American Control Conference,* 3107-3112.

Hessburg, T., Peng, H, Tomizuka, M. and Zhang, W.-B. (1991) An experimental study on lateral control of a vehicle. *PATH Research Report* UCB-IT-PRR-91-17, UC Berekley.

Hovanessian, S.A. (1988) Introduction to Sensor Systems. Artech House.

Interstate hazard analysis, (1994).

Jula, H., Kosmatopoulos, E. and Ioannou, P.A. (1998) Collision avoidance analysis for lane changing and merging. *Submitted to IEEE Trans, on Vehicular Technology.*

Ioannou, P.A., Ahmed-Zaid, F. and Wuh, D. (1992) A time headway autonomous Intelligent Cruise Controller: Design and simulation. *Technical Report,* USC-SCT 92-11-01, Los Angeles.

Ioannou, P.A. and Chien, C.C. (1993) Autonomous Intelligent Cruise Control. *IEEE Trans. on Vehicular Tech.* **42**:4, 657-672.

Ioannou, P.A. and Xu, T. (1994) Throttle and Brake Control. *IVHS Journal* **1**:4, 345-377.

Ioannou, P.A. et al (1994) Activity D: Lateral and longitudinal control analysis. *Final Report.*

Ioannou, P.A. (1997) Control and sensor requirements and issues in AHS. *Automated Highway Systems* (P. A. Ioannou, ed.), pp. 195-212, Plenum Press, New York.

Jurie, F. et al. (1993) High speed vehicle guidance based on vision. *Intelligent Autonomous Vehicles IFAC Workshop,* 203-208.

Kaltwasser, J. and Kassubek, J. (1994) A new cooperative optimized access for inter-vehicle communication. *Vehicle Navigation & Information Systems Conference Proceedings,* 144-148.

Kanaris, A., Grammagnat, A. and Ioannou, P.A. (1997) Inter vehicle spacing – User's manual. USC Center for Advanced Transportation Technologies, Los Angeles.

Kanaris, A., Ioannou, P.A. and Ho, F.-S., (1997) Spacing and capacity evaluations for different AHS concepts. *Proc. of American Control Conference,* 2036-2040.

Kanellakopoulos, I. and Tomizuka, I.(1997) Commercial trucks and buses in Automated Highway Systems. *Automated Highway Systems* (P. A. Ioannou, ed.), pp. 213-246, Plenum Press, New York.

Kogure, G., Katahara, S. and Aoki, M. (1993) Iris motion and blink detection using video image sequence. *26th ISATA Intern. Symp. Automotive Technology and Automation,* 375.

Milestone 2 Report (1996) Cooperative Vehicle Concept Description. *Appendix B,* National Automated Highway System Consortium (NAHSC).

Milestone 2 Report (1996) Mixed traffic throughput analysis. *Appendix I,* National Automated Highway System Consortium (NAHSC).

Milestone 2 Report (1996) Hard-braking safety analysis method and detailed results. *Appendix J,* National Automated Highway System Consortium (NAHSC).

Mitzui, K. et al (1994) Vehicle-to-vehicle 2-way communication & ranging system using spread spectrum technique. *Vehicle Navigation & Information Systems Conference Proceedings,* 153-158.

Nathoo, N.S. and Healey, A.J. (1978) Coupled vertical-lateral dynamics of a pneumatic tired vehicle: part I- A mathematical model. *ASME J. Dyn. Syst. Meas. Control* **100**, 311-318.

Pipes, L.A. (1953) An operational analysis of traffic dynamics. *J. of Applied Physics* **24**, 271-281.

Peng, H. and Tomizuka, M. (1990) Vehicle lateral control for highway automation. *Proc. American Control Conference,* 788-793.

Raza, H., Xu, Z., Yang, B. and Ioannou, P.A. (1997) Modeling and control design for a computer-controlled brake system. *IEEE Trans. on Control Systems Technol.* **5**:3, 279-296.

Scolfield, J. et al. (1993) Progress towards impaired driver attentiveness detection system. Prometheus-Future Systems, auto tech.

Sheikholeslam, S. and Desoer, C.A. (1991) Longitudinal control of a platoon of vehicles with no communication of lead vehicle information. *Proc. American Control Conference,* 3102-3106.

Shladover, S.E. (1977) Longitudinal control of automated guideway transit vehicles within platoons. *ASME J. Dyn. Syst. Meas. Control* **100**, 302-310.

Shladover, S.E. (1997) Reasons for operating AHS vehicles in platoons. *Automated Highway Systems* (P. A. Ioannou, ed.), pp. 11-27, Plenum Press, New York.

Skipper, J. and Wierwille, W. (1986) Drowsy driver detection using discriminant analysis. *Human Factors* **28**:5, 527.

Smith, S.M. and Brady, J.M. (1993) A scene segmenter: visual tracking of moving vehicles. *Intelligent Autonomous Vehicles, IFAC Workshop,* 117-124.

Stevens, W.B. (1997) Evolution to an Automated Highway System. *Automated Highway Systems* (P. A. Ioannou, ed.), pp. 109-124, Plenum Press, New York.

Swaroop, D. and Hedrick, J.K. (1996) String stability of interconnected systems. *IEEE Trans. on Automatic Control* **41**:3, 349-357.

Swaroop, D. Hedrick, J.K., Chien, C.C. and Ioannou, P.A. (1994) A comparison of spacing and headway control laws for automatically controlled vehicles, *J. of Vehicle System Dynamics* **23**, 597-625.

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

Trassoudaine, L. et al, (1993) Visual tracking by a multisensorial approach. *Intelligent Autonomous Vehicles, IFAC Workshop,* 111-116.

Tyler, J.S. (1964) The characteristics of model following systems as synthesized by optimal control. *IEEE Transactions on Automatic Control* **9**, 485-498.

Walker, R.J. and Harris, C.J. (1993) A Multi-Sensor Fusion System for a Laboratory Based Autonomous Vehicle. *Intelligent Autonomous Vehicles IFAC Workshop,* 105-110.

Wang, J.-S. and Knipling, R.R. (1993) Lane change/merge: Problem size assessment and statistical description. *USDOT Techinical Report (Draft),* Information Management Consultants Inc., McLean, Virginia.

Ward, J.D. (1997) Step by step to an Automated Highway System-and beyond. *Automated Highway Systems* (P. A. Ioannou, ed.), pp. 73-91, Plenum Press, New York.

Zhang, S. and Thomas, B.T. (1993) Knowledge-based vehicle navigation in complex road networks. *Intelligent Autonomous Vehicles IFAC Workshop,* 209-214.

This page intentionally left blank

8 TRAFFIC CONTROL

Markos Papageorgiou

8.1 Introduction

Traffic Congestion

Transportation has always been a crucial aspect of human civilization, but it is only in the second half of this century that the phenomenon of traffic congestion has become predominant due to the rapid increase in the number of vehicles and in the transportation demand in virtually all transportation modes. Traffic congestion appears when too many vehicles attempt to use a common transportation infrastructure with limited capacity. In the best case, traffic congestion leads to queueing phenomena (and corresponding delays) while the infrastructure capacity ("the server") is fully utilized. In the worst (and far more typical) case, traffic congestion leads to a degraded use of the available infrastructure (reduced throughput that may even lead to a fatal gridlock) with excess delays, reduced safety, and, recently, increased environmental pollution.

The Need for Traffic Control

The emergence of traffic (i.e. many interacting vehicles using a common infrastructure) and subsequently traffic congestion (whereby demand exceeds the infrastructure capacity) have opened new innovation needs in the transportation area. The energy crisis in the 1970's, the increased importance of environmental concerns, and the limited economic and physical resources are among the most important reasons why a brute-force approach (i.e., the continuous expansion of the available transportation infrastructure) cannot continue to be the only answer to the ever increasing transportation and mobility needs of modern societies. The efficient, safe, and less polluting transportation of persons and goods calls for an optimal utilization of the available infrastructure via suitable application of a variety of traffic control measures. This trend is enabled by the rapid developments in the areas of communications and computing, but it is quite evident that the efficiency of traffic control directly depends on the efficiency and relevance of the employed control methodologies. This chapter will provide an overview of advanced traffic control strategies for three particular areas: Urban road networks, freeway networks, and route guidance and information systems.

The Control Loop

Figure 8.1 illustrates the basic elements of a control loop. The traffic flow behaviour in the (road or freeway or mixed) network depends on some external quantities that are classified into two groups:

- *Control inputs* that are directly related to corresponding control devices such as traffic lights, variable message signs, etc; the control inputs may be selected from an admissible control region subject to technical, physical, and operational constraints.
- *Disturbances*, whose values cannot be manipulated, but may possibly be measurable (e.g. demand) or detectable (e.g, incident) or predictable over a future time horizon.

The network's output or performance is measured via suitable indices, such as the total time spent by all vehicles in the network over a time horizon. The task of the *surveillance* is to enhance and to extend the information provided by measurement devices (e.g. loop detectors) as required by the subsequent control strategy and the human operators. The kernel of the control loop is the *control strategy*, whose task is to specify in real time the control inputs, based on available measurements/estimations/predictions, so as to achieve the pre-specified goals (e.g. minimization of total time spent) despite the influence of various disturbances. If this task is undertaken by a human operator, we have a manual control system. In

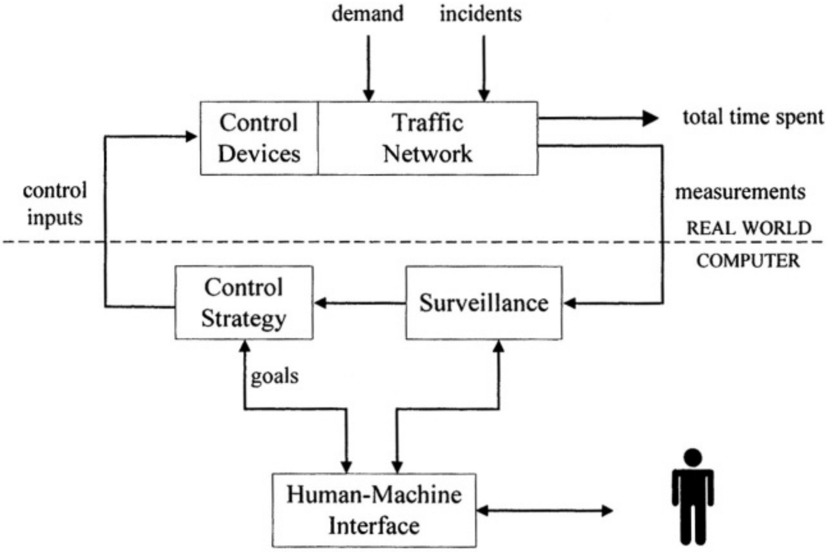


Figure 8-1. The control loop.

an automatic control system, this task is undertaken by an algorithm (the control strategy). The relevance and efficiency of the control strategy largely determines the efficiency of the overall control system. Therefore, whenever possible, control strategies should be designed with care, via application of powerful and systematic methods of optimization and automatic control, rather than via questionable heuristics (Papageorgiou, 1998). Traffic control strategies for urban road and freeway networks is the main focus of this chapter.

Discrete-Time Representation

For the needs of this chapter we will use a discrete-time representation of traffic variables with discrete time index k = 0, 1, 2, ... and time interval T. A *traffic volume* or *flow* q(k) (in veh/h) is defined as the number of vehicles crossing a corresponding location during the time period [kT, (k+1)T], divided by T. *Traffic density* $\rho(k)$ (in veh/km) is the number of vehicles included in a road segment of length Δ at time kT, divided by Δ . *Mean speed* v(k) (in km/h) is the average speed at time kT of all vehicles included in a road segment.

A Basic Property

We consider a traffic network (Figure 8-2) that receives demands $d_i(k)$ (in veh/h) at its origins i = 1, 2, ... and we define the total demand $d(k) = d_1(k) + d_2(k) + \dots$ We assume that d(k), k = 0, ..., K, is independent of any control measures taken in the network. We define exit flows $s_i(k)$ at the network destinations i = 1, 2, ..., and the total exit flow $s(k) = s_1(k) + s_2(k) + \dots$ We wish to apply control measures so as to minimize the total time spent T_s in the network over a time horizon K, i.e.

$$T_s = T\sum_{k=0}^{K} N(k) \tag{8.1}$$

where N(k) is the total number of vehicles in the network at time k. Due to conservation of vehicles

$$N(k) = N(k-1) + T[d(k) - s(k)] \tag{8.2}$$

hence

$$N(k) = N(0) + T\sum_{\kappa=0}^{k-1}[d(\kappa) - s(\kappa)]. \tag{8.3}$$

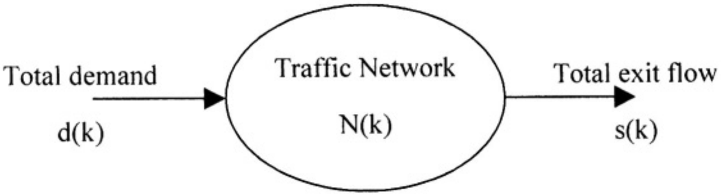


Figure 8-2. A traffic network.

Substituting (8.3) in (8.1) we obtain

$$T_s = T\sum_{k=0}^{K}[N(0) + T\sum_{\kappa=0}^{k-1} d(\kappa) - T\sum_{\kappa=0}^{k-1} s(\kappa)]. \qquad (8.4)$$

The first two terms in the outer sum of (8.4) are independent of the control measures taken in the network, hence minimization of T_s is equivalent to maximization of the following quantity

$$S = T^2\sum_{k=0}^{K}\sum_{\kappa=0}^{k-1} s(\kappa) = T^2\sum_{k=0}^{K-1}(K-k)s(k). \qquad (8.5)$$

Thus, minimization of the total time spent in a traffic network is equivalent to maximization of the time-weighted exit flows. In other words, the earlier the vehicles are able to exit the network (by appropriate use of the available control measures) the less time they will have spent in the network.

8.2 Road Traffic Control

Basic Notions

Traffic lights at intersections is the major control measure in road networks. Traffic lights were originally installed in order to guarantee the safe crossing of antagonistic streams of vehicles and pedestrians. With steadily increasing traffic demands, it was soon realized that once traffic lights exist, they may lead (under equally safe traffic conditions) to more or less efficient network operations, hence there must exist an optimal control strategy leading to minimization of the total time spent by all vehicles in the network.

Although the corresponding optimal control problem may be readily formulated for any road network, its real-time solution and realization in a control loop like the one of Figure 8-1 faces a number of apparently insurmountable difficulties:

- The red-green switchings of traffic lights call for the introduction of binary variables, which renders the optimization problem combinatorial.
- The size of the problem for a whole network is very large.
- Many unpredictable and hardly measurable disturbances (incidents, illegal parking, pedestrian crossings, intersection blocking, etc.) may perturb the traffic flow.
- Measurements of traffic conditions are mostly local (via loop detectors) and highly noisy due to various physical effects.
- There are tight real-time constraints, e.g. decision making within 2s for advanced control systems.

The combination of these difficulties renders the solution of a detailed optimal control problem infeasible for more than one intersection. Therefore, proposed control strategies for road traffic control introduce a number of simplifications of different kinds or address only a part of the related traffic control problems.

An intersection consists of a number of approaches and the crossing area. An approach may have one or more lanes but has a unique, independent queue. Approaches are used by corresponding traffic streams (veh/h). A *saturation flow* s is the average flow crossing the stop line of an approach when the corresponding stream has right of way (r.o.w.) and the upstream demand (or the waiting queue) is sufficiently large. Two *compatible* streams can safely cross the intersection simultaneously, else they are called *antagonistic*. A *signal cycle* is one repetition of the basic series of signal combinations at an intersection; its duration is called *cycle time* c. A *stage* (or *phase*) is a part of the signal cycle, during which one set of streams has r.o.w. (Figure 8-3). Constant *lost times* of a few seconds are necessary between stages to avoid interference between antagonistic streams of consecutive stages (Figure 8-4).

There are four possibilities for influencing traffic conditions via traffic lights operation:

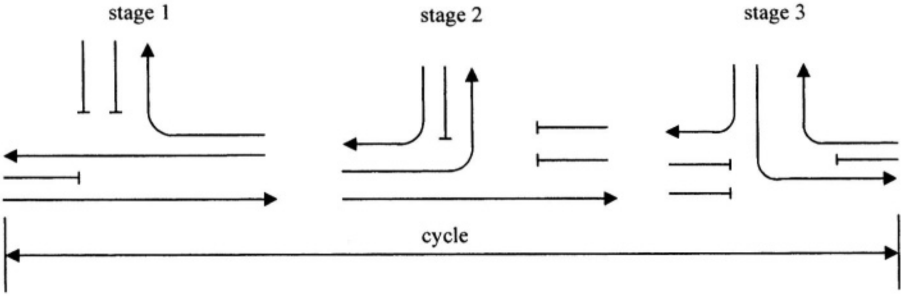


Figure 8-3. Example of signal cycle.

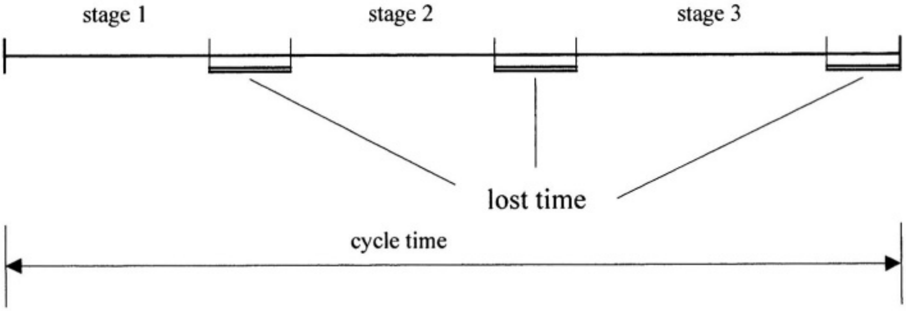


Figure 8-4. Cycle time and lost times.

- *Stage specification*: For complex intersections involving a large number of streams, the specification of the optimal number and constitution of stages is a non-trivial task that can have a major impact on intersection capacity and efficiency.
- *Split*: This is the relative green duration of each stage (as a portion of the cycle time) that should be optimized according to the demand of the involved streams.
- *Cycle time*: Longer cycle times typically increase the intersection capacity because the proportion of the constant lost times becomes accordingly smaller; on the other hand, longer cycle times may increase vehicle delays in undersaturated intersections due to longer waiting times during the red phase.
- *Offset*: This is the time difference between cycles for successive intersections that may give rise to a "green wave" along an arterial; clearly, the specification of offset should ideally take into account the possible existence of vehicle queues.

Control strategies employed for road traffic control may be classified according to the following characteristics:

- *Fixed-time strategies* are derived off-line by use of appropriate optimization codes based on historical constant demands for each stream for a given time-of-day; *traffic-responsive strategies* make use of real-time measurements (typically one or two loops per link) to calculate in real time the suitable signal settings.
- *Isolated strategies* are applicable to single intersections while *coordinated strategies* consider an urban zone or even a whole network comprising many intersections.
- Some strategies are only applicable to *undersaturated* traffic conditions, whereby vehicle queues are only created during the red phases and are dissolved during the green phases; other strategies are adapted also for *oversaturated* conditions with partially increasing queues that in some cases may even reach the upstream intersection.

Isolated Intersection Control

Fixed-time strategies. Isolated fixed-time strategies are only applicable to undersaturated traffic conditions. *Stage-based strategies* under this class determine the optimal splits and cycle time so as to minimize the total delay or maximize the intersection capacity. *Phase-based* strategies determine not only optimal splits and cycle time but also the optimal staging, which may be an important feature for complex intersections.

Well-known examples of stage-based strategies are SIGSET and SIGCAP proposed by Allsop (1971; 1976). Assuming m prespecified stages, SIGSET and SIGCAP specify the splits $\lambda_1,...,\lambda_m$ and the cycle time c. Note that

$$\lambda_0 + \lambda_1 + ... + \lambda_m = 1 \qquad (8.6)$$

holds by definition, where $\lambda_0 = L/c$, and L is the total lost time. In order to avoid queue building, the following capacity constraint must hold for each stream j

$$s_j \sum_{i=1}^{m} \alpha_{ij} \lambda_i \geq d_j \qquad \forall j \tag{8.7}$$

where s_j and d_j are the saturation flow and the demand, respectively, of stream j; α_{ij} is 1 if stream j has r.o.w. at stage i, and 0 else. Inequality (8.7) requires that the demand d_j of stream j should not be higher than the maximum possible flow assigned to this stream. Finally, a maximum-cycle and m minimum-green constraints are also taken into account.

A nonlinear total delay function derived by Webster (1958) for undersaturated conditions is used in SIGSET as an optimization objective. Thus SIGSET solves a linearly constrained nonlinear programming problem to minimize the total intersection delay for given stream demands d_j. On the other hand, SIGCAP may be used to maximize the intersection's capacity as follows. Assume that the real demand is not d_j as in (8.7) but $\mu \cdot d_j$ with $\mu \geq 1$. SIGCAP replaces d_j in (8.7) by $\mu \cdot d_j$ and maximizes μ under the same constraints as SIGSET, which leads to a linear programming problem.

Note that, for reasons mentioned earlier, capacity maximization always leads to the maximum allowable cycle time. Clearly, SIGCAP should be used for intersections with high demand variability in order to prevent oversaturation, while SIGSET may be used under sufficient capacity margins by replacing d_j in (8.7) by $p_j d_j$, where $p_j \leq 1$ are pre-specified margin parameters.

Phase-based approaches (Improta and Cantarella, 1984) solve a similar problem, suitably extended to consider different staging combinations. Phase-based approaches consider the compatibility relations of involved streams as pre-specified and deliver the optimal staging, splits, and cycle time, so as to minimize total delay or maximize the intersection capacity. The resulting optimization problem is of the binary-mixed-integer-linear-programming type which calls for branch-and-bound methods for an exact solution. The related computation time is naturally much higher than for stage-based approaches, but this is of minor importance as calculations are performed off-line.

Traffic-responsive strategies. Isolated, traffic-responsive strategies make use of real-time measurements provided by loop detectors that are located at the upstream end of each approach, to specify splits and cycle time for given staging. One of the simplest strategies under this class is the *vehicle-interval method* that is applicable to two-stage intersections. Minimum-green durations are assigned to both stages. If no vehicle passes the related detectors during the minimum green of a stage, the strategy proceeds to the next stage. If a vehicle is detected, a critical interval (CI) is created, during which any detected vehicle leads to a green prolongation that allows the vehicle to cross the intersection. If no vehicle is detected during CI, the strategy proceeds to the next stage, else a new CI is created, and so forth, until a pre-specified maximum-green value is reached. An extension of the method also considers the traffic demand on the antagonistic approaches to decide whether to proceed to the next stage or not.

A more sophisticated version of this kind of strategies was proposed by Miller (1963) and is included in the control tool MOVA (Vincent and Young, 1986). Miller's strategy answers every T seconds (e.g. T=2) the question: Should the switching to the next stage take place now, or should this decision be postponed by T? To answer this question, the strategy calculates (under certain simplified conditions) the time wins and losses caused in all approaches if the decision is postponed by $\kappa \cdot T$ seconds. The corresponding total time wins $J_\kappa, \kappa = 1, 2, \ldots$, are combined in a single criterion $J = \max\{J_\kappa; \kappa = 1, 2, \ldots\}$, and if $J < 0$, the switching takes place immediately, else the decision is postponed until the next time step.

A comparative field evaluation of these simple algorithms is presented by De la Bretegue and Jezeguel (1979).

Fixed-Time Coordinated Control

The most popular representatives of this class of strategies are outlined below. By their nature, fixed-time strategies are only applicable to undersaturated traffic conditions.

MAXBAND. The first version of MAXBAND was developed by Little, 1966, see also Little, *et al.,* 1981. MAXBAND considers a two-way arterial with n signals S_1, ..., S_n (intersections) and attempts to specify the corresponding offsets so as to maximize the number of vehicles that can travel within a given speed range without stopping at any signal (green wave), see Figure 8-5. Splits are considered in MAXBAND as given (in accordance with the secondary street demands), hence the problem consists in placing the known red durations (see the horizontal lines of each signal S_i) of the arterial's signals so as to maximize the inbound and outbound bandwidths $\bar{b}$ and b, respectively. For an appropriate problem formulation, it is necessary to introduce some binary decision variables, which leads to a binary-mixed-integer-linear-programming problem. The employed branch-and-bound solution method benefits from a number of nice properties of this particular problem to reduce the required computational effort. Little (1966) extended the basic MAXBAND method via incorporation of some cycle constraints to make it applicable also to networks of arterials.

MAXBAND has been applied to several road networks in North America and beyond. A number of significant extensions have been introduced in the original method in order to consider a variety of new aspects (see Gartner, 1991) such as: time of clearance of existing queue, left-turn movements, and different bandwidths for each link of the arterial (MULTIBAND) (Gartner, *et al.,* 1991).

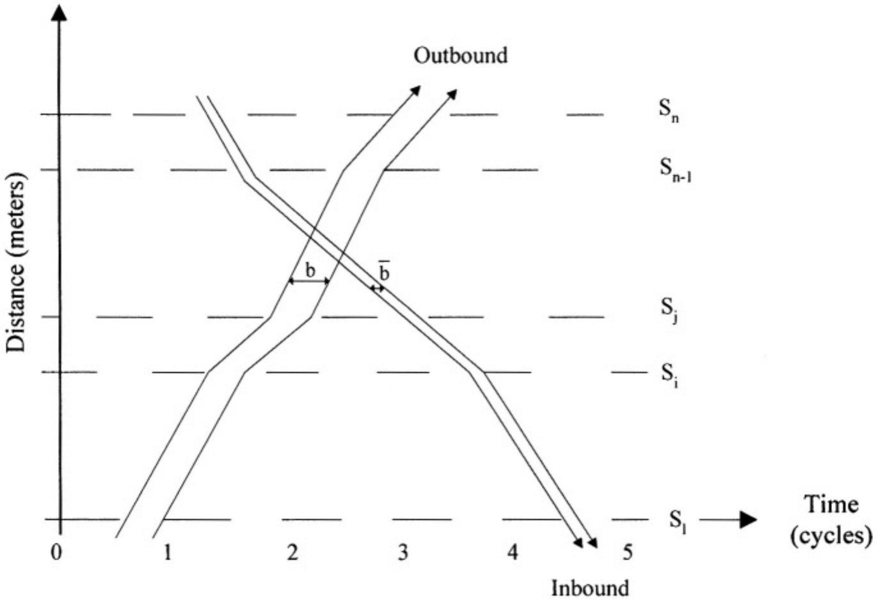


Figure 8-5. A maximum band along an arterial (after Little, *et al.*, 1965).

TRANSYT. TRANSYT was first developed by Robertson (1969) but was substantially extended and enhanced later. It is perhaps the most known and most frequently applied road control strategy, and it is often used as a reference method to test improvements enabled by real-time strategies. First field implementations of TRANSYT indicated savings of some 16% of the average travel time through the network.

Figure 8-6 depicts the method's basic structure: The initial signal settings include the pre-specified staging, the minimum green durations for each stage of each intersection, and the initial choice of splits, offsets, and cycle time. A unique cycle time c or c/2 is considered for all network intersections. The network and traffic flow data comprise the network's geometry, the saturation flows, the link travel times, the constant and known turning rates for each intersection, and the constant and known demands. The traffic model consists of nodes (intersections) and links (connecting streets). The concept of "platoon dispersion" (first-order, time-delay system) is used to model flow progression along a link. Oversaturated conditions cannot be described. The method proceeds as follows: For given values of the decision variables (control inputs), i.e. of splits, offsets, and cycle time, the dynamic network model calculates the corresponding performance index, e.g. the total number of vehicle stops. A heuristic "hill-climb" optimization algorithm introduces small changes to the decision variables and orders a new model run, and so forth, until a (local) minimum is found.

Drawbacks of fixed-time strategies. The main drawback of fixed-time strategies is that their settings are based on historical rather than real-time data. This may be a

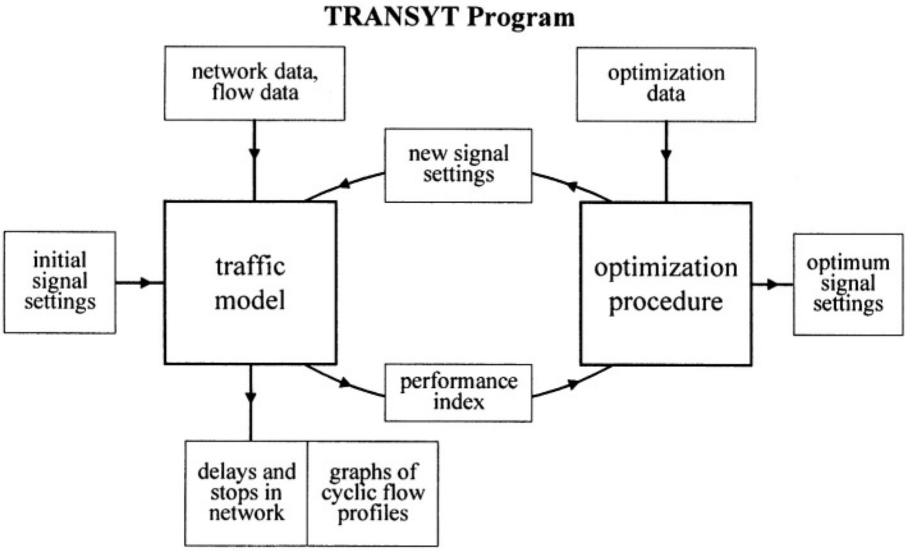


Figure 8-6. Structure of TRANSYT (after Robertson, 1969).

rude simplification because:

- Demands are not constant, even within a time-of-day.
- Demands may vary at different days, e.g. due to special events.
- Demands change in the long term leading to "aging" of the optimized settings.
- Turning movements are also changing in the same ways as demands; in addition, turning movements may change due to the drivers' response to the new optimized signal settings, whereby they try to minimize their individual travel times (Van Vuren, 1991).
- Incidents and farther disturbances may perturb traffic conditions in a non-predictable way.

For all these reasons, traffic-responsive coordinated strategies, if suitably designed, are potentially more efficient, but also more costly, as they require the installation, operation, and maintenance of a real-time control system (measurements, communications, central control room, local controllers).

Coordinated Traffic-Responsive Strategies

SCOOT. SCOOT was first developed by Hunt, *et al.* (1982) and has been extended later in several respects. It is the traffic-responsive version of TRANSYT and has been applied to over 40 cities in the United Kingdom and elsewhere. SCOOT utilizes traffic volume and occupancy (similar to traffic density) measurements from the upstream end of the network links. It runs in a central control computer and employs a philosophy similar to TRANSYT. More precisely, SCOOT includes a network model that is fed with real measurements (instead of historical values) and is run repeatedly in real time to investigate the effect of incremental changes of splits, offsets, and cycle time. If the changes turn out to be beneficial (in terms of a performance index) they are submitted to the signal controllers, see Luk, 1984, for a comparative field evaluation.

Advanced methods. More recently, a number of advanced traffic-responsive strategies have been developed: *OPAC* (Gartner, 1983), *PRODYN* (Farges, *et al.,* 1983), *CRONOS* (Boillot, *et al.,* 1991), *COP* (Sen and Head, 1997). These strategies do not consider explicitly splits, offsets, or cycles. Based on pre-specified staging, they consider in real time the optimal specification of the next few switching times τ_i, i = 1, 2, ..., over a future time horizon H, starting from the current time t and the currently applied stage. To obtain the optimal switching times, these methods solve in real time a dynamic optimization problem employing realistic, dynamic traffic models that include binary variables to reflect the impact of red/green phases on traffic flow. Several constraints, e.g. for maximum and minimum splits, are included. The performance index to be minimized is the total time spent by all vehicles.

The *rolling horizon* procedure is employed for real-time application of the results. Hereby, the optimization problem is solved over a time horizon H (e.g. 60 s), but results are applied only for a much shorter roll period h (e.g. 4 s), after which new measurements are collected and a new optimization problem is solved over an equally long time horizon H, and so forth. The rolling horizon procedure avoids myopic control actions while embedding a dynamic optimization problem in a traffic-responsive environment.

The basic problem faced by these strategies is due to the presence of binary variables that require employment of exponential-complexity algorithms for a global minimization. In fact, OPAC employs complete enumeration (assuming integer switching times) while PRODYN and COP employ dynamic programming. Due to the exponential complexity of these solution algorithms, the control strategies (though conceptually applicable to a whole network) are not real-time feasible for more than one intersection. Hence we end up with a number of decentralized (by intersection) optimal strategies, whose actions may be coordinated heuristically by a superior control layer (see e.g. Kessaci, *et al.,* 1990). On the other hand, CRONOS employs a heuristic global optimization method with polynomial complexity which allows for simultaneous consideration of several intersections, albeit for the price of specifying a local (rather than the global) minimum.

Store-and-forward based approaches. Store-and-forward modelling of traffic networks was first suggested by Gazis and Potts (1963) and has since been used in various works notably for road traffic control (Gazis, 1964; D'Ans and Gazis, 1974; Singh and Tamura, 1974; Michalopoulos and Stephanopoulos. 1977a; 1977b; Lim, *et al.,* 1981; Davison and Özgüner, 1983; Park, *et al.,* 1984; Rathi, 1988; Kim and Bell, 1992). The main idea when using store-and-forward models for road traffic control is to introduce a model simplification that enables the mathematical description of the traffic flow process without use of binary variables. This is of paramount importance because it opens the way to the application of a number of highly efficient optimization and control methods (such as linear programming, quadratic programming, nonlinear programming, and multivariable regulators) with polynomial complexity, which, on its turn, allows for coordinated control of large-scale networks in real time.

The critical simplification is introduced when modelling the outflow u_i of a stream i. Assuming sufficient demand on the link, the outflow u_i at discrete time k is set

$$u_i(k) = (g_i(k)/c)s_i \qquad (8.8)$$

where $g_i(k)$ is the green time duration for this stream and s_i is the corresponding saturation flow. If the time step T is equal to the cycle time c, Figure 8-7 illustrates that $u_i(k)$ in (8.8) is equal to the *average* flow during the corresponding cycle, rather than equal to $s_i$during the green phase and equal to zero during the red phase. In other words, (8.8) suggests that there is a continuous (uninterrupted) outflow from each network link (as long as there is sufficient demand). The consequences of this simplification are:

- The time step T of the discrete-time representation cannot be shorter than the cycle time c, hence real-time decisions cannot be taken more frequently than at every cycle.
- The oscillations of vehicle queues in the links due to green/red-commutations are not described by the model.
- The effect of offset for consecutive intersections cannot be described by the model.

Despite these consequences, the appropriate use of store-and-forward models may lead to efficient coordinated control strategies for large-scale networks as demonstrated in simulation studies in some of the aforementioned references. A recent control strategy of this class (*TUC*) derives a multivariable regulator based on store-and-forward modelling (Diakaki, *et al.,* 1999). TUC has been implemented and is currently operational in a part of Glasgow's (Scotland) urban network with very successful results.

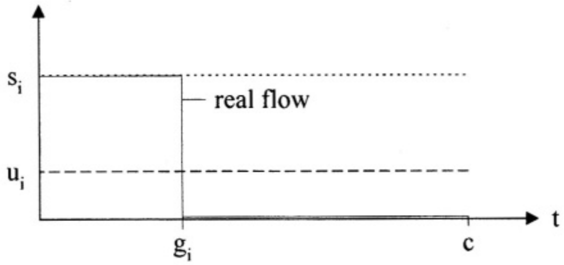


Figure 8-7. Simplified modelling of link outflow u_i.

Integrated Urban-Freeway Traffic Control

Modern metropolitan traffic networks include both urban roads and freeways and employ a variety of control measures such as signal control, ramp metering (see section 8.3), variable message signs, and route guidance (see section 8.4). Traditionally, control strategies for each type of control measure are designed and implemented separately, which may result in antagonistic actions and lack of synergy among different control strategies. However, modern traffic networks that include various infrastructure types, are perceived by the users as an entity, and all included control measures, regardless of their type or location, ultimately serve the same goal of higher network efficiency. Integrated control strategies should consider all control measures simultaneously towards a common control objective. Despite some preliminary works on this subject (see Capelle, 1979, for an early status report), the problem of control integration is quite difficult due to its high dimensions that reflect the geographical extension of the traffic network (Van Aerde and Yagar, 1988; Mahmassani and Jayakrishnan, 1991; Reiss, *et al.,* 1991; Kim and Bell, 1992; Chang, *et al.,* 1993; Papageorgiou, 1994; Elloumi, *et al.,* 1996). For this reason, it appears that store-and-forward modelling (at least for the urban road part) might be the only feasible way to design and operate in real time a unique integrated control strategy, The aforementioned Glasgow implementation covers in fact control measures of various types (signal control, ramp metering, variable message signs) via partial interconnection of three feedback strategies (Diakaki, *et al.,* 1999).

8.3 Freeway Traffic Control

Motivation

Freeways had been originally conceived so as to provide virtually unlimited mobility to road users, without the annoyance of flow interruptions by traffic lights. The rapid increase of traffic demand, however, lead soon to increasingly severe congestions, both *recurrent* (occurring daily during rush hours) and *non-recurrent* (due to incidents). The increasingly congested freeways within and around metropolitan areas resemble the urban traffic networks before introduction of traffic lights: Chaotic conditions at intersections, long queues, degraded infrastructure utilization, reduced safety. At the present stage, responsible authorities have not fully realized that freeways and freeway networks are limited-capacity facilities whose capacity is strongly underutilized on a daily basis due to the lack of systematic and comprehensive traffic control systems.

The control measures that are typically employed in freeway networks are:

- *Ramp metering*, activated via installation of traffic lights at on-ramps or freeway interchanges.
- *Lane control*, that comprises a number of possibilities including variable speed limits, congestion warning, tidal flow, keep-lane instructions, etc.

- *Driver information and guidance systems*, either by use of roadside variable message signs or via two-way communication with equipped vehicles (see section 8.4).

Ramp metering is the most direct and efficient way to control and upgrade freeway traffic. Various positive effects are achievable if ramp metering is appropriately applied:

- Increase in mainline throughput due to avoidance or reduction of congestion.
- Increase in the served volume due to avoidance of blocked off-ramps or freeway interchanges.
- Utilization of possible reserve capacity on parallel arterials.
- Efficient incident response
- Improved traffic safety due to reduced congestion and safer merging.

Some recent studies have demonstrated that efficient ramp metering strategies (employing optimal control algorithms) may provide spectacular improvements (50% reduction of total time spent) in large-scale freeway networks (Kotsialos, *et al.*, 2000; Mangeas, *et al.*, 2000).

Fixed-Time Ramp Metering Strategies

Fixed-time ramp metering strategies are derived off-line for particular times-of-day, based on constant historical demands, without use of real-time measurements. They are based on simple static models. A freeway with several on-ramps and off-ramps is subdivided into sections, each containing one on-ramp. We then have

$$q_j = \sum_{i=1}^{j} \alpha_{ij} r_i \qquad (8.9)$$

where q_j is the mainline flow of section j, r_i is the on-ramp volume of section i, and $\alpha_{ij} \in [0, 1]$ expresses the (known) portion of vehicles that enter the freeway in section i and do not exit the freeway upstream of section j. To avoid congestion

$$q_j \le q_{cap,j} \qquad \forall j \qquad (8.10)$$

must hold, where $q_{cap,j}$ is the capacity of section j. Further constraints are

$$r_{j,min} \le r_j \le \min\{r_{j,max}, d_j\} \qquad (8.11)$$

where d_j is the demand at on-ramp j. This approach was first suggested by Wattleworth (1965). Other similar formulations may be found in Yuan and Kreer, 1971; Tabac, 1972; Wang, 1972; Wang and May, 1973; Cheng, *et al.*, 1974; Schwarz and Tan, 1977.

As an objective criterion, one may wish to maximize the number of served vehicles (which is equivalent to minimising the total time spent)

$$\sum_j r_j \rightarrow \text{Max} \tag{8.12a}$$

or to maximize the total travel distance

$$\sum_j \Delta_j q_j \rightarrow \text{Max} \tag{8.12b}$$

(where Δ_j is the length of section j), or to balance the ramp queues

$$\sum_j (d_j - r_j)^2 \rightarrow \text{Min.} \tag{8.12c}$$

These formulations lead to linear programming or quadratic programming problems that may be readily solved by use of broadly available computer codes. An extension of these methods that renders the static model (8.9) dynamic by introduction of constant travel times for each section was suggested by Papageorgiou (1980).

The drawbacks of fixed-time ramp metering strategies are identical to the ones discussed under road traffic control. In addition, fixed-time ramp metering strategies may lead (due to the absence of real-time measurements) either to overload of the mainstream flow (congestion) or to underutilization of the freeway. In fact, ramp metering is an efficient but also delicate control measure. If ramp metering strategies are not accurate enough, then congestion may not be prevented from forming, or the mainstream capacity may be underutilized (e.g. due to groundlessly strong metering).

Reactive Ramp Metering Strategies

Reactive ramp metering strategies are employed at a tactical level, i.e. in the aim of keeping the freeway traffic conditions close to pre-specified set values, based on real-time measurements.

Local ramp metering. Local ramp metering strategies make use of traffic measurements in the vicinity of a ramp to calculate suitable ramp metering values. The *demand-capacity strategy*, quite popular in North America, reads

$$r(k) = \begin{cases} q_{cap} - q_{in}(k-1) & \text{if} \quad o_{out}(k) \le o_{cr} \\ r_{min} & \text{else} \end{cases} \tag{8.13}$$

where (Figure 8-8) q_{cap} is the freeway capacity downstream of the ramp, q_{in} is the freeway flow measurement upstream of the ramp, o_{out} is the freeway occupancy measurement downstream of the ramp, o_{cr} is the critical occupancy (at which the freeway flow becomes maximum), and r_{min} is a pre-specified minimum ramp flow value. The strategy (8.13) attempts to add to the measured upstream flow $q_{in}(k-1)$ as much ramp flow r(k) as necessary to reach the downstream freeway capacity q_{cap}. If,

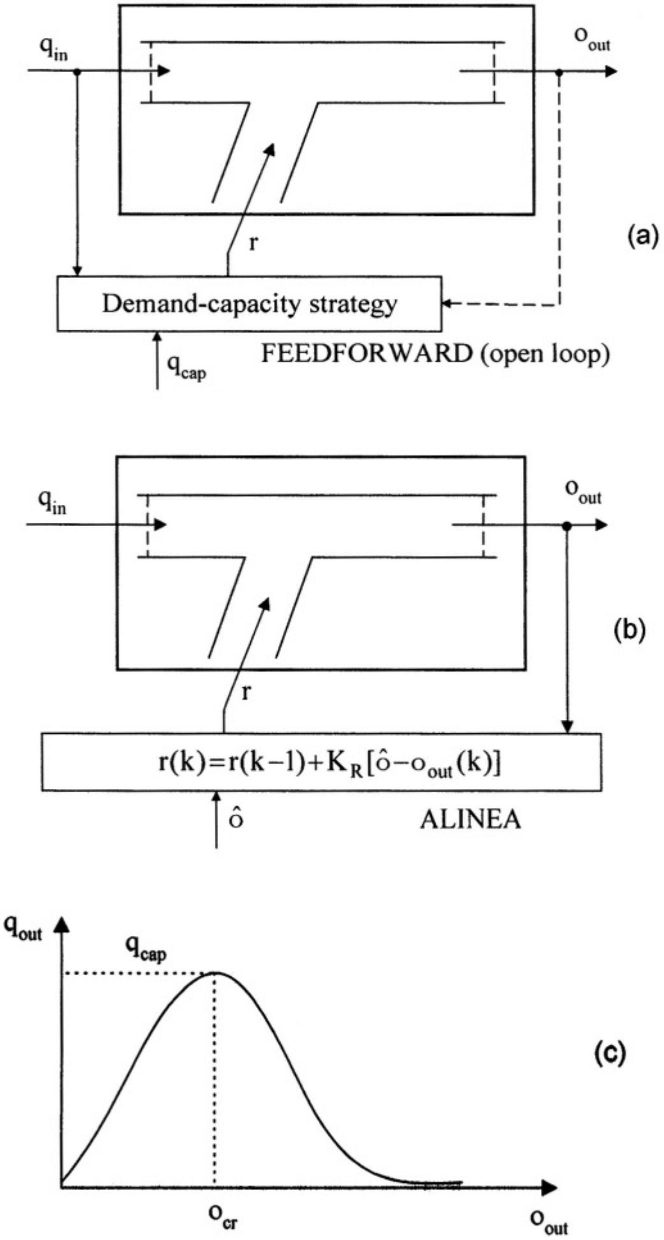


Figure 8-8. Local ramp metering strategies:
(a) Demand-capacity, (b) ALINEA, (c) the fundamental diagram.

however, for some reason, the downstream measured occupancy $o_{out}(k)$ becomes overcritical (i.e. a congestion may form), the ramp flow r(k) is reduced to the minimum flow r_{min} to avoid or to dissolve the congestion.

Comparing the control problem in hand with Figure 8-1, it becomes clear that the ramp flow r is a control input, the downstream occupancy o_{out} is an output, while the upstream freeway flow q_{in} is a disturbance. Hence, (8.13) does not really represent a closed-loop strategy but an open-loop disturbance-rejection policy (Figure 8-8a) which is generally known to be quite sensitive to various further non-measurable disturbances.

The *occupancy strategy* (Masher, *et al.,* 1975) is based on the same philosophy as the demand-capacity strategy, but it relies on occupancy-based estimation of q_{in}, which may, under certain conditions, reduce the corresponding implementation cost.

An alternative, closed-loop ramp metering strategy (*ALINEA*), suggested by Papageorgiou, *et al.* (1991), reads

$$r(k) = r(k-1) + K_R[\hat{o} - o_{out}(k)] \qquad (8.14)$$

where $K_R > 0$ is a regulator parameter and $\hat{o}$ is a set (desired) value for the downstream occupancy (typically, but not necessarily, $\hat{o} = o_{cr}$ may be set, in which case the downstream freeway flow becomes close to q_{cap}, see Figure 8-8c). In field experiments, ALINEA has not been very sensitive to the choice of the regulator parameter K_R.

Note that the demand-capacity strategy reacts to excessive occupancies o_{out} only after a threshold value (o_{cr}) is exceeded, and in a rather crude way, while ALINEA reacts smoothly even to slight differences $\hat{o} - o_{out}(k)$, and thus it may prevent congestion by stabilizing the traffic flow at a high throughput level. It is easily seen that at a stationary state (i.e. if q_{in} is constant), $o_{out}(k) = \hat{o}$ results from (8.14), although no measurements of the inflow q_{in} are explicitly used in the strategy.

The set value $\hat{o}$ may be changed any time, and thus ALINEA may be embedded into a hierarchical control system with set values of the individual ramps being specified in real time by a superior coordination level or by an operator.

All control strategies calculate suitable ramp volumes r. In the case of traffic-cycle realization of ramp metering, r is converted to a green-phase duration g by use of

$$g = (r/r_{sat}) \cdot c \qquad (8.15)$$

where c is the fixed cycle time and r_{sat} is the ramp's saturation flow. The green-phase duration g is constrained by $g \in [g_{min}, g_{max}]$, where $g_{min} > 0$ to avoid ramp closure, and $g_{max} \leq c$. In the case of an one-car-per-green realization, a constant-duration green phase permits exactly one vehicle to pass. Thus, the ramp volume r is controlled by varying the red-phase duration between a minimum (zero) and a maximum value.

Note that ALINEA is also applicable directly to the green or red-phase duration, by combining (8.14) and (8.15)

$$g(k) = g(k-1) + K_R' \, [\hat{o} \, - o_{out}(k)] \tag{8.16}$$

where $K_R' = K_R c / r_{sat}$. Note also that the values r(k–1) or g(k–1) used on the right-hand side of (8.14) or (8.16), respectively, should be the *bounded* values of the previous time step (i.e. after application of the g_{min} and g_{max} constraints) in order to avoid the wind-up phenomenon in the regulator.

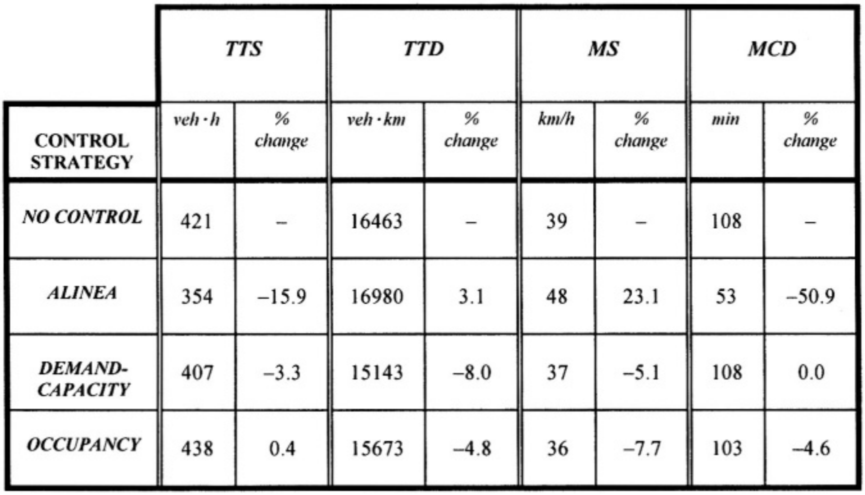
If the queue of vehicles on the ramp becomes excessive, interference with surface street traffic may occur. This may be detected with suitably placed detectors (on the upstream part of the on-ramp), leading to an override of the regulators decisions to allow more vehicles to enter the freeway and the ramp queue to diminish.

Note that the above specifications and constraints apply in the same way to any ramp metering strategy.

Comparative field trials have been conducted in various countries to assess and compare the efficiency of local ramp metering strategies, see e.g. Papageorgiou, *et al.,* 1998. One of these trials took place at the on-ramp Brançion of the clockwise direction of the Boulevard Périphérique (ringway) in Paris. Several ramp metering strategies were applied over a period of one month each, and 13 typical days (without incidents) per strategy were selected for comparison. The evaluation criteria included total travel time (TTT) on the mainstream; total waiting time (TWT) at the ramp; total time spent (TTS = TTT + TWT); total travel distance (TTD); mean speed (MS = TTD/TTS); and mean congestion duration (MCD), which is the accumulated period of time during the morning peak in which the measured occupancy is higher than o_{cr}. Table 8-1 displays an extract of the comparative results for the period 7:00 a.m. to 10:00 a.m. It can be seen that ALINEA leads to the best improvement of all evaluation criteria.

Multivariable regulator strategies. Multivariable regulators for ramp metering pursue the same goals as local ramp metering strategies: They attempt to operate the freeway traffic conditions near some pre-specified set (desired) values. While local ramp metering is performed independently for each ramp, based on local measurements, multivariable regulators make use of all available mainstream measurements $o_i(k)$, i= 1, ..., n, on a freeway stretch, to calculate simultaneously the ramp volume values $r_i(k)$, i = 1, ..., m, for all controllable ramps included in the same stretch (Papageorgiou, *et al.*, 1990). This provides potential improvements over local ramp metering because of more comprehensive information provision and because of coordinated control actions. Multivariable regulator approaches to ramp metering have been reported by Yuan and Kreer, 1968; Kaya, 1972; Knapp, 1972; Isaksen and Payne, 1973; Payne, *et al.,* 1973; Athans, *et al.,* 1975; Looze, *et al.,* 1978; Cremer, 1978; Goldstein and Kumar, 1982; Papageorgiou, 1984; Benmohamed and Meerkov, 1994. The multivariable regulator strategy *METALINE* may be viewed as a generalisation and extension of ALINEA, whereby the metered

Table 8-1. Comparative field results of local ramp metering.



CONTROL STRATEGY	*TTS*		*TTD*		*MS*		*MCD*	
	veh·h	*% change*	*veh·km*	*% change*	*km/h*	*% change*	*min*	*% change*
NO CONTROL	421	–	16463	–	39	–	108	–
ALINEA	354	–15.9	16980	3.1	48	23.1	53	–50.9
DEMAND-CAPACITY	407	–3.3	15143	–8.0	37	–5.1	108	0.0
OCCUPANCY	438	0.4	15673	–4.8	36	–7.7	103	–4.6

on-ramp volumes are calculated from (bold variables indicate vectors and matrices)

$$\mathbf{r}(k) = \mathbf{r}(k-1) - \mathbf{K}_1[\mathbf{o}(k) - \mathbf{o}(k-1)] + \mathbf{K}_2[\hat{\mathbf{O}} - \mathbf{O}(k)] \qquad (8.17)$$

where $\mathbf{r} = [r_1 \ldots r_m]^T$ is the vector of m controllable on-ramp volumes, $\mathbf{o} = [o_1 \ldots o_n]^T$ is the vector of n measured occupancies on the freeway stretch, $\mathbf{O} = [O_1 \ldots O_m]^T$ is a subset of $\mathbf{o}$ that includes m occupancy locations for which pre-specified set values $\hat{\mathbf{O}} = [\hat{O}_1 \ldots \hat{O}_m]^T$ may be given. Note that for control-theoretic reasons the number of set-valued occupancies cannot be higher than the number of controlled on-ramps. Typically one bottleneck location downstream of each controlled on-ramp is selected for inclusion in the vector $\mathbf{O}$. Finally, $\mathbf{K}_1$ and $\mathbf{K}_2$ are the regulator's constant gain matrices that must be suitably designed, see Papageorgiou, *et al.,* 1990; Diakaki and Papageorgiou, 1994, for details.

Field trials and simulation results comparing the efficiency of METALINE versus ALINEA lead to the following conclusions:

- While ALINEA requires hardly any design effort, METALINE application calls for a rather sophisticated design procedure that is based on advanced control-theoretic methods (LQ optimal control).
- For urban freeways with a high density of on-ramps, METALINE was found to provide no advantages over ALINEA (the later implemented independently at each controllable on-ramp) under recurrent congestion.
- In the case of non-recurrent congestion (e.g. due to an incident), METALINE performs better than ALINEA due to more comprehensive measurement information.

Some system operators hesitate to apply ramp metering because of the concern that congestion may be conveyed from the freeway to the adjacent street network. In fact, a ramp metering application designed to avoid or reduce congestion on freeways may have both positive and negative effects on the adjacent road network traffic. It is easy to see, based on notions and statements made earlier, that, if an efficient control strategy is applied for ramp metering, the freeway throughput will be generally increased. More precisely, ramp metering at the beginning of the rush hour may lead to on-ramp queues in order to prevent congestion to form on the freeway, which may temporarily lead to diversion towards the urban network. But due to congestion avoidance or reduction, the freeway will be eventually enabled to accommodate a higher throughput, thus attracting drivers from urban paths and leading to an improved overall network performance. This positive impact of ramp metering on both the freeway and the adjacent road network traffic conditions was confirmed in a specially designed field evaluation in the Corridor Périphérique in Paris, see Haj-Salem and Papageorgiou, 1992.

Nonlinear Optimal Ramp Metering Strategies

Prevention or reduction of traffic congestion on freeway networks may dramatically improve the infrastructure efficiency in terms of throughput and total time spent. Congestion on limited-capacity freeways forms because too many vehicles attempt to use them in a non-coordinated (uncontrolled) way. Once congestion is built up, the outflow from the congestion area is reduced and the off-ramps and interchanges covered by the congestion are blocked, which may in some extreme cases even lead to fatal gridlocks. Reactive ramp metering strategies may be helpful to a certain extent, but, first they need appropriate set values, and, second, their character is more or less local. What is needed for freeway networks or long stretches is a superior coordination level that calculates in real time optimal set values from a proactive, strategic point of view. Such an optimal control strategy should explicitly take into account:

- Demand predictions over a sufficiently long time horizon.
- The current traffic state both on the freeway and on the on-ramps.
- The limited storage capacity of the on-ramps.
- The ramp metering constraints discussed earlier.
- The nonlinear traffic flow dynamics, including the infrastructure's limited capacity.
- Any incidents currently present in the freeway network.

Based on this comprehensive information, the control strategy should deliver set values for the overall freeway network over a future time horizon so as

- to respect all present constraints
- to minimize an objective criterion such as the total time spent in the whole network including the on-ramps.

Such a comprehensive dynamic optimal control problem may be formulated and solved with moderate computation time by use of suitable solution algorithms.

The nonlinear traffic dynamics may be expressed by use of suitable dynamic models in the form

$$\mathbf{x}(k+1) = \mathbf{f}[\mathbf{x}(k), \mathbf{r}(k), \mathbf{d}(k)] \tag{8.18}$$

where the state vector **x** comprises all traffic densities and mean speeds of 500-m long freeway sections, as well as all ramp queues; the control vector **r** comprises all controllable ramp volumes; the disturbance vector **d** comprises all on-ramp demands. The ramp metering constraints are given by (8.11) while the queue constraints read

$$l_i(k) \le l_{i,max} \tag{8.19}$$

where l_i are queue lengths. The total time spent in the whole system over a time horizon K may be expressed

$$T_s = T\sum_{k=0}^{K}[\sum_{i=1}^{n}\rho_i(k)\cdot\Delta_i + \sum_{i=1}^{m} l_i(k)]. \tag{8.20}$$

Thus, for given current (initial) state **x**(0) from corresponding measurements, and given demand predictions **d**(k), k = 0, ..., K–1, the problem consists in specifying the ramp flows **r**(k), k = 0, ..., K–1, so as to minimize the total time spent (8.20) subject to the nonlinear traffic flow dynamics (8.18) and the constraints (8.11) and (8.19).

This problem or variations thereof was considered and solved in various works (Blinkin, 1976; Papageorgiou and Mayr, 1982; Papageorgiou, 1983; Bhouri, *et al.,* 1990; Bhouri, 1991; Stephanedes and Chang, 1993; Zhang, *et al.,* 1996; Chen, *et al.,* 1997; Kotsialos, *et al.,* 2000; Mangeas, *et al.,* 2000). Although simulation studies indicate substantial savings of travel time and substantial increase of throughput, advanced control strategies of this kind have not been implemented in the field as of yet.

Integrated Freeway Network Traffic Control

As mentioned earlier, modern freeway networks may include different types of control measures. The corresponding control strategies are usually designed and implemented independently, thus failing to exploit the synergistic effects that might result from coordination of the respective control actions. An advanced concept for integrated freeway network control results from suitable extension of the optimal control approach outlined above. More precisely, the dynamic model (8.18) of freeway traffic flow may be extended to enable the inclusion of further control measures, beyond the ramp metering rates **r**(k). Formally **r**(k) is then replaced in (8.18) by a general control input vector **u**(k) that comprises all implemented control measures of any type. Such an approach was implemented in the integrated freeway network control tool *AMOC* (Kotsialos, *et al.,* 1999) where ramp metering and route guidance (see section 8.4) are considered simultaneously with promising results, see also Moreno-Banos, *et al.,* 1993; Ataslar and Iftar, 1998.

Lane Control

Lane control may include one or a combination of the following actions:

- Variable speed limitation
- Changeable message signs with indications for “keep lane”, or congestion warning, or environmental warning (e.g. information about the pavement state)
- Incident warning
- Reversable flow lanes (tidal flow).

There are many freeway stretches, particularly in Germany and in The Netherlands, employing a selection of these measures. It is generally thought that control measures of this kind lead to a homogenization of traffic flow (i.e. more homogeneous speeds of cars within a lane and of average speeds of different lanes) which is believed to reduce the risk of falling into congestion at high traffic densities and to increase the freeway’s capacity. Very few systematic studies have been conducted to quantify the impact of these control measures (see e.g. Zackor, 1972; Smulders, 1990) and corresponding validated mathematical models are currently lacking. This is one the reasons why the corresponding control strategies of operating systems are of a heuristic character (e.g. Bode and Haller, 1983; Forner and Schmoll, 1984; Zackor and Balz, 1984.)

8.4 Route Guidance and Driver Information

Introduction

Freeway, urban, or mixed traffic networks include a large number of origins and destinations with multiple paths connecting each origin-destination pair. Fixed direction signs at bifurcation nodes of the network typically indicate the direction that is shortest in absence of congestion. However, during rush hours, the travel time on many routes changes substantially due to traffic congestion and alternative routes may become competitive. Drivers who are familiar with the traffic conditions in a network (e.g. commuters) optimize their individual routes based on their past experience, thus leading to the celebrated user-equilibrium conditions (Wardrop, 1952). But daily varying demands, changing environmental conditions, exceptional events (sport events, fairs, concerts, etc.) and, most importantly, incidents may change the traffic conditions in a non-predictable way. This may lead to an underutilization of the overall network’s capacity, whereby some links are heavily congested while capacity reserves are available on alternative routes. Route guidance and driver information systems (RGDIS) may be employed to improve the network efficiency via direct or indirect recommendation of alternative routes (see Hall, 1993, for a critical view).

A first classification of RGDIS distinguishes *pre-trip* from *en-route* advice. Pre-trip communication possibilities include the internet, phone services, mobile devices, and radio. These communication devices may be consulted by a potential road user to make a rational decision regarding:

- The effectuation or postponement of the intended trip
- The choice of transport mode (car, bus, underground, etc.)
- The choice of the departure time
- The (initial) path choice.

If the road user has decided to complete the trip by car, she may continue to receive information or advice via appropriate *en-route* devices such as radio services (RDS-TMC), road-side variable message sings (VMS), or special in-car equipment, in order to make sensible routing decisions at bifurcation nodes of the network. While radio broadcasting services and VMS have been in use for more than 25 years (and their number is steadily increasing) (see Stathopoulos, 1991; see also Saxton and Schenck, 1977, for an early overview), individual route guidance systems employing in-car devices and two-way communication with control centers are in their infancy (some experimental or early operational systems exist in some countries) (Henry, *et al.,* 1991).

At this point, it is appropriate to distinguish among two alternative policies (which in some cases may be combined) of providing en-route information versus explicit route recommendation. Many operators (particularly of VMS-based systems) prefer the provision of real-time information. Also the majority of drivers (according to some questionnaire results) seem to prefer this option that enables them to make their own decisions, rather than having to follow recommendations by an anonymous system. It should be emphasized, however, that pure information provision has a number of partially significant drawbacks:

- The translation of provided information into routing decisions requires the knowledge of the network which may not be present for all drivers.
- Although the control center disposes over complete information about the traffic conditions in the whole network, only a tiny part of this information can be conveyed to the users due to space limitations on the VMS. In some cases, only information about the traffic conditions on the downstream links of a bifurcation node are provided. Clearly, this information is not sufficient for a rational route decision for drivers with longer trips through the network, who may be eventually trapped into a severe congestion due to myopic decision making.
- Even if it would be possible to provide more comprehensive information, the drivers would have to make a route decision within a couple of seconds, i.e. after looking at the VMS and before reaching the bifurcation.
- There is no possibility for the operator or a control strategy to actively influence traffic conditions, as decisions are left with the drivers.

On the other hand, route guidance systems are constrained by the requirement not to suggest routes that would disbenefit complying drivers, else the credibility and eventually the impact of the whole system may be jeopardized. Moreover, route guidance systems call for a genuine control strategy that can optimize the network traffic conditions, e.g. by avoiding traffic congestion on alternative recommended routes due to drivers' overreaction.

Travel Time Display

A particular type of driver information system that is gaining increasing momentum due to its relative simplicity and its popularity with drivers is the display (on VMS) of travel times for well-defined stretches downstream of the VMS. This information is readily comprehensible by the drivers, and it may either provide a basis for route choice decisions or simply reduce the drivers' stress, particularly in congested traffic conditions.

For example, some 350 VMS are installed on the Boulevard Périphérique of Paris and on all approaches that lead to this ringway (Lamboley and Baudez, 1994; Haj-Salem, *et al.*, 1995). The displayed message is the current travel time on the ringway from the particular VMS location to two significant downstream freeway intersections, at distances of approximately 3 km and 6 km, respectively. A similar system, providing travel times on the two downstream freeway links of each bifurcation node, is operational in the dense freeway network around Paris.

The calculation of instantaneous travel times for a freeway stretch including several loop detectors is quite simple. The stretch is subdivided into a number of segments with lengths Δ_i ($\Delta_i = 500$ m in Paris), whereby each segment includes one detector station. A detector station may either directly provide mean speed measurements or it may provide flow and occupancy measurements, from which the mean speed v_i can be deduced with sufficient accuracy. The travel time τ_i of segment i is then given by

$$\tau_i = \Delta_i / v_i . \qquad (8.21)$$

The *instantaneous travel time* on a freeway stretch is defined to be the travel time of a virtual vehicle that travels the stretch under the assumption that the currently prevailing traffic conditions will not change during the trip. Based on this definition and (8.21), the instantaneous travel time τ of the whole freeway stretch, consisting ofN segments, may be calculated from

$$\tau = \sum_{i=l}^{N} \tau_i = \sum_{i=l}^{N} \Delta_i / v_i . \qquad (8.22)$$

This formula works quite well in practice but it has some drawbacks, notably, if the mean speed v_i in a segment becomes temporarily very low, then (8.22) delivers an unrealistically high travel time for the whole freeway stretch. Alternative formulas have been suggested and evaluated based on real traffic data, see Grol, *et al.,* 1997; Papageorgiou, 1999.

Clearly, any instantaneous travel time formula based only on current traffic measurements will induce a systematic estimation error if the traffic conditions in the stretch are rapidly changing, e.g. during congestion growth or dissipation. This error grows with the length of the stretch under question and may, under certain conditions, reach unacceptable levels for freeway stretches longer than, say, 10 km. What is needed in this case is a predictive scheme that delivers *predicted travel times* that come closer to the travel times that will be experienced by the drivers during their trip. Predictive travel times may be calculated:

- Based on historical information
- Via suitable extrapolation methods
- By employment of dynamic traffic flow models in real time

or via a combination of the above.

Route Guidance Strategies

Introduction. A route guidance system may be viewed as a traffic control system in the sense of Figure 8-1. Based on real-time measurements, sufficiently interpreted and extended within the surveillance block, a control strategy decides about the routes to be recommended (or the information to be provided) to the road users. This, on its turn, has an impact on the traffic flow conditions in the network, and this impact is reflected in the performance indices. Because of the real-time nature of the operation, requirements of short computation times are relatively strict.

Route guidance strategies may be classified according to various aspects:

- *Reactive* strategies are based only on and react to current measurements without the real-time use of mathematical models or other predictive tools; *predictive* strategies attempt to predict traffic conditions sufficiently far in the future (typically by use of mathematical models) in order to improve the quality of the provided recommendations.
- *Iterative* strategies run several model simulations in real time, each time with suitably modified route guidance, to ensure (at convergence) that the control goal (see below) will be achieved as accurately as possible; iterative strategies are by nature predictive. *One-shot* strategies may either be reactive, in which case they typically perform simple calculations based on real-time data, or they may be predictive, whereby they run one single time a simulation model to increase the relevance of their recommendations.
- Route guidance strategies may aim at either *system optimal* or *user optimal* traffic conditions. In the first case, the control goal is the minimisation of a global objective criterion (e.g. the total time spent) even for the price of recommending routes that are sometimes more costly than the regular routes. In the second case, every recommended route should not be more costly than the regular route, even for the price of sub-optimality with respect to the global objective criterion. Under a more strict definition, user optimal conditions imply *equal* cost on all utilized alternative routes connecting any two nodes in the network.

One-shot strategies. Most one-shot strategies are of the reactive type. Particularly for dense networks, with relatively short links, many bifurcations, and a high number of alternative routes connecting any two nodes, reactive strategies may be highly efficient in establishing user-optimal conditions on the basis of current traffic measurements. This is because reactive routing recommendations in this kind of network may be modified at downstream bifurcation nodes if traffic conditions change substantially.

Most reactive strategies are decentralized, i.e. they conduct their calculations at each bifurcation node independently of other nodes. Simple feedback regulators of the P (proportional) or PI (proportional-integral) types have been proposed by Messmer and Papageorgiou (1994). A P-regulator calculates splitting rates β_{nj} as follows

$$\beta_{nj}(k) = \beta_{nj}^{N} - K\,\Delta\tau_{nj}(k) \tag{8.23}$$

where $\beta_{nj}(k) \in [0, 1]$ is the portion of the flow arriving at bifurcation node n and destined to node j that is routed through the main direction at time k; β_{nj}^{N} is the nominal splitting of drivers (in absence of route guidance); K > 0 is a regulator parameter; $\Delta\tau_{nj}(k)$ is the instantaneous travel time difference between the main and the alternative route from node n to node j. Note that $\beta_{nj}(k)$ resulting from (8.23) is eventually truncated if it exceeds the range [0, 1]. The regulator (8.23) assigns more or less traffic to the alternative direction according to the sign and value of the current travel time difference $\Delta\tau(k)$ among both directions thus aiming at equalizing both corresponding travel times, in accordance with the user-optimum requirements. For K sufficiently high, an all-or-nothing (or bang-bang) strategy results from (8.23) whereby all vehicles are sent to the currently shortest direction. It should be noted that these simple regulators are not very sensitive to varying compliance rates of drivers (Pavlis and Papageorgiou, 1999). A first operational system employing decentralized P-regulators in the traffic network of Aalborg, Denmark, was reported by Mammar, *et al.* (1996). Multivariable regulators for central route guidance systems have also been suggested (Papageorgiou, 1990; Papageorgiou and Messmer, 1991), as well as heuristic feedback schemes (Berger and Shaw, 1975; Sarachik and Özgüner, 1982; Bolelli, *et al.,* 1991; Hawas and Mahmassani, 1995), and more recent automatic control concepts (Kachroo and Özbay, 1999).

One-shot strategies may employ in real time a more or less sophisticated mathematical model of the network traffic flow. Based on the current traffic state, the current control inputs, and predicted future demands, the model is run once, in order to provide information about the future traffic conditions under the current route guidance settings. A regulator is then used to control the predicted future, rather than the current, traffic conditions. Such control schemes are known as IMC (Internal Model Control) strategies in automatic control theory. They are preferable to reactive regulators when the traffic network has long links with a limited number of bifurcation nodes. A control scheme of this kind was applied to the Scottish highway network employing P-regulators (Messmer, *et al.,* 1998) or a heuristic expert system (Papageorgiou, *et al.,* 1994; Morin 1995).

Iterative strategies. Iterative strategies may aim at establishing either system optimal or user optimal conditions. For a system optimum, the pursued procedure has already been outlined in section 8.3. A macroscopic network traffic model may be written in the general form (8.18), where the control inputs are the splitting rates $\beta(k)$. The corresponding optimal control problem, aiming at minimizing the total time spent (8.20) under the constraints $0 \le \beta(k) \le 1$, may be solved by use of the same numerical algorithms as the optimal ramp metering or the integrated control problem (Papageorgiou, 1990; Messmer and Papageorgiou, 1994; 1995), see also Charbonnier, *et al.,* 1991; Lafortune, *et al.,* 1993; Wie, *et al.,* 1995.

There are several iterative procedures suggested towards establishing user optimal conditions (Mahmassani and Peeta, 1994; Ben-Akiva, *et al.,* 1997; Wisten and Smith, 1997). The typical structure of iterative strategies is as follows:

(a) Set the initial path assignments or splitting rates (control inputs)
(b) Run a simulation model over a time horizon H.
(c) Evaluate the travel times on alternative paths; if all travel time differences are sufficiently small, stop with the final solution.
(d) Modify the path assignments or splitting rates appropriately to reduce travel time differences; go to (b).

The simulation models employed by different algorithms in step (b) may be microscopic, macroscopic, or mesoscopic. *Microscopic* models address and describe the movement of each individual vehicle in the traffic flow in dependence of the movement of the adjacent vehicles, both in the longitudinal (car-following behaviour) and in the lateral (lane-changing behaviour) sense. Each vehicle has a pre-specified destination and its path is decided pre-trip or en-route according to the routing decisions of the algorithm. *Macroscopic* models may be expressed in the form (8.20). They describe the traffic flow as a fluid with particular characteristics via the aggregate traffic variables traffic density, flow, and mean speed (see section 8.1). Generally, the evolution of traffic density ρ is described via the conservation-of-vehicles equation, while the mean speed v is deduced from an empirical (static or dynamic) equation in dependence of the traffic density. Finally the traffic flow is by definition $q = \rho v$. To describe the routing behaviour, macroscopic models include partial densities and partial flows by destination. Partial flows are assigned at bifurcation nodes to downstream links according to the splitting rates provided from step (d). *Mesoscopic* models describe the evolution of mean speed macroscopically, but they also consider individual vehicles (or “vehicle packets”) which, however, are moved in the network according to the macroscopic mean speed (without employment of microscopic rules). The traffic flows at section boundaries are deduced from individual vehicle crossings, while traffic densities may be calculated directly from vehicle counts within a section. The reason why individual vehicles are introduced in mesoscopic models is in order to describe the routing behaviour. Thus, as in microscopic models, each vehicle has a pre-specified destination, and its path is decided pre-trip and en-route according to the routing decisions of step (d), without the need to introduce partial densities and flows as in macroscopic models (Mahmassani and Peeta, 1994; Ben-Akiva, *et al.,* 1997; Wisten and Smith, 1997).

The modification of path assignments or splitting rates in step (d) of the algorithm is typically effectuated in a functionally decentralized way, i.e. each splitting rate or path assignment portion is changed independently of any other. The sign and magnitude of the individual changes depend on the sign and magnitude of the corresponding travel time differences, in a similar way as in (8.23), with the significant difference that the travel time differences are here predictive (calculated in step (c)) rather than instantaneous.

The real-time implementation of iterative algorithms for route guidance purposes employs the same rolling horizon procedure outlined in section 8.2 in order to reduce the sensitivity with respect to predicted demands and modeling inaccuracies. No field implementation of an iterative route guidance procedure has been reported as yet. Main reasons for this are the relatively recent interest in RGDIS, but certainly also the code complexity of the corresponding algorithms.

8.5 Future Directions

The Theory-Practice Gap

As in many other engineering disciplines, only a small portion of the significant methodological advancements have really been exploited in the field. It is beyond our scope to investigate and discuss the reasons behind this theory-practice gap, but administrative inertia, little competitive pressure in the public sector, the complexity of traffic control systems, limited realization of the improvement potential behind advanced methods by the responsible authorities, and limited understanding of practical problems by some researchers may have a role in this. Whatever the reasons, the major challenge in the coming decade is the deployment of advanced and efficient traffic control strategies in the field.

More precisely, the majority of small and big cities even in industrialized countries, are still operating old-fashioned fixed-time signal control strategies, often even poorly optimized or maintained. Even when modern traffic-responsive control systems are installed in terms of hardware devices, the employed control strategies are often naïve, poorly tested and optimized, thus failing to exploit the possibilities provided by the relatively expensive hardware infrastructure.

Regarding freeway networks, the situation is even worse. Operational control systems of any kind are the exception rather than the rule. With regard to ramp metering, the main focus is not on improving efficiency but on secondary objectives of different kinds. The responsible traffic authorities and the decision makers are far from realizing the fact that advanced real-time ramp metering systems (employing optimal control algorithms) have the potential of changing dramatically the traffic conditions on today's heavily congested (hence underutilized) freeways with spectacular improvements that may reach 50% reduction of the total time spent.

With regard to driver information and route guidance systems, there is an increasing interest and an increasing number of operational systems employing variable message signs, but once more, the relatively expensive hardware infrastructure is not exploited to the degree possible, as implemented control strategies are typically naïve.

On the side of the research community, any effort should be made to inform the road authorities, the political decision makers, and the general public about the substantial improvements achievable via implementation of modern traffic control methods and tools. At the same time, it should be emphasized that many methodological works presented at conferences and technical journals address practical problems and concerns only in a limited way. In some cases, proposed traffic control strategies are not even thoroughly and properly tested via simulation, despite the meanwhile high number of available traffic simulators of various kinds. This poses a burden to real implementation of the methods, and perhaps the best way for researchers to familiarize themselves with the practical requirements and constraints is to get occasionally involved in real implementations.

Road Traffic Control Strategies

The number of developed signal control strategies is much higher than what could be included or mentioned in section 8.3. The need and trend is clearly towards traffic-responsive coordinated strategies. Two avenues may be identified as promising in this respect:

(a) Advanced signal control strategies, such as OPAC, PRODYN, CRONOS, and COP, have clear limitations regarding the network extent, to which they can be directly applied. The price to be paid if these systems run in a completely decentralized way (e.g. independently at each intersection) is currently not fully analyzed nor understood. This kind of thorough analysis is necessary in order to develop efficient (though probably heuristic) coordinating layers that reduce the negative impact of decentralization.

(b) Store-and-forward based concepts seem, more than 3 decades after their original conception, to offer a promising background for the development of signal control strategies that are traffic-responsive, coordinated (for large-scale networks), and can cope, under certain conditions, with oversaturation and, most importantly, with the imminent inaccuracies of traffic measurements in an urban road environment. In addition, this approach seems ideal for the design of (even more challenging) integrated traffic control strategies involving further traffic systems (freeways) and control measures.

Freeway Traffic Control Strategies

With regard to ramp metering, the most important methodological developments are well advanced. The most promising area for control strategies is the design and testing of hierarchical control structures for very large-scale freeway networks. Control hierarchies should include short-term demand predictions, optimal control algorithms for the coordinated calculation of set values network-wide, and reactive feedback strategies for implementation of the optimal control decisions.

The integrated control of freeway networks involving both ramp metering and route guidance measures is currently in a very preliminary phase with some very promising results, but a lot more developments are required to produce integrated control strategies that are efficient, but also applicable in real time to large-scale networks.

Finally lane control systems may prove much more useful than at present if their impact is studied more carefully and thoroughly so as to open the way to the design of efficient control strategies. This is perhaps one of the least studied areas within traffic control.

Driver Information and Route Guidance Systems

Although the subject of DIRGS is relatively new, a substantial amount of work has been devoted to it and a number of methods and tools have already emerged, but further developments, either completely new or combinations of already suggested methods, are possible and desirable. One-shot methods appear particularly attractive for real-time applications because they are simple, with negligible computational effort. More experience is required regarding their efficiency level and the topological and traffic conditions that are most suitable for their application. The surplus efficiency provided by iterative approaches should be further investigated, and possible combinations (e.g. in order to reduce the computational effort of iterative strategies) should be attempted.

8.6 References

Allsop, R.B. (1971) SIGSET: A computer program for calculating traffic capacity of signal-controlled road junctions. *Traffic Engineering & Control* **12**, 58-60.

Allsop, R.B. (1976) SIGCAP: A computer program for assessing the traffic capacity of signal-controlled road junctions. *Traffic Engineering & Control* **17**, 338-341.

Ataslar, B. and Iftar, A. (1998) A decentralised control approach for transportation networks. *Preprints of the 8th IFAC Symposium on Large Scale Systems,* Patras, Greece, Vol. 2, 348-353.

Athans, M., Houpt, P.K., Looze, D., Orlhac, D., Gershwin, S.B. and Speyer, J.L. (1975) Stochastic control of freeway corridor systems. *Proc. 1975 IEEE Conference on Decision and Control,* 676-685.

Ben-Akiva, M., Bierlaire, M., Bottom. J., Koutsopoulos, H. and Mishalani, R. (1997) Development of a route guidance generation system for real-time application. *Preprints 8th IFAC/IFIP/IFORS Symposium on Transportation Systems,* Chania, Greece, 433-439.

Benmohamed, L. and Meerkov, S.M. (1994) Feedback Control of Highway Congestion by a Fair On-Ramp Metering. *Proc. 33rd IEEE Conference on Decision and Control,* Vol. 3, Lake Buena Vista, Florida, 2437-2442.

Berger, C.R. and Shaw, L. (1975) Diversion control of freeway traffic. *6th IFAC World Congress,* Part IIIC, Paper 4.3, Boston, Massachusetts.

Bhouri, N. (1991) *Commande d'un système de trafic autoroutier: Application au Boulevard Périphérique de Paris.* Ph.D. Dissertation, Université de Paris-Sud, Centre d'Orsay, France.

Bhouri, N.; Papageorgiou, M. and Blosseville, J. M. (1990) Optimal control of traffic flow on periurban ringways with application to the Boulevard Périphérique in Paris. *Preprints of the 11th IFAC World Congress,* Tallinn, Estonia, Vol. 10, 236-243.

Blinkin, M. (1976) Problem of optimal control of traffic flow on highways. *Automation and Remote Control* **37**, 662-667.

Bode, K.-R. and Haller, W. (1983) Geschwindigkeitssteuerung auf der A7 zwischen den Autobahndreiecken Hannover-Nord und Walsrode. *Strassenverkehrstechnik* **27**, 145-151.

Boillot, F., Blosseville, J.M., Lesort, J.B., Motyka, V., Papageorgiou, M. and Sellam, S. (1992) Optimal signal control of urban traffic networks. *6th IEE Intern. Conference on Road Traffic Monitoring and Control,* London, England, 75-79.

Bolleli, A., Mauro, V. and Perono, E. (1991) Models and strategies for dynamic route guidance - Part B: A decentralized, fully dynamic, infrastructure supported route guidance. *Proc. DRIVE Conference,* Brussels, Belgium, 99-105.

Capelle, D.G. (1979) Freeway/corridor systems. *Proc. Intern. Symposium on Traffic Control Systems,* Berkeley, California, Vol. 1, 33-35.

Chang, G.-H., Ho, P.-K. and Wei, C.-H. (1993) A dynamic system-optimum control model for commuting traffic corridors. *Transportation Research* **1C**, 3-22.

Charbonnier, C., Farges, J.L. and Henry, J.-J. (1991) Models and strategies for dynamic route guidance - Part C: Optimal control approach. *Proc. DRIVE Conference,* Brussels, Belgium, 106–112.

Chen, O.J., Hotz, A.F. and Ben-Akiva, M.E. (1997) Development and evaluation of a dynamic ramp metering control model. *Preprints 8th IFAC/IFIP/IFORS Symposium on Transportation Systems,* Chania, Greece, 1162-1168.

Cheng, I.C., Gruz, J.B. and Paquet, J.G. (1974) Entrance ramp control for travel rate maximization in expressways. *Transportation Research* **8**, 503-508.

Cremer, M. (1978) A state feedback approach to freeway traffic control. *Preprints 7th IFAC World Congress,* Helsinki, Finland, 1575-1582.

D' Ans, G.C. and Gazis, D.C. (1976) Optimal control of Oversaturated store-and-forward transportation networks. *Transportation Science* **10**, 1-19.

Davison, E.J. and Özgüner, Ü. (1983) Decentralized control of traffic networks. *IEEE Trans. on Automatic Control* **28**, 677-688.

De la Bretegue, L. and Jezeguel, R. (1979) Adaptive control at an isolated intersection - A comparative study of some algorithms. *Traffic Engineering and Control* **20**, 361-363.

Diakaki, C. and Papageorgiou, M. (1995) *Design and Simulation Test of Integrated Corridor Control for M8 Eastbound Corridor in Glasgow.* Internal Report No. 1995-3. Dynamic Systems and Simulation Laboratory, Technical University of Crete, Chania, Greece.

Diakaki, C., Papageorgiou, M. and McLean, T. (1999) Application and evaluation of the integrated traffic-responsive urban corridor control strategy IN-TUC in Glasgow. *Preprint CD-ROM of the TRB (Transportation Research Board) 78th Annual Meeting,* Washington, D.C., Paper No. 990310.

Elloumi, N., Haj-Salem, H. and Papageorgiou, M. (1996) Integrated control of traffic corridors - Application of an LP-methodology. *4th Meeting of the EURO Working Group on Transportation Systems,* Newcastle, UK.

Farges, J.-L., Henry, J-J. and Tufal, J. (1983) The PRODYN real-time traffic algorithm. *4th IFAC Symposium on Transportation Systems,* Baden Baden, W. Germany, 307-312.

Forner, K. and Schmoll, U. (1984) Situationsabhängige Geschwindigkeitsbeeinflussung auf der Bundesautobahn A8 bei Stuttgart. *Strassenverkehrstechnik* **28**, 1-4.

Gartner, N.H. (1983) OPAC: A demand-responsive strategy for traffic signal control. *Transportation Research Record* **906**, 75-84.

Gartner, N.H. (1991) Road traffic control: Progression methods. In *"Concise Encyclopedia of Traffic & Transportation Systems",* M. Papageorgiou, Editor, Pergamon Press, Oxford, UK, 391-396.

Gartner, N.H., Assmann, S.F., Lasaga, F. and Hom, D.L. (1991) A multiband approach to arterial traffic signal optimization. *Transportation Research* **25B**, 55-74.

Gazis, D.C. (1964) Optimum control of a system of oversaturated intersections. *Operation Research* **12**, 815-831.

Gazis, D.C. and Potts, R.B. (1963) The oversaturated intersection. *Proc. 2nd Intern. Symposium on Traffic Theory,* London, UK, 221-237.

Goldstein, N.B. and Kumar, K.S.P. (1982) A decentralized control strategy for freeway regulation. *Transportation Research* **16B**, 279-290.

Grol, H.J.M. van, Manfredi, S., Danech-Pajouh, M. (1991) *On-line Network Stale Estimation and Short Term Prediction.* Deliverable 5.2 of DACCORD, Transport Telematics Programme, Brussels, Belgium.

Haj-Salem, H., Cohen, S., Sididki, E. and Papageorgiou, M. (1995) Field trial results of VMS travel time display on the Corridor Périphérique in Paris. *Proc. 4th Intern. ASCE Conference on Applications of Advanced Technologies in Transportation Engineering,* Capri, Italy, 368-372.

Haj-Salem, H. and Papageorgiou, M. (1995) Ramp metering impact on urban corridor traffic: Field results. *Transportation Research* **29A**, 303-319.

Hall, R.W. (1993) Non-recurrent congestion: How big is the problem? Are traveller information systems the solution? *Transportation Research* **1C**, 89-103.

Hawas, Y. and Mahmassani, H. S. (1995) A decentralized scheme for real-time route guidance in vehicular traffic networks. *Proc. 2nd World Congress of Intelligent Transport Systems,* Yokohama, Japan, 1956–1963.

Henry, J.-J., Charbonnier, C. and Farges, J.L. (1991) Route guidance, individual. In *"Concise Encyclopedia of Traffic and Transportation Systems",* M. Papageorgiou, Editor, Pergamon Press, Oxford, UK, 417-422.

Hunt, P.B., Robertson, D.L. and Bretherton, R.D. (1982) The SCOOT on-line traffic signal optimisation technique. *Traffic Engineering & Control* **23**, 190-192.

Improta, G. and Cantarella, G.E. (1984) Control systems design for an individual signalised junction. *Transportation Research* **18B**, 147-167.

Isaksen, L. and Payne, H.J. (1973) Suboptimal control of linear systems by augmentation with application to freeway traffic regulation. *IEEE Trans. on Automatic Control* **18**, 210-219.

Kachroo, P. and Özbay, K. (1999) *Feedback Control Theory for Dynamic Traffic Assignment.* Springer Verlag, New York.

Kaya, A. (1972) Computer and optimization techniques for efficient utilization or urban freeway systems. *Proc. 5th IFAC World Congress,* Paris, France, paper 12.1.

Kessaci, A., Farges, J-L. and Henry, J-J. (1990) Upper level for real time urban traffic control systems. *Preprints 11th IFAC World Congress,* Tallinn, Estonia, Vol. 10, 226-229.

Kim, K.-J. and Bell, M.G.H. (1992) Development of an integrated traffic control strategy for both urban signalised and motorway networks. *Proc. 1st Meeting of the EURO Working Group on Urban Traffic and Transportation,* Landshut, Germany.

Knapp, C.H. (1972) Traffic estimation and control at bottlenecks. *Proc. IEEE International Conference on Cybernetics and Society,* Washington, D.C., 469-472.

Kotsialos, A., Papageorgiou, M. and Messmer A. (1999) Optimal co-ordinated and integrated motorway network traffic control. *14th Intern. Symposium on Transportation and Traffic Theory,* Jerusalem, Israel, to appear.

Kotsialos, A., Papageorgiou, M., Mangeas, M. and Haj-Salem H. (2000) Coordinated and integrated control of motorway networks via nonlinear optimal control. *Transportation Research* **C**, submitted for publication.

Lafortune, S., Sengupta, R., Kaufman, D.E. and Smith, R.L. (1993) Dynamic system-optimal traffic assignment using a state space model. *Transportation Research* **27B**, 451-473

Lamboley, C. and Baudez, G. (1994) Information based on travel time: A service for managing urban traffic and informing road-users. *Proc. 1st World Congress on Applications of Transport Telematics and IVHS,* Paris, France, Vol. 1, 181-190.

Lim, J.H., Hwang, S.H., Suh, I.H. and Bien, Z. (1981) Hierarchical optimal control of oversaturated urban networks. *International Journal of Control* **33**, 727-737.

Little, J.D.C. (1966) The synchronisation of traffic signals by mixed-integer-linear-programming. *Operations Research* **14**, 568-594.

Little, J.D.C., Kelson, M.D. and Gartner, N.H. (1981) MAXBAND: A program for setting signals on arteries and triangular networks. *Transportation Research Record* **795**, 40-46.

Looze, D.P., Houpt, P.K., Sandell, N.R. and Athans, M. (1978) On decentralized estimation and control with application to freeway ramp metering. *IEEE Trans. on Automatic Control* **23**, 268-275.

Luk, J.Y.K. (1984) Two traffic-responsive area traffic control methods: SCAT and SCOOT. *Traffic Engineering & Control* **25**, 14-22.

Mahmassani, H. and Jayakrishnan, R. (1991) System performance and user response under real-time information in a congested traffic corrid or. *Transportation Research* **25A**, 293-307.

Mahmassani, H. S. and Peeta, S. (1994) Network performance under system optimal and user equilibrium dynamic assignments: Implications for advanced traveller information systems, *Transportation Research Record* **1408**, 83-93.

Mammar, S., Messmer, A., Jensen, P., Papageorgiou, M., Haj-Salem, H. and Jensen, L. (1996) Automatic control of variable message signs in Aalborg. *Transportation Research* **4C**, 131-150.

Mangeas, M., Haj-Salem, H., Kotsialos, A. and Papageorgiou, M. (2000) Application of a nonlinear coordinated and integrated control strategy to the Ile-de-France motorway network. *Transportation Research* **C**, submitted for publication.

Masher, D.P., Ross, D.W., Wong, P.J., Tuan, P.L., Zeidler and Peracek, S. (1975) *Guidelines for Design and Operating of Ramp Control Systems.* Standford Research Institute Report NCHRP 3-22, SRI Project 3340. SRI, Menid Park, California.

Messmer, A. and Papageorgiou, M. (1994) Automatic control methods applied to freeway network traffic. *Automatica* **30**, 691-702.

Messmer, A. and Papageorgiou, M. (1995) Route diversion control in motorway networks via nonlinear optimization. *IEEE Trans. on Control Systems Technology* **3**, 144-154.

Messmer, A., Papageorgiou, M. and Mackenzie, N. (1998) Automatic control of variable message signs in the interurban Scottish highway network. *Transportation Research* **6C**, 173-187.

Michalopoulos, P.G. and Stephanopoulos G. (1977a) Oversaturated signal systems with queue length constraints - I: Single intersection. *Transportation Research* **11**, 413-421.

Michalopoulos, P.G. and Stephanopoulos, G. (1977b) Oversaturated signal systems with queue length constraints - II: System of intersections. *Transportation Research* **11**, 423-428.

Miller, A.J. (1963) A computer control system for traffic networks. *2nd Intern. Symposium on Traffic Theory,* London, UK, 200-220.

Moreno-Banos, J.C., Papageorgiou, M. and Schäffner, C. (1993) Integrated optimal flow control in traffic networks. *European Journal of Operations Research* **71**, 317-323.

Morin, J.-M. (1995) Aid-to-decision for variable message sign control in motorway networks during incident condition. *Proc. 4th ASCE Intern. Conference on Applications of Advanced Technologies in Transportation Engineering,* Capri, Italy, 378-382.

Papageorgiou, M. (1980) A new approach to time-of-day control based on a dynamic freeway traffic model. *Transportation Research* **14B**, 349-360.

Papageorgiou, M. (1983) *Application of Automatic Control Concepts to Traffic Flow Modeling and Control.* Springer Verlag, New York.

Papageorgiou, M. (1984) Multilayer control system design applied to freeway traffic. *IEEE Trans. on Automatic Control* **29**, 482-490.

Papageorgiou, M. (1990) Dynamic modeling, assignment, and route guidance in traffic networks. *Transportation Research* **24B**, 471-495.

Papageorgiou, M. (1994) An integrated control approach for traffic corridors. *Transportation Research* **3C**, 19-30.

Papageorgiou M. (1998) Automatic control methods in traffic and transportation. In *"Operations Research and Decision Aid Methodologies in Traffic and Transportation Management",* P. Toint, M. Labbe, K. Tanczos, G. Laporte, Editors, Springer Verlag, New York, 46-83.

Papageorgiou, M. (1999) *Comments on Annex B of D10.3 and a Suggestion for a New Travel Time Estimator.* Unpublished Working Paper.

Papageorgiou, M., Blosseville, J.-M and Hadj-Salem, H. (1990) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris - Part II: Coordinated on-ramp metering. *Transportation Research* **24A**, 361-370.

Papageorgiou, M., Gower, P., Messmer, A. and Morin, J.M. (1994) Control strategies for variable message signs. *1st World Congress on Applications of Transport Telematics & IVHS,* Paris, France, 1229-1236.

Papageorgiou, M., Hadj-Salem, H., Blosseville, J.-M. (1991) ALINEA: A local feedback control law for on-ramp metering. *Transportation Research Record* **1320**, 58 - 64.

Papageorgiou, M., Haj-Salem, H., Middelham, F. (1998) ALINEA local ramp metering: Summary of field results, *Transportation Research Record* **1603**, 90-98.

Papageorgiou, M. and Mayr, R. (1982) Optimal decomposition methods applied to motorway traffic control. *International Journal of Control* **35**, 269-280.

Papageorgiou, M. and Messmer, A. (1991) Dynamic network traffic assignment and route guidance via feedback regulation. *Transportation Research Record* **1306**, 49-58.

Park, E.S., Lim, J.H., Suh, I.H. and Bien, Z. (1984) Hierarchical optimal control of urban traffic networks. *International Journal of Control* **40**, 813-829.

Pavlis, Y. and Papageorgiou, M. (1999) Simple decentralized feedback strategies for route guidance in traffic networks. *Transportation Science,* to appear.

Payne, H.J., Meisel, W.S. and Teener, M.D. (1973) Ramp control to relieve freeway congestion caused by traffic disturbances. *Highway Research Record* **469**, 52-64.

Rathi, A.K. (1988) A control scheme for high traffic density sectors. *Transportation Research* **22B**, 81-101.

Reiss, R.A., Gartner, N.H. and Cohen, S.L. (1991) Dynamic control and traffic performance in a freeway corridor: A simulation study. *Transportation Research* **25A**, 267-276.

Robertson, D.I. (1969) TRANSYT method for area traffic control. *Traffic Engineering & Control* **10**, 276-281.

Sarachik, P.E. and Özgüner, Ü. (1982) On decentralized dynamic routing for congested traffic networks. *IEEE Trans. on Automatic Control* **27**, 1233-1238.

Saxton, L. and Schenck, C. (1977) Diversion and corridor control systems in Western Europe. In *"World Survey on Current Research and Development on Roads and Road Transport",* International Road Federation, Washington, D.C., 692-723.

Schwartz, S.C. and Tan, H.H, (1977) Integrated control of freeway entrance ramps by threshold regulation. *Proc. IEEE Conference on Decision and Control,* 984-986.

Sen, S. and Head, L. (1997) Controlled optimization of phases at an intersection. *Transportation Science* **31**, 5-17.

Singh, M.G. and Tamura, H. (1974) Modelling and hierarchical optimisation of oversaturated urban traffic networks. *International Journal of Control* **20**, 269-280.

Smulders, S. (1990) Control of freeway traffic flow by variable speed signs. *Transportation Research* **24B**, 111-132.

Stathopoulos, A. (1991) Route guidance, collective, In *"Concise Encyclopedia of Traffic and Transportation Systems",* M. Papageorgiou, Editor, Pergamon Press, Oxford, UK, 412-416.

Stephanedes, Y. and Chang, K.-K. (1993) Optimal control of freeway corridors. *ASCE Journal of Transportation Engineering* **119**, 504-514.

Tabac, D. (1972) A linear programming model of highway traffic control. *Proc. 6th Annual Princeton Conference on Information Science and Systems,* Princeton, New Jersey, 568-570.

Van Aerde, M. and Yagar, S. (1998) Dynamic integrated freeway/traffic signal networks: Problems and proposed solutions. *Transportation Research* **22A**, 435-443.

Van Vuren, T. (1991) Signal control and traffic assignment. In *"Concise Encyclopedia of Traffic and Transportation Systems",* M. Papageorgiou, Editor, Pergamon Press, Oxford, UK, 468-473.

Vincent, R.A. and Young, C.P. (1986) Self-optimizing traffic signal control using microprocessor: The TRRL "MOVA" strategy for isolated intersections. *Traffic Engineering and Control* **27**, 385-387.

Wang, C.F. (1972) On a ramp-flow assignment problem. *Transportation Science* **6**, 114-130.

Wang, J.J. and May, A.D. (1973) Computer model for optimal freeway on-ramp control. *Highway Research Record* **469**, 16-25.

Wardrop, J.G. (1952) Some theoretical aspects of road traffic research. *Proc. Inst. Civil Engineers,* Part II **1**, 325-362.

Wattleworth, J.A. (1965) Peak-period analysis and control of a freeway system. *Highway Research Record* **157**, 1-21.

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

Wie, B., Tobin, R., Bernstein, D. and Friesz, T. (1995) Comparison of system optimum and user equilibrium dynamic traffic assign ment with schedule delays. *Transportation Research* **3C,** 389-411.

Wisten, M.B. and Smith, M.J. (1997) Distributed computation of dynamic traffic equilibria. *Transportation Research* **5C,** 77-93.

Yuan, L.S., Kreer, J.B. (1968) An optimal control algorithm for ramp metering of urban freeways. *Proc. 6th IEEE Annual Allerton Conference on Circuit and System Theory.* University of Illinois, Allerton, Illinois.

Yuan, L.S. and Kreer, J.B. (1971) Adjustment of freeway ramp metering rates to balance entrance ramp queues. *Transportation Research* **5**, 127-133.

Zackor, H. (1972) Beurteilung verkehrsabhängiger Geschwindigkeitsbeschränkungen auf Autobahnen. *Forschung Strassenbau und Strassenverkehrstechnik* **128**, 1-61.

Zackor, H. and Balz, W. (1984) Verkehrstechnische Funktionen des Leitsystems. *Strassenverkehrstechnik* **28**, 5-9.

Zhang, H., Ritchie, S. and Recker, W. (1996) Some general results on the optimal ramp metering control problem. *Transportation Research* **4C**, 51-69.

This page intentionally left blank

9 CONTINUOUS SPACE MODELLING

Tönu Puu and Martin Beckmann

9.1 Introduction

Since the advent of linear programming and nonlinear programming modelling of transportation and location has meant that space is treated in terms of subscripted variables where the subscripts represent points in space enumerated in an arbitrary manner. The geometric image of what is being modelled is lost in this process.

The older tradition of visualizing space as a 2-dimensional continuum persists, however, not only in theoretical geography, but also in some operations research models (Kantorowich 1942, Beckmann 1952). By its appeal to geometric intuition this approach permits results to be more readily checked against reality.

Moreover optimization in a continuous medium is a more highly developed technique compared to discrete optimization. But, as this chapter will show, the difficulties can still be formidable. A distinction between personal and goods transportation although basically of an economic nature, is forced upon us also by purely mathematical considerations (see section 9.4).

While continuous space modelling turns out to produce explicit answers only in rather special cases, the approach is valuable nevertheless in strengthening our intuition and enabling a deeper understanding of the general nature of transportation and location problems and their solution.

The reader may still ask, why continuous analysis of spatial phenomena, if more "realistic" models with finite sets of locations, linked by a graph or network, will do? Several answers can be given. In a general perspective, the case is quite similar to that of the nature of matter. The consideration of matter as an infinitely divisible fluid and as atomistic are outlooks that both go back to Greek antiquity. Over the history of the natural science both have been pursued in parallel. In hindsight it would be impossible to say, and meaningless to ask, whether modelling hydrodynamics in terms of the Navier-

Stokes equations has been more or less fruitful than for instance quantum mechanics. The outlooks have been complements rather than substitutes, and now and then there has even been unexpected cross-fertilization between them.

The recent history of dynamical systems provides a good example. From the 17th Century on, differential equations have been "the" tools for modelling causal processes of change over time. During the last two centuries, the partial differential equations, involving both space and time, were the glory of analysis, and provided closed form solutions to such phenomena as the progress of electromagnetic waves, the diffusion of heat, or the formation of charged potentials.

When economists started to study dynamic phenomena by models such as the “cobweb" market dynamics, the Cournot duopoly, or the multiplier-accelerator model of the business cycle, discrete iterated processes were chosen, which could make a "simplistic" or "home made" impression when compared to the accumulated knowledge of both ordinary and partial differential equations. Mathematicians had only been interested in such models at an abstract level, as iterated maps in for instance complex analysis, without any connotation of dynamic processes at all. However, while economists, knowledgable in mathematical physics, more and more shifted over to modelling continuous time processes, modern mathematics took a leap in the reverse over to discrete models. The tendency today is to pass from partial to ordinary differential equations, and further from the latter to iterated maps on so called Poincaré sections. Knowledge is accumulating about models of maps, whereas partial differential equation theory hardly makes any progress at all. No doubt this has much to do with the emergence of the digital computer. So, it is even more difficult to say what is the most fruitful in the long run in foresight than it is in hindsight.

As for the issue of realism, it all is a matter of resolution. On the level of a detailed city map no doubt every piece of land is occupied for a particular purpose, a residence, an industry or a piece of infrastructure, whereas atomistic vehicles only move on given kinked tracks on some grid of roads or streets. However, in a coarser resolution it makes sense to speak of fractional densities in land use for residences, industry, and infrastructure. We can also conceive of the flow of vehicles, where some enter and some leave all along the route, as being continuously changing in volume, and we can even bundle the kinked tracks together in continuous curves.

Flows and metrics (for actual distance, travel time, or generalized cost) are concepts which themselves do not discriminate between the discrete and the continuous. It is when we assume flows as continuous and smooth, and the metric as isotropic (direction-independent), that modelling becomes constrained. As we said, this is a matter of resolution and focus. It is actually surprising how rare very regular networks, such as a Manhattan or a ring-radial, are, and how good, due to random irregularities, an isotropic approximation is, once we consider any substantial geographical area. The equidistant points almost lie on a circle with the origin at the centre. The metric however is not Euclidean, as there always remains a detour factor, which is larger the more sparse the network is. The process of investing in the network reduces detours, subject to decreasing returns. At first a new short-cut, strategically placed, makes a big difference, but in an already dense network any new short-cut matters little. Considering generalized cost in terms of travel time, rather than distance alone, we should also take into account that road investments can be for increasing the capacity of already existent stretches of the network, so reducing delays due to congestion. But this does not change the conclusion that the isotropic approximation for local transportation cost is fairly reasonable.

If we want to say something in particular in favor of continuous modelling, it is related to the visual and intuitive pertaining to geometrical shape, as said earlier. Matrices for incidence, distance, cost, or flow, for a discrete set of locations, are abstract and have nothing visual to them. Classical spatial economic theory associated with the names of von Thünen, Launhardt, Weber, Christaller and Lösch, on the contrary, had a very strong visual content which was favorable to intuitive understanding. The shortcoming was that the theory was tied to a linear format due to the assumption of an Euclidean metric. To free the theory from the limitations of this linearity and yet keep the visual and geometric content may be said to be the very purpose of any programme for a modern continuous analysis of the space economy.

Shortest Paths

In modelling travellers' choices the basic assumption is that of rational behavior, in line with general economic theory. Concretely this means that a trip is demanded when and only when its utility exceeds its cost and that routes are chosen that minimize transportation cost from origin to destination. In general transportation cost is composed of various elements, including time and money cost. Often it is simply time that constitutes the relevant cost. Both time and money cost will depend on, in fact increase, with the physical length of the route from origin to destination. To measure this length we must specify the elements of distance, i.e. the metric of the space in which routes are embedded. (See section 9.5).

When nothing further is said, in "continuous space modelling of transportation", the metric assumed is that of 2-dimensional Euclidean space. A shortest path is then a straight line segment joining origin and destination.

This is no longer the case when two features are considered that are important in certain scenarios:

1) Topographic conditions that influence local and/or directional transportation cost.

2) Traffic congestion that raises transportation cost.

Consider first the "isotropic" case where local transportation cost depends only on location (x,y) and not on the direction. Let a route between origin $a = (x(s_0), y(s_0))$ and destination $b = (x(s_1), y(s_1))$ be paremeterized as $(x(s), y(s))$. Transportation cost for this route is then:

$$I(a,b)=\int_{s_0}^{s_1} k(x(s),y(s))\sqrt{x'^2+y'^2}ds \tag{9.1}$$

where $x'=dx/ds$ and $y'=dy/ds$ are the derivatives of the coordinate functions.

When local transportation cost depends also on direction $k(x,y,x',y')$ we have instead

$$I(a,b)=\int_{s_0}^{s_1} k(x(s),y(s),x'(s),y'(s))\sqrt{x'^2+y'^2}ds \tag{9.2}$$

To find a shortest path between *a* and *b* is a classical problem in the calculus of variations. To minimize (9.1) with respect to the *functions* $x(s)$ and $y(s)$, the functions $x(s)$ and $y(s)$must satisfy the Euler equations:

$$\frac{\partial k}{\partial x}\sqrt{x'^2+y'^2}-\frac{d}{ds}\left(k\frac{x'}{\sqrt{x'^2+y'^2}}\right)=0 \tag{9.3}$$

$$\frac{\partial k}{\partial y}\sqrt{x'^2+y'^2}-\frac{d}{ds}\left(k\frac{y'}{\sqrt{x'^2+y'^2}}\right)=0 \tag{9.4}$$

Whenever we can parameterize the route by one of the coordinates, as $y(x)$ or $x(y)$, we have just one single equation:

$$\frac{\partial k}{\partial y}\sqrt{1+y'^2}-\frac{d}{dx}\left(k\frac{y'}{\sqrt{1+y'^2}}\right)=0 \tag{9.5}$$

In the following application we assume circular symmetry and therefore use polar coordinates r,ω so that $x=r\cos\omega$ and $y=r\sin\omega$. Suppose we can parameterize the route by the angular coordinate ω which is always possible whenever there are no radial segments in the route.

Then we must minimize

$$I=\int_{\omega_0}^{\omega_1} k(r)\sqrt{r^2+r'^2}d\omega \tag{9.6}$$

where the path is $r(\omega)$ and $r' = dr/d\omega$. The appropriate Euler equation reads:

$$\frac{\partial}{\partial r}\left(k\sqrt{r^2+r'^2}\right)-\frac{d}{d\omega}\left(\frac{\partial}{\partial r'}\left(k\sqrt{r^2+r'^2}\right)\right)=0 \tag{9.7}$$

As the integration variable ω does not appear explicitly we obtain the first integral

$$k\sqrt{r^2+r'^2}-r'\frac{\partial}{\partial r'}\left(k\sqrt{r^2+r'^2}\right)=c \tag{9.8}$$

where c is an arbitrary integration constant. The expression simplifies to

$$k\frac{r^2}{\sqrt{r^2+r'^2}}=c \tag{9.9}$$

determining the angular derivative r' for each distance r. This procedure is appropriate whenever $k(r)$ is a function ofthe radius vector alone, such as the cases dealt with below: $k = 1 / r$, when the routes become logarithmic spirals, and $k = r$, when the routes become equilateral hyperbolas.

Iso-vectures

From a given origin a, consider the shortest paths to destinations in various directions.

The set of points that can be reached at distance r - or more precisely by an expenditure of transportation cost I- is a closed curve called an iso-vecture (Launhardt, 1885). When transportation costs are constant or a function of only the radial distance from the origin a, then the shortest paths are straight lines and the iso-vectures are circles.

A similar construction yields shortest paths to a destination b and iso-vectures describing sets oforigins at a given distance or a given expenditure in transportation cost to the point ofdestination. When transportation costs are isotropic, i.e. independent or direction, the shortest paths intersect the iso-vectures at right angles.

In an efficient continuous flow field (see below), for any two points on the same flow line, the segment ofthe flow linejoining them is a shortest path.

Market Areas

In the case of goods transportation the object considered is not trips but trade between points of supply and points of demand.

Transportation of a good from a single origin to destinations continuously distributed in the two-dimensional plane was first studied by W Launhardt (1885). At the single point as origin, a producer is located supplying a good to customers located throughout a two-dimensional market area. Quantities demanded depend on the local price which rises by the amount of transportation costs from the origin. In a homogeneous environment, this means that prices are a linear function of distance from the supplier

$$p(r) = p_M + kr \tag{9.10}$$

where $p(r)$ is local price, p_M is price at the supplier, the "mill price", k is the rate of transportation cost, the freight rate, and r is distance from the supplier.

With demand a linear function of price (customarily assumed in location theory) and constant customer density, quantity sold decreases linearly with distance. At a critical distance R, the market radius, it falls to zero. With a linear demand function for quantity q

$$q(p) = a - p \tag{9.11}$$

This market radius is

$$R = \frac{a - p_M}{k} \tag{9.12}$$

using (9.11).

Transportation Demand. Regardless of the shape of the demand function $q(p)$ there exists a simple relationship between volume and ton-miles of transportation demanded and the freight rate k. Total sales of a producer, and accordingly total volume of transportation, V, originating from the supplier are

$$V = \int_0^R q(p_M + kr) 2\pi\rho r dr \tag{9.13}$$

where ρ is the density of customers, assumed to be the same everywhere. Let a be the maximal price at which quantity demanded is zero. Transforming variables

$$V = 2\pi\rho\int_{0}^{R} q(p_M + kr)r\,dr = \frac{2\pi\rho}{k^2}\int_{p_M}^{a} q(t)(t - p_M)dt = \frac{\text{const}}{k^2} \tag{9.14}$$

In economic parlance, the elasticity of demand for transportation is 2.

When ton-miles are considered

$$T = 2\pi\rho\int_{0}^{R} f(p + kr)r^2 dr = \frac{2\pi\rho}{k^3}\int_{p_M}^{a} f(t)(t - p_M)^2 dt \tag{9.15}$$

Their elasticity is therefore 3.

As Launhardt has observed (1885): Railways reduced the freight rate for overland transportation to one tenth. This made possible a hundredfold increase in transportation volume and a thousandfold increase in ton-mileage.

Transportation demand is not independent of the system of spatial pricing used by firms. Two different systems are widely practised: f.o.b. or mill pricing, with transportation charges added to a mill price p_M and c.i.f. or uniform delivered prices p_U, where prices charged to buyers are the same at all buyer locations.

With linear demand functions $q(p) = a - p$ under mill pricing ton miles of transportation demanded are

$$T_M = \int \left(a - (p_M + kr)\right)2\pi\rho r^2 dr \tag{9.16}$$

for a given market radius R

Under uniform pricing ton-miles of transportation demanded are

$$T_U = (a - p_U)\int_{0}^{R} 2\pi\rho r^2 dr \tag{9.17}$$

With

$$p_U = p_M + k\bar{r} \qquad \bar{r} = \frac{\int r^2 2\pi\rho dr}{\int r 2\pi\rho dr} \tag{9.18}$$

total quantity sold is easily shown to be the same. Here $\bar{r}$ is average distance, averaged over customers, not sales.

Comparing transportation demands

$$T_U - T_M = \left(a - p_M - k\bar{r}\right)\int r\rho(r)dr - \int\left(a - p_M - kr\right)r\rho(r)dr$$
$$\propto k\left(\int r^2\rho(r)dr\int\rho(r)dr - \left(\int r\rho(r)dr\right)^2\right) \tag{9.19}$$

where we have written $2\pi\rho r = \rho(r)$.

But

$$\int r^2\rho dr\int\rho dr > \left(\int r\rho dr\right)^2 \tag{9.20}$$

by the Cauchy-Schwartz inequality. Thus uniform pricing generates the greater ton-mileage of transportation. The reason is that uniform pricing does not cause quantity demanded per capita to decrease with distance.

This inequality holds for all demand functions provided total quantities sold are equalized (Beckmann 1985).

9.2 Continuous Origins, Point Destination

The case of point destinations with continuous origins is mathematically similar to that of point origins and continuous destinations. The differences are in the economic applications.

Supply Areas

When demand is located in a single point and suppliers are distributed continuously they occupy a supply area.

The best studied example is that of the agricultural hinterland supplying food to a city (von Thünen 1826). Farmers receive the market price in the city minus transportation cost. The limit of the supply area is reached at the distance R where transportation costs exhaust the product price

$$p = kR \tag{9.21}$$

When yields ρ per area are constant, the quantity supplied is proportional to the square of the market radius.

$$Q = \pi\rho R^2 \tag{9.22}$$

and the price at which this is supplied is determined by (9.21). Thus

$$p = \frac{1}{k}\sqrt{\frac{Q}{\pi\rho}} \tag{9.23}$$

When quantity Q is defined by a linear demand function is given

$$Q = a - p \tag{9.24}$$

the price is determined by an equilibrium condition

$$a - p = \pi\rho\left(\frac{p}{k}\right)^2 \tag{9.25}$$

or

$$p + \frac{\pi\rho}{k^2}p^2 = a \tag{9.26}$$

which shows market price to increase with transportation cost.

Monocentric City

A single destination for travellers from a continuously extended two-dimensional area is the format of mono-centric city models. The central business district (CBD) as a point destination is assumed to contain all shops and all work places. Density of population or households $\rho(r)$ is either assumed or shown to be decreasing with distance. Each household generates a constant number b of trips to the CBD per period. The volume of traffic entering a circle of radius r is then

$$b\int_r^R \rho(r)2\pi r dr \tag{9.27}$$

where R is the radius of the city, determined by the maximal length of work and shopping trips chosen by households. At the center one has a traffic volume

$$V = 2\pi b \int_0^R \rho(r) r dr \tag{9.28}$$

It has been empirically observed (Colin Clark) and theoretically determined by Alan Wilson that density $\rho(r)$ falls exponentially with distance.

$$\rho(r) = a \mathrm{e}^{-\alpha r} \tag{9.29}$$

Total traffic into the CBD is then

$$V = 2\pi p b a \int_0^R \mathrm{e}^{-\alpha r} r dr \tag{9.30}$$

or simply b times city population P

$$V = bP = a \int_0^R \mathrm{e}^{-\alpha r} 2\pi r dr \tag{9.31}$$

regardless of the density distribution of trip origins.

Now let the frequency of trips per household itself depend on distance from the CBD, say exponentially

$$b(r) = b \mathrm{e}^{-\beta r} \tag{9.32}$$

because, say, shopping trips are made less frequently at longer distances.

The traffic volume to the CBD is then

$$\begin{aligned} V = \int_0^R 2\pi a b \mathrm{e}^{-(\alpha+\beta)r} r dr &= \frac{2\pi a b}{(\alpha+\beta)^2} \int_0^{(\alpha+\beta)R} \mathrm{e}^{-t} t dt \\ &= \frac{2\pi a b}{(\alpha+\beta)^2}\left(1-\left(1+(\alpha+\beta)\right)R \mathrm{e}^{-(\alpha+\beta)R}\right) \end{aligned} \tag{9.33}$$

For very extended cities with $R >> 1$ this is approximately

$$V \approx \frac{2\pi a b}{(\alpha+\beta)^2} \tag{9.34}$$

decreasing as the rates of decline α of population density and β of trip frequency increase. Notice that for $R >> 1$ the ratio of traffic volume to city population

$$P \approx \frac{2\pi a}{\alpha^2} \qquad (9.35)$$

is approximately

$$\frac{V}{P} = \frac{b}{\left(1+\frac{\beta}{\alpha}\right)^2} \qquad (9.36)$$

which is larger, the smaller is β relative to α.

9.3 Continuous Origins and Destinations: One Good

Although travellers and goods will both seek shortest paths, the motivation for their movement differs fundamentally. Traders send goods to locations yielding the highest return of price over transportation cost and procure goods from sources at the lowest sum of price and transportation cost. Travellers' choice of destination is guided by the utility of achieving trip purposes after deduction of transportation cost. The mathematical models to be used in this and the following section will differ sharply.

Spatial Market Equilibrium

When a particular good is both produced and consumed over a continuously extended region, but local supply does not everywhere match local demand, a continuously extended spatial market is needed to organize the distribution of this good. Exports from and imports into the region can be ignored, provided aggregate supply equals aggregate demand in the region. In fact, rather than using supply and demand as separate entities it is sufficient to consider their difference, say,

local excess demand = local demand - local supply

Excess demand should be defined as a density = excess demand per area. The commodity movements are best described by a field of flow vectors $\phi(x,y)$ at locations (x, y) say, whose direction are those of the commodity movements and whose lengths (absolute value) equal the density of flow. (Beckmann 1952).

Now local excess demand q must be covered by local net imports, expressed as the negative flow divergence

$$-\nabla \cdot \phi = q \tag{9.37}$$

where the divergence is defined in component terms as

$$\nabla \cdot \phi = \frac{\partial \phi_1}{\partial x} + \frac{\partial \phi_2}{\partial y} \tag{9.38}$$

This is an equation stating conservation, or the matching of supply and demand. Another condition of spatial equilibrium is that profits have been competed away. This shows up in the determination of the direction of flow. To recover transportation cost, flow must flow in the direction of increasing prices, and the absolute value of the price gradient must equal the rate of transportation cost. This rate is assumed to be isotropic (the same for all directions). Thus

$$k\frac{\phi}{|\phi|} = \nabla p \tag{9.39}$$

where $p = p(x,y)$ is the local price; and in the absence of flows

$$|\nabla p| \leq k \qquad \text{whenever} \qquad \phi = 0 \tag{9.40}$$

An efficient continuous flow field has then the direction of a gradient field. The potential lines for the gradients are then curves of equal price - so-called iso-price lines or iso-tims - for the commodity in question. The efficiency conditions that guide the flows can thus be interpreted in economic terms as conditions for economic arbitrage: transportation costs must be covered exactly by the gain in price between points of purchase and sales.

A gradient field can have no spiral singularities. The only admissible singularities are sources and sinks of possibly infinite densities and saddle points.

It is remarkable that the equilibrium conditions for a spatial market with given excess demands results also from minimizing total transportation cost

$$\min_{\phi} \iint k|\phi| dx dy \tag{9.41}$$

subject to a transportation program

$$\nabla \cdot \phi + q = 0 \tag{9.42}$$

This is an example of welfare maximization under perfect competition, a well known proposition in economic theory.

When excess demand is allowed to depend on local price

$$q(x,y)=q(x,y,p) \tag{9.43}$$

transportation cost minimization is modified to a utility maximization problem in terms of a consumers' surplus function as utility

$$u(q)=\int_0^q p(t,x,y)dt \tag{9.44}$$

where $p(q,x,y)$ is the inverse of the excess demand function $q(p,x,y)$. The functional to be maximized now without constraint is

$$\iint\left(u(\nabla\cdot\phi,x,y)-k|\phi|\right)dxdy \tag{9.45}$$

yielding once more equation (9.39)

$$k\frac{\phi}{|\phi|}=\nabla p$$

as well as

$$p=\frac{\partial u}{\partial q} \tag{9.46}$$

under sthe constraint (9.42)

$$\nabla\cdot\phi+q=0$$

Here as always in transportation planning and traffic routing, the objective is the maximization of net benefits. Cost minimization is a simplification that arises when traffic demand, here $q(x,y)$ is assumed independent of individual travel costs, or prices p.

The integral of minimal transportation cost may be transformed, using Gauss' integral theorem

$$\iint k|\phi|dxdy = \iint \phi \cdot \nabla p dxdy \tag{9.47}$$
$$= \oint p\phi_n ds - \iint p\nabla \cdot \phi dxdy$$
$$= \iint pq dxdy$$

Thus, since the boundary integral vanishes, minimum transportation cost is seen to equal the aggregate value difference of the good between import and export locations. (cf. Beckmann Puu 1985)

The equilibrium model for the exchange of a good by local exports and imports is appropriate for commodities only. In transporting persons it is not meaningful to allow, say, arrivals from any origin to satisfy demand at a destination, as is legitimate in the goods case.

Extension to Several Goods

When continuous spatial markets are considered for several goods $m = 1, \ldots M,$ then total goods traffic through a point (x,y) is the aggregate of flows in absolute terms

$$|\phi(x,y)| = \sum_{m=1}^{M} |\phi_m(x,y)| \tag{9.48}$$

The vector sum of commodity flows is relevant for the movement of empty vehicles since for total vehicle movement to balance

$$\sum_m \phi_m + \psi = 0 \tag{9.49}$$

so that

$$\nabla \cdot \sum_m \phi_m \tag{9.50}$$

represents the vehicle net outflow. An efficient movement ψ, of empty vehicles is achieved by

$$\min_\psi \iint k|\psi| \tag{9.51}$$

subject to

$$\nabla \cdot \left(\psi - \sum_m \phi_m \right) = 0 \tag{9.52}$$

The efficient movement of empty tank ships was the first empirical transportation problem treated by linear programming (Koopmans 1947).

9.4 Transportation of Persons: Continuous Origins and Destinations

The Problem

The transportation of people as compared to the transportation of commodities becomes considerably more complex, because it necessarily involves not one, and not a finite number, but an infinity (even a continuum) of flow fields. Recall that a commodity flow for one spatially extended market is represented by a single flow and collects the local excess supplies on its way and again discharges them at locations of excess demand. The flow lines themselves obey the appropriate Euler equations for minimal cost routing in the continuous field. On the other hand in the continuous plane where each location needs to interact with each other location there is a unique flow field only if we specify a single point of origin, or alternatively a single point of destination. Once both origins and destinations are continuously spread out there is one flow field for each point of origin, or again alternatively each point of destination.

Several complications arise by this. First, given an origin, the discharge of trips at various destinations can still be related to the divergence of that given flow, but the origin itself becomes a singularity, a source of infinite strength, and we need particular measures to deal with the singularity. Second, at any point of the space there is an infinite crisscross of flow lines passing through it in various directions and having different volumes of flow. Each of those flow lines obeys the appropriate Euler equation. For each point of space we might want to evaluate the total traffic passing. This is an issue faced by Angel and Hyman (1976), Puu (1979) and Vaughan (1987). Those traffic measures are interesting for several purposes.

Obviously the load of traffic itself determines the cost of transportation through a point in an urban area by means of congestion effects and resulting delays which are important components in transportation cost. Therefore the traffic pattern resulting from the choice of optimal routes feeds back as a determinant for the choice of routing of the individual trips. We can imagine that if transportation cost is invariant over space, then people would communicate along straight line segments. If people are equally distributed in a closed urban area, then, by the mere geometry of the bounded area, many more trips would pass the center than the periphery, and so traffic would be concentrated to the central parts. Accordingly, if we assume network capacity to be more or less constant over the urban area this would create an immense congestion in the center, and the commuters once they realize this high cost of transfer in the central parts would in the next round avoid the center. This would then mean less concentration of traffic in the center and decrease the tendency of avoiding it. Finally, we could imagine a traffic equilibrium, where the congestion created by the individual route choices is exactly that which agrees with the perception of congestion which leads to this choice of routing.

It must be admitted that the derivation of an exact traffic distribution is analytically tough, and that it has been solved only for very special simplified cases. What the equilibrium traffic distribution exactly looks like in general, and even more whether it is always stable so that we do not just deal with an endless chain of repercussions, are still open questions.

On the other hand the derivation of traffic distributions is an important issue because it is implicit in the solution to problems such as the optimal distribution of limited funds for road capacity in space, and ultimately the design of the optimal city (see section 9.5).

Some Simple Cases

As mentioned we need a lot of simplifications to be able to analyze the traffic distribution problem. The first assumption is that our region is a unit disk with a constant population density which we normalize to unity. Each point in the region needs to interact with each other point and the number of trips generated is supposed to be the product of population densities in the origin and the destination locations. This is according to the gravity or the entropy hypothesis dealt with elsewhere in this book. In both hypotheses there is also an adverse dependence on distance or on transportation cost. We will choose the gravity hypothesis as it is easier to analyse.

We also need a principle for the choice of routes. As already indicated, we follow the general spirit of this chapter and assume that there is a given local isotropic transportation cost in each location, and then we choose the routes by minimizing the path integral of the cost along the route.

Suppose we fix a location of origin denoted by coordinates ξ, η. Then we have a unique flow field of minimum cost paths from that particular origin to all the other locations of the region. Denote the optimal flow field originating from the point ξ, η by $\phi = (\phi_1(x,y), \phi_2(x,y))$.

If we take the gravity hypothesis and denote local population density by $p(x,y)$, then we have the sink density:

$$\nabla \cdot \phi = -\frac{p(\xi,\eta)p(x,y)}{I^{\varepsilon}} \tag{9.53}$$

The factors in the numerator are the population densities in the origin $(\xi,\eta)$and destination (x,y) locations, and the denominator is some power ε of transportation cost along the minimum cost routes as defined by the appropriate Euler equations. Equation (9.53) then renders a new differential equation for flow volume $|\phi|$ which as we will see is a component in traffic.

Let us now take the radially symmetric case and moreover suppose that we have $k(r)=r^n$, where n is a positive or negative power. As Wardrop (1969) has shown, the optimal routes can then be found as geodesics on a cone or a cylinder obtained by a conformal map from the plane region. If the power is positive the routes obtained in this way only apply to a $1/(1+n)$:th sector of the disk, the rest are broken radials. Even most of these cases become so messy that we only retain two illustrative cases: Linear routes over the disk, given $n = 0$, and logarithmic spirals, given $n = -1$.

Case I: Linear Routes. If the routes are linear, then the flow lines radiate from the points of origin ξ, η and we can introduce a new polar coordinate system τ, θ for any given origin

$$x=\xi+\tau\cos\theta \tag{9.54}$$

$$y=\eta+\tau\sin\theta \tag{9.55}$$

From these expressions we easily obtain

$$\tau=\sqrt{(x-\xi)^2+(y-\eta)^2} \tag{9.56}$$

by Pythagoras's Theorem. The flow angle of the linear routes is already defined in equations (9.54)-(9.55) so the direction field accordingly becomes:

$$\frac{\phi}{|\phi|}=(\cos\theta,\sin\theta)=\nabla\tau \tag{9.57}$$

where the equality to the gradient is easily obtained differenting (9.56) with respect to x and y. As it is implicit in the choice of linear routes that the local transportation cost is constant the cost for a trip is just proportionate to the distance function τ, and so we can put $I=\tau$. Moreover assuming unit population density as indicated we have $p(\xi,\eta)\equiv p(x,y)\equiv 1$. So, from (9.53) we get

$$\nabla \cdot \phi = -\frac{1}{\tau^{\varepsilon}} \tag{9.58}$$

However, with the linear routes we have from (9.57)

$$\phi = |\phi| \nabla \tau \tag{9.59}$$

We can directly apply the definition of the divergence operator to (9.59), so that

$$\nabla \cdot \phi = \nabla |\phi| \cdot \nabla \tau + |\phi| \nabla^2 \tau = -\frac{1}{\tau^{\varepsilon}} \tag{9.60}$$

Now, by the formula for the total derivative

$$\nabla |\phi| \cdot \nabla \tau = \frac{d|\phi|}{d\tau} \tag{9.61}$$

and we easily from (9.56) evaluate the Laplacian as

$$\nabla^2 \tau = \nabla \cdot \nabla \tau = \frac{1}{\tau} \tag{9.62}$$

Therefore the condition on the divergence (9.60) just results in an ordinary differential equation

$$\frac{d|\phi|}{d\tau} + \frac{|\phi|}{\tau} + \frac{1}{\tau^{\varepsilon}} = 0 \tag{9.63}$$

which is readily solved and yields

$$|\phi| = \left(\frac{1}{2-\varepsilon}\right)\left(\frac{T^{2-\varepsilon} - \tau^{2-\varepsilon}}{\tau}\right) \tag{9.64}$$

where T is an arbitrary constant of integration. As this constant is raised to the same power as the distance measure it too has the dimension of a distance. In particular if we assume that we deal with the case of an insulated area with no traffic going across its boundary then T becomes the distance from the point of origin ξ, η to the boundary curve in the direction θ, because this makes the traffic intensity $|\phi|$ go to zero at the boundary points.

The above expression (9.64) obviously does not work for the special case $\varepsilon = 2$. In that case we have:

$$|\phi| = \frac{\ln T - \ln \tau}{\tau} \tag{9.65}$$

We have just solved part of the traffic derivation problem. What we got is the traffic through a given point that started in the point ξ, η only. In order to arrive at a measure of all the traffic we have to consider all the different points of origin. We hence have to integrate over all the points ξ, η to obtain

$$f(x,y) = \iint_S |\phi| d\xi d\eta \tag{9.66}$$

as the measure of traffic in the point *x,y,* where *S* denotes the region studied. To evaluate the integral we note that it is wise to revert to the coordinates τ, θ defined above. As the Jacobian is τ, so that we have the substitution $d\xi d\eta = \tau d\tau d\theta$, we can get rid of the denominator in the above expression for the integrand. Hence from (9.64):

$$f(x,y) = \frac{1}{2-\varepsilon} \iint_S \left(T^{2-\varepsilon} - \tau^{2-\varepsilon}\right) d\tau d\theta \tag{9.67}$$

for $\varepsilon \neq 2$. If $\varepsilon = 2$ we likewise have from (9.65)

$$f(x,y) = \iint_S (\ln T - \ln \tau) d\tau d\theta \tag{9.68}$$

It is now time to reap the full profits from the assumption of circular symmetry, i.e. that we deal with a region that is the unit disk. In that case it is relatively easy to evaluate the inner integral in (9.67). We get

$$f(x,y) = \frac{1}{(2-\varepsilon)(3-\varepsilon)} \int_0^{2\pi} \left((U+V)^{3-\varepsilon} - U^{3-\varepsilon} - V^{3-\varepsilon}\right) d\theta \tag{9.69}$$

where

$$U = \sqrt{1 - r^2 \sin^2 \theta} - r\cos\theta \tag{9.70}$$

$$V = \sqrt{1 - r^2 \sin^2 \theta} + r\cos\theta \tag{9.71}$$

are the line segments of the chord of the unit boundary circle through the point *x, y*, located at the distance $r = \sqrt{x^2 + y^2}$ from the origin.

Again (9.69) holds if $\varepsilon \neq 2$. For the case $\varepsilon = 2$ we get from (9.68)

$$f(x,y) = \int_0^{2\pi} \left((U+V)\ln(U+V) - U\ln U - U\ln V\right) d\theta \tag{9.72}$$

We realize that the resulting traffic intensity after integration with respect to the remaining variable θ will only depend on the radius vector *r*. This verifies that the traffic distribution indeed has circular symmetry as expected.

The remaining integral (9.69) can be evaluated numerically. It can also be obtained as a closed form solution for two cases: with $\varepsilon = 1$ and with $\varepsilon = 0$. The first case is the easier one. From (9.70)-(9.71) we have the integrand

$$(U+V)^2 - U^2 - V^2 = 2UV = 2\left(1 - r^2\right) \tag{9.73}$$

and so from (9.69)

$$f(x,y) = \frac{1}{2}\int_0^{2\pi} 2\left(1 - r^2\right) d\theta = 2\pi\left(1 - r^2\right) \tag{9.74}$$

which is a paraboloid turned upside down, with a maximum of traffic in the center, successively reduced to zero at the unit circle boundary. This is natural as in any convex area most straight line routes pass through the center.

The case with $\varepsilon = 0$ is slightly tougher. We then from (9.70)-(9.71) have the integrand

$$(U+V)^3 - U^3 - V^3 = 3UV(U+V) = 6\left(1 - r^2\right)\sqrt{1 - r^2 \sin^2 \theta} \tag{9.75}$$

Accordingly we get

$$f(x,y) = \left(1 - r^2\right)\int_0^{2\pi} \sqrt{1 - r^2 \sin^2 \theta}\, d\theta \tag{9.76}$$

This integral is four times the so called complete elliptic integral ofthe second kind, usually denoted $E(r)$, so finally

$$f(x,y)=4(1-r^2)E(r) \tag{9.77}$$

A numerical expression for the elliptic integral is most easily stated in terms of the power series

$$E(r)=\frac{\pi}{2}\left(1-\left(\frac{1}{2}\right)^2\frac{r^2}{1}-\left(\frac{3}{8}\right)^2\frac{r^4}{3}-\ldots o(r^6)\right) \tag{9.78}$$

Compared to the previous unit exponent the concentration to the center is actually the same, equal to 2π, as we see from (9.77)-(9.78) as compared to (9.74). However, traffic decreases somewhat more steeply with distance from the center before it goes down to zero at the unit circle boundary.

Total traffic created, i.e. the integral of (9.77) over the unit disk, can be calculated to $128\pi/45\approx 8.94$ as compared to $\pi^2\approx 9.87$ for the case with a distance dependence (see Puu 1979). It might seem odd that more traffic is created when there is no adverse effect from distance. The result is, however, not self evident, because distance dependence of the gravity type, though being an obstacle to long distance trips, creates many more local trips. In fact communication "within a point itself" goes to infinity, which has been taken as an absurdity in the gravity model, but it is harmless in this context.

Case II: Spiral Routes. The other case where it is relatively easy to evaluate traffic in closed form is when the optimal routes are logarithmic spirals, or rather half spirals extending over an angle ofhalfthe disk. They arise when we have local transportation cost that is reciprocal to the distance from the center. As the cost goes to infinity at the center this represents avoiding a congested center of the region. The case is also of interest in principle as it illustrates a different mode ofproblem solution.

With $k = 1/r$ the Euler equation (9.9) becomes

$$\frac{r}{\sqrt{r^2+r'^2}}=c \tag{9.79}$$

which implies $r'/r=\text{constant}$ and therefore has the obvious solution

$$r(\omega)=a\,e^{b\omega} \tag{9.80}$$

where $b=1/c^2-1$ and a is a new integration constant. With polar coordinates

$$x = r\cos\omega \tag{9.81}$$

$$y = r\sin\omega \tag{9.82}$$

and using ω as the path parameter we get the direction field

$$x' = r'\cos\omega - r\sin\omega = br\cos\omega - r\sin\omega \tag{9.83}$$

$$y' = r'\sin\omega + r\cos\omega = br\sin\omega + r\cos\omega \tag{9.84}$$

where the primes denote differentiation with respect to ω. The last result is obvious because from (9.80) we have $r' = abe^{b\omega} = br$. From equations (9.83)-(9.84) we get the following expression for the derivative of arc length with respect to the angular parameter

$$\frac{ds}{d\omega} = \sqrt{x'^2 + y'^2} = r\sqrt{1+b^2} \tag{9.85}$$

The unit flow field is therefore obtained through dividing the derivatives (9.83)-(9.84) by (9.85). Thus:

$$\frac{\phi}{|\phi|} = (\cos\theta, \sin\theta) \tag{9.86}$$

$$= \left(\frac{b}{\sqrt{1+b^2}}\cos\omega - \frac{1}{\sqrt{1+b^2}}\sin\omega, \frac{b}{\sqrt{1+b^2}}\sin\omega + \frac{1}{\sqrt{1+b^2}}\cos\omega \right)$$

where θ as before denotes the angle of the route. Note that, unlike the case of linear routes, θ is not an invariant along the route. There however exists a spatial invariant, the scalar product

$$\cos(\theta - \omega) = (\cos\theta, \sin\theta)\cdot(\cos\omega, \sin\omega) = \frac{b}{\sqrt{1+b^2}} \tag{9.87}$$

as can easily be calculated from the expression (9.86) by scalar multiplication with the vector $(\cos\omega, \sin\omega)$. Observe that as ω is the angular coordinate for any position vector, whereas θ is the angle for the route, the statement in (9.87) that the expression $\cos(\theta - \omega)$ be equal to $b / \sqrt{1+b^2}$, i.e. a constant, says that any given route crosses all circles of various radii under the same angle to the radius vector.

In principle we could now, as we know the route directions, go on as before, solve for $|\phi|$ from its appropriate differential equation, and then integrate with respect to all the points of origin. In the present case it however is not so easy to solve that equation, so therefore we illustrate a different strategy of actually circumventing having to solve any differential equation at all.

Consider the condition on the divergence of flow (9.53) given an interaction which we again take as dependent on distance τ traversed and given $p(x,y) \equiv p(\xi,\eta) \equiv 1$.

$$\nabla \cdot \phi = -\frac{1}{\tau^{\varepsilon}} \tag{9.88}$$

Unlike the case of linear routes distance is no longer identical with cost, but we keep the distance dependence for computational reasons.

Given the spiral routes we can easily integrate the distance along any optimal route:

$$\begin{aligned} \tau &= \int_{\omega_0}^{\omega_1} \sqrt{r^2 + r'^2}\, d\omega = \sqrt{1+b^2} \int_{\omega_0}^{\omega_1} a\, e^{b\omega}\, d\omega \\ &= \frac{\sqrt{1+b^2}}{b} a\, e^{b\omega} \Big|_{\omega_0}^{\omega_1} = \frac{\sqrt{1+b^2}}{b} (r_1 - r_0) \end{aligned} \tag{9.89}$$

since $a \exp(b\omega_i) = r_i$. To make things simple we take $\varepsilon = 1$, so that from (9.88) and (9.89) we have

$$\nabla \cdot \phi = -\frac{b}{\sqrt{1+b^2}} \frac{1}{r_1 - r_0} \tag{9.90}$$

From our previous result (9.87) we can also write

$$\nabla \cdot \phi = -\cos(\theta - \omega) \frac{1}{r_1 - r_0} \tag{9.91}$$

Now let us divide the unit disk S in an inner disk S_i and an outer ring S_o by a circle C with radius $0 < r < 1$. Suppose we integrate the divergence of the flow field over he outer ring. From Gauss's Integral Theorem we get:

$$\iint_{S_o} \nabla \cdot \phi\, dxdy = \oint_C \phi \cdot \mathbf{n} ds \tag{9.92}$$

where the right hand line integral is taken along the boundary circle that separates the inner disk S_i from the outer ring S_o. We integrate the outward normal component of the flow on the boundary circle. There is also the outer boundary circle of the whole region S, but we already assumed that there would be no traffic with the exterior of the insulated disk, so this other line integral vanishes. As to the right hand integrand we note that $\mathbf{n} = (\cos\omega, \sin\omega)$, so

$$\phi\cdot\mathbf{n} = |\phi|(\cos\theta,\sin\theta)\cdot(\cos\omega,\sin\omega) = |\phi|\cos(\theta-\omega) \tag{9.93}$$

Now parameterizing the circle C by the angle ω we can express arc length as $ds = rd\omega$. Hence:

$$\oint_C \phi\cdot\mathbf{n}ds = -r\int_0^{2\pi}|\phi|\cos(\theta-\omega)d\omega \tag{9.94}$$

The minus sign comes from the fact that by integrating over the angle ω we traverse the boundary in the negative direction. Using the result in the preceding double integral we get

$$\iint_{S_o}\nabla\cdot\phi dxdy = -r\int_0^{2\pi}|\phi|\cos(\theta-\omega)d\omega \tag{9.95}$$

Suppose now that we had started from integrating $\sec(\theta-\omega)\nabla\cdot\phi$, an expression which will be given an intuitive interpretation below, instead of just $\nabla\cdot\phi$. Then the (9.95) would become

$$\iint_{S_o}\sec(\theta-\omega)\nabla\cdot\phi dxdy = -r\int_0^{2\pi}|\phi|d\omega \tag{9.96}$$

Finally, integrating (9.96) over all points of origin in the inner disk S_i we get

$$\begin{aligned}\iint_{S_i}\iint_{S_o}\sec(\theta-\omega)\nabla\cdot\phi dxdyd\xi d\eta &= r\int_0^{2\pi}\left(\iint_{S_i}|\phi|d\xi d\eta\right)d\omega \\ &= 2\pi rf(x,y)\end{aligned} \tag{9.97}$$

We hence arrived at a direct estimate oftraffic without ever solving the differential equation for $|\phi|$. At least this expression is correct when all routes are outward, starting in the inner disk and ending in the outer ring.

The method is more general and works in a large number of cases than might be apparent. What we do in the left hand side is to "count" trips in terms ofthe divergence $\nabla \cdot \phi$. In the right hand side we then get traffic intensity at the distance r multipliedby the perimeter of the circle at that distance.

Angel and Hyman(1976) introduced this procedure as "cordon crossing". The idea is that we imagine a circle or "cordon" at a given distance from the center and count the number of trips crossing that cordon. In order to arrive at the average density of crossing (i.e. traffic) we obviously have to divide by the perimeter of the circle (the length of the cordon).

As shown in Puu (1979), however, we cannot just count the number of trips in terms of $\nabla \cdot \phi$. From the above formula we see that we also have to use weights and integrate $\sec(\theta - \omega)\nabla \cdot \phi$ instead of just $\nabla \cdot \phi$. This makes intuitive sense. If the cordon is a thin ring instead of just a circle, then the length of any trip through it and hence the load oftraffic on the ring caused by it depends on the angle ofincidence. A radial route implies minimal load, whereas a tangential route implies maximal load. In the limit as the ring collapses to a circle, $\sec(\theta - \omega)$ is exactly the weight we need. The corrected procedure leads to a measure in terms of weighted cordon crossing.

In terms of our spiral routes we had $\nabla \cdot \phi = -\cos(\theta - \omega) / (r_1 - r_0)$ from (9.91),and therefore the appropriate integrand is as handy as $\sec(\theta - \omega)\nabla \cdot \phi = 1/(r_1 - r_0)$.

Before putting this to use in (9.97) our previous formula we have to consider the following facts. First, we should not integrate for the destination points over an angle larger than π, because no optimal spiral routes extend over a larger angle, a clockwise route then being better than a counter-clockwise or vice versa. The integration over only half the complete circle also means that we have to double the traffic measure obtained, because otherwise we take care of counter clockwise trips only. There are equally many clockwise. Second, we only discuss trips originating in the inner disk and ending in the outer ring. There are also equally many trips in the reverse direction. Therefore we have to double the measure once more.

In summary we have

$$4\iint_{S_1}\iint_{S_0} \frac{1}{r_1 - r_0} dxdyd\xi d\eta = 2\pi r f(x,y) \tag{9.98}$$

Reverting to polar coordinates we have the substitutions $dxdy = r_1 dr_1 d\omega_1$ and $d\xi d\eta = r_0 dr_0 d\omega_0$. Accordingly

$$4\int_0^{2\pi}\int_0^{r}\int_0^{\pi}\int_r^{1}\frac{r_0 r_1}{r_1 - r_0}dr_1 d\omega_1 dr_0 d\omega_0 = 2\pi r\, f(x,y) \tag{9.99}$$

The integrand is however independent of the angular coordinates so we can evaluate those two integrals to equal $2\pi^2$. Hence

$$f(x,y) = \frac{4\pi}{r}\int_0^{r}\int_r^{1}\frac{r_0 r_1}{r_1 - r_0}dr_1 dr_0 \tag{9.100}$$

This can finally be evaluated in closed form. We now see the reason why we used distance and not transportation cost to determine the adverse effect in the gravitation law. Had we taken the more logical alternative then the difference in the denominator of the integrand would have become a difference in the logarithms of the radii, and then no closed form solution would be at hand. Anyhow, the final result is:

$$f(x,y) = \frac{4\pi}{3}\frac{r(r-1) - r^3\ln r + \left(r^3 - 1\right)\ln(1-r)}{r} \tag{9.101}$$

Unlike the two previous distributions derived in (9.74) and (9.77) there is no peak load in the center of the region. The center is avoided by the spiral routes and therefore peak load is at an intermediate distance between center and periphery. Peak load also is much lower than in the previous cases, but as it is spread out over a whole ring rather than concentrated in a point total traffic generated is again $\pi^2 \approx 9.87$ just as in the case of linear routes and unit distance elasticity of interaction.

The direct method illustrated in this section is applicable to a wide range of cases and is therefore of much more interest than might be imagined from the case illustrated. Logarithmic spirals are particularly simple as they are monotonously inward or outward and so cross each cordon circle only once. It is, however, possible to deal with cases where routes originating in the outer ring just pass through the inner disk and again end in the outer ring (as in the case of linear routes), or routes which originate in the inner disk, pass through the outer ring and re-enter. We only have to count the crossings by such trips twice. The method is useful for simulations, since very few cases are possible to treat in terms of closed form solutions.

As expected, and unlike the previous distributions which result in a central concentration of traffic, the present distribution represents a distribution of traffic which has a maximum intensity at an intermediate distance from center to periphery. There is very small load in the center, which is in contradiction with the assumed distribution of transportation costs for the choice of the logarithmic routes (according to which the center was infinitely congested). The reverse was the case with linear routes which arose under an assumption of constant transportation costs, but led to considerable congestion in the center.

Therefore the case leading to traffic equilibrium is intermediate, but unlike these extreme cases it is so complex as to defy closed form solution. Anyhow, we get a feeling for the type of analysis by studying these extreme and ultimately simplified cases.

9.5 Transportation Cost Metrics

General Considerations: Real Metric Spaces

It is now important that we realize the following fact: Any system of transportation possibilities as embodied in a network can be fully represented by a suitable transportation cost metric.

The most natural metric is the Euclidean, which is implicit in all classical location and land use analysis. The connections are then always straight lines, and the constant distance (cost) loci become concentric circles. We should note that it is impossible to literally build a physical network in this manner, because the roads would have to go everywhere in all directions. So the entire region would have to be paved with roads, and there would have to be an infinite number of junctions, where an infinity of roads meet. The Euclidean distance function is, as we know:

$$\tau = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{9.102}$$

where x_i, y_i are the locations of the points to be connected.

A general distance metric $\tau(x_1, y_1, x_2, y_2)$ is any nonnegative function, such that: (i) The distance is zero if and only if $(x_1, y_1) = (x_2, y_2)$, (ii) The distance from x_1, y_1 to x_2, y_2 equals the distance from x_2, y_2 to x_1, y_1, (iii) The distance from x_1, y_1 to x_3, y_3 does not exceed the sum of the distances from x_1, y_1 to x_2, y_2 and from x_2, y_2 to x_3, y_3 for any x_2, y_2. (The triangle inequality.)

Another example, which in contrast to the Euclidean Metric is possible to construct in reality, at least in finite mesh, is the rectangular grid or Manhattan Metric. The distance along such an idealized regular network is the sum of the horizontal and vertical distances taken separately.

A purely mathematical consequence of adding the horizontal and vertical distances separately is that we can move farther - in an Euclidean sense - if we move a certain network distance East-West or South-North, than if we move in a diagonal direction. As a matter of fact all points of equal network distance are found to lie on a square tilted 45 degrees. A family of such concentric tilted squares is the equivalent to the circles of the Euclidean metric.

Note that, with a finite grid width, we do not reach all the points, just equally spaced points on it. Moreover, it should be noted that, with a finite mesh density the density of reachable dots on the tilted squares can only be increased in finite steps. The smaller the mesh, the more points can be reached - and in the limit of vanishing mesh size all points are reachable.

The form of the distance function for the Manhattan Metric is:

$$\tau = |x_1 - x_2| + |y_1 - y_2| \tag{9.103}$$

It is worth noting that the Manhattan Metric belongs to the same general family as the Euclidean. The family name is Minkowski Metrics, and the general form is:

$$\tau = \left(|x_1 - x_2|^\gamma + |y_1 - y_2|^\gamma\right)^{\frac{1}{\gamma}} \tag{9.104}$$

For the Euclidean case we have $\gamma = 2$, for the Manhattan $\gamma = 1$. When $\gamma < 1$, the isodistance loci become pointed in the horizontal and vertical directions. They could mimic a case with efficient trunk line routes in the East-West and North-South directions, and a system of minor subsidiary roads that make points outside the network accessible, though considerably less efficiently, and therefore at higher cost.

Common to all those metrics is a strong directional dependence confined to two directions at right angles. Instead of a square grid we could, of course, also think of a triangular grid, based on three directions intersecting at angles of 60 degrees. The constant network distance loci have a hexagonal shape.

The metric for a triangular network can formally be written as:

$$\tau = \left(|u| + |v| + |w|\right) \tag{9.105}$$

where

$$u = (x_1 - x_2) + \sqrt{3}(y_1 - y_2) \tag{9.106}$$

$$v = (x_1 - x_2) - \sqrt{3}(y_1 - y_2) \tag{9.107}$$

$$w = 2(x_1 - x_2) \tag{9.108}$$

The distance loci (or iso-vectures) in this case become hexagonal. If we wish we can also make a Minkowski like metric out of this hexagonal structure, by putting:

$$\tau = \left(|u|^{\gamma} + |v|^{\gamma} + |w|^{\gamma}\right)^{\frac{1}{\gamma}} \tag{9.109}$$

In particular, with $\gamma = 2$ we get the Euclidean circles back.

Regular networks are of interest as idealized cases when regions with uniformly distributed population and economic activity have to be provided with means of transportation. There is a theorem, known already to Launhardt (1889), that at junctions in an optimally designed network the roads of equal capacity and equal traffic load must meet at equal angles. Consequently, the regular networks come out as natural candidates.

In the literature much has been written about network design when we consider transport in terms of collection or discharge between a center and its surroundings. The appropriate design for such a network is the ring-radial type. Mosler (1987) dealt extensively with those.

Historically this has been the typical layout for city regions with a central market place and a round city wall (by the way a solution to the isoperimetric problem because a minimum perimeter also provided for maximum defence efficiency). The natural communication system for such a city was in terms of radials and orthogonals to those, i.e. more or less concentric rings, as in a cobweb. In modern times it was the successive walls, moats, and fortresses of growing cities that provided the space for the spacious ring-roads which more or less solved the traffic problem of the modern city.

Like any other network, the ring-radial one has a finite mesh density, but there is no harm in finding out its geometric properties, i.e. its metric, by assuming the mesh infinitely dense.

Unlike the Euclidean, the Manhattan, and the other uniform networks discussed, the ring-radial system has a given spatial layout, so we cannot just put the point of departure at the origin of the coordinate system, because this would result in a simplified special case.

Hence, we have to specify both endpoints of a path, say r_1, θ_1 and r_2, θ_2 in polar coordinates. It is now obvious that the optimal path may follow two different principles. It can follow one radial from the point of origin right into the centre, and another radial right out to the destination. In that case the distance is:

$$\tau = r_1 + r_2 \tag{9.110}$$

The route can also traverse only the radial difference

$$|r_1 - r_2| \tag{9.111}$$

but then it must be combined with a path on a ringroad corresponding to the angular difference $|\theta_1 - \theta_2|$. The distance traversed in the latter transit will be

$$\min(r_1,r_2)|\theta_1-\theta_2| \tag{9.112}$$

because the angular transit will be shortest at the smallest radius. Moreover, we have to check that the angular difference is less than π, otherwise we have to take the complementary angle. This means going clockwise or counter-clockwise, whichever is shortest. The distance by the combined radial and angular transit is:

$$\tau=|r_1-r_2|+|\theta_1-\theta_2|\min(r_1,r_2) \tag{9.113}$$

Equating the distance expressions we get:

$$|\theta_1-\theta_2|=2 \tag{9.114}$$

For an angular difference smaller than 2, the shortest route is a combined ring-radial transit, for larger difference the transit is purely radial.

It should be emphasised that the different metrics can be combined or nested: For instance a square network of main roads of finite mesh with capillary Euclidean distance feeding lines as already mentioned, or a droplike metric of airports with an Euclidean or a Manhattan metric. See Puu (1997) for details. The different modes easily combine to a new metric.

Refraction of Traffic. A combination of metrics is also present when traffic moves to a major highway from the surrounding area on which a Euclidean metric is assumed. Transportation cost per mile k_0 in this area will exceed transportation cost per mile k_1 on the highway.

$$k_1<k_0 \tag{9.115}$$

The question is in what direction, that is at what angle, should traffic approach the highway.

This seemingly marginal problem has generated a literature in economic theory, starting with Wilhelm Launhardt (1887, p. 20-21) and continued by Tord Palander (1935), Heinrich von Stackelberg (1938), August Lösch (1940, 1954) and Karl Mosler (1987).

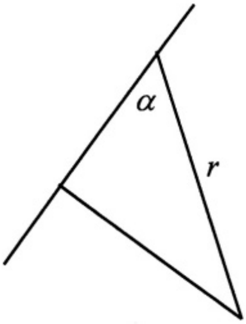


Figure 9.1 Refraction of traffic

When traffic enters at an angle α the saving in transportation cost is (Fig. 9.1) $r(k_1 \cos\alpha + k_0 \sin\alpha - k_0)$.

Maximizing with respect to α requires

$$0 = -k_1 \sin\alpha + k_0 \cos\alpha \tag{9.116}$$

$$\tan\alpha = \frac{k_0}{k_1} \tag{9.117}$$

As the ratio k_0 / k_1 increases, α approaches a right angle.

According to Lösch the feeder lines to the Great Northern Railroad in North Dakota have this angular orientation (Lösch 1954).

Further Remarks. The three types of metrics discussed, the square, triangular, and ring-radial, share the property of having a strong directional dependence. In a continuous setting it is not impossible to deal with such ansisotropy, though the approach is much more powerful without a dependence on direction. It therefore is a good thing that most road networks in reality if extended over larger areas than a city centre tend to have such a strong stochastic element that any systematic directional influence from a strict regularity vanishes. The isotropic approximation, i.e. an "Euclidean" metric therefore becomes pretty good.

Computational Aspects: Virtual Metric Space

The metrics inherent in a transportation network is one thing. Another is the transformation ofthe real geographical space to a virtual space for analytical or computational purposes. Wardrop (1969) considered the transformation ofthe region to another (flat) and Tobler (1963) to another (curved) space, Wardrop using complex analytical functions, and Tobler using differential geometry. The purpose was to "straighten out" the optimal routes to geodesics, or even lines in Wardop's case.

Conformal Mapping. We will first illustrate the case by the Wardrop issue which links neatly to the direct traffic computations discussed above.

Wardrop considered cases where $k = r^n$, i.e. local transportation cost is a power function of the distance from the origin. We already considered two cases, n=0 and n=–1, where the routes became straight lines and logarithmic spirals respectively.

We can now consider the class of functions in full generality. With a power function, and written in polar coordinates we get the transportation cost integral (9.6) to be minimized as:

$$I = \int_{\omega_0}^{\omega_1} r^n \sqrt{r^2 + r'^2} d\omega \tag{9.118}$$

Suppose we introduce the new coordinate transformation

$$R = r^{n+1} \tag{9.119}$$

$$\Omega = (n+1)\omega \tag{9.120}$$

Then we can easily compute the derivative $R' = dR / d\Omega = r^n r'$ by differentiating the above expressions (9.119) and (9.120). Substituting into the integral (9.118) and multiplying $r^n = R^{n/(n+1)}$ squared into the square root expression we easily obtain:

$$I = \frac{1}{n+1} \int_{\Omega_0}^{\Omega_1} \sqrt{R^2 + R'^2} d\Omega \tag{9.121}$$

which in the new coordinates is an expression for finding minimum cost routes with just a unitary (constant) local transportation cost. The solutions for the optimal routes in the space

$$X = R\cos\Omega = r^{n+1}\cos((n+1)\omega) \tag{9.122}$$

$$Y = R\sin\Omega = r^{n+1}\sin((n+1)\omega) \tag{9.123}$$

are accordingly straight lines. Observe that we, unlike Wardrop, do not introduce any complex functions; we just consider the result of the operation in the real plane. The coordinate transformation projects radii vectors by a power function at the same time as it multiplies angles. Optimal routes which are straight lines in the virtual space accordingly have mapped images that are curved in the original space.

It is, however, easier to deal with the straight lines in the virtual space. There is one particular reason for this: The mapping is conformal, and therefore it does not change angles between intersecting curves. In particular this applies to the angles under which the optimal routes intersect various circles of constant radius. So, if we want to compute the angle $\sec(\theta-\omega)\nabla\cdot\phi$ to be used in counting the weighted cordon crossings it does not matter if we use $\sec(\Theta-\Omega)\nabla\cdot\phi$ instead, where Θ is the angle of the straight line route in the virtual space, and Ω is the angle of the radius vector in that space as already defined. As we have $(\Theta-\Omega)=(\theta-\omega)$, we can make things easy for us and compute the angle in the virtual space where optimal routes become straight lines and circles remain circles (see Puu 1979).

There are however two things to be noted. First, the Jacobian of the map $(x,y)\Rightarrow(X,Y)$ becomes

$$\frac{\partial(X,Y)}{\partial(x,y)} = \frac{1}{((n+1)r^n)^2} \tag{9.124}$$

Therefore, even if population is uniformly distributed on the original space, the distributions are no longer equal in the virtual space of analysis. They simply have to be modified by the Jacobians of the mapping. If $n > 0$, then population in the virtual space is moved closer to the center because the mapping packs the area that way. The overall density is also decreased because the map also enlarges a sector of the disk to the whole disk. The question is what becomes of the whole disk if each sector of a given angle is enlarged to the whole disk. The solution is in terms of multilayer covering of the disk in terms of a Riemann surface. See Puu (1979).

In this context Angel and Hyman (1976) make an important observation: For an isotropic local transportation cost function we can always apply a map such that the optimal routes become geodesics in a virtual space, or we can apply a map so as to make a nonuniform density of origins or destinations uniform, but we cannot do both at once. Any map to attain one of the goals automatically brings a given distortion to the other. So, defining a virtual space by the coordinate transformation $(x,y) \Rightarrow (X,Y)$, automatically results in a distortion ofthe population densities at origin and destination to be used at the simulations.

The second observation is the following. If the power n is positive, then the map $(x,y) \Rightarrow (X,Y)$ by multiplying angles enlarges only a sector of the disk to the entire disk. If for instance $n = 1$ then half the disk is enlarged to the whole disk. Looking at the situation in the original space we find that all the straight line routes in the virtual (X,Y)-space are squeezed into a half disk in (x,y)-space. There is a deeper reason behind this: There are no continuous routes that extend over an angle larger than half the disk. The higher n is, the smaller the sector into which the continuous optimal routes are squeezed.

But we must have connections also between points in the disk separated by larger angles, so how is this solved? The answer is that there are alternative routes consisting of broken radials which replace the smooth curves for angles that are too large. Actually, we would have found out this from the outset if we had considered more advanced results from the calculus of variations, kinetic foci and the Jacobi conditions. See Puu (1979) for details. Then we would have found that sufficiency conditions fail for smooth routes covering too large angles.

As a consequence the traffic distributions are found by piecing two things together: Linear routes in for small angle connections in the virtual space, and radial ones for large angle connections.

Stereographic Projection. To illustrate Tobler's case consider a transportation cost function

$$k(x,y) = \frac{4}{4+r^2} \tag{9.125}$$

where

$$r = \sqrt{x^2+y^2} \tag{9.126}$$

This is a bit like the case of logarithmic spirals, with maximum local transportation cost in the supposedly congested center, though the cost does not go to infinity there. The transportation cost along a parameterized route $x(s)$, $y(s)$ is

$$I=\int_{s_0}^{s_1}\frac{4}{4+x^2+y^2}\sqrt{x'^2+y'^2}ds \tag{9.127}$$

and is not obviously dealt with. Neither does it help to revert to polar coordinates. Instead we can introduce the following coordinate transformation into threespace:

$$u=\frac{4x}{4+r^2} \tag{9.128}$$

$$v=\frac{4y}{4+r^2} \tag{9.129}$$

$$w=\frac{2r^2}{4+r^2} \tag{9.130}$$

It is easy to find that

$$u^2+v^2+(w-1)^2\equiv 1 \tag{9.131}$$

for any values of x,y. Hence all points to which the plane is mapped lay on a sphere with unit radius and center in the point $u = v = 0$, $w = 1$.

The plane is hence mapped onto the surface of a unit sphere through a so called stereographic projection. Geometrically we can imagine the sphere as placed laying tangentially on the origin in x,y-space. Then the points are mapped by a ray from the north pole through the sphere to the plane. This mapping is one to one, except the north pole itself on the sphere which corresponds to any point at infinity.

The most interesting thing to remember about the stereographic projection is that it maps great circles on the sphere to circles on the plane. This is interesting because the shortest paths on the sphere are the geodesic great circles, and should they turn up as optimal routes, which they will, then the optimal routes in the original space obviously become circular arcs.

The parameterized curve $x(s)$, $y(s)$ through the coordinate transformation $x,y \Rightarrow u,v,w$ as defined maps to a parameterized curve $u(s)$, $v(s)$, $w(s)$ embedded in the unit sphere. We easily obtain the derivatives from (9.128)-(9.130):

$$u' = \frac{16-4x^2+4y^2}{\left(4+r^2\right)^2}x' - \frac{8xy}{\left(4+r^2\right)^2}y' \tag{9.132}$$

$$v' = \frac{16+4x^2-4y^2}{\left(4+r^2\right)^2}y' - \frac{8xy}{\left(4+r^2\right)^2}x' \tag{9.133}$$

$$w' = \frac{16x}{\left(4+r^2\right)^2}x' + \frac{16y}{\left(4+r^2\right)^2}y' \tag{9.134}$$

We can now calculate the arc length element on the unit sphere from (9.132)-(9.134)

$$\sqrt{u'^2+v'^2+w'^2} = \frac{4}{4+r^2}\sqrt{x'^2+y'^2} \tag{9.135}$$

This calculation is messy but elementary. All the products of x', y' vanish and only the squares remain. We note that the infinitesimal arc in our original parameter on the sphere equals the product of local transportation cost and the infinitesimal arc in our original space.

Hence the transportation cost integral we have to minimize becomes

$$I = \int_{s_0}^{s_1} \frac{4}{4+r^2}\sqrt{x'^2+y'^2}ds = \int_{s_0}^{s_1} \sqrt{u'^2+v'^2+w'^2}ds \tag{9.136}$$

so that transportation cost in our original space indeed equals arc length on the virtual space which now is the unit sphere. Shortest paths on the unit sphere are geodesics, more specifically great circles.

If we consider a field of radiating paths in various directions from a certain point of origin they hence become a set of great circles on the unit sphere. The characteristic of such a bundle of great circles is that they all meet again in the point of the sphere opposite to the point of origin.

As we noted, the images of great circles on the unit sphere by the stereographic projection become circular arcs. Accordingly, in the original plane space the optimal routes from a given point are a set of circular arcs which again meet in another point. Such a field in physics is known from electrostatics as the dipole.

Hence, by using a curved virtual space in the spirit of Tobler we managed to solve the routing problem in a very simple manner without carrying out any calculations at all, just by the knowledge of the character of shortest routes on a sphere.

9.6 Concluding Remarks: Road Investment

Perhaps Continuous Space Modelling (CSM) is best at providing an overall picture of the direction and volume of traffic flows generated by a given distribution of traffic sources and sinks. To translate this into concrete proposals for road investments, say, would require supplementary studies in a narrower and presumably discrete context.

Put another way, CSM can lead us to an understanding of the economics of a situation, while falling short of supplying engineering answers. Advances in computer image technology may however lead us to this goal in the future.

Still it is possible to formulate a few issues concerning road investments in the continuous space format.

We consider two stylized scenarios. The first deals with increasing the capacity of an existent road network, the second with increasing the density of the network.

If capacity is the issue, which is particularly appropriate in a congested urban area, then local transportation cost *k* will depend on the ratio of traffic *f*, as defined above, to road capacity, for which the areal density of road capital *c* can be taken as a proxy:

$$k\left(\frac{f}{c}\right) \tag{9.137}$$

A possible functional form will be

$$k\left(\frac{f}{c}\right)=k_0\left(1-\frac{f}{c}\right)^{-\gamma} \tag{9.138}$$

which implies that when traffic f goes to capacity c then $k \to \infty$ as speed goes to a total stand still.

Minimizing total transportation cost:

$$T=\iint k\left(\frac{f}{c}\right)f\,dxdy \tag{9.139}$$

subject to a total road investment budget:

$$\iint c dxdy = C \tag{9.140}$$

leads to a Lagrange problem

$$\text{Max} - \iint k\left(\frac{f}{c}\right) f\, dxdy + \mu\left(C - \iint c dxdy\right) \tag{9.141}$$

which is solved by

$$k'\left(\frac{f}{c}\right)\frac{f^2}{c^2} - \mu = 0 \tag{9.142}$$

implying

$$\frac{f}{c} = \text{constant} \tag{9.143}$$

Capacity should then be made proportional to traffic flow throughout. Notice, however, the hidden assumptions in (9.139) that capacity costs the same per unit in all locations and that transportation cost is the same function of flow/capacity everywhere.

In the second scenario, increasing the density of the network rather than the capacity of an existent network, we can specify the local transportation cost function in the following way: Disregarding congestion, and hence taking cost as proportionate to the distance traversed, we note that whatever the density ofthe network it is always at least necessary to traverse the Euclidean distance, normalized to unity. If the network is sparse, adding a new short-cut saves the average necessity of detour-taking to a considerable degree.

However, the denser the network already is, the less will be saved in terms of detours. Accordingly there are decreasing returns from road investments, and we can specify local transportation cost in terms of this type of investment b as

$$k = k(b) \tag{9.144}$$

where $k(b)$ is a decreasing concave function with a positive infimum as $b \rightarrow \infty$, for instance

$$k = 1 + \frac{1}{b} \tag{9.145}$$

The equivalent to (9.139) now reads

$$T = \iint f\; k(b)\, dxdy \tag{9.146}$$

Maximizing (9.146) with respect to b subject to the budget constraint

$$\iint b dx dy = B \tag{9.147}$$

now leads to the Lagrange problem

$$\text{Max} - \iint k(b) f dx dy + \mu\left(B - \iint b dx dy\right) \tag{9.148}$$

which is solved by

$$-k'(b) f = \mu \tag{9.149}$$

or

$$b = (k')^{-1}\left(-\frac{\mu}{f}\right) \tag{9.150}$$

Thus with $k = 1 + 1/b$ as in (9.145):

$$\frac{1}{b^2} = \mu \tag{9.151}$$

or

$$b = \sqrt{\frac{f}{\mu}} \tag{9.152}$$

Road capital should be distributed in proportion to the square root of the measure of traffic rather than to traffic. In Puu (1979) this rule was tested for the sparsely populated parts of Sweden where congestion was of little importance and the necessity of detour taking was considerable. The fit to the square root rule was astonishingly good, $b \propto f^{0.46}$ with $R^2 = 0.65$.

This illustrates that one might get some general information by CSM about such issues as the distribution of road capital. Still, this is when the traffic distribution is a given datum. The tough part of analysis still remains as we noted in section 9.4.

Changing the distribution of road capital in terms of road density and/or road capacity or density automatically changes the distribution of local transportation cost. As a consequence the choice of routes and the traffic distribution itself changes. The difficulty, however, remains even in discrete modelling where most evaluations of road investment programs by that approach take traffic as a given datum.

These are highly stylized scenarios. The modelling ofhow transportation cost actually depends on local conditions in order to arrive at realistic functions $k(f/c)$ and $k(b)$ is an engineering job. CSM can point research in this direction - but will not do this job as well.

Bibliography

Angel, S. and G. M. Hyman (1970) ”Urban Velocity Fields”, *Environment and Planning* 2, 211-224.

Angel, S. and G. M. Hyman (1971) ”Urban Travel Time”, *Papers andProceedings of the Regional Science Association* 26, 85-99.

Angel, S. and G. M. Hyman (1972) ”Urban Spatial Interaction”, *Environment and Planning* 4, 99-118.

Angel, S. and G. M. Hyman (1972) ”Urban Transportation Expenditures”, *Papers and Proceedings of the Regional Science Association* 27, 105-123.

Angel, S. and G. M. Hyman (1976) *Urban Fields -A Geometry of Movement for Regional Science* (Pion, London).

Beckmann, M. J. (1952) ”A Continuous Model ofTransportation”, *Econometrica* 20, 643-660.

Beckmann, M. J. (1953) ”The Partial Equilibrium of a Continuous Space Market”, *Weltwirtschaftliches Archiv* 71, 73-89.

Beckmann, M. J. (1968) *Location Theory* (Random House, New York).

Beckmann, M. J. (1985) "Spatial price policy and the demand for transportation", *Journal of Regional Science* 25, 367-371

Beckmann M. J. and T. Puu (1985) *Spatial Economics - Potential, Density, andFlow* (North-Holland)

Bunge, W. (1966) *Theoretical Geography* (Gleerups, Lund).

Koopmans, T. C. (1949) "Optimum Utilization of the Transportation System", *Econometrica* 17: Suppl. 136-146.

Launhardt, W. (1885) *Mathematische Begründung der Volkswirtschaftslehre* (B. B. Teubner, Leipzig. Reprinted (1945) by Scientia Verlag, Aalen.

Launhardt, W. (1887) *Theorie des Trassierens* (Schmorl und von Seefeld, Hannover)

Lösch, A. (1954) *The Economics ofLocation* (Yale University Press, New Haven, Conn.)

Mosler, K. C. (1976) *Optimale Transportnetze - Zur Bestimmung ihres kostengünstigsten Standorts bei gegebener Nachfrage* (Springer, Berlin).

Mosler, K. C. (1987) *Continuous Location of Transportation Networks* (Springer-Verlag)

Palander, T. F. (1935) *Beiträge zur Standortstheorie* (Almqvist & Wiksell, Uppsala).

Puu, T. (1977) ”A Proposed Definition of Traffic Flow in Continuous Transportation Models”, *Environment and Planning* 9, 559-567.

Puu, T. (1978) ”On Traffic Equilibrium, Congestion Tolls and the Allocation of Transportation Capacity in a Congested Urban Area”, *Environment andPlanning* 10, 29-36.

Puu, T. (1978) ”Towards a Theory of Optimal Roads”, *Regional Science and Urban Economics* 8, 225-248.

Puu, T. (1978) ”On the Existence of Optimal Paths and Cost Surfaces in Isotropic Continuous Transportation Models”, *Environment andPlanning* 10, 1121-1130.

Puu, T. (1979) *The Allocation ofRoad Capital in Two-Dimensional Space* (North-Holland)

Puu, T. (1987) *Mathematical Location and Land Use Theory - An Introduction* (Springer-Verlag)
von Stackelberg, H. (1938) "Das Brechungsgesetz des Verkehrs", *Jahrbücher für Nationalökonomie undStatistik* 148: 680-696.

Tobler, W. R. (1963) "Geographic Area and Map Projections", *Geographical Review* 53, 59-78

Vaughan, R (1987) *Urban Spatial Traffic Patterns* (Pion Ltd.)

Wardrop, J. G. (1966) ”Journey Speed and Flow in Central Urban Areas”, *Traffic Engineering and Control* 9, 528-532.

Wardrop, J. G. (1969) ”Minimum Cost Paths in Urban Areas”, *Strassenbau- und Strassenverkehrstechnik* 86, 184-190.

Warntz, W. (1967) ”Global Science and the Tyranny of Space”, *Papers and Proceedings ofthe Regional Science Association* 19, 7-19.

Werner, C. (1966) *Zur Geometrie von Verkehrsnetzen* (Abhandlungen des 1. Geografischen Instituts der FU, Berlin).

Williams, H. C. W. L. and J. D. Orthuzar Salas (1976) ”Some Generalizations and Applications of the Velocity Field Concept”, *Transportation Research* 10, 65-73.

Zitron, N. R. (1974) ”A Continuous Model of Optimal-Cost Routes in a Circular City”, *Journal ofOptimization Theory andApplications* 14, 291-303

Zitron, N. R. (1978) ”A Critical Conditions for the Cost Density in the Circular City Model", *Journal ofOptimization Theory andApplications* 24, 507-512

Acknowledgements

We acknowledge gratefully the assistance ofMs. Gunvor Albihn in typing some earlier drafts of this article and the hospitality of the *Center for Economic Studies (CES)* at the University of Munich, under the directorship of H.W. Sinn, in enabling there to complete the final version of this chapter. We also wish to express our gratitude to *The Swedish Transport and Communication Research Board* who made the necessary visit in Munich for one of the authors financially possible.

10 LOCATION MODELS IN TRANSPORTATION

Mark S. Daskin and Susan H. Owen

10.1 Introduction

Transportation as a derived demand and the relation to location choices

The demand for transportation services is a derived demand (Manheim, 1979). In other words, the demand for such services arises from the need to move goods and/or people from one location to another and not from some inherent desirability of transportation in and of itself. Transportation is necessary only because people and/or goods are not located where they need to be when they need to be there.

For example, the need for either public or private commuter transportation in the morning and evening rush hours arises from the fact that an individual's residence and his/her work location are rarely the same. Therefore, people need to travel from their homes to their workplaces in the early morning and vice versa in the late afternoon or evening. Similarly, the need to ship automobiles arises from the fact that they are produced in a limited number of assembly plants – typically only one or two plants produce a particular vehicle – and are sold throughout the country. Thus, cars need to be shipped from assembly plants to dealers for sale to customers. As a final example, only a few locations throughout the country are amenable to growing grapefruits. Thus, there is a need to transport the fruit from farms in Florida and California to locations throughout the remainder of the country.

Since the demand for transportation services arises from the spatial mismatch between where people are and where they want to be or where products are manufactured (or grown) and where they are sold (or used in subsequent manufacturing processes), it is important to study the factors that influence location decisions. Clearly, location decisions determine, in part, what transport facilities will be used. Note, however, that the existence or lack of good transport services may also influence location decisions. For example, the presence of a railroad line with at-grade crossings may dictate the need for additional emergency medical service

(EMS) bases in a community to ensure that residences on both sides of the tracks are adequately served even if trains are blocking the roadway.

While transportation systems and facility location decisions clearly interact, a complete analysis of the factors influencing location choices is beyond the scope of this chapter. Thus, we have elected not to examine residential location choices. The reader interested in this topic is referred to the seminal work of Lerman (1976) and subsequent research on the topic. Instead, the focus of this chapter will be on models that can be used to assist in locating public and private sector facilities. We use the word "assist" deliberately. Location decisions are strategic in nature. As such, there are typically many factors that affect these decisions, some of which are difficult or impossible to quantify. Mathematical models of the sort outlined below are therefore valuable tools that *assist* decision-makers; they cannot be considered a replacement for decision-making expertise and strong management.

The remainder of this chapter is organized as follows: We conclude this section with a discussion of four typical facility location contexts in which the models we will be discussing might be of use. In section 10.2 we present a taxonomy of facility location problems. This classification scheme helps us to further limit the scope of the chapter and to identify areas of emerging research. In section 10.3, we outline a number of classical facility location problems and formulate them using integer linear programs. Section 10.4 discusses several solution algorithms for these models. In section 10.5 we discuss limitations of the basic models and extensions to these models. Finally conclusions and directions for future work are outlined in section 10.6.

Examples of facility location decision contexts

The location of ambulance bases has been one topic of considerable attention in the location literature (Eaton and Daskin, 1980; Eaton *et al.*, 1985). The problem is typically to determine the *number* and *location* of ambulance bases needed to ensure adequate service to the population at hand. Adequate service is often defined in terms of whether or not an ambulance can travel from its base to the location of an emergency within some specified time limit (e.g., five minutes). Sometimes more complex models are employed which consider travel from the base to the emergency site as well as travel from the emergency location to the hospital. Emergency services have also led to the incorporation of stochastic demand and service time elements into location models (Fitzsimmons, 1973).

The location of non-emergency public facilities has also been the subject of considerable attention within the location literature. Maze *et al.* (1981, 1982) study the location of bus garages. In such problems the objective is often to minimize the total deadhead time of buses, or the time a bus spends driving between its garage and the beginning and ending points of its route. The problem is often complicated by the presence of multiple bus types. A number of constraints typically limit the possible assignments. These constraints include the following: limits on the overall garage capacity in different locations and at different times of the day, restrictions on the types of vehicles that can be assigned to each garage (since maintenance is performed on the buses at the garages and not all garages will be equipped to maintain all vehicle types), and minimum and maximum numbers of vehicles of each type that can be assigned to each garage.

In the private sector, facility location models have been used to locate production plants, warehouses and trans-shipment facilities. For example, in locating rail ramps (locations at which finished automobiles are transferred from rail cars to truck for final delivery to dealers) one needs to determine how many rail ramps to have throughout the network and where they should be. The objectives typically consider both cost and customer service. This problem is further complicated by the presence of multiple products and by the need to utilize the rail and truck fleets efficiently.

Finally, there are many cases in which facility location and network design interact. For example, the number and location of airline hubs determines the configuration of the network to a large degree (Flynn and Ratick, 1988; Kuby and Gray, 1993). Indeed, the location of hub facilities has become a subfield in its own right and is discussed in section 10.5 below.

10.2 A taxonomy of facility location problems

In this section we present a taxonomy of facility location models. We note that location problems and the models that represent them can be classified in a variety of ways. The scheme we present is similar to that outlined by Daskin (1995) and by Brandeau and Chiu (1989) and Krarup and Pruzan (1990). In the context of this taxonomy, we highlight those topics that will be the focus of the remainder of this chapter and identify areas of location research that are beyond its scope.

Continuous vs. network location models

In some situations, it is appropriate to allow facilities to be located anywhere within a particular space. Sometimes demands are also modeled as being continuously distributed through the space in question. Such models are *continuous models.* These models are most appropriate when one simply wants to get a feel for how many facilities might be needed and roughly where they should be located. The Weber problem (Weber, 1929), that of finding the center of gravity of a set of discrete points, is the prototype continuous location problem. A significant portion of continuous location research, particularly that which attempts to incorporate vehicle routing costs into location models, is based on geometric probability theory (Kendall and Moran, 1963)

Continuous location models present both computational and practical difficulties. From a computational perspective, continuous models often result in non-linear formulations that are difficult to solve, particularly when more than one facility is being located. From a practical perspective, such models often suggest locations that are infeasible (e.g., locating a distribution center in the middle of a large lake). While a number of authors (e.g., Aneja and Parlar, 1994; Batta, Ghose and Palekar, 1989; Larson and Sadiq, 1983; Hansen, Peeters and Thisse, 1982; and Hurter, Schaefer and Wendell, 1975) have developed continuous models with forbidden regions or barriers to travel, such models pose even greater computational burdens.

In contrast to continuous location models, *network location models* assume that facilities and demands are located on the nodes or links of a network. In most cases, demands are assumed to be located on the nodes of the network. A key question that often arises is whether or not the objective function is degraded if one limits the search for facility locations to the network nodes. Hakimi (1964) was the first to show that for one common network location problem – the P-median problem discussed below – locating facilities only on the nodes does not degrade the solution. Limiting the search for facility sites to the network nodes is so computationally advantageous that we often do so even when this will degrade the solution. Often this limitation is further justified by the presence of a finite set of sites at which the decision-maker is willing to consider locating facilities.

Discrete location problems assume that demands are located at discrete points in a space and that facilities are also to be located at discrete points. A distance metric (e.g., the Euclidean metric) is used to compute the distances between these facility sites and demand nodes. Many of the models and algorithms developed for network location problems can be applied to discrete location problems as well.

This chapter will concentrate on discrete and network location problems.

Network vs. tree location problems

Within the class of network problems, problems on trees are often considered as a special case. There are several reasons for this. First, and perhaps most importantly, many location problems that are NP-hard on a general network can be solved in polynomial time when posed on a tree. Second, the polynomial tree algorithms can sometimes be used as the building blocks for (heuristic) algorithms for problems on more general graphs. Third, some transportation networks, particularly those in developing countries, and a significant number of telecommunication networks have a tree structure. Finally, a number of problems that are seemingly unrelated to facility location can be modeled as location problems on trees or other specially structured graphs. For example, Watson (1996) modeled the benefits of rationalizing the number of different steel coils an auto manufacturer purchases as a location problem on a line.

To retain generality, the focus of this chapter is on location problems on general networks and not on the specialized algorithms that arise in locating on trees.

Static vs. dynamic problems

All of the classical location models discussed in section 10.3 below are static models, assuming that decisions are made at one time and that the cost, distance and demand data are independent of time. In reality, model inputs change over time and decisions are made over an extended horizon. Decisions made in one period often preclude future options. Thus, opening a facility in year X may mean that the facility must remain open throughout the remainder of the planning period.

While our focus will be on static models, we will briefly discuss issues related to dynamic location problems in section 10.5

Deterministic vs. stochastic

Not only are classical location models static, but they are also deterministic in that they assume that (1) all model inputs are known with certainty and (2) there is no randomness underlying the process by which demands arise and customers are served. A significant branch of the literature has focused on ways of relaxing these two assumptions. Section 10.3 summarizes the traditional deterministic models; section 10.5 outlines a number of extensions of these models that account for both uncertainty in problem inputs as well as randomness associated with the underlying demand or service processes.

Single vs. multi-objective

Finally, most models use a single objective. However, since facility location problems are strategic in nature, there are likely to be a number of different constituents interested in the location decisions. These constituents often have different and sometimes conflicting objectives. Sometimes conflicting objectives simply seem to be inherent to the problem at hand. Erkut and Neuman (1989) argue that the location of undesirable facilities is an inherently multi-objective problem. For example, in locating solid waste repositories, we would like to minimize the impact of these facilities on residential areas. We would also like to minimize the cost of hauling waste to the sites. These two objectives conflict since much of the solid waste in an urban area is generated at residential units.

In section 10.3 we summarize traditional single objective models while section 10.5 outlines a number of the issues associated with multi-objective problems.

10.3 Classical location models

In this section we outline three classes of objectives that are typically used in facility location modeling. Within each class a number of different objective functions and models are identified. Key properties of these models are also discussed.

Covering models

The simplest class of facility location problems arises from the notion of coverage. Covering models are appropriate when there is some critical service distance (or time or cost). Demands that can be served within this distance are said to be covered, while those that are not are uncovered. Typical covering problems include: finding the minimum number of facilities needed to cover all demand nodes (the set covering problem); finding the location of a fixed number of facilities to maximize the number of covered demands (the maximal covering problem); and finding the locations of a fixed number of facilities to minimize the maximum distance between a demand node and the facility assigned to cover its demands (the P-center problem).

Perhaps the earliest references to location problems are Biblical and relate to this class of models. In Numbers 35:9-15 as well as Exodus 21:13, Deuteronomy 4:41-43 and Deuteronomy 19:1-13, the Bible commands the Israelites to create cities of refuge to which an individual who inadvertently kills another person can flee from retribution from the family of the deceased. They are commanded (Numbers 35:9-15) to identify 6 such cities, three on the west side of the Jordan river and three on the east side. Later (Deuteronomy 4:41-43), the Bible identifies the three cities on the east side of the Jordan, but since "Their precise location is not known," (Plaut, 1981, p. 1344) we cannot infer much about the access that they provided to different populated areas. The commandments regarding the cities to the west of the Jordan are more explicit (Deuteronomy 19:1-13). The Israelites were to "survey the distances, and divide into three parts the territory of the country ... so that any manslayer may have a place to flee to ... Otherwise, when the distance is great, the blood-avenger, pursuing the manslayer in hot anger, may overtake him and kill him...." (Plaut, 1981, p. 1467). Clearly the cities were to be dispersed evenly throughout the countryside and were to be made accessible to the populace.

The placement of the cities of refuge is actually a dynamic location problem since Deuteronomy goes on to say "And when the Lord your God enlarges your territory ... then you shall add three more towns to those three." (Plaut, 1981, p. 1467). We note that there is some controversy about whether this suggests an additional three cities (for a total of nine) or whether it refers to the three cities east of the Jordan River. In either case, the cities of refuge were to be located so that the maximum distance someone would have to travel to reach one would be less than the distance a pursuer was likely to travel "in hot pursuit" after learning of the death of his/her relative. As indicated above, problems in which there is a distance (or time or cost) beyond which service is unacceptable and below which it is acceptable are *covering problems.* In the modern era, such problems typically arise in the location of emergency service facilities including ambulances, fire stations, and police units.

The simplest location covering problem is the *set covering problem* (Toregas *et al.*, 1971). Here the objective is to find the locations of the minimum number of facilities so that all demands are covered within the acceptable distance. To formulate this problem we define the following inputs and sets:

I = the set of demand Points, indexed by i

J = the set of candidate facility locations, indexed by j

d_{ij} = distance between demand node i and candidate site j

D_c = coverage distance

N_i = $\left\{ j \middle| d_{ij} \le D_c \right\}$

= the set of all candidate locations that can cover demand node i

and the following decision variable

$$X_j = \begin{cases} 1 & \text{if we locate at candidate site } j \\ 0 & \text{if not} \end{cases}$$

With this notation, the set covering problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{j \in J} X_j \qquad (10.1)$$

$$\text{Subject to} \quad \sum_{j \in N_i} X_j \ge 1 \qquad \forall i \in I \qquad (10.2)$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (10.3)$$

The objective function (10.1) minimizes the number of selected facilities. Constraint (10.2) ensures that each demand node is covered by at least one selected site. Constraint (10.3) is an integrality constraint. We note that the objective function can be generalized by including a site-specific cost as a coefficient of the decision variable. The problem would then be that of finding the minimum cost set of facility sites that cover all demand nodes at least once.

The set covering problem is NP-hard even when all facility costs are identical. Fortunately, the linear programming relaxation of the set covering problem as posed above often results in an all-integer solution. When the LP relaxation is fractional, typically only a few branches in a branch and bound algorithm are needed to obtain an optimal all-integer solution. Often a variety of row and column reduction rules (see Daskin (1995) for a discussion of such rules) can be used to reduce the size of the problem considerably. Two such rules warrant particular discussion. First, let $M_j = \left\{ i \middle| d_{ij} \le D_c \right\}$. In other words, M_j is the set of demand nodes covered by a facility at candidate site *j*. If $M_k \subset M_j$, then candidate site *j* *dominates* candidate site *k* and we can set $X_k = 0$. Second, if $N_i \subset N_h$, then any facility that covers node *i* also covers node *h*. In this case, we can eliminate row *h* from the constraint matrix.

Finally, we note that the optimal solution to the set covering problem when facilities can be located on the links as well as the nodes will often be better (i.e., it will require fewer facilities) than the solution to the problem when facilities are restricted to the nodes. This is illustrated by the simple network shown in Figure 10-1. If the coverage distance is 10 and facilities can only be located on the nodes, then two facilities are needed: one at A and one at either B or C. If we can locate on the links as well as the nodes, a facility located ten units from node A and two units to the left of node B would cover all three demand nodes. In this simple example, we can halve the number of facilities by locating on a network link. Church and Meadows (1979) show how the node set can be augmented by a finite set of network intersection points so that locating on the set of nodes or network intersection points will always be as good as locating anywhere on the links.

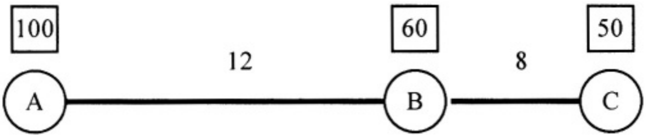


Figure 10-1. Example network

One problem with the set covering model is that it often dictates locating more facilities than can be afforded. Typically many of the facilities that are identified cover a small fraction of the total demand. For example, if we try to cover the 150 largest cities in the United States (with a combined population of over 58 million people) within 250 miles, 18 facilities are needed. If we relax the requirement that all demands be covered and require only 98% of the demands to be covered, we can reduce the number of required facilities by 16.7% to 15 facilities. We can halve the number of required facilities and use only 9 sites if we could get by with covering only 86% of the total demand. Another problem with the set covering model is that it fails to discriminate between large and small demand nodes. Thus, covering New York City (with a population of 7.32 million people) is just as important in the set covering problem as is covering Pasadena, Texas (with a population under 120,000).

These two problems have led researchers to consider relaxing the requirements of the set covering model. As indicated above, one way to reduce the number of required facilities is to relax the constraint that all demands be covered. The *maximal covering problem* (Church and ReVelle, 1974) finds the locations of a specified number of facilities that maximize the number of covered demands. There is a subtle but important distinction between the problem statements for the set covering and maximal covering problems. The set covering problem requires that all demand *nodes* be covered while the maximal covering problem maximizes the number of covered *demands*. As such, the maximal covering problem distinguishes between large and small demand nodes.

To formulate the maximal covering problem we define the following two additional inputs and one decision variable:

h_i = demand at node i

P = the number of facilities to locate

$$Z_i = \begin{cases} 1 & \text{if demand node } i \text{ is covered} \\ 0 & \text{if not} \end{cases}$$

With this additional notation, the maximal covering problem may be formulated as follows:

$$\text{Maximize} \quad \sum_{i \in I} h_i Z_i \tag{10.4}$$

$$\text{Subject to} \quad \sum_{j \in N_i} X_j - Z_i \geq 0 \qquad \forall i \in I \tag{10.5}$$

$$\sum_{j \in J} X_j = P \tag{10.6}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{10.7}$$

$$Z_i \in \{0,1\} \qquad \forall i \in I \tag{10.8}$$

The objective function (10.4) maximizes the number of covered demands. Constraint (10.5) states that demands at node *i* cannot be counted as covered unless we locate at one of the candidate sites which covers node *i*. Constraint (10.6) states that we are to locate *P* facilities. Constraints (10.7) and (10.8) are standard integrality constraints. We note that constraints (10.8) can be relaxed to simple upper bounding constraints. Also, by solving the maximal covering problem for values of *P* from 1 up to the number needed for full coverage, we can trace out the tradeoff curve between coverage and the number of facilities used – a proxy for the cost of the system.

The maximal covering problem is also NP-hard (Megiddo, Zemel and Hakimi, 1983), but it can generally be solved effectively using heuristics of the sort outlined in section 10.4 or using Lagrangian relaxation embedded within a branch and bound algorithm (Daskin, 1995; Daskin and Owen, 1998; Galvão and ReVelle, 1996). As in the case of the set covering problem, the objective function can be improved if we allow location on network links. For example, if the coverage distance is 6 in figure 10-1, the objective function for the maximal covering problem with P=1 and facilities restricted to the nodes would be 100 (with the facility located at node A). However, if we can locate on the links as well as the nodes, the facility could be located between nodes A and B and the objective function would increase to 160. As before, if we restrict candidate sites to the node set augmented by the set of network intersection points defined by Church and Meadows (1979), the solution will be as good as that obtained by allowing sites to be anywhere on the links.

A second way to relax the requirements of the set covering problem is to increase the coverage distance. The *P-center problem* (Hakimi, 1964, 1965) locates a specified number of facilities (*P*) so that the maximum distance between a demand node and the nearest facility is minimized. As in the case of the set covering model, this model considers demand nodes and not the demand levels at the nodes.

The P-center problem comes in a variety of flavors. If candidate sites are restricted to the set of nodes, we have the *vertex P-center problem,* while the problem in which facilities can be anywhere on the network is termed the *absolute P-center problem.* Both versions can be either *weighted* (if the demand nodes have different weights and the objective function is defined in terms of the maximum demand-weighted distance) or *unweighted* (if all demand nodes have identical weights).

To formulate the weighted vertex P-center problem we need to introduce two additional decision variables.

W = the maximum distance between a demand node and the facility to which it is assigned

$$Y_{ij} = \begin{cases} 1 & \text{if demand node } i \text{ is assigned to a facility at node } j \\ 0 & \text{if not} \end{cases}$$

With these additional variables, the P-center problem can be formulated as follows:

$$\begin{array}{llll}
\text{Minimize} & W & & (10.9) \\
\text{Subject to} & \sum\limits_{j \in J} X_j = P & & (10.10) \\
& \sum\limits_{j \in J} Y_{ij} = 1 & \forall i \in I & (10.11) \\
& Y_{ij} - X_j \leq 0 & \forall i \in I,\ j \in J & (10.12) \\
& W - \sum\limits_{j \in J} h_i d_{ij} Y_{ij} \geq 0 & \forall i \in I & (10.13) \\
& X_j \in \{0,1\} & \forall j \in J & (10.14) \\
& Y_{ij} \in \{0,1\} & \forall i \in I,\ j \in J & (10.15)
\end{array}$$

The objective function (10.9) minimizes the maximum demand-weighted distance between a demand node and the facility to which it is assigned. Constraint (10.10) stipulates that P facilities are to be located. Constraint (10.11) requires that each demand node be assigned to exactly one facility. Constraint (10.12) is a linkage constraint stating that demands can only be assigned to open facilities. Constraint (10.13) defines the maximum demand-weighted distance in terms of the assignment variables. Finally, (10.14) and (10.15) are integrality constraints. Constraints (10.15) can be relaxed to simple upper bounding constraints.

For fixed values of P, the vertex P-center problem can be solved in $O\left(N^P\right)$ time since we can enumerate each possible set of candidate locations in this time.

Clearly, even for moderate values of *N* and *P,* such enumeration is not realistic and more sophisticated approaches are required. For variable values of P, the problem is NP-hard.

Assuming integer distances, the unweighted vertex or absolute P-center problem is most often solved using a binary search over a range of coverage distances (Handler and Mirchandani, 1979; Handler, 1990). For each coverage distance, a set covering problem is solved.

Comments on covering models

The binary nature of coverage – either a candidate site covers a node within the coverage distance or it does not – is both an advantage and a liability for models within this class. On the positive side, this characteristic results in constraints that often have zeros or ones as coefficients, contributing to the relative ease with which such models can often be solved. On the other hand, the binary nature of coverage often means that there are multiple alternate optima for a problem. In the case of the P-center problem a single demand node/facility site pair defines the maximum distance and many sets of facility locations and demand assignments may result in the optimal objective function value. This can lead to excessive branching if branch and bound algorithms are not designed appropriately. Similar problems exist for the set covering problem. Plane and Hendrick (1979), Daskin and Stern (1981), Benedict (1983), and Daskin, Hogan and ReVelle (1988) discuss a number of hierarchical extensions of covering models that select from among the alternate optima a solution that optimizes a secondary objective (e.g., maximizing the number of demand nodes that are covered more than once).

Most models within the class of covering models are very sensitive to network coding. For example, adding a demand node to a set covering problem (perhaps with a very small demand) may alter the locations selected and may dictate the need for an additional facility. In an effort to circumvent this sort of problem, Daskin and Owen (1998) recently introduced two new covering models. The *partial set covering problem* finds the locations of the minimum number of facilities needed to cover a specified fraction of the demand nodes or the total demand. The *partial covering P-center problem* finds the locations of a specified number of facilities so that the maximum distance between a demand node within the "service set" and the nearest facility is minimized. The total number of nodes or demands within the service set must equal or exceed some user-specified value.

Average distance models

Covering models are appropriate for cases in which adequate service is defined by the maximum time or distance between a demand and the facility serving it. In many cases, however, the total (or average) distance for all nodes is more important. For example, in shipping goods from plants to distribution centers, an activity that often uses truckload shipments, the total distance traveled between plants and distribution centers is likely to be of greater concern than is the maximum distance. In this section, we discuss two classic problems within the class of average distance models.

The *P-median model* (Hakimi, 1964, 1965) finds the locations of *P* facilities to minimize the demand-weighted total distance between demand nodes and the facilities to which they are assigned. This model may be formulated as follows:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij} \tag{10.16}$$

$$\text{Subject to} \quad \sum_{j \in J} X_j = P \tag{10.17}$$

$$\sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{10.18}$$

$$Y_{ij} - X_j \leq 0 \qquad \forall i \in I, j \in J \tag{10.19}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{10.20}$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I, j \in J \tag{10.21}$$

The objective function (10.16) minimizes the demand-weighted total distance. Note that for a fixed total demand, this is equivalent to minimizing the demand-weighted average distance. Constraints (10.17) through (10.19) are identical to (10.10) through (10.12) of the P-center problem. Constraints (10.20) and (10.21) are integrality constraints. Again, constraints (10.21) can be relaxed to simple non-negativity constraints.

Like the P-center problem, the P-median problem can be solved in polynomial time for fixed values of P, but is NP-hard for variable values of P (Garey and Johnson, 1979). Unlike the covering problems outlined above, however, at least one optimal solution to the problem consists of locating only on the nodes of the network (Hakimi, 1964). This can readily be shown by assuming that a facility is located optimally between nodes A and B. Let H_A be the demand served by the facility that enters the link via node A and define H_B similarly in terms of node B. Assume without loss of generality that $H_A \geq H_B$. Then, moving the facility closer to A will not increase the objective function. Thus, we can convert any solution in which a facility is located on a link to a nodal solution without degrading the objective value.

The P-median problem fails to account for the fixed costs associated with locating facilities. This is rectified by the *fixed charge location problem.* To formulate this problem we define the following three additional inputs:

f_j = fixed cost of locating a facility at candidate site *j*

C_j = capacity of a facility at candidate site *j*

α = cost per unit demand per unit distance

With this notation, the capacitated fixed charge location problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{j \in J} f_j X_j + \alpha \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij} \tag{10.22}$$

$$\text{Subject to} \quad \sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{10.23}$$

$$Y_{ij} - X_j \leq 0 \qquad \forall i \in I, j \in J \tag{10.24}$$

$$\sum_{j \in J} h_i Y_{ij} - C_j X_j \leq 0 \qquad \forall i \in I \tag{10.25}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{10.26}$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I, j \in J \tag{10.27}$$

The objective function (10.22) minimizes the sum of the fixed facility location costs and the total demand-weighted distance multiplied by the cost per demand per unit distance. Constraints (10.23) and (10.24) are assignment and linkage constraints. Constraint (10.25) states that the total demand assigned to a facility must not exceed the capacity of the facility. Constraints (10.26) and (10.27) are standard integrality constraints. If (10.27) is enforced as a binary constraint, the model assumes facilities are singly sourced. Relaxing this constraint allows demands to be assigned to multiple facilities. We also note that constraint (10.24) is not needed in this integer programming formulation as constraint (10.25) will force demands to be assigned only to open facilities. However, including constraint (10.24) in the formulation significantly strengthens the linear programming relaxation of the model. If constraint (10.25) is removed, we are left with the *uncapacitated fixed charge location problem.* In this case, demands can always be singly sourced, even if (10.27) is relaxed.

Undesirable facility location models

The covering and average distance models discussed above assume that locating facilities as close as possible to demands is desirable. For many facilities this is the case. However, for undesirable facilities (e.g., prisons, power plants, and solid waste repositories) at least one objective involves locating facilities far from demand nodes. In this section we summarize two such models.

The *maxisum location problem* seeks the locations of P facilities such that the total demand-weighted distance between demand nodes and the facilities to which they are assigned is maximized. This model may be formulated as follows:

$$\text{Maximize} \quad \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij} \tag{10.28}$$

Subject to

$$\sum_{j \in J} X_j = P \tag{10.29}$$

$$\sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{10.30}$$

$$Y_{ij} - X_j \leq 0 \qquad \forall i \in I, j \in J \tag{10.31}$$

$$\sum_{k=1}^{m} Y_{i[k]_i} - X_{[m]_i} \geq 0 \qquad \forall i \in I,\ m = 1,\ldots,N-1 \tag{10.32}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{10.33}$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I, j \in J \tag{10.34}$$

This formulation is identical to that of the P-median problem with two notable exceptions. First, the objective is to maximize the demand-weighted total distance and not to minimize it. Second, constraint (10.32) has been included. In the absence of constraint (10.32) demands will be assigned to the most remote facility to maximize the demand-weighted total distance. Constraint (10.32) ensures that demands are assigned to the nearest selected facility. In this constraint, $[k]_i$ is the index of the k^{th} farthest candidate location from demand node *i*. Constraint (10.32) then states that if the m^{th} closest facility to demand node *i* is opened then demand node *i* must be assigned to that facility or to a closer facility.

Finally, we outline the *P-dispersion model.* Whereas all of the models previously presented deal with the distance between demand nodes and candidate sites, the P-dispersion model is concerned only with the distance between sites. The objective is to maximize the minimum distance between any pair of sites. Such a model is useful in locating military bases (e.g., nuclear weapon silos) as well as franchise outlets. In the latter case, the objective is a proxy for minimizing the competition between outlets.

To formulate this model we require the following additional input (M) and decision variable (D):

M = a large constant (e.g., $\max\limits_{i \in I, j \in J} \{d_{ij}\}$)

D = the minimum separation distance between any pair of facilities

With this notation, the P-dispersion model may be formulated as follows:

Maximize

$$D \tag{10.35}$$

Subject to

$$\sum_{j \in J} X_j = P \tag{10.36}$$

$$D + \left(M - d_{ij}\right) X_i + \left(M - d_{ij}\right) X_j \leq 2M - d_{ij} \qquad \forall i, j \in J, i < j \tag{10.37}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (10.38)$$

The objective function (10.35) maximizes the distance between the two closest facilities. Constraint (10.36) requires that P facilities are located. Constraint (10.37) defines the minimum separation between any pair of facilities in terms of the location choices. If either X_i or X_j is zero, the constraint will not be binding. If both are equal to 1, the constraint is equivalent to $D \le d_{ij}$. Constraint (10.38) is a standard integrality constraint.

10.4 Solution algorithms for traditional models

In this section, we briefly outline three broad classes of algorithms that have been applied to location problems such as those formulated in section 10.3 above. Note that these approaches are not limited to location problems. The general ideas have been applied to a large number of combinatorial optimization problems. We illustrate the development of these algorithms using the P-median model.

Greedy heuristics

The simplest algorithm for solving many location problems is a greedy adding algorithm. In this approach, we find the best site at which to locate the first facility using total enumeration. This can clearly be done in an amount of time that is proportional to the number of candidate facility sites. The location of the second facility is then identified by enumerating all possible locations for that facility, *holding the location of the first facility fixed.* Each subsequent facility is located in an identical manner. Figure 10-2 is a flowchart of this approach.

Just as one can add facilities in a greedy manner (always doing what is best *given* the locations of the facilities that have already been included in the emerging solution), we can also drop facilities in a greedy manner. In other words, we can begin the algorithm with facilities tentatively located at all candidate sites. We then remove facilities one at a time. In each case, we remove the facility whose elimination from the current set will result in the smallest possible increase in the demand-weighted average distance, the P-median objective. We continue in this manner until exactly P facilities remain in the solution.

Improvement heuristics

While greedy algorithms can generate a feasible solution quite quickly, they often do not perform particularly well. A number of algorithms have been devised to improve upon a feasible solution generated from a greedy algorithm, randomly or by some other means.

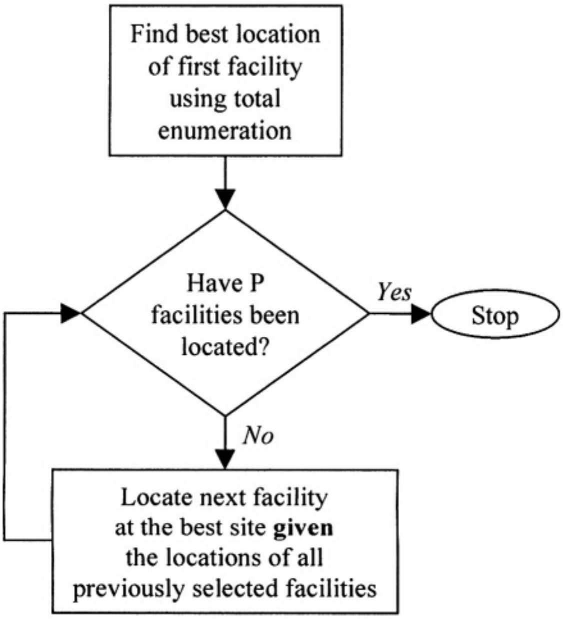

Find best location of first facility using total enumeration
Have P facilities been located?
Yes
Stop
No
Locate next facility at the best site **given** the locations of all previously selected facilities

Figure 10-2. Greedy algorithm

Teitz and Bart (1968) proposed the use of an exchange or substitution algorithm for the P-median problem. Their algorithm is similar in concept to the two-opt algorithm for the traveling salesman problem (Lin and Kernighan, 1973). The basic idea of an exchange algorithm is to replace a node in the current set of selected sites with a node not in that set. If the replacement results in an improved objective value, the change is accepted as the new incumbent solution; otherwise, the change is not made and the solution reverts to the original incumbent solution. Generally, the procedure continues until it is impossible to find an exchange that will improve the objective function.

In implementing an exchange algorithm, there are a number of options. For example, (1) one can accept the first exchange that improves the objective function, or (2) once we identify a node in the solution set whose removal will improve the solution, we can search for the best possible replacement for that node, or (3) we can search for the best possible exchange at each iteration of the algorithm. If option (1) or (2) is adopted, there are additional implementation choices related to initializing the indices over which we search for the next exchange. For example, should we search for the next removal and replacement nodes by returning to index 1 or should we continue from the current index values. In short, there are many ways to implement an exchange algorithm and the details will affect the solution that is attained. In addition, the exchange algorithm can be executed after the greedy algorithm has terminated or after each new node is added during the execution of the greedy algorithm.

While the exchange (or substitution) algorithm is similar to many other improvement algorithms used as heuristics for combinatorial optimization problems, the neighborhood search algorithm (Maranzana, 1964) exploits the spatial-nature of location problems. In the neighborhood search algorithm, we again begin with an incumbent solution. Demand nodes are then assigned to the nearest facility. The set of nodes assigned to a facility constitutes a "neighborhood" around that facility. Within each neighborhood, we then solve a 1-median problem optimally. (This can readily be done by total enumeration, if necessary.) Facilities are then relocated to the optimal 1-median locations within each neighborhood. If any facility sites change as a result of this procedure, new neighborhoods can be defined and the algorithm is repeated. The procedure continues until no change in the facility sites or neighborhoods is possible. Figure 10-3 is a flowchart of this algorithm.

Note that the demand-weighted total distance can decrease as a result of either the reassignment of demands to facilities (the formation of neighborhoods) or the relocation of facility sites. Also note that while the exchange algorithm relocates one facility at a time, the neighborhood search algorithm may result in multiple facility relocations in each iteration. Finally, it is worth noting that the neighborhood search algorithm considers exchanges of facilities *only* within a neighborhood. As such, it is a more restricted search than the exchange algorithm outlined above.

Hansen and Mladenovic (1997) present a variable neighborhood search algorithm for solving the P-median problem. In their context, a neighborhood is not the set of nodes assigned to a facility; rather, the neighborhood at a distance of k from a current *solution* is the set of solutions that can be obtained from the current solution by substituting k nodes not in the solution for k nodes that are in the solution. The algorithm performs an intensive local search (similar to the exchange algorithm outlined above) on the current solution and then diversifies the search by randomly selecting a solution from a neighborhood at a distance of k from the current best solution. The process continues, incrementing k, until some exogenously specified maximum value of k is attained. The algorithm compares very well with conventional heuristics as well as tabu search. Finally, we note that Densham and Rushton (1981) propose a number of data structures designed to expedite the search for good exchanges for P-median problems.

Lagrangian relaxation

While greedy adding or dropping algorithms coupled with improvement heuristics can find feasible solutions to many location problems relatively quickly, these procedures do not provide bounds on the quality of the solutions found. Theoretical,

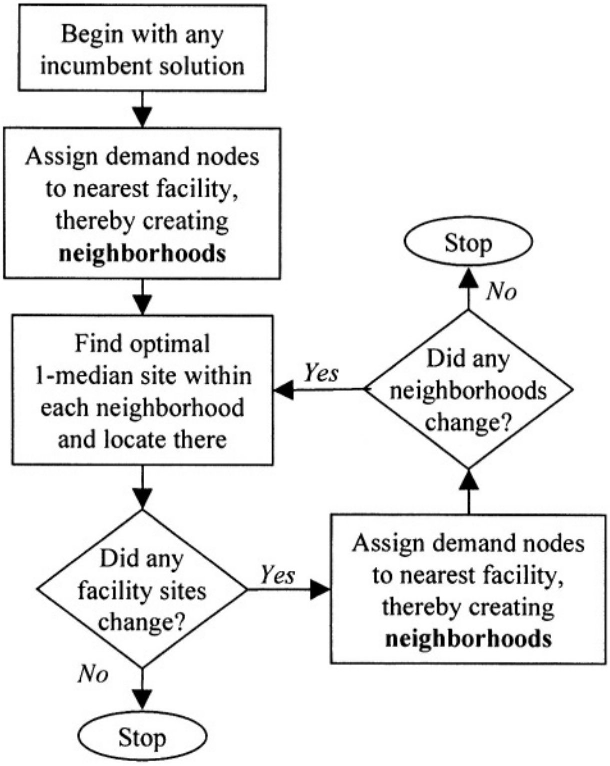

Begin with any incumbent solution
Assign demand nodes to nearest facility, thereby creating **neighborhoods**
Find optimal 1-median site within each neighborhood and locate there
Did any facility sites change?
Yes
No
Stop
Assign demand nodes to nearest facility, thereby creating **neighborhoods**
Did any neighborhoods change?
Yes
No
Stop

Figure 10-3. Neighborhood search algorithm

worst-case bounds exist for some algorithms, but in practice such bounds are often of only theoretical interest as the solutions obtained by the algorithms are often much better than the worst-case bounds would suggest. Thus, there is a need for a means of evaluating the quality of heuristic solutions. Lagrangian relaxation provides both upper and lower bounds on the objective function value. When embedded in a branch and bound procedure, Lagrangian relaxation can be used in place of linear programming to obtain optimal solutions to integer linear programming formulations of location problems.

The basic idea behind Lagrangian relaxation is to relax one or more of the constraints that make the problem difficult to solve by multiplying the constraint by a Lagrange multiplier and bringing the constraint into the objective function. We then alternate between solving the relaxed problem for the original decision variables with fixed values of the Lagrange multipliers and updating the Lagrange multipliers. In each iteration, the Lagrangian objective function provides a lower bound on the optimal objective value for a minimization problem (an upper bound for a maximization problem). The solution to the Lagrangian problem is likely to be infeasible for the original problem since we have relaxed one or more constraints. However, if we can convert the infeasible Lagrangian solution into a feasible solution for the original problem, the objective function value corresponding to that solution will provide an upper bound on the problem's objective function for a minimization problem (or a lower bound for a maximization problem).

We would like the relaxed problem to have two properties. First, for any fixed values of the Lagrange multipliers, we would like to be able to solve the relaxed problem very quickly. This property is essential since we will need to solve the relaxed problem hundreds (or thousands) of times for different multiplier values. Ideally, we should not have to resort to advanced optimization techniques to solve the relaxed problem. In the best of all worlds, the relaxed problem can be solved by inspection or by sorting the coefficients of different terms in the objective function. This will be illustrated in the Lagrangian relaxation for the P-median problem discussed below.

The second desirable property relates to integer linear programming problems. If the solution to the relaxed problem is guaranteed to be all-integer (even when the integrality constraints on the relaxed problem are themselves relaxed), then we can show that the bounds obtained from the Lagrangian procedure will be no tighter than those obtained from the linear programming (LP) relaxation of the original problem. However, if the solution to the relaxed problem is not guaranteed to be all-integer, but an all-integer solution can readily be found, then the bounds from the Lagrangian algorithm can be tighter than those found by solving the LP relaxation of the original problem. Thus, we would like it to be easy to find an integer solution to the Lagrangian relaxation, but we would like the relaxation to be such that its LP relaxation does not result in an all-integer solution. For the relaxation discussed below, this second property does not hold. Nevertheless, the relaxation proves to be very powerful and effective at solving the P-median problem.

The reader interested in additional background on Lagrangian relaxation is referred to the seminal paper on the topic by Fisher (1981) as well as the tutorial paper on the topic (Fisher, 1985).

To illustrate Lagrangian relaxation, we consider relaxing constraints (10.18) to obtain the following relaxation of the P-median problem:

$$\begin{aligned} \underset{\lambda}{\text{Max}}\ \underset{X,Y}{\text{Min}} \quad \mathrm{L} &= \sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij} + \sum_{i \in I} \lambda_i \left(1 - \sum_{j \in J} Y_{ij} \right) \\ &= \sum_{i \in I} \sum_{j \in J} \left(h_i d_{ij} - \lambda_i \right) Y_{ij} + \sum_{i \in I} \lambda_i \end{aligned} \tag{10.39}$$

$$\text{Subject to} \quad \sum_{j \in J} X_j = P \tag{10.40}$$

$$Y_{ij} - X_j \leq 0 \qquad \forall i \in I, j \in J \qquad (10.41)$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (10.42)$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I, j \in J \qquad (10.43)$$

Note that the Lagrangian function L in (10.39) is minimized with respect to the location and assignment variables (X_j and Y_{ij}) and is maximized with respect to the Lagrange multipliers (λ_i). The largest value of this function over all iterations is a lower bound on the objective function for the P-median problem.

To solve this problem for fixed values of the Lagrange multipliers, we note that if $(h_i d_{ij} - \lambda_i)$ is less than 0, we would like to set $Y_{ij} = 1$; otherwise we would like to set $Y_{ij} = 0$. Setting $Y_{ij} = 0$ is not a problem, as this will not violate any of the constraints in the problem. However, setting $Y_{ij} = 1$ will violate constraint (10.19) unless we have $X_j = 1$. If $X_j = 0$, we must set $Y_{ij} = 0$ no matter what the sign of the coefficient of Y_{ij} is. Thus, we need to determine which locations to pick to minimize (10.39). We do this by computing the value (in the Lagrangian problem) associated with locating at node *j*. This is given by $V_j = \sum_{i \in I} \min(0, h_i d_{ij} - \lambda_i)$ for each candidate site *j*. We then sort the V_j values from smallest (most negative) to largest. Then we set $X_j = 1$ for the locations with the *P* smallest V_j values and $X_j = 0$ for all other candidate sites. Finally, we set $Y_{ij} = 0$ if $X_j = 0$ and $Y_{ij} = 1$ if $X_j = 1$ and $(h_i d_{ij} - \lambda_i)$ is less than 0.

The solution to the Lagrangian problem is likely to violate one or more of the relaxed assignment constraints (10.18). However, the Lagrangian solution can be converted into a feasible solution to the P-median problem by locating at those *P* sites for which $X_j = 1$ and then assigning demands to the nearest selected facility. The objective function value for this solution is an upper bound on the P-median problem. Clearly, the smallest such value over all iterations is the best upper bound. We also note that the bound may be improved by applying one or both of the improvement heuristics described above to the solution. This can be done at each iteration or at selected points in the algorithm. For example, we might elect to use an improvement algorithm only when the objective function corresponding to the feasible solution (constructed as outlined above) represents an improvement over the best upper bound known so far.

At each iteration, the Lagrange multipliers are updated using subgradient optimization. Readers interested in the details of this approach are referred to the general papers by Fisher (1981, 1985) and the text by Daskin (1995).

Comparison of approaches

In this section, we summarize the results of applying the different algorithms outlined above to a problem with 150 demand nodes representing the 150 most populous cities in the United States. The demand at each city is assigned to be the city's population. Great circle distances are used in this example.

Table 10.1 summarizes the objective function values for this problem for 8 values of *P*, the number of facilities to locate, and six different algorithms. In addition, the table shows the number of Lagrangian iterations used, the number of branch and bound nodes explored and the solution time in seconds using a Pentium computer running at 133 MHz. The problems were solved using the SITATION software (Daskin, 1995) modified to run under Windows 95 (© Microsoft Corporation). All default Lagrangian parameters were used except that we required a provably optimal solution in all cases. Table 10.2 gives the percent error for each of the different algorithms and the different number of facilities sited. The greedy algorithm averaged 7 percent above the optimal values for these test runs. The exchange algorithm solutions averaged within 1 percent of the optimal values. Solutions from the neighborhood search algorithm averaged 1.6 percent and 3.9 percent above the optimal values, depending on whether the neighborhood search procedure was performed after each facility was added to the solution or only after all facilities had been added to the solution.

Note that the Lagrangian algorithm required an average of 10 seconds, 4.5 branch and bound nodes and 556 iterations to find the optimal solutions. The longest execution time was just over half a minute, requiring 15 branch and bound nodes and 1,934 Lagrangian iterations. Thus, while the relaxation outlined above will result in bounds that are no better than those that can be obtained by solving the LP relaxation of the P-median problem, the instances discussed here can be solved quite quickly on a relatively slow personal computer. This has generally been our experience for small to moderate problem sizes for the P-median problem as well as for the other classical problems outlined in section 10.3.

Table 10.1 – Objective function value and related results for P-median problem

ALGORITHM	P=3	P=5	P=8	P=10	P=12	P=15	P=18	P=20
Greedy Adding	319.72	227.03	158.82	135.55	114.22	91.35	76.83	68.90
Exchange								
Each Iteration	315.57	206.06	154.07	129.26	107.84	83.55	70.90	64.55
Last Iteration Only	315.57	206.06	152.90	127.66	107.06	83.55	71.52	63.99
Neighborhood								
Each Iteration	324.35	206.10	154.07	133.67	108.30	83.77	70.90	64.55
Last Iteration Only	319.72	206.51	158.39	133.67	112.09	87.13	74.74	66.81
Lagrangian with Branch and Bound								
Objective Function	315.57	206.06	152.90	127.13	105.53	83.55	70.90	63.69
Iterations	419	228	70	1,934	422	829	121	422
Branch and Bound Nodes	3	3	1	15	3	7	1	3
Time (seconds)	8.45	4.56	1.48	31.30	9.06	15.54	3.18	7.69

Table 10.2 – Percent error for P-median problem

ALGORITHM	P=3	P=5	P=8	P=10	P=12	P=15	P=18	P=20	Average Error
Greedy Adding	1.3%	10.2%	3.9%	6.6%	8.2%	9.3%	8.4%	8.2%	7.0%
Exchange									
Each Iteration	0.0%	0.0%	0.8%	1.7%	2.2%	0.0%	0.0%	1.3%	0.7%
Last Iteration Only	0.0%	0.0%	0.0%	0.4%	1.4%	0.0%	0.9%	0.5%	0.4%
Neighborhood									
Each Iteration	2.8%	0.0%	0.8%	5.1%	2.6%	0.3%	0.0%	1.3%	1.6%
Last Iteration Only	1.3%	0.2%	3.6%	5.1%	6.2%	4.3%	5.4%	4.9%	3.9%
Lagrangian with Branch and Bound	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

10.5 New directions and models

Limitations of traditional models

The classical models outlined in section 10.3 represent the building blocks of many location models used in practice. However, they are simply that: building blocks. Real-world problems are significantly more complex than are the simple models formulated above. In this section we outline a number of the deficiencies associated with traditional models and then summarize recent attempts to address these limitations.

Multiple objectives. Facility location decisions have long-term strategic impacts. As such, they are likely to affect many facets of an organization's operations and will impact many players and interest groups. Thus, in determining "optimal" facility siting plans, it is essential that the multitude of operational impacts and interest groups be represented. In doing so, we will invariably have to consider multiple objectives and not simply a single decision criterion.

The need to consider multiple objectives is most apparent in locating obnoxious or undesirable facilities such as waste disposal sites, prisons, and power plants (Ratick and White, 1988; Erkut and Neuman, 1989, 1992; Erkut and Verter, 1995; Verter and Erkut, 1995). In these cases, as indicated above, efficiency related objectives will tend to push the facility sites toward population centers, while the undesirable nature of the facilities will tend to result in a more remote and dispersed set of sites. In addition, economies of scale may dictate using relatively few facilities, while equity issues and the need to ensure that a few communities do not bear an undue burden will drive the solution toward using many smaller facilities (Ratick and White, 1988).

Stochastic inputs. The future is never known with certainty. This is particularly true of the long-term future. Thus, while traditional models assume that input parameters are known with certainty, most of these inputs should be treated as being subject to uncertainty. Future demands can only be predicted using forecasting methods and should therefore be considered uncertain. Similarly, the costs of constructing and operating facilities will change in unpredictable ways over time. Transport costs may not be known with certainty. Travel times will change with the unknown levels of future congestion and even distances may change as the network is improved or modified with time.

Dynamic decisions. Just as the inputs to location problems should be treated as stochastic or uncertain, so too should location decisions be treated as dynamic. In the real world, decisions made one day are re-evaluated and altered as the operating environment evolves. Facilities that are initially opened with limited capacity may be expanded over time. Alternatively, facilities that are opened near the beginning of a planning period may be closed as the location of demand changes or as transport and operating costs change. New facilities may be opened. Demands that are initially assigned to one facility may, in future time periods, be reassigned to another facility as demands increase and facilities reach their capacity or as transport costs change.

Location models need to account for these dynamic changes in the siting plan. Also, recognizing the uncertain nature of future demands and costs, location models should be designed to help planners identify good initial solutions that preclude as few future options as possible. In that way, good solutions for the present will be identified and, as new information is obtained, the models can be exercised again to indicate desirable changes in the configuration of facilities.

Vehicle routing considerations. All of the classical facility location problems outlined above assume either that service is delivered from a facility by a vehicle traveling to a single customer and then returning to the facility location or that customers travel individually to the facility for service. For example, the demand-weighted distance in the P-median and fixed charge location models is calculated as the distance from a facility to a demand node multiplied by the demand at the node. No consideration is given to routes that visit multiple customers. In fact, however, such routes are common in less-than-truckload (LTL) shipping. The presence of multi-customer service routes can significantly alter the transport cost component of location models. Therefore, they can also impact the number and location of facilities to be sited.

Interaction of the network with facility locations. Finally, all of the facility location models outlined above attempt to find optimal facility locations *given* the configuration of the network. In many cases, it is important to determine the network configuration and facility locations simultaneously. For example, in studying airline hub and spoke networks, we must simultaneously determine the location of the hubs, the assignment of non-hub airports to the hubs (i.e., the links to be used in the spoke part of the network), and the connectivity of the hubs. Similar problems arise in telecommunication, power transmission, and computer networks.

In developing countries, funds spent on facilities are often fungible with funds for other development purposes. For example, we may have a choice between using funds to build a school, expand a hospital, or add a new road. The ability to move funds from one activity to another suggests that facility location models must account for alternative uses of these limited resources. It is particularly important that such models account for the possibility of improving the network since network modifications may do more to reduce the average demand-weighted distance than will additional facilities.

In the sections below, we outline a number of models that have been used to address some of the complications discussed above.

Multi-objective models

As indicated above, most real-world facility location problems entail multiple conflicting objectives. Thus, the goal of a facility location analysis should not be to determine a single "optimal" siting plan. Rather, the goal should be to determine a set of non-inferior or non-dominated location plans. A plan *k* will consist of a set of facility locations and demand-to-facility assignments. Plan *k* is *non-inferior* or *non-dominated* if there is no other plan that does at least as well as plan *k* for all objectives and better than plan *k* for at least one objective.

An analyst examining a real-world location problem often develops an approximation to the non-inferior set. There are three reasons for doing so. First, the complete non-inferior set may be very large. Second, it may be time consuming or difficult to identify all solutions in the non-inferior set. Third, an approximation of the non-inferior set may be all that is required for the decision-maker to distinguish those areas of the objective space – the space indicating the possible levels of attainment of different objectives – in which she wants to concentrate future analyses from those areas which she is willing to exclude from future consideration.

Once the set of non-inferior plans has been identified either completely or approximately, the analyst must determine an appropriate means of conveying the information to the decision-maker. The information to be related includes both the tradeoffs between possible levels of attainment for the different objectives (the *objective space*) and the actual siting plans themselves (the *decision space*). Once this information is conveyed, the decision-maker may ask for additional analyses to be done or he/she may be in a position to choose between the siting plans. Additional analyses may include: (1) further exploration of the available alternatives in some portion of the objective space if only an approximation to the non-inferior set has been provided, (2) evaluation of non-modeled objectives for some or all of the solutions in the non-inferior set, and (3) inclusion of additional constraints. Additional constraints may restrict the facility sites to locations at which the agency or firm already owns property, or may exclude or include specific sites from/in the solution.

Flynn and Ratick (1988) use a multi-objective approach in designing essential air services for rural areas after airline deregulation. Belardo *et al.* (1984) use a multi-objective approach to locate multiple types of emergency equipment to be used in the event of an offshore oil spill. Finally, Current, ReVelle, and Cohon (1988) suggest a multi-objective approach to identifying the tradeoff between the distance of a path and the number of people impacted by travel on the path. The objectives are to minimize the path length and to minimize the number of people located within a covering distance of nodes on the path. This can be used in finding paths for hazardous material transport. In a related paper (Current, ReVelle and Cohon, 1985) they examine the problem in which the covered demand is to be maximized. Current, Min, and Schilling (1990) review multi-objective facility location models.

To illustrate the issues involved in multi-objective analysis, consider the problem of trading off coverage and demand-weighted distance. Such a problem may arise if it is important to both minimize the total cost of providing service (as measured by the demand-weighted distance) and maximize the number of customers that receive service within a specified standard (as measured by the number of covered demands). Increasing the number of covered demands may be possible only at the expense of increasing the average cost or average distance.

For problems in which the total demand is fixed and is independent of the level of service, maximizing the number of covered demands is equivalent to minimizing the number of uncovered demands. Thus, using the notation defined above, the multi-objective problem can be defined as follows:

$$\text{Minimize} \quad (V_1, V_2) = \left(\sum_{i \in I} \sum_{j \in J} h_i d_{ij} Y_{ij}, \sum_{i \in I} \sum_{j \notin N_i} h_i Y_{ij} \right) \tag{10.44}$$

$$\text{Subject to} \quad \sum_{j \in J} X_j = P \tag{10.45}$$

$$\sum_{j \in J} Y_{ij} = 1 \qquad \forall i \in I \tag{10.46}$$

$$Y_{ij} - X_j \leq 0 \qquad \forall i \in I, j \in J \tag{10.47}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{10.48}$$

$$Y_{ij} \in \{0,1\} \qquad \forall i \in I, j \in J \tag{10.49}$$

Note that the objective function (10.44) is no longer a scalar quantity, but has been replaced by a vector quantity. The first objective, V_1, is to minimize the demand-weighted total distance, or the objective from the P-median problem. The second objective, V_2, is to minimize the total uncovered demand. The constraints, as written above, are identical to those of the P-median problem. (For the complete derivation of the second objective in this form, the reader is referred to section 8.2 of Daskin (1995).)

The second term of the objective function can be rewritten as $\sum_{i \in I} \sum_{j \in J} h_i \hat{a}_{ij} Y_{ij}$ where $\hat{a}_{ij}$ is 1 if a facility located at candidate site j cannot cover demands at node i and is 0 otherwise. In this form, the two objectives are mathematically equivalent with the only difference being that $\hat{a}_{ij}$ replaces d_{ij} in the objective function. This observation suggests that the problem can be solved (approximately) by weighting the two objectives (Cohon, 1978). The resulting single objective is $\sum_{i \in I} \sum_{j \in J} h_i \left[\beta d_{ij} + (1-\beta)\hat{a}_{ij}\right] Y_{ij}$ where β is the relative weight placed on the P-median objective and $0 \leq \beta \leq 1$. By varying β, we can trace out the approximate tradeoff curve between the demand-weighted average distance and the number of uncovered demands.

Using this approach for the 150 node data set described in section 10.4 above and locating 8 facilities with a coverage distance of 300 miles, we obtain the tradeoff curve shown in Figure 10-4. Note that the solution to the P-median problem (point A) results in an average distance of 152.9 miles, but 17 percent of the demands are not covered within 300 miles. At the other extreme, the maximal covering model (solution F) leaves only 8.7 percent of the demand uncovered, but does so at the expense of increasing the average distance by over 18 percent to 181.1 miles. Solutions B, C, D, and E represent compromise solutions that result in a greater average distance than that attainable through the P-median solution but with better coverage.

Note that the weighting method outlined above will not find all non-inferior solutions. In fact, it will be unable to find many of the non-inferior solutions. For example, solutions within a triangle defined by solutions D, E and the hypothetical solution G in Figure 10-4 will not be found since the weighting method essentially minimizes a linear objective function that is parallel to the line segment DE. Similar triangles in which the weighting method cannot find solutions can be defined for each pair of adjacent solutions in Figure 10-4. Solutions within these triangles are known as *duality gap* solutions.

To find some of these other solutions, we can exclude one or more of the sites found in the solutions represented by the tradeoff curve shown in Figure 10-4 and resolve the problem. (This will not guarantee that new solutions will be found, but tends to work well in practice.) Houston, Texas is found in all of the solutions shown in Figure 10-4 except solution F. Excluding Houston from the set of candidate solutions and resolving the problem results in four additional non-inferior solutions (as well as one inferior solution). The new tradeoff curve is shown in Figure 10-5. The four new non-inferior solutions are shown with an (*) symbol.

Finally, we may be interested in the number of solutions in which a particular city is found. Figure 10-6 shows that only 22 cities are used in the 10 solutions shown in Figure 10-5. Note that with 8 cities per solution, there could have been as many as 80 different cities represented by the 10 solutions. New York is in 9 of the 10 solutions, while Los Angeles, Fort Wayne and Topeka are in 7 of the 10 solutions. This sort of information can be quite useful to decision-makers.

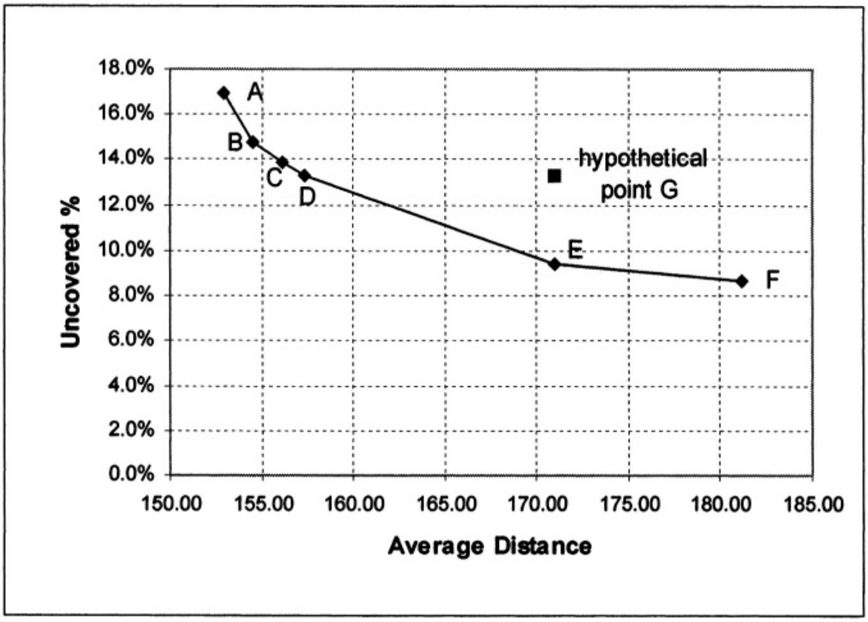


Figure 10-4. Example tradeoff curve with no duality gap solutions

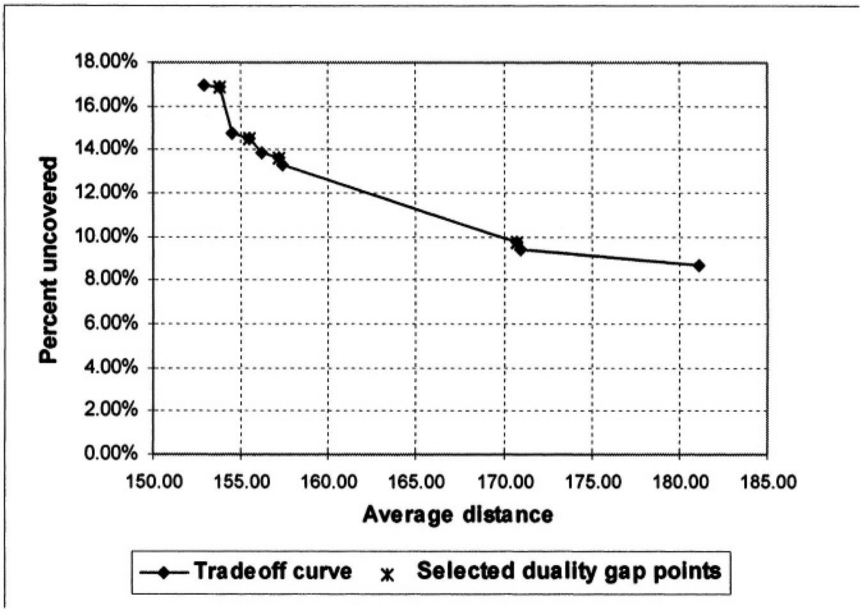


Figure 10-5. Example tradeoff curve with duality gap solutions attained by excluding Houston, Texas from the candidate solution set

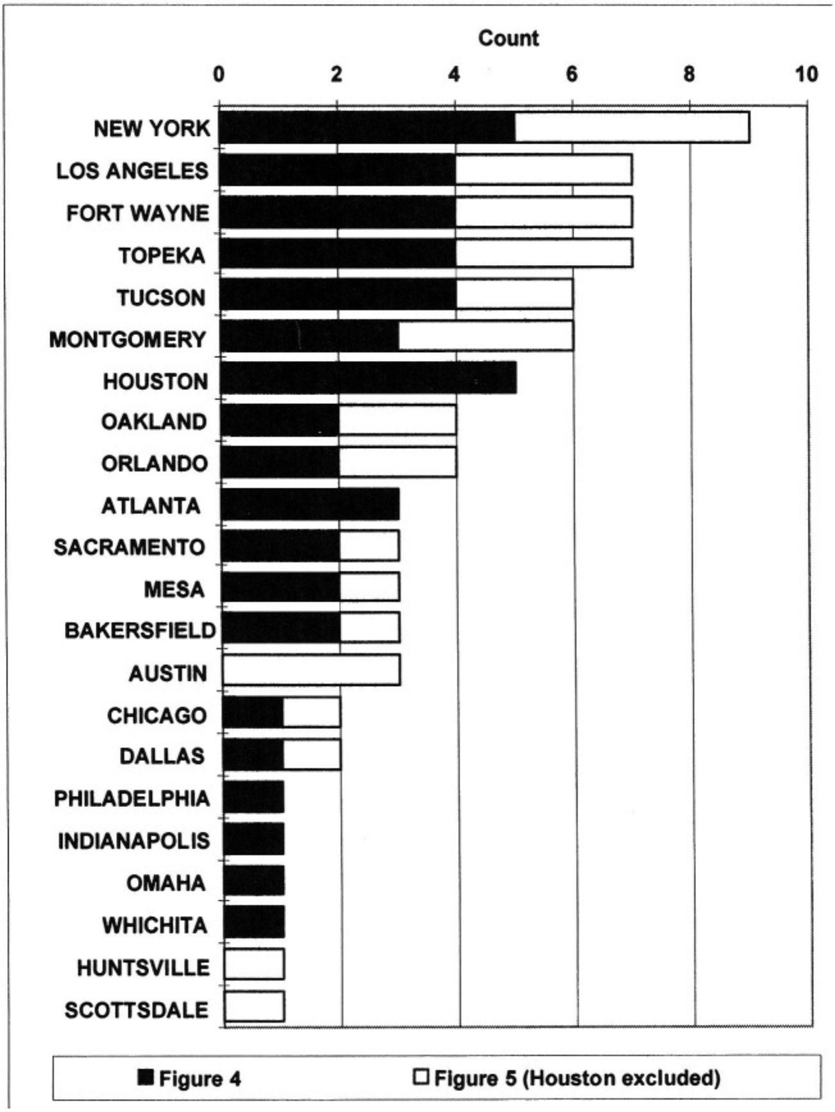


Figure 10-6. Frequency of city usage in solutions

Hierarchical objective models are closely related to multi-objective models. The key difference is that instead of trying to find a tradeoff between the objectives, one tries to optimize a primary objective and then, from among the alternate optima for the primary objective, seeks to optimize a secondary (or tertiary) objective. Daskin and Stern (1981) propose a hierarchical objective approach to locating emergency medical services or ambulances. Their primary objective is that of the set covering model, to which they append a secondary objective of maximizing a measure of backup coverage. In essence, this approach involves maximizing the sum of the slack variables in constraint (10.2). Plane and Hendrick (1977) propose a hierarchical objective location problem for locating fire stations in Denver. Their primary objective is also to minimize the number of required facilities, while their secondary objective is to maximize the number of existing facilities in the solution. Benedict (1983) outlines a number of other hierarchical objective covering models.

Stochastic inputs

While the traditional models outlined in section 10.3 treat problem inputs as known deterministic constants, in reality the inputs to almost all practical facility location models are uncertain. Thus it is important to consider the inherent stochasticity associated with these input parameters.

Manne (1961) was among the first to incorporate stochastic inputs into a previously deterministic model. He examined the problem of capacity expansion with uncertain demand, and found that the problem was mathematically equivalent to one with certainty but with a lower interest rate. As the uncertainty in demand increased, the optimal level of capacity expansion also increased. Recently, Bean, Higle and Smith (1992) extended this model to the case of non-stationary demands.

Since Manne's early work, the literature on stochastic modeling has grown tremendously. Research focusing specifically on stochastic location models can be divided into three branches. One branch treats selected inputs as uncertain and then builds stochastic programming models or chance constrained models based on those uncertain inputs. A second branch of the literature embeds queuing models into location models to account for the uncertainty in customer arrival times and facility service times. These models are most appropriate when locating emergency service stations such as ambulance bases and fire stations. Finally, scenario planning approaches are taking hold within the location literature as a third way of accounting for uncertain future events. While a full review of the stochastic location literature is beyond the scope of this chapter, we will briefly describe some of the major developments that have occurred along each of the three branches.

Probabilistic location models

Within the location literature, the incorporation of uncertainty dates back to a stochastic set covering model proposed by Chapman and White (1974). At about the same time, Carbone (1974) extended the P-median model to consider instances with random demand values. He seeks to find the minimum value of K such that

$$P\left(\sum_i \sum_j h_i d_{ij} Y_{ij} \leq K\right) = \alpha .$$

In an approach similar to that suggested by Chapman and White, Daskin (1982, 1983) formulates a probabilistic extension of the maximal covering problem in which facilities are treated as being either busy or available. Facilities are assumed to be busy with probability ρ. This results in the maximum expected covering model whose objective is to maximize the number of demands that are covered by an available facility. ReVelle and Hogan (1989a) formulate a similar model in which they maximize the number of demands that are covered at least b times, where b is the number of coverages needed to ensure that a demand is covered by an available facility with probability β. ReVelle and Hogan (1989b) propose a similar extension of the set covering model in which they minimize the number of facilities required to ensure that all demands are covered with probability β when facility availability is accounted for. Daskin, Hogan and ReVelle (1988) summarize these and other related models.

Marianov and ReVelle (1992) extend the expected covering model to incorporate multiple vehicle types housed within each facility. The authors define demands to be covered only if there is a sufficient number of available fire engines and fire trucks within the coverage distance of the demand node.

Marianov and ReVelle (1996) use an M/G/s-loss queuing model to compute the minimum number of facilities needed to ensure coverage of a node with at least probability α. This number is then used as the required number of facilities that must be located within a given distance of a node for the node to be covered. The model also incorporates stochastic travel times.

In a somewhat different vein, Louveaux (1986) adopts a stochastic programming with recourse approach to the P-median problem. Demands and travel times are treated as random variables and the assignment variables are part of the recourse function.

Queuing-based location models

The approaches described above typically use probabilistic analyses to (a) determine the number of facilities needed to cover demands with a specified probability or (b) determine the incremental expected coverage that can be provided by having one additional available facility cover a demand node. In the first case, the objective is typically to minimize the number of facilities needed to cover all demands or to maximize the number of covered demands. In the second case, the objective is to maximize the total expected coverage. Neither of these approaches explicitly accounts for the queuing interactions that occur in a spatially distributed queuing system.

Larson (1974) proposes a hypercube queuing model to account for different server availabilities. For a system with N servers he defines 2^N states with each state representing a unique combination of busy and idle servers. Thus, for example, state (0,1,0,0) represents the state in which servers 1, 3, and 4 are idle, but server 2 is busy. These states are used to develop a system with 2^N simultaneous linear equations for analyzing the performance of spatially distributed emergency response systems. Alternatively, N simultaneous non-linear equations can be used in an approximate approach (Larson, 1975). Batta, Dolan and Krishnamurthy (1989) use Larson's hypercube model to examine the validity of the assumptions in Daskin's expected covering model and find that some assumptions (e.g., the implicit assumption of server independence) are violated in the multi-server case. Using Larson's queuing model in a location context is difficult due to the large number of equations that need to be solved in evaluating a potential configuration of facility sites.

Berman, Larson, and Chiu (1985) model a single facility location problem as an M/G/1 queue. They show that if demands are lost when the server is busy, then at least one node of the graph is an optimal solution to the problem and that node corresponds to the Hakimi median. In this case, the objective is to minimize the sum of the expected travel time to demands that are served plus a penalty term for unserved demands. When demands can queue for service, the objective is to minimize the sum of the time in the queue and the travel time to the demands. The authors show that in this case the optimal location begins at the Hakimi median for sufficiently small overall demand values, moves away from that location as the demands increase, and then returns there for high demand values.

Mirchandani and Odoni (1979) extend the P-median problem to include stochastic travel times. The travel times are assumed to come from discrete distributions. The state of the system can then be described by a vector of link travel times corresponding to a realization of the random variables associated with each of the link travel times. The state of the system is assumed to change according to a Markov process. The authors show that the Hakimi (node optimality) property extends to this case as long as the utility function for travel time is convex and non-increasing. Mirchandani (1980) discusses further extensions along these lines. Weaver and Church (1983) develop computational methods for solving such problems.

Berman and Odoni (1982) and Berman and LeBlanc (1984) also consider the case in which travel times change according to a Markov process. Each state of the Markov process corresponds to a unique combination of link travel times. The key difference between this and the earlier work of Mirchandani and Odoni is that facilities can be relocated (at a cost) in response to changes in the state of the system.

Berman and Odoni show that the Hakimi (node optimality) property holds for this problem when relocation costs are concave functions of the travel times. They also propose a node exchange heuristic for solving the problem when only one facility is to be located. Berman and Leblanc extend the heuristic to the case of multiple facilities. Berman (1985) considers a similar problem for a single facility in the presence of state-dependent travel times and Poisson demands.

Scenario-based location models

The third major approach to handling uncertain problem inputs is to use scenarios (van der Heijden, 1994; Vanston *et al.,* 1977). While probabilistic and queuing-based approaches tend to focus on short-term stochasticity, scenario-based approaches are able to deal with longer-term uncertainty in the problem inputs. That is, in the probabilistic and queuing-based approaches the system is likely to be in any of a number of different states at different times during the operation of the located facilities. Thus, ambulances will be alternately busy and idle during the course of any single day during the operation of an emergency medical system. Similarly, each of the possible 2^N states in a hypercube queuing model is likely to be visited at some point in time during the operation of the system being modeled. On the other hand, in some scenario planning models only *one* of the possible future scenarios will be realized, and we are uncertain as to which one it will be.

A scenario represents a possible set of future conditions. In a location context, these conditions will include demand values, fixed costs, operating costs, travel times and capacities. Vanston *et al.* (1977) propose a 12-point approach to generating alternate scenarios. They recommend the use of 3 to 6 scenarios, one of which should be "the one believed to be the most probable." Other scenarios "should be chosen according to the degree to which they provide maximum value to the planning process" (p. 161). They apply the approach to the generation of alternate scenarios for assessing national policies concerning portable fuels.

The notion of *regret* plays a critical role in analyzing scenarios. The regret associated with a scenario *k* is the difference between the objective function value associated with a compromise solution and the value of the optimal solution for scenario *k*. Two common objectives that are used in scenario planning are minimizing the expected regret and minimizing the maximum regret over all scenarios. In addition, we sometimes optimize the expected behavior of the system or minimize the worst-case performance of the system.

To illustrate the notion of regret, let $\hat{V}_k$ be the optimal value of the P-median problem under scenario *k*. Index the assignment variables (Y_{ij}), the demands (h_i), and the distances (d_{ij}) by *k* as well to indicate that they are scenario-dependent. Let *K* be the set of all scenarios. To minimize the maximum regret, we need to solve the following optimization problem.

$$\text{Minimize} \qquad U \tag{10.50}$$

$$\text{Subject to:} \qquad \sum_{j\in J} X_j = P \tag{10.51}$$

$$\sum_{j\in J} Y_{ijk} = 1 \qquad \forall i \in I, k \in K \tag{10.52}$$

$$Y_{ijk} - X_j \leq 0 \qquad \forall i \in I, j \in J, k \in K \tag{10.53}$$

$$U - \left(\sum_{i\in I} \sum_{j\in J} h_{ik}\, d_{ijk}\, Y_{ijk} - \hat{V}_k \right) \geq 0 \qquad \forall k \in K \tag{10.54}$$

$$X_j \in \{0,1\} \qquad \forall j \in J \tag{10.55}$$

$$Y_{ijk} \in \{0,1\} \qquad \forall i \in I, j \in J, k \in K \tag{10.56}$$

The objective function (10.50) minimizes the maximum regret over all scenarios. Constraints (10.51) and (10.55) are identical to (10.17) and (10.20) respectively. Constraints (10.52), (10.53) and (10.56) are direct analogs of (10.18), (10.19) and (10.21) with the scenario subscript k appended to the assignment variables. Constraint (10.54) defines the maximum regret (U) as the largest difference over all scenarios between the objective function for scenario k using the compromise locations ($\sum_{i\in I} \sum_{j\in J} h_{ik}\, d_{ijk}\, Y_{ijk}$) and the best that can be done under scenario k ($\hat{V}_k$).

Sheppard (1974) was one of the first to propose the use of scenarios in facility location planning. He suggests finding a solution that minimizes the expected cost over all scenarios. Schilling (1982) extends the maximal covering location problem to incorporate scenarios. He maximizes the number of covered demands over all possible scenarios. In the model he proposes, some facilities must be common to all scenarios, while others can be located in a scenario-specific manner. Schilling examines the tradeoff between the number of common facilities and the maximum percentage decrease over all scenarios from the optimal coverage. He then suggests that building sites common to all scenarios in this tradeoff analysis first will allow decision-makers to gather additional information regarding the evolving future scenario before specifying the remaining location sites. Daskin, Hopp and Medina (1992) show that this approach can lead to the selection of the worst possible sites under certain conditions.

Serra and Marianov (1997) use scenarios to model different travel time and demand conditions over the course of a day. They attempt to find locations that minimize the maximum average travel time over seven defined scenarios as well as locations that minimize the maximum regret over these scenarios. Carson and Batta (1990) also use scenarios to describe demand conditions at different times of the day in locating an ambulance on the State University of New York's Amherst campus. Jornsten and Bjorndal (1994) formulate an uncapacitated dynamic fixed charge facility location model using a scenario planning approach. Their objective is to minimize the expected cost across all scenarios and time periods. Scenario and policy aggregation is used to solve the model.

Serra, Ratick and ReVelle (1996) use scenarios to describe uncertain future demands. They too adopt a minimax regret approach. Ghosh and McLafferty (1982) use a similar approach to locate retail stores. Kouvelis and Yu (1996), in a recent text, adopt a robustness approach to a variety of discrete optimization problems including the 1-median problem on a tree. More recently, Averbakh (1997) and Averbakh and Berman (1997a, 1997b, 1997c) have employed the notion of scenarios and regret in locating facilities. The focus of their work is on the development of polynomial time algorithms for specially structured instances of these problems.

The expected regret objective is unsatisfactory for two reasons. First, it requires planners to identify probabilities associated with the scenarios. There is considerable debate within the literature about whether doing so is appropriate or warranted. Second, and perhaps more importantly, this objective tends not to place sufficient weight on extreme scenarios. On the other hand, the minimax regret approach tends to place too much weight on extreme scenarios as it allows a single scenario to define the siting plan.

To remedy this problem of extreme scenarios, Daskin, Hesse and ReVelle (1997) formulate the α-reliable P-minimax regret model. The model minimizes the maximum regret over an endogenously determined subset of the scenarios, referred to as the reliability set. The total probability associated with the reliability set must be at least α. Owen (1998) formulates additional demand- and scenario-based models that address the problem of having to specify scenario probabilities. The demand-based models maximize the number of demands that are adequately served in at least *m* of the scenarios, where *m* is an input. Adequate service can be defined in a number of ways, but is here defined as a function of the best service that a node can receive under the optimal locations for the scenario in question. In the author's scenario-based models, the objective is to maximize the number of scenarios in which a target objective is attained. The target objective is scenario-dependent. These models are solved using either branch and bound in a commercial optimization package or using a specially designed evolutionary algorithm (Owen and Daskin, 1998).

Dynamic decisions

In an early paper on dynamic location problems, Ballou (1968) states, "... the effect of the future time dimension cannot be neglected in location analysis." (p. 271). He develops an algorithm for solving a single facility location problem in which demands change deterministically over time. Sweeney and Tatham (1976) extend these results and show that Ballou's approach is suboptimal. They develop an optimal approach that involves considering not only the best solution at period *t*, but the best R_t solutions. The solutions are found using a variant of Benders' decomposition. Wesolowsky (1973) considers a problem similar to that modeled by Ballou but in a plane. Tapiero (1971) also examines planar location problems and formulates a transportation, location and inventory model using control theory. Necessary conditions are presented but computational results are not provided. Ratick *et al.* (1987) adopt a different approach to a joint location/inventory problem. Motivated by a problem involving the storage, loading and unloading of coal at ports, they propose a linked set of network models. Osleeb and Ratick (1990) use a similar model to explore just-in-time inventory policies.

Drezner and Wesolowsky (1991) consider the problem of when to move a single facility located on a plane in response to linearly growing demands. Campbell (1990) also models the case of increasing demand and considers location, relocation, and transport costs. He develops lower and upper bounds on the optimal objective function values. The heuristic solution algorithm he proposes allows one relocation per unit time.

Wesolowsky and Truscott (1973) use dynamic programming to model multi-facility location and relocation decisions over time. Their approach breaks down when either the number of new facilities or the number of facilities to be relocated is large in any period. Scott (1971) also uses dynamic programming to solve a multi-period problem. He simplifies the problem by assuming that the optimal static siting plan for the last period will be optimal for the dynamic case as well. The problem then becomes one of sequencing the construction of these facilities, with one new facility being added in each period. Drezner (1995a) considers a similar problem in which one new facility is added at each point in time. In period *j,* he needs to solve a problem similar to the j-median problem. The problem is modeled in AMPL (Fourer, Gay, and Kernighan, 1993).

Most dynamic models simply extend a static model by adding a temporal subscript to the location and assignment variables and linking the location decisions in one period with those of the subsequent or preceding period. The classical work of Van Roy and Erlenkotter (1982) is typical of this approach. A temporal subscript is added to an uncapacitated facility location problem. New facilities may be opened (and then remain open for the duration of the planning period) and existing facilities may be closed (and remain closed for the remainder of the planning period). The problem is solved using a dual ascent algorithm (similar to that developed by Erlenkotter, 1978) embedded in a branch and bound procedure. Roodman and Schwarz (1975) use integer linear programming with a temporal subscript appended to the location and allocation variables to consider the optimal sequence of facility closings in the face of declining demand. The problem is solved using branch and bound.

Schilling (1980) develops a multi-objective approach to dynamic facility location. He extends the maximal covering problem to a multi-period context by allowing the coverage sets and locations to be time dependent. Facilities must remain open once they are initially opened. His vector optimization problem is to maximize the coverage in each of *T* periods. Each period represents a different objective. Gunawardane (1982) also considers dynamic extensions of the set covering and maximal covering problems and allows facilities to be opened and closed over time. Min (1988) also considers a dynamic location problem in a multi-objective context. His work deals with expanding and relocating libraries in Columbus. Objectives for this application include population coverage and accessibility to transport routes and parking lots.

Ratick, Du and Moser (1992) consider a rather different dynamic location problem. Rather than planning for future conditions that evolve over time, they need to plan for cyclic conditions that change monthly and repeat on an annual basis. Their model is concerned with the location and use of different forms of dredging equipment to ensure that a channel depth remains available with a given probability. The problem is formulated as a mixed integer chance constrained problem.

Finally, we refer the reader to Owen and Daskin (1998) for a more complete review of both stochastic and dynamic facility location models.

Integrated vehicle routing and facility location models

The models outlined above all assume that demands are served directly from the facilities being located. Such an assumption is valid for either truckload shipping or for many emergency services. However, we need to modify the models above to account for more complex and sometimes more realistic shipment patterns. For example, in the case of less-than-truckload (LTL) shipping, vehicles will depart from one of the facilities being located, visit a number of different customers and return to the facility. In hub location problems, commodities flow from origins to destinations through the facilities (hubs) whose location we are trying to find. Often commodities move between hubs at a discounted transport cost.

In this section, we outline models that extend the basic formulations of section 10.3 to include more complex vehicle routing.

Location-routing models

There is a vast literature on vehicle routing models. Most such models are based on the traveling salesman problem. The reader is referred to Lawler *et al.* (1985) for an excellent introduction to this problem. For reviews of related models the reader is referred to Golden and Assad (1988) and Fisher (1995). Early work on solving vehicle routing problems focused on a variety of construction and improvement heuristics (e.g., Clarke and Wright, 1964; Gillet and Miller, 1974). This was followed by the development of optimization-based algorithms (e.g., Fisher and Jaikumar, 1981). The focus of much of the work today is on (a) integrating vehicle routing with other components of logistics, including inventory management (Chien, Balakrishnan and Wong, 1989) and facility location and (b) the application of modern heuristics to vehicle routing problems, including tabu search (Gendreau, Hertz and Laporte, 1994; Xu and Kelly, 1996; Duhamel, Potvin and Rousseau, 1997; and Taillard *et al.,* 1997), genetic algorithms, and simulated annealing.

A complete review of such models is clearly beyond the scope of this chapter. Bodin *et al.* (1983) reviewed the state of the art in vehicle routing at that time and cited over 600 references and the literature has continued to explode in this field in the 15 intervening years. Instead, the focus of this chapter will be on (a) integrated models of facility location and vehicle routing, (b) an innovative use of facility location models to solve a vehicle routing problem, and (c) hub location problems.

Integrated location-routing models

Webb (1968) and Eilon, Watson-Gandy and Christofides (1971) were among the first to point out that modeling distribution cost as the cost of a round trip from a facility to a customer may significantly misrepresent the actual costs and may, as a consequence, result in the selection of sub-optimal facility sites when multi-stop tours are used. Perl and Daskin (1985) formulated an integrated integer linear programming problem for the warehouse location routing problem. They recognized that the problem involves three inter-related and fundamental decisions: where to locate facilities, how to allocate customers to facilities, and how to route vehicles through the customers allocated to a facility. They solved the problem by iteratively applying a set of three heuristics. Each heuristic attacked a pair of the three fundamental decisions. Using another approach, Srivastava (1993) develops savings (Clarke and Wright, 1964) and clustering heuristics to solve the problem. He found that these heuristics generally opened the same number of facilities as did an optimal algorithm and that the total costs from the heuristics were generally within 5% of the optimal total cost.

Laporte (1988) reviews location-routing models and discusses applications, formulations and solution approaches. He classifies formulations into two categories, those that use a three-index notation and those that use a two-index notation. The three-index notation involves assignment variables that represent whether vehicle k goes directly from customer i to customer j on a route. The two-index formulations handle problems with a symmetric cost or distance matrix. Variables now represent whether or not edge (i, j) is in the solution, or in capacitated cases, the number of times the edge appears in the optimal solution. He then identifies a graph extension and transformation of the cost matrix that allows the two-index approach to be applied to asymmetric problems. The graph extension involves generating multiple copies of each candidate facility location, where the number of copies is equal to the number of possible routes emanating from the facility.

Laporte, Nobert, and Taillefer (1988) consider three variants of location-routing problems, including (1) capacity constrained vehicle routing problems, (2) cost constrained vehicle routing problems, and (3) cost constrained location-routing problems. The authors examine multi-depot, asymmetrical problems and develop an optimal solution procedure that enables them to solve problems with up to 80 nodes. The solution procedure involves transforming the associated graph and formulating the problem as a variant of the traveling salesman problem. A specialized branch and bound algorithm is then used to solve the problem optimally.

In dynamic versions of location-routing problems, one must determine the optimal sequence of depot, vehicle, and route configurations over a given time horizon. Laporte and Dejax (1989) present an exact and an approximate solution method for this problem extension. The optimal solution method requires representing the dynamic location-routing problem as a network and then solving the integer linear program associated with the network. This method is appropriate only for small-scale problem instances. The heuristic method utilizes approximations of system costs and a directed graph formulation in determining a global solution. This method can be used for large-scale problems if certain conditions are met.

Berger (1997) formulates a location routing problem for perishable commodities as a variant of a fixed charge facility location problem. Her model is applicable in cases in which the routes are constrained to be short (since the commodity is perishable) and the vehicle does not have to return to the original depot within the time window. As such, she does not model the cost or time associated with the return of the vehicle to the depot from the final customer served on the route. To formulate this problem, she defines the following notation:

$$a_{ik} = \begin{cases} 1 & \text{if customer i is on path k} \\ 0 & \text{if not} \end{cases}$$

c_{jk} = cost of serving path k from a facility at candidate site j

α = the relative cost of routing compared to facility location

f_j = fixed cost of locating a facility at candidate site j

P_j = set of all paths k such that the assignment of path k to a facility at candidate site j is feasible in terms of the time constraints

$$Y_{jk} = \begin{cases} 1 & \text{if feasible path k is served from a facility at site j} \\ 0 & \text{if not} \end{cases}$$

With these definitions, she formulates the problem as follows:

$$\text{Minimize} \quad \sum_{j \in J} f_j X_j + \alpha \sum_{j \in J} \sum_{k \in P_j} c_{jk} Y_{jk} \tag{10.57}$$

$$\text{subject to} \quad \sum_{j \in J} \sum_{k \in P_j} a_{ij} Y_{jk} = 1 \qquad \forall i \in I \tag{10.58}$$

$$X_j - Y_{jk} \geq 0 \qquad \forall i \in I; \forall j \in J \qquad (10.59)$$

$$X_j \in \{0,1\} \qquad \forall j \in J \qquad (10.60)$$

$$Y_{jk} \in \{0,1\} \qquad \forall j \in J; \forall k \in P_j \qquad (10.61)$$

The similarity between this formulation and the uncapacitated fixed charge problem (10.22)-(10.24), (10.26) and (10.27) should be clear. The objective function (10.57) minimizes the combined facility location and routing costs. Constraint (10.58) stipulates that each customer *i* must be assigned to exactly one path. Constraint (10.59) states that paths can only be assigned to open facilities. Finally, constraints (10.60) and (10.61) are standard integrality constraints.

There are two problems with this formulation. First, the number of possible paths is astronomical. For example, for a problem with 100 nodes, even if we knew that paths could not include more than 5 customers, there would be nearly 80 million possible paths (only a fraction of which would be feasible in terms of a time or distance constraint on the paths). If we had 20 candidate locations, this would result in nearly 1.6 billion assignment variables (Y_{jk}). This problem was handled using column generation. Second, the linear programming relaxation of this formulation is very weak. Customers are typically assigned to multiple paths. Constraint (10.59) then requires only a fraction of a facility to be opened. To improve the formulation, Berger introduced the following constraint as a replacement for (10.59):

$$X_j - \sum_{k \in P_j} a_{ij} Y_{jk} \geq 0 \qquad \forall i \in I; \forall j \in J \qquad (10.62)$$

This constraint requires the location variable to be greater than or equal to the sum of the path assignment variables for any customer assigned to that facility. Since, for each opened facility, it is likely that at least one customer will be assigned only to routes emanating from that facility, this constraint significantly tightens the formulation. Branching only on the location variables, Berger solved the problem for 323 of 324 test problems in which the number of customers, number of candidate locations, maximum route length, and relative routing/location weight were varied. Only 29 of the 323 problems had a gap exceeding 1% of the lower bound. The average gap was under 0.5% and the maximum gap was under 4%.

Applications of location-routing problems have also been studied in the literature. List *et al.* (1991) survey recent research related to hazardous materials transportation, including work in the areas of risk analysis, routing/scheduling, and facility location. They illustrate the evolution of models from simple, single objective formulations to complex, multi-objective or probabilistic models. ReVelle, Cohon, and Shobrys (1991) examine the problem of storing spent fuel rods from commercial nuclear reactors. To solve this problem, a model is needed which sights storage facilities, assigns reactors to the facilities, and selects routes for shipping the fuel rods. The nature of the problem requires consideration of multiple objectives, considering transportation costs as well as perceived risk. Several methods are combined to derive optimal solutions for this complex location-routing problem.

A location-based heuristic for the vehicle routing problem

The intersection between the location and routing literatures has generally been limited to integrated location-routing models of the form outlined above. In a notable exception, Bramel and Simchi-Levi (1995) present a framework for modeling capacitated vehicle routing problems as capacitated concentrator location models with the single sourcing constraint imposed to identify good customer groups for the vehicle routing problem. The fixed cost associated with any candidate location is the stem cost of going from a depot whose location is known to the center of a group of customers to be served on a route. The cost of going to and from each customer from the center of the route is the connection cost in the fixed charge location problem. A nearest insertion algorithm is used to build vehicle routes through the points associated with each facility (or concentrator or group of customers). They show that this model is asymptotically optimal in the sense that as the number of customers increases the relative error between the solution produced by their algorithm and the optimal solution goes to zero. Finally, they indicate how their approach can be used to attack a variant of the inventory routing problem.

Hub location models

Finally, we consider hub location problems first introduced by O'Kelly (1986a, 1986b), in which one locates a number of hubs through which origin-destination flows are to be routed. The objective is generally to minimize the demand-weighted total transport cost, though other objectives, including covering and center objectives, have been formulated (Campbell, 1994; Daskin, 1995). The advantage of using a hub network as opposed to a fully connected network is twofold. First, fewer links need to be constructed (O'Kelly, 1986b). Second, there may be significant economies of scale associated with flows between hubs (O'Kelly, 1986a).

Figure 10.7 illustrates a simple hub and spoke network. Nodes A through I are spoke nodes, while nodes X, Y, and Z are hub nodes. Each spoke is connected to a single hub while the network of hub nodes is a fully connected graph. This is the typical configuration assumed in most of the literature, though some models relax this set of assumptions.

O'Kelly (1987) formulates the hub location problem with these assumptions as a quadratic assignment problem. He notes that even in the absence of hub or link capacities, it may not be optimal to assign each spoke to the nearest hub as is the case in most uncapacitated facility location problems. To see why this is so, consider Figure 10.7 and assume that the flows between node A and nodes Z, G, H, and I are high while the flow between node A and nodes B through F and nodes X and Y are relatively small. In this case, even though node A is closer to hub X, it may be better to assign node A to hub Z, particularly if the inter-hub transport cost is not significantly discounted. If the inter-hub transport cost is sufficiently discounted (e.g., if the inter-hub transport cost is effectively zero), then assignment of spoke nodes to the nearest hub will always be optimal. O'Kelly (1987) proposes two heuristics for solving the hub location problem. In the first, each spoke node is assumed to be assigned to the nearest hub, while in the second, every possible combination of assigning a spoke node to its nearest and second nearest hub is considered for each possible configuration of hubs.

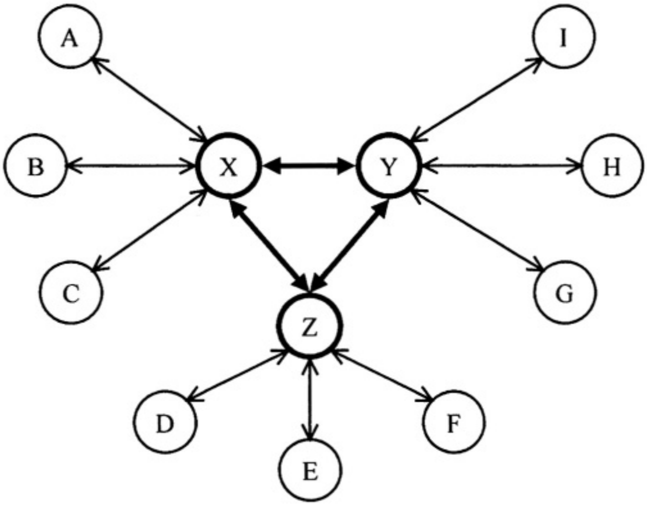


Figure 10-7. Example hub and spoke network with spoke nodes A through I and hub nodes X, Y and Z

Skorin-Kapov and Skorin-Kapov (1994) compare the performance of the O'Kelly heuristics with that of a new tabu search method for the hub location problem. In the tabu search heuristic, equal importance is given to the problems of locating the P hubs and allocating each non-hub node to a single hub. The tabu search method dominates both of the O'Kelly heuristics on a test set of problems from the literature.

Ernst and Krishnamoorthy (1997) also examine the use of modern heuristic techniques for solving hub location problems. They reformulate the uncapacitated, single allocation P-hub median problem as a mixed integer LP with fewer variables and constraints than previously reported. Using a simulated annealing algorithm, the authors find bounds comparable to those obtained with the tabu search method of Skorin-Kapov and Skorin-Kapov, both in terms of solution quality and time.

O'Kelly and Miller (1994) and Campbell (1994) review hub location models. In addition, a recent issue of *Location Science* (volume 4, number 3, October, 1996) was devoted largely to such models.

Integrated network design and facility location models

Hub location models can be viewed as integrating facility location and network design decisions. This is particularly true when (1) spoke nodes are not necessarily assigned to the nearest hub and (2) the graph associated with the hub nodes is not a complete graph. In these cases, important network design problems arise in the context of hub location problems.

There have been relatively few papers, however, which directly examine the impact of network design on facility location and vice versa. One of the first such papers is that of Berman, Ingco and Odoni (1992) which considers the impact of arc reductions and arc additions on the P-median objective when the locations of the facilities are known. Reductions are actions that a planner can take to reduce the travel time or cost associated with an existing link, while link additions actually change the topography of the network. For network reductions, the authors consider two cases. In the first, the total number of units by which the network arcs can be reduced is specified and the problem is to allocate the reduction over the arcs in the most beneficial manner. In the second case, there is a unit cost per reduction associated with each link and a total budget for reducing the network. The problem is thus to allocate the available budget over the links. Both trees and general graphs are considered.

Peeters and Thomas (1995) perform a large number of computational experiments to determine the impact of different network topologies on the location of facilities and the objective function value for the P-median problem. They conclude, not surprisingly, that there is a "relationship between the shape of the network and the optimal location."

Finally, Melkote (1996) generalizes three classic location problems -- the uncapacitated fixed charge location problem, the maximal covering problem, and the capacitated fixed charge location problem -- to incorporate link selection. For the three problem classes he identifies, integrated design/location models are formulated. Using approaches similar to those of Balakrishnan, Magnanti and Wong (1989), he identifies ways of formulating the uncapacitated network design/facility location problem that lead to relatively tight linear programming relaxations. The fixed charge problems are solved using a conventional branch and bound package, while he develops a specially structured heuristic to obtain solutions for large-scale maximal covering problems.

10.6 Conclusions

The area of facility location modeling has attracted transportation researchers, as well as others, for at least three reasons. First, the problems are of significant strategic importance in both the public and private sectors. Second, the problems are methodologically challenging. Almost all problems of interest are NP-hard on a general network. When the basic models are extended to incorporate such issues as multiple objectives, dynamic decision making, uncertain inputs, vehicle routing, or network design, the models become increasingly relevant to real-world applications and increasingly difficult to solve. Finally, though we have not discussed this work in detail, facility location models have been applied in a variety of non-location contexts. To list only a few examples, Chung (1986) reviews a number of such applications of the maximal covering location problem. Watson (1996) applied P-median modeling approaches to the problem of rationalizing steel coil purchases at an auto manufacturer. Daskin, Jones and Lowe (1990) and Hsu *et al.* (1995) apply covering models to the problem of selecting a set of standard tools in a flexible manufacturing context.

Research in this area is likely to continue at a vigorous pace for some time to come because of its practical importance and methodological value. In particular, we envision increased attention being devoted to multi-objective problems, particularly for public sector decisions. Improved methods for identifying non-inferior (or nearly non-inferior) solutions are likely to be developed using modern heuristics. Such techniques will undoubtedly be embedded in software that can run quickly on personal computers and enables decision-makers to rapidly modify the problem as they explore alternative courses of action.

Increased attention will certainly be focused on better ways of incorporating stochastic and dynamic problem characteristics into location models. The long-term strategic nature of the underlying decisions coupled with the inherent uncertainty associated with the data inputs will demand this sort of attention. In this regard, we expect that scenario planning will become an increasingly important tool in the armada of approaches that location planners bring to bear on real-world problems.

Most attempts at integrating facility location and vehicle routing have treated the problems as if they occur on the same time scale. In fact, location decisions are long-term, while routing decisions can be changed, and often are changed, on a daily basis. Thus, there is a need for models that identify facility locations that are robust with respect to a range of vehicle routing decisions.

Finally, the interaction between network design and facility location is relatively unexplored. This area will become increasingly important as a variety of transportation networks (e.g., intelligent highway and transport systems) and communication and data transmission networks are developed.

10.7 References

Aneja, Y. P. and Parlar, M. (1994) Algorithms for Weber facility location in the presence of forbidden regions and/or barriers to travel. *Transportation Science* **28**, 70-76.

Averbakh, I. (1997) *On the complexity of a class of robust location problems.* Working Paper, Western Washington University.

Averbakh, I. and Berman, O. (1997a) *Algorithms for the robust 1-center problem.* Working Paper.

Averbakh, I. and Berman, O. (1997b) *Minimax regret p-center location on a network with demand uncertainty. Location Science* **5**, 247-254.

Averbakh, I. and Berman, O. (1997c) Minimax regret robust median location on a network under uncertainty. Working Paper.

Ballou, R. H. (1968) Dynamic warehouse location analysis. *Journal of Marketing Research* **5**, 271-276.

Batta R., Dolan J. M. and Krishnamurthy, N. N. (1989) The maximal expected covering location problem: Revisited. *Transportation Science* **23**, 277-287.

Batta, R., Ghose, A. and Palekar, U. S. (1989) Locating facilities on the Manhattan metric with arbitrary shaped barriers and convex forbidden regions, *Transportation Science* **23**, 26-36.

Bean, J. C., Higle, J. L. and Smith, R. L. (1992) Capacity expansion under stochastic demands, *Operations Research* **40**, S210-S216.

Belardo, S., Harrald, J., Wallace, W. A. and Ward, J. (1984) A partial covering approach to siting response resources for major maritime oil spills. *Management Science* **30**, 1184-1196.

Benedict, J. M. (1983) Three hierarchical objective models which incorporate the concept of excess coverage to locate ems vehicles or hospitals. M.S. thesis, Department of Civil Engineering, Northwestern University, Evanston, 1L 60208.

Berger, R. T. (1997) Location-Routing Models for Distribution System Design. Ph.D. dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, 1L60208.

Berman, O. (1985) Locating a facility on a congested network with random lengths. *Networks,* **15**, 275-294.

Berman, O., Ingco, D. I. and Odoni, A. R. (1992) Improving the location of minisum facilities through network modification. *Annals of Operations Research* **40**, 1-16.

Berman O., Larson, R. C. and Chiu, S. S. (1985) Optimal server location on a network operating as an M/G/1 queue. *Operations Research* **33**, 746-771.

Berman, O. and LeBlanc, B. (1984) Location-relocation of mobile facilities on a stochastic network. *Transportation Science* **18**, 315-330.

Berman, O. and Odoni, A. R. (1982) Locating mobile servers on a network with Markovian properties. *Networks* **12**, 73-86.

Bodin, L., Golden, B. L., Assad, A., and Ball M. (1983) The state of the art in the routing and scheduling of vehicles and crews. *Computers and Operations Research* **10**, 63-212.

Bramel, J. and Simchi-Levi, D. (1995) A location based heuristic for general routing problems. *Operations Research* **43**, 649-660.

Brandeau, M. L. and Chiu, S. S. (1989) An overview of representative problems in location research. *Management Science* **35**, 645-674

Campbell, J. F. (1990) Locating transportation terminals to serve an expanding demand. *Transportation Research* **24B**, 173-192.

Campbell, J. F. (1994) Integer programming formulations of discrete hub location problems. *European Journal of Operational Research* **72**, 387-405.

Carbone, R. (1974) Public facilities location under stochastic demand. *1NFOR* **12**, 261-270.

Carson, Y. M. and Batta, R. (1990) Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. *Interfaces* **20**:5, 43-49.

Chapman, S. C. and White, J. A. (1974) Probabilistic formulations of emergency service facilities location problems. Paper presented al the 1974 ORSA/T1MS Conference, San Juan, Puerto Rico.

Chien, T. W., Balakrishnan, A. and Wong, R. T. (1989) An integrated inventory allocation and vehicle routing problem. *Transportation Science* **23**, 67-76.

Church, R. L. and Meadows, M (1979) Location modeling utilizing maximum service distance criteria. *Geographical Analysis* **11**, 358-373.

Church, R. L. and ReVelle, C. (1974) The maximal covering location problem. *Papers of the Regional Science Association* **32**, 101-118.

Clarke, G. and Wright, J. W. (1964) Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research* **12**, 568-581.

Cohon, J. L. (1978) *Multiobjeclive Programming and Planning.* Academic Press, New York, NY.

Current, J., Min, H. and Schilling, D. (1990) Multiobjective analysis of facility location decisions. *European Journal of Operational Research* **49**, 295-307.

Current, J., Ratick, S. and ReVelle, C. (1997) Dynamic facility location when the total number of facilities is uncertain: A decision analysis approach. Forthcoming in the *European Journal of Operational Research.*

Current, J., ReVelle, C. and Cohon, J. (1985) The maximum-covering/shortest-path problem: A multiobjective network design and routing problem. *European Journal of Operational Research* **21**, 189-199.

Current, J., ReVelle, C. and Cohon, J. (1988) The minimum-covering/shortest-path problem. *Decision Sciences* **19**, 490-503.

Daskin, M. S. (1982) Application of an expected covering model to emergency medical service system design. *Decision Sciences* **13**, 416-439.

Daskin, M. S. (1983) A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science* **17**, 48-70.

Daskin, M. S., Hesse, S. M. and ReVelle, C. S. (1997) α-reliable P-minimax regret: A new model for strategic facility location modeling. *Location Science* **5**, 227-246.

Daskin, M. S., Hogan, K. and ReVelle, C. (1988) Integration of multiple, excess, backup, and expected covering models. *Environment and Planning B* **15**, 15-35.

Daskin, M. S., Jones, P. C. and Lowe, T. J. (1990) Rationalizing tool selection in a flexible manufacturing system for sheet metal products. *Operations Research* **38**, 1104-1115.

Daskin, M. S. and Owen, S. H. (1998) Two new location covering problems: The partial covering *P*-center problem and the partial set covering problem. Forthcoming in *Geographical Analysis.*

Daskin, M. S. and Stern, E. (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment, *Transportation Science* **15**, 137-152.

Densham, P. J. and Rushton, G. (1992) A more efficient heuristic for solving large P-median problems. *Papers in Regional Science: The Journal of the RSAI* **71**, 307-329.

Drezner, Z. (1995) Dynamic facility location: The progressive P-median problem. *Location Science* **3**, 1-7.

Drezner, Z. and Wesolowsky, G. O. (1991) Facility location when demand is time dependent. *Naval Research Logistics* **38**, 763-777.

Duhamel, C., Potvin, J. Y. and Rousseau, J. M. (1997) A tabu search heuristic for the vehicle routing problem with backhauls and time windows. *Transportation Science* **31**, 49-59.

Eaton, D. and Daskin, M. S. (1980) A multiple model approach to planning emergency medical service vehicle deployment. *Proceedings of the International Conference on Systems Science in Health Care* ed. C. Tilquin, pp. 951-959, Pergamon Press, Montreal, Canada.

Eaton, D., Daskin, M. S., Simmons, D., Bulloch, B. and Jansma, G. (1985) Determining emergency medical service vehicle deployment in Austin, Texas. *Interfaces* **15**:1,96-108.

Eilon, S., Watson-Gandy, C. D. T. and Christofides, N. (1971) *Distribution Management: Mathematical Modelling and Practical Analysis.* Griffin, London.

Erkut, E. and Neuman, S. (1989) Analytical models for locating undesirable facilities. *European Journal of Operational Research* **40**, 275-291.

Erkut, E. and Neuman, S. (1992) A multiobjective model for locating undesirable facilities. *Annals of Operational Research* **40**, 209-227.

Erkut, E. and Verter, V. (1995) Hazardous materials logistics. *Facility Location: A Survey ofApplications and Methods* ed. Z. Drezner, pp. 467-506, Springer-Verlag, New York.

Erlenkotter, D. (1978) A dual-based procedure for uncapacitated facility location. *Operations Research* **26**, 992-1009.

Ernst, A. T. and Krishnamoorthy, M. (1996) Efficient algorithms for the uncapacitated single allocation *p*-hub median problem. *Location Science* **4**, 139-154.

Fisher, M. L. (1981) The Lagrangian relaxation method for solving integer programming problems. *Management Science* **27**, 1-18.

Fisher, M. L. (1985) An applications oriented guide to Lagrangian relaxation. *Interfaces* **15**:2, 10-21.

Fisher, M. L. (1995) Vehicle routing. *Handbooks in Operations Research and Management Science: Network Routing* eds M. O. Ball, T. L. Magnanti, C. L. Monma and G. L. Nemhauser, pp. 1-33. Elsevier Science Publishers, North Holland, Amsterdam.

Fisher, M. L, and Jaikumar, R. (1981) A generalized assignment heuristic for vehicle routing. *Networks,* **11**, 109-124.

Fitzsimmons, J. A. (1973) A methodology for emergency ambulance deployment. *Management Science* **19**, 627-636.

Flynn, J. and Ratick, S. (1988) A multiobjective hierarchical covering model for the essential air services program. *Transportation Science* **22**, 139-147.

Fourer, R., Gay, D. M. and Kernighan, B. W. (1993) *AMPL: A Modeling Language For Mathematical Programming.* Scientific Press, San Francisco, CA.

Galvão, R. D. and ReVelle, C. (1996) A Lagrangean heuristic for the maximal covering location problem. *European Journal of Operational Research* **88**, 114-123.

Garey, M. R. and Johnson, D. S. (1979) *Computers and Intractability: A Guide to the Theory ofNP-Completeness.* W. H. Freeman and Co., New York.

Gendreau, M., Hertz, A. and Laporte, G. (1994) A tabu search heuristic for the vehicle routing problem. *Management Science* **40**, 1277-1290.

Ghosh, A. and McLafferty, S. L. (1982) Locating stores in uncertain environments: A scenario planning approach. *Journal of Retailing* **58**:4, 5-22.

Gillet, B. E. and Miller, L. R. (1974) A heuristic algorithm for the vehicle dispatch problem. *Operations Research* **22**, 340-350.

Golden, B. L., and Assad, A. A. (eds.) (1988) *Vehicle Routing: Methods and Studies.* Studies in Management Science and Systems, North Holland, Amsterdam.

Gunawardane, G. (1982) Dynamic versions of set covering type public facility location problems. *European Journal ofOperational Research* **10,** 190-195.

Hakimi, S. (1965) Optimum location of switching centers in a communications network and some related graph theoretic problems. *Operations Research* **13**, 462-475.

Hakimi, S. (1964) Optimum location of switching centers and the absolute centers and medians of a graph. *Operations Research* **12,** 450-459.

Handler, G. Y. (1990) P-center problems. *Discrete Location Theory* eds. P. B. Mirchandani and R. L. Francis, pp. 305-347, John Wiley, New York.

Handler, G. Y. and Mirchandani, P. B. (1979) *Location on Networks.* M.I.T. Press, Cambridge, MA.

Hansen, P. and Mladenovic , N. (1997) Variable neighborhood search for the p-median. *Location Science* **5**, 207-226.

Hansen, P., Peeters, D. and Thisse, J. F. (1982) An algorithm for the constrained Weber problem. *Management Science* **28**, 1285-1295.

Hertz, J. H. (1964) *The Pentateuch and Haftorahs.* Soncino Press, London.

Hurter, A. P., Schaefer, M. K. and Wendell, R. E. (1975) Solutions of constrained location problems. *Management Science* **22**, 51-56.

Hsu, V. N., Daskin, M. S., Jones, P. C. and Lowe, T. J. (1995) Tool selection for optimal part production: A Lagrangian relaxation approach. *IIE Transactions* **27**, 417-426.

Jornsten, K. and Bjorndal, M. (1994) Dynamic location under uncertainty. *Studies in Regional and Urban Planning* **3**, 163-184.

Kendall, M. G. and Moran, P. A. P. (1963) *Geometrical Probability.* Griffin's Statistical Monographs and Courses, Charles Griffin and Company Ltd., London.

Kouvelis, P. and Yu, G. (1996) *Robust Discrete Optimization and Its Applications.* Kluwer Academic Publishers, Boston, MA.

Kuby, M. J. and Gray, R. G. (1993) The hub network design problem with stop feeders: The case of Federal Express. *Transportation Research* **27A**, 1-12.

Krarup, J. and Pruzan, P. M. (1990) Ingredients of location analysis. *Discrete Location Theory* eds. P.B. Mirchandani and R. L. Francis, pp. 1-54, Wiley, New York.

Laporte, G. (1988) Location-routing problems. *Vehicle Routing: Methods and Studies* eds. B. L. Golden and A. A. Assad, pp. 163-197, Elsevier Science Publishers, North Holland, Amsterdam.

Laporte, G. and Dejax, P. J. (1989) Dynamic location-routing problems. *Journal of the Operational Research Society* **40**:5, 471-482.

Laporte, G., Nobert, Y. and Tallefer, S. (1988) Solving a family of multi-depot vehicle routing and location-routing problems. *Transportation Science* **22,** 161-172.

Lawler, E. L., Lenstra, J. K., Rinnoy Kan, A. H. G. and Shmoys, D. B. (eds.) (1985) *The Traveling Salesman Problem: A Guided Tow of Combinatorial Optimization.* John Wiley and Sons, Inc., New York.

Larson, R. C. (1974) A hypercube queuing model for facility location and redistricting in urban emergency services, *Computers and Operations Research* **1**, 67-95.

Larson, R. C. (1975) Approximating the performance of urban emergency service systems, *Operations Research* **23**, 845-868.

Larson, R. C. and Sadiq, G. (1983) Facility location with the Manhattan metric in the presence of barriers to travel. *Operations Research* **31**, 652-669.

Lerman, S. (1976) Location, housing, auto ownership and mode to work: A joint choice model. *Transportation Research Record,* **610**, Washington, D.C.

Lin, S. and Kernighan, B. W. (1973) An effective heuristic algorithm for the traveling salesman problem. *Operations Research* **21**, 498-516.

List, G. F., Mirchandani, P. B., Turnquist, M. A. and Zografos, K. G. (1991) Modeling and analysis for hazardous materials transportation: Risk analysis, routing/scheduling and facility location. *Transportation Science* **25**, 100-114.

Louveaux, F. V. (1986) Discrete stochastic location models. *Annals of Operations Research* **6**, 23-34.

Manheim, M. L. (1979) *Fundamentals of Transportation Systems Analysis.* M.I.T. Press, Cambridge, MA.

Manne, A. S. (1961) Capacity expansion and probabilistic growth. *Econometrica* **29**, 632-649.

Maranzana, F. E. (1964) On the location of supply points to minimize transport costs. *Operational Research Quarterly* **15**, 261-270.

Marianov, V. and ReVelle, C. S. (1992) A probabilistic fire-protection siting model with joint vehicle reliability requirements. *Papers in Regional Science: The Journal of the RSAI* **71**, 217-241.

Marianov, V. and ReVelle, C. S. (1996) The queueing maximal availability location problem: A model for siting of emergency vehicles. *European Journal of Operational Research* **93**, 110-120.

Maze, T. H., Khasnabis, S., Kapur, K. C. and Kutsal, M. (1982) *A methodology for locating and sizing fixed facilities and the Detroit case study. Final report,* Urban Mass Transportation Administration, Grant Number MI-11-0004, Wayne State University, Detroit, MI.

Maze, T. H., Khasnabis, S., Kapur, K. and Poola, M. S. (1981) Proposed approach to determine the optimal number, size and location of bus garage additions. *Transportation Research Record* **781**, Washington, D.C.

Megiddo, N., Zemel, E. and Hakimi, S. L. (1983) The maximal coverage location problem. *SIAM Journal of Algebraic and Discrete Methods* **4**, 253-261.

Melkote, S. (1996) Integrated Models of Facility Location and Network Design. Ph.D. dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208.

Min, H. (1988) The dynamic expansion and relocation of capacitated public facilities: A multi-objective approach. *Computers and Operations Research* **15**, 243-252.

Mirchandani, P. B. (1980) Locational decisions on stochastic networks. *Geographical Analysis* **12**, 172-183.

Mirchandani, P. and Odoni, A. (1979) Location of medians on stochastic networks. *Transportation Science* **13**, 85-97.

O'Kelly, M. E. (1986a) Activity levels at hub facilities in interacting networks. *Geographical Analysis* **18**, p343-356.

O'Kelly, M. E. (1986b) The location of interacting hub facilities. *Transportation Science* **20**, 92-106.

O'Kelly, M. E. (1987) A quadratic integer program for the location of interacting hub facilities. *European Journal of Operational Research* **32**, 393-404.

O'Kelly, M. E. and Miller, H. M. (1994) The hub network design problems: A review and synthesis. *The Journal of Transport Geography* **2**, 31 -40.

Osleeb, J. and Ratick, S. (1990) A dynamic location-allocation model for evaluating the spatial impacts of just-in-time planning. *Geographical Analysis* **22**, 50-69.

Owen, S. H. (1998) Scenario Planning Approaches to Facility Location: Models and Solution Methods. Ph.D. dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208.

Owen, S. H. and Daskin, M. S. (1998) Strategic facility location: A review, *European Journal of Operational Research* **111**, 423-447.

Owen, S. H. and Daskin, M. S. (1998) Strategic facility location via evolutionary programming. Working paper, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 60208.

Peeters, D. and Thomas, 1. (1995) The effect of spatial structure on location-allocation results. *Transportation Science* **29**, 366-373.

Perl, J. and Daskin, M. S. (1985) A warehouse location-routing model. *Transportation Research,* **19B**, 381-396.

Plane, D. R. and Hendrick, T. E. (1977) Mathematical programming and the location offire companies for the Denver fire department. *Operations Research* **25**, 563-578.

Plaut, W. G. (ed.) (1981) *The Torah: A Modern Commentary.* Union of American Hebrew Congregations, New York.

Ratick, S. J., Du, W. and Moser, D. A. (1992) Development of a reliability based dynamic dredging decision model. *European Journal of Operational Research* **58**, 318-334.

Ratick, S., Osleeb, J., Kuby, M. and Lee, K. (1987) Interperiod Network Storage Location-Allocation (1NSLA) Models. *Spatial Analysis and Location Allocation Models* eds. G. Rushton and A. Ghosh, pp. 269-301. Van Nostrand Reinhold Company, New York.

Ratick, S, J. and White, A. L. (1988) A risk-sharing model for locating noxious facilities. *Environment and Planning* **B**, 15, 165-179.

ReVelle, C., Cohon, J. and Shobrys, D. (1991) Simultaneous siting and routing in the disposal of hazardous wastes. *Transportation Science* **25**, 138-145.

ReVelle, C. S. and Hogan, K. (1989a) The maximum availability location problem. *Transportation Science* **23**, 192-200.

ReVelle, C. S. and Hogan, K. (1989b) The maximum reliability location problem and **α-reliable** P-center problem: Derivatives of the probabilistic location set covering problem. *Annals of Operations Research* **18**, 155-174.

Roodman, G. M. and Schwarz, L. B. (1975) Optimal and heuristic facility phase-out strategies. *A11E Transactions* **7**, 177-184.

Schilling, D. A. (1980) Dynamic location modeling for public sector facilities: A multicriteria approach. *Decision Sciences* **11**, 714-724.

Schilling, D. A. (1982) Strategic facility planning: The analysis ofoptions. *Decision Sciences* **13**, 1-14.

Scott, A. J. (1971)Dynamic location-allocation systems: Some basic planning strategies. *Environment and Planning* **3**, 73-82.

Serra, D. and Marianov, V. (1997) The P-median problem in a changing network: The case of Barcelona. Submitted for publication to *Location Science.*

Serra, D., Rattick, S. and ReVelle, C. (1996) The maximum capture problem with uncertainty. *Environment and Planning B* **23**, 49-59.

Sheppard, E. S. (1974) A conceptual framework for dynamic location-allocation analysis. *Environment and Planning A* **6**, 547-564.

Skorin-Kapov, D. and Skorin-Kapov, J. (1994) On tabu search for the location of interacting hub facilities. *European Journal of Operational Research* **73**, 502-509.

Sweeney, D. J. and Tatham, R. L. (1976) An improved long-run model for multiple warehouse location. *Management Science* **22**, 748-758.

Srivastava, R. (1993) Alternate solution procedures for the location-routing problem. *OMEGA* **21**, 497-506.

Taillard, É., Badeau, P., Gendreau, M., Guertin, F. and Potvin, J. Y. (1997) A tabu search heuristic for the vehicle routing problem with soft time windows. *Transportation Science* **31**, 159-169.

Tapiero, C. S. (1971) Transportation-location-allocation problems over time. *Journal of Regional Science* **11**, 377-384.

Teitz, M. B. and Bart, P. (1968) Heuristic methods for estimating generalized vertex median of a weighted graph. *Operations Research* **16**, 955-961.

Toregas, C., Swain, R., ReVelle, C. and Bergman, L. (1971) The location of emergency service facilities. *Operations Research* **19**, 1363-1373.

van der Heijden, K. (1994) Probabilistic planning and scenario planning. *Subjective Probability,* eds. G. Wright and P. Ayton, pp. 549-572, Wiley, New York.

Van Roy, T. J. and Erlenkotter, D. (1982) A dual-based procedure for dynamic facility location. *Management Science* **28**, 1091 -1105.

Vanston, J. H., Frisbie, W. P., Lopreato, S. C. and Poston, D. L. (1977) Alternate scenario planning. *Technological Forecasting and Social Change* **10**, 159-180.

Watson, M. (1996) A Standardization Analysis Process Applied to Steel Coils in the Automotive Industry. Ph.D. dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, 1L 60208.

Weaver, J. and Church, R. (1983) Computational procedures for location problems on stochastic networks. *Transportation Science* **17**, 168-180.

Webb, M. H. J. (1968) Cost functions in the location of depot for multiple-delivery journeys. *Operational Research Society* **19**, 3 11-320.

Weber, A. (1929) *Uber den Standort der Industrien (Alfred Weber's Theory ofthe Location ofIndustries).* University of Chicago, Chicago, IL.

Wesolowsky, G. O. (1973) Dynamic facility location. *Management Science* **19**, 1241-1248.

Wesolowsky, G. O. and Truscott W. G. (1976) The multiperiod location-allocation problem with relocation of facilities. *Management Science* **22**, 57-65.

Xu, J. and Kelly, J. P. (1996) A network flow-based tabu search heuristic for the vehicle routing problem. *Transportation Science* **30**, 379-393.

This page intentionally left blank

11 NETWORK EQUILIBRIUM AND PRICING

Michael Florian and Donald Hearn

11.1 Introduction

Traffic equilibrium models are commonly in use for the prediction of traffic patterns on transportation networks that are subject to congestion phenomena. Even though their application in various transportation planning contexts has increased dramatically over the past twenty five years, due to the development of efficient solution algorithms and the increasing power of various computing platforms, they are based on concepts that have been stated more than seventy years ago. The idea of traffic equilibrium originated as early as 1924, when Knight gave a simple and an intuitive description of a postulate of traffic behavior under congested conditions, as follows:

> *"Suppose that between two points there are two highways, one of which is broad enough to accommodate without crowding all the traffic which may care to use it, but is poorly graded and surfaced, while the other is a much better road, but narrow and quite limited in capacity. If a large number of trucks operate between the two termini and are free to choose either of the two routes, they will tend to distribute themselves between the roads in such proportions that the cost per unit of transportation, or effective returns per unit of investment, will be the same for every truck on both routes. As more trucks use the narrower and better road, congestion develops, until a certain point it becomes equally profitable to use the broader but poorer highway."*

Some 28 years later, Wardrop (1952) stated two principles which formalize this notion of equilibrium and introduced the alternative behavior postulate of the minimization of the total travel costs. His first principle states that "the journey times on all routes actually used are equal and less than those which would be experienced by a simple vehicle of any unused route". Under certain assumptions, another interpretation of this principle is that the routes actually used are the shortest in time under prevailing traffic conditions and their perception by the travellers. Wardrop's first principle of equilibrium of route choice, which is identical to the notion postulated by Knight, become accepted over the past forty years as a sound behavioral principle to describe the spreading of trips over alternative routes. The traffic flows that satisfy this principle are usually referred to as "user optimal" flows, since each user chooses the route that is perceived to be the best. On the other hand, the "system optimal" flows are characterized by Wardrop's second principle which states that "the average journey time is minimum".

The first mathematical model of network equilibrium was formulated by Beckmann, McGuire and Winsten (1956). This seminal contribution was the starting point for other research and then application of such route choice models. The purpose of this chapter is to present the elements of the network equilibrium models used in transportation planning, review their mathematical properties, most commonly used solution methods and outline past and current applications.

11.2 Model Formulation - Deterministic Models

The network models that are most commonly used are steady state models, in spite of the fact that all traffic phenomena are temporal. One considers a given period of time for which the demand for travel is quantified and then one seeks to determine the flow pattern which results from the interaction of the demand and the performance of the transport infrastructure available.

A deterministic network equilibrium model of route choice may be formulated by using the following notation. The transportation network consists of nodes n, $n \in N$, which represent origins and destinations of traffic and intersections and arcs a, $a \in A$, which represent the road network. The number of vehicles on link a is v_a, $a \in A$ and the cost of travelling on a link is given by a user cost function $s_a(v)$, $a \in A$, where v is the vector of link flows over the entire network. These cost functions may model the time delay for travel on that arc, in which case it is referred to as a volume-delay function, however it may model other costs such as tolls or fuel consumption. The vector user cost function $s(v)$ is assumed to be monotone (strictly monotone)

$$\left(s(v^1) - s(v^2)\right) \left(v^1 - v^2\right) \geq (>)0, \; v^1 \neq v^2 \tag{11.1}$$

continuous and differentiable. The origin to destination demands g_i, $i \in I$, where I is the set of origin-destination (O-D) pairs, are distributed over directed paths k, $k \in K_i$, where K_i is the set of paths for O-D pair i and it is assumed that $K_i \neq \phi$. Also $K = \cup_{i \in I} K_i$. The flows on paths k, h_k, satisfy conservation of flow and nonnegativity constraints

$$\sum_{k \in K_i} h_k = g_i, \; i \in I \quad \text{and} \quad h_k \geq 0, \; k \in K. \tag{11.2}$$

The corresponding link flows v_a are given by

$$v_a = \sum_{k \in K} \delta_{ak} h_k, \; a \in A \tag{11.3}$$

where $\delta_{ak} = \begin{cases} 1 & \text{if link } a \text{ belongs to path } k \\ 0 & \text{otherwise} \end{cases}$

The link costs are additive in the sense that the cost of a path, $s_k(v)$, is the sum of the user costs on the links of the path

$$s_k(v) = \sum_{a \in A} \delta_{ak} s_a(v) , \quad k \in K. \tag{11.4}$$

If $u_i (= u_i(v))$, $i \in I$, are the costs of shortest paths for O-D pairs i

$$u_i = \min_{k \in K_i} s_k(v) , \quad i \in I \tag{11.5}$$

the demands for travel g_i, $i \in I$, are given by functions $G_i(u)$, where u is the vector of least cost travel times for all the O-D pairs of the network

$$g_i = G_i(u) \geq 0 , \quad i \in I. \tag{11.6}$$

The vector of demand functions $G(u)$ is assumed to be strictly monotone decreasing

$$\left(G(u^1) - G(u^2)\right)\left(u^1 - u^2\right) < 0 , \tag{11.7}$$

continuous and bounded from above.

A network equilibrium model that satisfies Wardrop's user optimal principle is formulated by stating that

$$s_k(v^*) - u_i^* \begin{cases} = 0 & \text{if } h_k^* > 0 \\ \geq 0 & \text{if } h_k^* = 0 \end{cases} \quad k \in K_i , \quad i \in I \tag{11.8}$$

over the feasible set (11.2), (11.3). It is relatively straightforward to show that (11.8) may be restated in the "complementarity" form

$$u_i^* \leq s_k(v^*) \text{ and } (s_k(v^*) - u_i^*)h_k^* = 0 , \quad k \in K_i, \quad i \in I \tag{11.9}$$

and that (11.8) and (11.5) are equivalent to

$$s_{k_1}^* \leq s_{k_2}^* , \quad if \ h_{k_1}^* > 0 , \quad k_1, k_2 \in K_i , \quad i \in I. \tag{11.10}$$

Another very useful restatement of Wardrop's first principle serves to convert the model to a variational inequality as done by Smith (1979) and Dafermos (1980). This is accomplished by noting that (11.8) is equivalent to

$$\{s_k(v^*) - u_i^*\}(h_k - h_k^*) \geq 0 , \quad k \in K_i, i \in I \tag{11.11}$$

above h_k, $k \in K$, in any feasible path flow. If $h_k^* > 0$ then $s_k(v^*) = u_i^*$, since h_k may be smaller than h_k^*. If $h_k^* = 0$, then (11.11) is satisfied when $s_k(v^*) - u_i^* \geq 0$. By summing (11.11) over $k \in K$ one obtains

$$\sum_{i \in I} \sum_{k \in K_i} s_k(v^*)(h_k - h_k^*) \geq \sum_{i \in I} u_i^*(g_i - g_i^*). \tag{11.12}$$

By using (11.3) and (11.4) a change of summation on the left-hand side yields

$$\sum_{a\in A} s_a(v^*)(v_a - v_a^*) \geq \sum_{i\in I} u_i^*(g_i - g_i^*). \tag{11.13}$$

Since the vector demand function $G(u)$ is strictly monotone decreasing, it is invertible. Let $w_i(g)$ denote the inverse of the demand function. By substituting for u_i one obtains

$$\sum_{a\in A} s_a(v^*)\,(v_a - v_a^*) - \sum_{i\in I} w_i(g^*)(g - g^*) \geq 0 \tag{11.14}$$

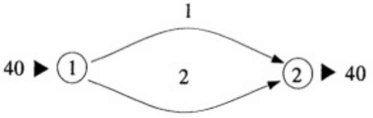
over the feasible set (11.2), (11.3), which may be rewritten in matrix notation as

$$s(v^*)(v - v^*) - w(g^*)(g - g^*) \geq 0. \tag{11.15}$$

It can be verified that (11.14) implies (11.8) by constructing a flow pattern which differs from the equilibrium flow on only one path k_1, $k_1 \in K_i$, for which $h_{k_1} = h_{k_1}^* + \delta\ , \quad 0 \leq |\delta| \leq h_{k_1}^*$.

The existence of a solution of the network equilibrium model is ensured by the continuity of the cost and demand functions and the fact that the feasible set is compact, if cycle flows do not occur and the demand functions are bounded from above (see Aashtiani and Magnanti (1981) and Dafermos (1980)).

The following example illustrates that a solution may not exist when the link cost functions are not continuous. The network consists of one O-D pair and two links as shown in Figure 11-1.



$$s_1(v_1) = 4\,v_1$$
$$s_2(v_2) = \begin{cases} 4\,v_2 & v_2 < 20 \\ 4\,v_2 + 10, & v_2 \geq 20 \end{cases}$$

Figure 11-1. Network with discontinuous cost functions

The demand from 1 to 2 is 40. When $v_1 = v_2 = 20$, the cost of link 2 is higher than the cost of link 1, but for $v_1 = 20 + \epsilon$ and $v_2 = 20 - \epsilon > 0$, the cost of link 1 is higher than the cost of link 2. In practical applications, the continuity requirement is usually satisfied.

It is important that a network model, which is used to predict traffic flows for different network and demand scenarios, yields unique link flows and origin to destination costs. If this were not so, the comparison of the different future situations would be difficult to carry out since differences between scenarios would depend on the nonuniqueness of the flows. Fortunately, for many applications, the network equilibrium models have unique flows and origin to destination demands. This is ensured when the link cost functions are strictly monotone and the demand functions (and their inverses) are strictly monotone decreasing, as assumed above.

To demonstrate this (see Smith (1979), Dafermos (1980), Aashtiani and Magnanti (1981)), suppose that there are two distinct solutions (v^1, g^1) and (v^2, g^2). By writing (11.15) once with $v = v^1$, $g = g^1$ and $v^* = v^2$, $g^* = g^2$ and once with $v^* = v^1$, $g^* = g^1$ and $v = v^2$, $g = g^2$, and adding the two inequalities one obtains

$$\left(s(v^1) - s(v^2)\right)\left(v^1 - v^2\right) - \left(w(g^1) - w(g^2)\right)\left(g^1 - g^2\right) \leq 0. \qquad (11.16)$$

By imposing (11.1) and (11.7) it follows that (11.16) is satisfied if and only if each term is equal to zero. Hence

$$\left(s(v^1) - s(v^2)\right)\left(v^1 - v^2\right) = 0\,, \qquad (11.17)$$

$$\text{and } \left(w(g^1) - w(g^2)\right)\left(g^1 - g^2\right) = 0\,, \qquad (11.18)$$

with the conclusion that $v^1 = v^2$ if the link cost functions are strictly monotone and $g^1 = g^2$ if the inverse demand functions are strictly monotone. Since u_i^*, $i \in I$ are the lengths of shortest paths for each O-D pair i based on the link costs $s_a(v^*)$, it follows that they are unique as well.

It is worthwhile to note that the path flows h_k^*, $k \in K$ are not unique, in general. Given the link flows v_a^*, $a \in A$, the corresponding path flows are given as the solution of the simultaneous linear equations

$$v_a^* = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} h_k^*\ , \quad a \in A. \qquad (11.19)$$

Since in most applications $|K| >> |A|$, the number of variables h_k^* far exceeds the number of constraints and the decomposition of link flows into path flows is not unique. The consequence of this property is that the analysis of the path flows h_k^*, $k \in K$ requires some care since they are not unique, but they contain nevertheless valuable information.

Another important property of the network equilibrium model is that it is stable. Roughly speaking, the equilibrium flows depend continuously upon the travel demands and link cost functions. Small changes in the travel demands result in small changes of the traffic flows. This was demonstrated by Dafermos and Nagurney (1984) for the model of this section and by Hall (1978) for the fixed demand variant of the network equilibrium model. This property is very desirable in applications, provided that the model is a suitable representation of the observed link flows.

Most of the applications of network equilibrium in practice have been achieved for simpler versions of the model (11.15) subject to (11.2), (11.3). They were facilitated by the fact that if the Jacobians $\nabla s(v)$, $\nabla w(g)$ of the cost functions $s(v)$ and inverse demand functions $w(g)$ are symmetric, that is

$$\frac{\partial s_a(v)}{\partial v_{\tilde{a}}} = \frac{\partial s_{\tilde{a}}(v)}{\partial v_a}, \quad \text{for all} \quad a,\, \tilde{a} \in A$$

and

$$\frac{\partial w_i(g)}{\partial g_{\tilde{i}}} = \frac{\partial w_{\tilde{i}}(g)}{\partial g_i}, \quad \text{for all} \quad i,\, \tilde{i} \in I.$$

Then, (11.15) is equivalent to a convex cost optimization problem since, by Green's Lemma, the vectors $s(v)$ and $w(g)$ can be viewed as gradients of the line integrals $\oint_0^v s(x)dx$ and $\oint_0^g w(y)dy$ respectively. The assumptions made on $s(v)$ and $w(g)$ imply that

$$Z(v,g) = \oint_0^v s(x)dx - \oint_0^g w(y)dy \tag{11.20}$$

is a convex function in (v,g) and the minimization of $Z(v,g)$ is equivalent to solving (11.15). If, furthermore, the link cost functions are separable, that is $s_a(v) = s_a(v_a)$, $a \in A$ and so are the inverse demand functions, $w_i(g) = w_i(g_i)$, $i \in I$, the strict monotonicity assumptions on $s(v)$ and $w(g)$ imply that $s_a(v_a)$ are strictly increasing and $g_i(u_i)$ are strictly decreasing and their Jacobians are diagonal matrices. The equivalent convex optimization problem becomes simply

$$\min\ Z(v,g) = \sum_{a \in A} \int_0^{v_a} s_a(x)dx - \sum_{i \in I} \int_0^{g_i} w_i(y)dy \tag{11.21}$$

subject to (11.2), (11.3).

In the case of fixed demand, the problem takes the classical form:

$$\min\ S(v) = \sum_{a \in A} \int_0^{v_a} s_a(x)dx \tag{11.22}$$

subject to (11.2), (11.3) with $g_i = \bar{g}_i$, $i \in I$, that is, constant demand.

11.3 Model Formulation - Stochastic Models

Stochastic network equilibrium models are based on the hypothesis that travelers make systematic errors in their perception of the travel costs. In the deterministic model it is taken for granted that the travelers have perfect knowledge of the link travel costs and, hence, of path costs. The probability density functions which are postulated to represent the systematic perception errors result in different models. Stochastic network equilibrium models are preferred for applications when the network is not subject to a high level of congestion and the choice of paths is not determined solely by the travel times or costs, but also by preference variations.

In order to formulate stochastic network equilibrium models it is necessary to introduce pr_k, the probability that an individual chosen from the population g_i will choose path k, $k \in K_i$. This is defined as

$$pr_k \triangleq pr_k(z_i)\,, \quad k \in K_i,\ i \in I \tag{11.23}$$

where $z_i = (z_1, ..., z_k, ...)$ is the vector of *perceived* travel times of all paths k for an O-D pair i. The perceived travel times on link a are assumed to be given by a probability density function

$$z_a \quad \sim \quad D(s_a,\, \theta s_a)$$

where s_a is the actual travel cost and θs_a is its variance, and θ is a constant. Thus the probability of choosing path k is given by

$$pr_k = Pr(z_k = \min_{k\prime \in K_i} \{z_{k\prime}\}),\ k \in K_i, \quad i \in I, \tag{11.24}$$

the probability that the path is perceived to be the shortest.

If $D(s_a, \theta s_a)$ is assumed to be the normal distribution, then the vector of perceived travel times (z_i) is a multivariate normally distributed. The weak law of large numbers implies that, on the average, the path flows h_k satisfy

$$pr_k = \frac{h_k}{g_i}, \quad k \in K_i,\ i \in I\,. \tag{11.25}$$

It is possible to show (see Sheffi and Powell (1982)) that the stochastic network equilibrium model is equivalent to solving the unconstrained optimization problem

$$\min_v \, -\sum_{i \in I} g_i E\,(\min_{k \in K_i}\{z_k | s^k(v)\}) + \sum_{a \in A} v_a s_a(v_a) - \sum_{a \in A} \int_0^{v_a} s_a(x)dx \tag{11.26}$$

subject only to nonnegativity constraints and that the minima of (11.26) coincide with solutions of the stochastic network equilibrium models. The objective (11.26) is not convex in general, but it can be shown that there is only one stationary point and, in the neighborhood of this point, the objective function is strictly convex in the flow variables. Hence the resulting link flows are unique.

An interesting special case of stochastic network equilibrium models occurs when the path probabilities are given by a logit function:

$$pr_k = \frac{\exp\left(-\theta s_k(v)\right)}{\sum\limits_{k' \in K_i} \exp\left(-\theta s_{k'}(v)\right)}\,, \quad k \in K_i,\ i \in I\,. \tag{11.27}$$

In this particular case, it is easy to show that the equivalent optimization problem is

$$\min_h \sum_{i \in I} \sum_{k \in K_i} h_k \ln h_k + \theta \sum_{a \in A} \int_0^{v_a} s_a(x)dx \tag{11.28}$$

subject to the usual constraints (11.2), (11.3).

When the link travel costs are constant, that is $s_a(v_a) = \bar{s}_a$, then the solutions of this model are dependent on the way that the network is represented since all paths are perceived to be independent, even if they share links (see the Chapter of Ben-Akiva and Bierlaire in this volume).

11.4 Solution Algorithms for Network Equilibrium Models

Deterministic symmetric models

As shown in Section 11.2, if the link cost functions and the demand functions are separable, a network equilibrium model which has a unique solution may be reformulated as an equivalent convex cost differentiable optimization problem. Since the feasible flows satisfy (11.2), (11.3), the conservation of flow and nonnegativity constraints and the only interaction between the arc flows for an origin or an origin-destination pair occurs in the objective function. This makes it possible to construct a wide range of algorithms for solving the problem, each based on a particular decomposition of the flows.

It is possible to classify the algorithms for the symmetric network equilibrium problem according to the way that the problem is decomposed, which may be by O-D pair i, by origin p or by using simplicial decomposition of the problem based on the extreme points of the feasible region (11.2), (11.3). The most commonly used algorithms are based on the linear approximation algorithm (Frank and Wolfe (1956)) and operate in the space of arc flows. The adaptation of this algorithm and some of its variants are described first. Then algorithmic approaches that were devised for the solutions in the space of path flows, which are referred to as path equilibration algorithms, are briefly discussed. The convergence properties of the algorithms are noted, but not proved in detail since they may be referenced in standard nonlinear programming texts indicated in the references.

One of the simplest convergent algorithms for minimizing a convex function subject to linear constraints is the linear approximation method. Bruynooghe, Gibert and Sakarovitch (1968) were the first to propose the method, however the later work of LeBlanc, Morlok and Pierskalla (1975) and Nguyen (1976) made this method popular in practice. Computer codes are widely available for solving both the fixed demand and the variable demand version of the network equilibrium models. We present next the details of the adaptation of this method to solve the fixed demand problem which is recalled to be

$$\min\ S(v) = \sum_{a \in A} \int_0^{v_a} s_a(x)dx$$

$$\text{subject to} \quad \sum_{k \in K_i} h_k = \bar{g}_i, \ i \in I \ , \ h_k \geq 0, \ k \in K$$

$$v_a = \sum_{k \in K} \delta_{ak} h_k, \ a \in A \ .$$

Given an initial feasible solution, a feasible direction of descent is generated by solving a subproblem which is obtained by a first order approximation of the objective function. The linearized approximation at an intermediate iteration l at a solution v^l

$$S(v^l) + \nabla S(v^l)\ (y - v^l). \tag{11.29}$$

By eliminating the constant terms $S(v^l)$ and $\Delta S(v^l)v^l$, the linearized subproblem simplifies to

$$\min \sum_{i \in I} \sum_{k \in K_i} \sum_{a \in A} s_a(v_a^l)\delta_{ak} y_k \tag{11.30}$$

$$\text{subject to} \quad \sum_{k \in K_i} y_k = \bar{g}_i \ , \ i \in I \tag{11.31}$$

$$y_k \geq 0 \ , \ k \in K \, . \tag{11.32}$$

By changing the order of summation in (11.30) and by using (11.4) the objective becomes

$$\min \sum_{i \in I} \sum_{k \in K_i} s_k(v^l) y_k \tag{11.33}$$

subject to (11.31), (11.32).

The solution of this problem is obtained by computing shortest paths for each O-D pair i and allocating the demand $\bar{g}_i$ to that path ("all-or-nothing" assignment).

This yields the arc flow vector

$$z_a^l = \sum_{k \in K} \delta_{ak}\, y_k^l \quad , \quad a \in A \tag{11.34}$$

and the direction of descent is ,

$$d_a^l = \left(z_a^l - v_a^l\right) \quad , \quad a \in A. \tag{11.35}$$

An iteration of the linear approximation algorithm is completed by finding the solution of

$$\min_{0 \leq \lambda \leq 1} \; S(v^l + \lambda d^l) \tag{11.36}$$

or, equivalently, by finding λ , $0 < \lambda < 1$ for which

$$\sum_{a \in A} s_a(v_a^l + \lambda d_a^l)\, d_a^l = 0 \tag{11.37}$$

unless the minimum of (11.36) is attained for $\lambda = 0$ or $\lambda = 1$.

The following algorithm results:

Linear approximation method

Step 0. Find v^1; $s^1 = s\left(v^l\right)$, $l = 1$.

Step 1. Perform an "all-or-nothing" assignment based on the current arc costs $s\left(v^l\right)$ and obtain y^l. $d^l = \left(y^l - v^l\right)$.

Step 2. Verify if a predetermined stopping criterion is satisfied. If it is, stop; otherwise continue to

Step 3. Find optimal step size λ^l by solving (11.37).

Step 4. Update arc flows $v^{l+1} = v^l + \lambda^l d^l$ and arc costs $s^{l+1} = s\left(v^{l+1}\right)$; set $l = l + 1$ and return to Step 1.

The algorithm generates paths at each iteration, but these are not kept. Hence the storage requirements are modest and do not increase with the number of iterations. Also, it is easy to obtain a lower bound on the value of the optimal objective function. Since $S(v)$ is a convex function and $\nabla S(v) = s(v)$, due to convexity

$$S(v^*) \geq S\left(v^{l'}\right) + s\left(v^{l'}\right)\left(y^{l'} - v^{l'}\right) \quad , \quad l' = 1, 2, \ldots l. \tag{11.38}$$

The right hand side of (11.38) provides a lower bound on $S(v^*)$ at each iteration. The best lower bound (BLB) at a current iteration l is

$$\mathrm{BLB} = \max_{l'=1,2,\ldots,l} S\left(v^{l'}\right) + s\left(v^{l'}\right)\left(y^{l'} - v^{l'}\right) . \tag{11.39}$$

As a consequence, a natural stopping criterion, denoted the relative gap (RGAP) is

$$\mathrm{RGAP} = \frac{S(v^l) - \mathrm{BLB}}{S(v^l)}.100. \tag{11.40}$$

Since $S(v^l) - \mathrm{BLB}$ is an estimate of the difference between an optimal solution and the current solution, the computations are terminated when $\mathrm{RGAP} \leq \epsilon_1 > 0$, where $\epsilon_1 > 0$ is a predetermined parameter.

Other stopping criteria that are used are a maximum number of iterations, $l^{\max}$, or thequantity $\{s(v^l)v^l - s(v^l)y^l\}$ which tends to zero as the optimum solution is approached. Hence, this stopping criterion is

$$\frac{\{s(v^l)v^l - s(v^l)y^l\}}{\sum_i \bar{g}_i} \leq \epsilon_2 \tag{11.41}$$

where $\epsilon_2 > 0$ is another predetermined parameter. The left-hand side of (11.41) has the physical interpretation of the difference between average trip costs on currently used paths and the average trip costs on current shortest paths. This quantity does not decrease monotonically with the number of iterations.

An intuitive interpretation of this algorithm is that the travelers adjust their route choice from congested routes to less congested routes until all routes are of about equal length. This explains its resemblance to many heuristic algorithms that had been suggested and used to solve this problem. On the other hand, the linear approximation algorithm exhibits slow convergence in the vicinity of the optimal solution due to the fact that its asymptotic rate of convergence is sublinear. This has motivated the development of variants of this algorithm which attempt to improve its rate of convergence. One of these variants is presented later in this section.

The variable demand network equilibrium model (11.21) may be solved by a *partial* linear approximation method, first suggested by Evans (1976), which employs the linearization of only some of the variables of the objective function. In (11.21), only the arc cost functions are linearized. The resulting subproblem at iteration l is

$$\min \sum_{i\in I} \sum_{k\in K_i} \sum_{a\in A} s_a(v_a^l)\delta_{ak} y_k - \sum_{i\in I} w_i(g_i^l) x_i \tag{11.42}$$

$$\text{subject to} \quad \sum_{k\in K_i} y_k - x_i = 0 \ , \quad i \in I \tag{11.43}$$

$$y_k \geq 0 \, , \ k \in K \, , \ x_i \geq 0 \, , \ i \in I \, . \tag{11.44}$$

This subproblem is solved by determining u_i^l, $i \in I$ to be the costs of the shortest paths based on the current link costs $s(v^l)$ and then simplifying (11.42) by using (11.43) and (11.44) to solve

$$\min \sum_{i \in I} \left(u_i^l - w_i(g_i^l)\right) x_i \tag{11.45}$$

$$\text{subject to} \qquad x_i \geq 0\,, \quad i \in I. \tag{11.46}$$

By applying the Karush-Kuhn-Tucker (see Luenberger (1965)) conditions to (11.45), (11.46), x_i^l are determined analytically as follows:

$$x_i^l = \begin{cases} G_i(u_i^l) & \text{if } G_i(u_i^l) \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{11.47}$$

The demands x_i^l are then assigned to the shortest paths in order to obtain y_a^l, $a \in A$ and the direction of descent is $d^l = \left\{ (y^l - v^l) \; ; \; (x^l - g^l) \right\}$.

Even though the solutions obtained with the linear approximation method are usually acceptable for the solution of the fixed or variable network equilibrium models for large scale problems, the slow convergence of the method in the neighborhood of the solution has motivated the development of several variants that improve its asymptotic rate of convergence. These include the adaptation of the PARTAN method (see LeBlanc, Helgason and Boyce (1985), Florian, Guélat and Spiess (1987) and Arezki and Van Vliet (1990)) and the restricted simplicial decomposition method which we describe in more detail below.

The motivation for exploring the PARTAN variant of the linear approximation method is that, for unconstrained minimization problems, the PARTAN algorithm is equivalent to the conjugate gradient algorithm. This algorithm alternates a regular iteration of the linear approximation algorithm with a direction generated by using every other solution, v^{l-1} and v^{l+1}. The solution at this alternate iteration is obtained by finding α^l, $\alpha^l \leq \alpha_{\max}^l$, which minimizes the objective function for the solution $v^{l-1} + \alpha(v^{l+1} - v^{l-1})$ where $\alpha_{\max}^l$ is the largest step size that maintains the nonnegativity of the path flow. It can be shown that, at a current iteration l, the largest step size that may be taken is given by the formula

$$\alpha_{\max}^l = \frac{1}{1 - \bar{\lambda}^l \bar{\lambda}^{l-1} \bar{\alpha}^{l-1}}$$

where $\bar{\lambda}^l = 1 - \lambda^l$ and $\bar{\alpha}^{l-1} = 1 - \alpha^{l-1}$. The algorithm may be stated as follows:

Linear approximation with PARTAN

Step 0. Find a feasible solution v^1 ; $s^1 = s(v^1)$; $y^0 = v^1$; $l = 1$.

Step 1. Find the linear approximation direction, $d^l = y^l - v^l$.

Step 2. If a predetermined stopping criterion is satisfied, stop; otherwise continue to

Step 3. Find optimal step size λ^l.

Step 4. Update arc flows $\tilde{v}^l = v^l + \lambda^l d^l$.

Step 5. If $l = 1$, then $v^{l+1} = \tilde{v}^l$; $s^{l+1} = s\left(v^{l+1}\right)$; $l = l + 1$ and return to step 1; otherwise, the PARTAN direction is $d_p^l = \left(\tilde{v}^l - v^{l-1}\right)$.

Step 6. Find the optimal PARTAN step size α^l as the solution of min $S(v^{l-1} + \alpha d_p^l)$, subject to $0 \leq \alpha \leq \alpha_{\max}^l$.

Step 7. Update arc flows $v^{l+1} = v^{l-1} + \alpha^l d_p^l$ and arc costs $s^{l+1} = s\left(v^{l+1}\right)$; $l = l + 1$ and return to Step 1.

The restricted simplicial decomposition algorithm (see Hearn, Lawphongpanich and Ventura (1985, 1987)) is an extension of the simplicial decomposition methods proposed by von Hohenbalken (1977) for solving nonlinear programs with pseudo-convex continuously differentiable objective functions and linear constraints. Its application to solving the fixed demand network equilibrium problem (11.22) subject to (11.2), (11.3) is as follows. Since the feasible region Θ is bounded there are a finite number of extreme points and every flow in Θ can be written as a convex combination of these extreme points. If Θ_y denotes a set of retained extreme points of Θ and Θ_v a set which is empty or contains the flows at a current iteration and q, $q \geq 1$, denotes an integer parameter which controls the number of extreme points at an iteration, the algorithm is as follows:

Restricted simplicial decomposition

Step 0. v^0 is a feasible solution; set $\Theta_y^0 = \phi$, $\Theta_v^0 = \left\{v^0\right\}$ and $l = 0$.

Step 1. Solve the subproblem

$$\min \sum_{a \in A} s_a\left(v_a^l\right) y_a$$

$$\text{subject to } \sum_{k \in K_i} y_k = \bar{g}_i, \quad i \in I \ , \quad y_k \geq 0, \quad k \in K$$

$$z_a = \sum_{k \in K} \delta_{ak} y_k$$

which is the same as Step 1 if the linear approximation method, which is solved by an "all-or-nothing" allocation of the demands $\bar{g}_i$ to shortest

paths for each O-D pair i.
Let the solution be z^l.

Step 2. If $s(v^l)\left(z^l - v^l\right) \geq 0$, stop; v^l is the optimal solution; otherwise, if $|\Theta_y^l| < q$ then $\Theta_y^{l+1} = \Theta_y \cup z^l$ and $\Theta_v^{l+1} = \Theta_v^l$. If $|\Theta_y^l| = q$ replace the element of Θ_y^l that has the minimal weight in the expression of v^l in the convex combination of elements of Θ^l with z^l to obtain Θ_y^{l+1} and let $\Theta_v^{l+1} = \left\{v^l\right\}$. Set $\Theta^{l+1} = \Theta_y^{l+1} \cup \Theta_v^{l+1}$.
Go to Step 3.

Step 3. (Master Problem). Let $v^{l+1} = \arg\min \left\{S(v)|v\right.$ belongs to the convex hull of $\left.\Theta^{l+1}\right\}$. $v^{l+1} = \sum_{j=1}^{m} \lambda_j z_j$, where $m = |\Theta^{l+1}|$ and $z_j \epsilon\, \Theta^{l+1}$.
Remove all elements z_j with $\lambda_j = 0$ from Θ_y^{l+1} and Θ_v^{l+1}. Set $l = l+1$ and return to Step 1.

The efficiency of the algorithm depends to a large extent on the solution of the master problem in Step 3. In order to achieve convergence, the master problem need not be solved exactly. It is only necessary to ensure that a sufficient decrease of the objective is achieved in successive iterations. This may achieved by approximating the objective of the master problem with a quadratic objective function. Under appropriate assumptions there exists an almost closed form solution of the quadratic approximation of *S*.

We turn our attention next to path equilibration algorithms for the symmetric fixed demand network equilibrium. In this approach the problem is decomposed by O-D pair i and a sequence of problems, for each O-D pair i is solved in the space of path flows. This general approach, which is equivalent to a Gauss-Seidel decomposition (or relaxation), is also known as "cyclic decomposition", since in a step of the algorithm a single O-D problem is solved, by keeping the flows of all other O-D pairs fixed. The algorithm terminates when there is no improvement in the solution for all O-D pairs i, which constitute a "cycle".

The subproblem solved for each O-D pair i is the fixed demand network equilibrium problem

$$\min \quad \sum_{a \in A} \int_{0}^{v_a^i + \bar{v}_p} s_a(x)dx \tag{11.48}$$

$$\text{subject to} \quad \sum_{k \in K_i} h_k = \bar{g}_i \ , \quad i \in I \tag{11.49}$$

$$h_k \geq 0\,, \quad k \in K_i \tag{11.50}$$

where $$\bar{v}_a = \sum_{i' \neq i} \sum_{k \in K_{i'}} \delta_{ak} h_k \tag{11.51}$$

and $$v_a^i = \sum_{k \in K_i} \delta_{ak} h_k\,. \tag{11.52}$$

The Gauss-Seidel solution strategy may be stated as follows:

Cyclic decomposition by O-D pair

Step 0. Given an initial solution, set $i = 0, \ i' = 0$.

Step 1. If $i' = |I|$, stop; otherwise set $i = i \mod |I| + 1$ and continue.

Step 2. If the current solution is optimal for the i_{th} subproblem (11.48)-(11.52), set $i' = i' + 1$ and return to Step 1; otherwise solve the i_{th} subproblem, update flows, set $i' = 0$ and return to Step 1.

The convergence of the Gauss-Seidel strategy is ensured since the objective function is convex and any local minimum is a global minimum as well.

Path equilibration algorithms used to solve (11.48)-(11.52) operate in the space of path flow and obtain a solution where all used paths are of equal cost. Since the number of paths grows exponentially with the network size $\left(|N|, \ |A|\right)$ path equilibration algorithms are usually implemented by using a restriction strategy, where the paths that carry flow are generated as required. Let $K_i^+ = \{k \in K_i | h_k > 0\}$ be the set of paths with positive flows. The simplest such algorithm, due to Dafermos (1971), finds the shortest path and longest path and transfers flow between these paths in order to equalize their cost. The algorithm may be stated as follows:

Path equilibration algorithm

Step 0. Find an initial solution v_a^i ; $s_a = s_a\left(v_a^i + \bar{v}_a\right)$; determine an initial K_i^+.

Step 1. Find k_1 such that $s_{k_1} = \min_{k \in K_i^+} s_k$ and
k_2 such that $s_{k_2} = \max_{k \in K_i^+} s_k$.

If $(s_{k_1} - s_{k_2}) \leq \epsilon$, go to Step 4; otherwise define the direction of descent $d_{k_1} = (h_{k_2} - h_{k_1})$ for path flow k_1 and $d_{k_2} = (h_{k_1} - h_{k_2})$ for path flow k_2.

Step 2. Find the Step size λ which

$$\min_{\lambda} \sum_{a \in A} \int_0^{v_a^i + \lambda y_a + \bar{v}_a} s_a(x)dx \tag{11.53}$$

$$0 \leq \lambda \leq \left(-\frac{h_{k_2}}{d_{k_2}}\right) \tag{11.54}$$

where

$$y_a = \delta_{ak_1} d_{k_1} + \delta_{ak_2} d_{k_2} \tag{11.55}$$

Step 3. $h_k = h_k + \lambda d_k$; for $k = k_1, k_2$; $v_a^i = v_a^i + \lambda y_a$; $s_a = s_a(v_a^i + \bar{v}_a)$.

Step 4. Compute the shortest path $\tilde{k}$ with cost $\tilde{s}_k = \min_{k \in K_1} s_{k_i}$.

If $\tilde{s}_k < \min_{k \in K_i^+} s_k$, then update k_i^+, $k_i^+ = k_i^+ \cup \tilde{k}$ and return to Step 1; otherwise stop.

This algorithm is just one of many path equilibration schemes possible. In order to generate a direction of descent for the subproblem (11.48)-(11.52) one may use the adaptation of the reduced gradient or the projected gradient algorithms. The algorithms used for the subproblem are well known nonlinear programming methods and hence are convergent.

Deterministic Asymmetric Models

For the simplicity of the exposition, only algorithms for the fixed demand network equilibrium problem are presented. That is, for finding v^* feasible which satisfies

$$s(v^*)\,(v - v^*) \geq 0, \quad \forall \text{ feasible } v. \tag{11.56}$$

A large class of algorithms for this problem, which are referenced to as relaxation methods, result when the cost function is modified at each iteration by fixing the interaction between blocks of variables and thus removing, at each iteration, the asymmetry of the cost functions. These algorithms include the nonlinear Jacobi method and the nonlinear Gauss-Seidel method. They are sometimes referred to as diagonalization methods, since the resulting Jacobians of the relaxed vector of cost functions are diagonal.

In order to describe a relaxation algorithm it is convenient to introduce a smooth function $\hat{s}(v,\tilde{v}) : \Theta \times \Theta \longrightarrow R^n$ with the property that $\hat{s}(v,v) = s(v)$ and $\nabla_v \hat{S}(v,\tilde{v})$ is positive definite and symmetric. Hence, if $v^{l+1} \cong v^l$, then v^{l+1} is a solution of the asymmetric network equilibrium model and is the unique solution of the variational inequality problem

$$\hat{s}\left(v^{l+1}, v^l\right)\left(v - v^{l+1}\right) \geq 0, \ \forall \text{ feasible } v, \tag{11.57}$$

which is obtained by solving the strictly convex differentiable optimization problem

$$v^{l+1} = \underset{v}{\operatorname{argmin}} \sum_{a \in A} \int_0^{v_a} \hat{s}_a(x)dx. \tag{11.58}$$

Different algorithms result from the choices made for the function $\hat{s}(v,\tilde{v})$. The nonlinear Jacobi method obtained for

$$\hat{s}_a\left(v, v^l\right) = s_a\left(v_1^l, \ldots, v_a, \ldots, v_{|A|}^l\right) \tag{11.59}$$

and the nonlinear Gauss-Seidel method results for

$$\hat{s}_a\left(v, v^l\right) = s_a\left(v_1^{l+1}, \ldots, v_{a-1}^{l+1}, v_a, \ v_{a+1}^l, \ldots, v_{|A|}^l\right) \tag{11.60}$$

An algorithm defined by (11.58) is globally convergent (see Dafermos (1983)) if

$$||\nabla_v s^{-\frac{1}{2}}\left(v^1, \ \tilde{v}^1\right) \ \nabla_{\tilde{v}} s\left(v^2, \tilde{v}^2\right) \ \nabla_v s^{-\frac{1}{2}}\left(v^3, \ \tilde{v}^3\right)||_2 < 1 \tag{11.61}$$

$$\text{for } v^i, \tilde{v}^i \text{ feasible}, \ i = 1, 2, 3.$$

This sufficient condition for the convergence of the relaxation methods is difficult to verify and rather restrictive. The intuitive interpretation of this condition is that the Jacobian of the vector of cost functions is weakly asymmetric. In summary, one way to state this class of relaxation algorithms is the following:

Relaxation algorithm

Step 0. Find a feasible solution v^1, $l = 1$.

Step 1. Determine v^{l+1} as the solution of (11.58).

Step 2 If $||v^{l+1} - v^l|| \leq \epsilon$, stop. Otherwise, continue to

Step 3. $l = l + 1$ and go to Step 1.

Among the algorithms that have been proposed for solving the asymmetric network equilibrium models one finds the simplicial decomposition method, gap descent methods, projection algorithms and a dual cutting plane method. These methods are not presented here but may be consulted by the interested reader at the indicated references.

Stochastic Symmetric Models

The solution algorithms for this problem employ a simulation in order to obtain a direction of descent for the objective function (11.26). In order to evaluate the objective function exactly, on exhaustive path enumeration for all O-D pairs of the network would be necessary. This is clearly prohibitive from a computational perspective. An algorithm that implements the basic step

$$v^{l+1} = v^l + \alpha_l d^l \tag{11.62}$$

where α_l, the step sizes at iteration l, satisfy

$$\sum_{l=1}^{\infty} \alpha_l = \infty \; ; \; \sum_{i=1}^{\infty} \alpha_i^2 < \infty \tag{11.63}$$

and d^l, the direction of descent, is determined by a Monte Carlo simulation (see Powell and Sheffi (1982)). This simulation is performed by sampling all links for a travel time realization, computing shortest paths and performing an "all-or-nothing" assignment on these paths; this procedure is repeated several times and then a flow vector $\hat{y}_a$ is obtained by averaging the link flows. The number of times that the procedure is repeated determines the variance of $\hat{y}_a$. The direction of descent d_a^l is $(\hat{y}_a - v_a)$.

It can be proved that when the step sizes satisfy (11.63), this algorithm converges to the unique solution of the stochastic network equilibrium model. Since $\alpha_l = \frac{1}{l}$ satisfies (11.63), this choice is often made, and the method if referred to as the "method of successive averages". This method lacks a natural stopping criterion and the descent in the values of the objective function is not monotone.

We turn our attention to the logit based stochastic network equilibrium model

$$\min_h \sum_{i \in I} \sum_{k \in K_i} h_k \ln h_k + \theta \sum_{a \in A} \int_0^{v_a} s_a(x)dx \tag{11.64}$$

subject to (11.2), (11.3).

This problem may be also solved by the method of successive averages, without requiring simulation in order to obtain a direction of descent. At each iteration l, the direction of descent is obtained by computing shortest paths based on current link costs and by performing an "all-or-nothing" assignment on these paths.

It can be shown theoretically and empirically, that the flow pattern that results from stochastic network equilibrium models tends towards the flow pattern obtained with the deterministic model as the network becomes congested. This may be recognized intuitively by inspecting the terms of the corresponding objective functions (11.63) and (11.64).

11.5 Combined Mode Models

The network equilibrium models presented so far in this Chapter do not distinguish different classes or different modes of traffic. In many applications the network equilibrium models are more complex and lead to more elaborate models that identify explicitly different modes, such as private car on public transit, or different classes of traffic which may correspond to different vehicle types or different socio-economic classes. Some examples of such models are presented next.

Suppose that the vehicles travelling on the network are subdivided into $|M|$ different types, $m \in M$ (see Dafermos (1972)). The link cost function on each link is different for each vehicle type and depends on the different types of vehicles that use the link, $s_a^m(v)$, $a \in A$, where v is the vector of flows (v_a^m) $a \in A$, $m \in M$. The demand for each vehicle type $\bar{g}_i^m$ is known. The corresponding deterministic network equilibrium model is given by the variational inequality

$$\sum_{m \in M} \sum_{a \in A} s_a^m(v^*)(v_a^m - v_a^{m^*}) \geq 0 \qquad (11.65)$$

$$\text{subject to} \quad \sum_{k \in K_i^m} h_k = \bar{g}_i^m \, , \quad i \in I, \; m \in M \qquad (11.66)$$

$$h_k \geq 0, \; k \in K_i^m, \; i \in I, \; m \in M \qquad (11.67)$$

$$v_a^m = \sum_{i \in I} \sum_{k \in K_i^m} \delta_{ak} h_k \, . \qquad (11.68)$$

Unless some simplifying assumptions are made regarding the link cost functions, the solution of this model requires an efficient algorithm for a very large scale variational inequality model.

One way to simplify this multi-class model is to induce symmetry and separability in the link cost functions. For instance, if one postulates that the user cost functions simplify to

$$s_a^m(v) = s_a \left(\sum_{m \in M} v_a^m \right) + t_a^m, \; a \in A, \; m \in M \qquad (11.69)$$

which implies that the travel time depends on the total number of vehicles on the link and that only a constant term for each link and class, t_a^m, differentiates the various classes of traffic. Then, with the appropriate manipulation, one obtains an equivalent convex cost minimization problem

$$\min \sum_{a\in A} \int_0^{v_a} s_a(v_a) + \sum_{a\in A} \sum_{m\in M} t_a^m v_a^m \tag{11.70}$$

$$\text{subject to} \quad \sum_{k\in K_i^m} h_k = g_i^m \ , \ i \in I, \ m \in M \tag{11.71}$$

$$h_k \geq 0 \, , k \in K \tag{11.72}$$

$$v_a^m = \sum_{k\in K_i^m} \delta_{ak} h_k \ , \ a \in A \, , \ m \in M \tag{11.73}$$

$$v_a = \sum_{m\in M} v_a^m \tag{11.74}$$

This model may be solved efficiently by an adaptation of the linear approximation method, and has been used extensively in applications. Other variants of (11.71) - (11.74) are possible as well.

Another example of a combined model is a two mode model of traffic (see Florian and Spiess (1983)) where one mode is the private car and the other mode is public transit. A mode choice function

$$G_i(u_i) = 1/\left(1 + \exp\left(\alpha + \beta(u_i^1 - u_i^2)\right)\right) \ , \ \ i \in I \tag{11.75}$$

gives the probability (or proportion) of trips that will use mode 1, which has a travel cost of u_i^1 and the competing mode 2 has a travel cost of u_i^2. A two mode network equilibrium model may be formulated by assuming that a “user optimal” route choice is made on both mode 1 and mode 2 (which may correspond to a transit mode).

$$s_k(v^*) - u_i^{m*} \begin{cases} = 0 & \text{if } h_k^* > 0 \\ \geq 0 & \text{if } h_k^* = 0 \end{cases} , \ k \in K_i^m \ , \ i \in I \ , \ m = 1,2 \tag{11.76}$$

subject to conservation of flow and nonnegativity constraints

$$\sum_{k\in K_i^m} h_k = g_i^m \ , \ \ i \in I, \ m = 1,2 \tag{11.77}$$

$$h_k \geq 0 \ , \ k \in K_i^m \ , \ i \in I, \ m = 1,2 \tag{11.78}$$

$$\text{and} \qquad u_i^m(v) = \min_{k\in K_i^m} s_k(v) \ , \ i \in I, \ m = 1,2 \tag{11.79}$$

The link cost functions $s_a^m(v)$, $m = 1, 2$; $a \in A$ are asymmetric and not separable, in general.

This model may be cast in the form of a variational inequality by carrying out the usual derivation by using (11.11). The resulting variational inequality is

$$\sum_{a \in A} s_a^1(v^*)\,(v_a^1 - v_a^{1^*}) + \sum_{a \in A} s_a^2(v_a^*)\,(v_a^2 - v_a^{2^*}) - \sum_{i \in I} w_i(g_i^{1^*})\,(g_i^1 - g_i^{1^*}) \geq 0 \quad (11.80)$$

where $w_i(g_i^{1^*})$ is the inverse of (11.75).

The Jacobi method may be used to obtain a solution to this model.

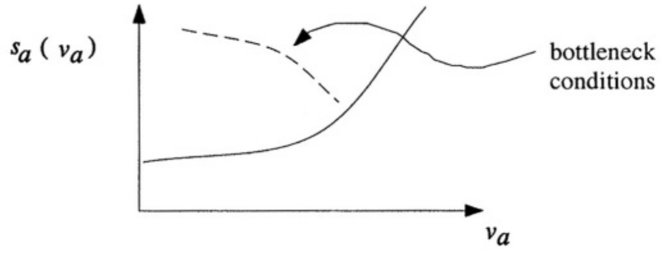
The two multi-class multi-mode models presented above are just two examples of the multitude of combined models which are formulated for particular transportation planning applications. Some of these models are so complex, that their solution is obtained by *ad hoc* equilibration procedures which are inspired by the method of successive averages.

11.6 Application and Validation of Network Equilibrium Models

The validation of network equilibrium models is reported in relatively few published empirical studies in spite of the fact that literally thousand of applications have been successfully carried out. The early studies of Florian and Nguyen (1976) on the urban network of the City of Winnipeg, Canada and Dow and Van Vliet (1979) on the urban network of the City of Leeds, England are examples of successful validation exercises. A practical problem which arises when applying the fixed demand network equilibrium models is the determination of the origin-destination matrix. Synthetic demand models or origin-destination surveys are used to determine the demand for travel with varying degrees of success. Even when survey data is available, it does not include information for all the trips taken during the peak hour and various adjustment methods are used to reconcile the differences between the flow predictions and the observed link counts. The process of calibrating a network equilibrium model involves the network representation, the calibration and allocation of user cost functions (known also as volume delay functions) to the links of the network as well as eventual adjustments of the origin-destination matrix. Often significant approximations are made in order to build the necessary data base for the application of the model. Yet, it may still be the best predictive tool available for evaluating the impact of network changes in the short and medium term. The link flows obtained in the validation studies mentioned above simulate the average hourly flows during the peak period quite well and the origin to destination travel times are satisfactorily reproduced as well.

The static nature of the network equilibrium model, which renders its solution to be efficient, is also one of its main drawbacks in the simulation of traffic flows.

The application of network equilibrium models is based implicitly on the assumption that the traffic is not subject to severe bottlenecks which may cause the traffic to back up and "spill back" to upstream links. When this assumption is not satisfied, monotone increasing user cost functions do not model properly the phenomenon of increased travel time and reduced flow on links which contain traffic bottlenecks as indicated in the graph below



The use of the network equilibrium model is not particularly demanding in computer expenditures. For networks of medium size (200 origins/destinations, 1 500 nodes and 4 000 links) the computation of the network equilibrium flows may require 10–30 sec. on a personal computer built with the Intel Pentium II, 450 Mhz processor. Large networks (3 000 origins/destinations, 15 000 nodes, 45 000 links) may require up to 30 minutes of computing on a similar personal computer. The ever increasing power of processors which are used to build personal computers and workstations will render the computation of equilibrium flows on even larger networks possible in elapsed times of the order of minutes.

11.7 Dynamic (or Time Dependent) Network Equilibrium Models

The time dependent (temporal or dynamic) formulations of network equilibrium models received little attention until the past years with the exception of the early contributions of Merchant and Nemhauser (1978a, 1978b). There are probably several reasons for the renewed interest in time dependent route choice models. The practical application of static network equilibrium models revealed their limitation in representing the temporal traffic flow behavior, as pointed out in the previous section, the renewed interest in Europe, Japan and North America in the design and implementation of Intelligent Transportation Systems (ITS) measures require network models which may evaluate such actions and hence must be dynamic; the increased power of widely available computing power raised the expectations of the complexity of network models which may be practicable.

Several conceptual frameworks and models were proposed for the modeling of time varying traffic and the time dependent version of network equilibrium models (see Cascetta and Cantarella (1991), Smith (1991), Friesz *et al* (1993). The formulation of temporal network equilibrium models require the introduction of the time dimension. In a continuous dynamic model this time dimension is infinite which makes the mathematical analysis more complex. A deterministic version of such a model is presented below.

The notation required expands all the variables used in the static version with the time index t. One considers a time period [0,*T*] for the consideration of trip departures and all the traffic is presumed to exit the network by time T^l. The time period under consideration is $\left[0,T^l\right]$.

The departure rates are given by demand functions $g_i(t)$, $i \in I$ which give rise to path flows $h_k(t)$, $k \in K$ for $t \in \left[0,T^l\right]$. The actual *experienced* travel time for a path k which carries flow generated at $t \in [0,T]$ is $s_k(t,h)$. It is assumed that the departure rates are continuous over time, which requires the use of the theory of Lebesgue integrals (measurable functions). $g(t)$ and $h(t)$ must be Lebesgue integrable (see Royden (1963)).

The feasible region of dynamic network equilibrium model is defined by the conservation of flow and nonnegativity constraints

$$\sum_{k \in K_i} h_k(t) = g_i(t),\ i \in I$$

$$h_k(t) \geq 0,\ k \in K\ \mathit{for\ almost\ all}\ t \in \left[0,T^l\right] \tag{11.81}$$

The term *for almost all* t means for all t up to a set of measure zero. It follows that two feasible path flow rates h' and h'' are equal to each other if they are equal up to a set of measure zero.

The generalization of Wardrop's user optimal principle to temporal flows may be stated as follows: *A feasible flow with path flow rates* h *and well defined actual path travel times* $s(h)$ *is a user optimal dynamic network equilibrium if, at any time* t, *the actual path travel time is the shortest up to a set ofmeasure zero for any positive path flow.* Therefore, the dynamic network equilibrium model is defined, up to a set of measure zero as:

$$s_h(t,h^*) - u_i^*(t)\begin{cases} = 0 & if\ h_k^*(t) > 0 \\ \geq 0 & if\ h_k^*(t) = 0 \end{cases},\ k \in K_i,\ i \in I\ \mathit{for\ almost\ all}\ t \tag{11.82}$$

$$\text{where} \quad u_i^*(t) = \min_{k \in K_i} s_k(t,h^*)\ \mathit{for\ almost\ all}\ \ t,\ i \in I \tag{11.83}$$

It is possible to show that the equilibrium conditions lead to a variational inequality which is

$$\int_0^T \sum_{i \in I} \sum_{k \in K_i} s_k(t,h^*)\left[h_k(t) - h_k(t^*)\right]dt \geq 0 \tag{11.84}$$

for almost all t.

While equation (11.84) is remarkably similar to the variational inequality formulation of the static network equilibrium model, there are many significant differences. Since the traffic flows are temporal it is reasonable to impose the First In-First Out (FIFO) condition on the flows. That is, a flow which starts later cannot arrive earlier by overtaking a flow which starts earlier. In order to satisfy this condition when the path costs are additive, that is they are the sum of the corresponding link costs, some rather stringent conditions must be satisfied by the link cost functions. The FIFO condition may be stated as

$$t' > t'' \quad \Rightarrow \quad t' + s_a(t') > t'' + s_a(t''), \;\; a \in A \tag{11.85}$$

It is equivalent to stating that, for a small time increment Δ_t, $\Delta_t > 0$

$$t + s_a(t) < t + \Delta_t + s_a(t + \Delta_t)$$

which, after dividing by Δ_t and taking the limit as $\Delta_t \rightarrow 0$ leads to

$$\frac{ds_a(t)}{dt} > -1 \tag{11.86}$$

The question then is which link cost functions satisfy the FIFO condition (11.86). Friesz *et al* (1993) showed that the condition is satisfied for linear function and a more recent study by Xu *et al* (1999) concludes that the FIFO condition will be satisfied for nonlinear arc cost functions which are not “too steep”.

Other than the difficulty posed by the FIFO condition, a solution algorithm for the dynamic network equilibrium model requires a time discretization and the adaptation of a general method for solving variational inequalities in order to compute a solution of (11.84). Conceptually, the solution algorithm may be viewed as consisting of two parts: the first is to determine $h_k(t)$ given $s_k(h,t)$; the second, referred to as *the network loading problem,* is to determine $s_k(h,t)$ given the path flow rates $h_k(t)$. A diagram which illustrates this view is given in Figure 11.2. The solution of the network loading problem is a rather complex undertaking since it requires the computation of the time varying link flows, travel costs and path travel costs by using the time varying path flow rates $h_k(t)$. A numerical approach is given in Wu *et al* (1997).

This formulation of the dynamic network equilibrium model does not impose capacity constraints on the link flows and occupancies. Hence, the queuing phenomena which lead to “spill back” when traffic bottleneck occur may not be properly represented.

In conclusion, the time dependent generalization of the network equilibrium model is complex mathematically and perhaps somewhat limited, due to the FIFO condition and the lack of capacity constraints. Current research efforts deviate from the pure analytical approach and resort to various simulation approaches to solve the equivalent of the network loading problem. As an alternative to analytical approaches to solve the network loading problem one can turn to simulation methods.

Traffic simulation methods may be characterized by the flow representation, the treatment of space and time and the model used for traffic dynamics. *Time-based models* advance the clock at fixed intervals while *event-based models* maintain an ordered list of events to process in continuous time. *Microscopic simulation models* consider individual vehicles in continuous space and by using very small clock intervals (which results in practically continuous time). Such models are based on the behaviour of individual drivers in the following of the downstream (leader) vehicle in its lane and lane changing behavior. *Mesoscopic* traffic simulation models may consider individual vehicles or packets (groups of vehicles which move together) which are moved in continuous space and either discrete or continuous time. *Macroscopic* traffic simulation models employ a fluid representation of the flow in discretized time and space and are based on approximate solutions of the partial differential equation which describes the propagation of traffic which respects the fundamental diagram of traffic flow (see Leutzbach (1988)). Examples of microscopic simulation models are AIMSUN2 (Barcelo et al (1996)), MITSIM (Yang and Koutsopoulos (1996)). DYNASMART (Jayakrishnan et al (1994)) and DYNAMIT (Ben-Akiva et al (1994)) are examples of mesoscopic simulation models, while METANET (Messmer et al (1995)) and FREEFLO (Payne (1979) are examples of macroscopic simulation models.

A mesoscopic approach to the simulation of traffic is probably the best choice for the solution of the network loading problem. It is less detailed than a microscopic simulation, which requires significant resources for a proper calibration, and more detailed than a macroscopic simulation, which does not offer a simple way to represent traffic by lane and node signal controls. The theoretical foundations of mesoscopic models may draw on traffic flow theory (such as DYNASMART, DYNAMIT) or on queuing theory, where the traffic flow is represented as a network of queues (such as DTASQ (Florian, Mahut and Tremblay (2001))). A dynamic traffic assignment model is constructed by combining a flow allocation algorithm to the used paths with a traffic simulation model for network loading. The theoretical properties of such hybrid models are not well understand but the empirical results are encouraging.

At the present time hybrid optimization simulation models may be the most promising avenue for the solution of medium size dynamic traffic assignment models. It is reasonable to expect the solution of networks of up to 2 000 nodes and 5 000 links in elapsed times of the order of one hour on a personal computer based on Intel Pentium IV processors running at 1 Ghz.

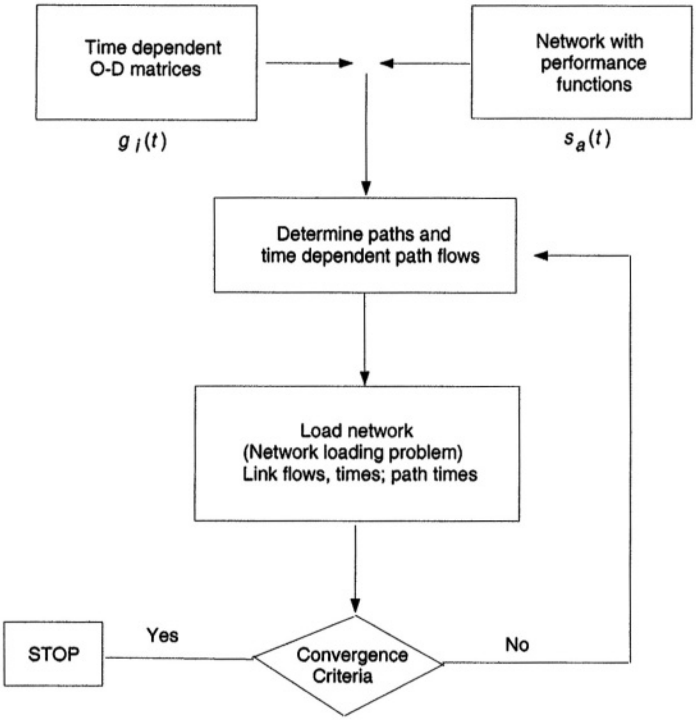

Time dependent O-D matrices
$g_i(t)$
Network with performance functions
$s_a(t)$
Determine paths and time dependent path flows
Load network (Network loading problem) Link flows, times; path times
STOP
Yes
Convergence Criteria
No

Figure 11-2. Dynamic traffic assignment and the network loading problem

11.8 Congestion Toll Pricing in Network Equilibrium Models

The Fixed Demand Case. Wardrop's second principle is usually referred to as the "system optimal" principle, since these are the flows that minimize the total cost of travel. The fixed demand version of this problem is

$$\min \sum_{a\in A} v_a s_a(v_a) \tag{11.87}$$

subject to (11.2), (11.3).

By applying the Karush-Kuhn-Tucker conditions to the system optimal problem one obtains that all the used paths have equal marginal costs, $m_k(v)$, that is

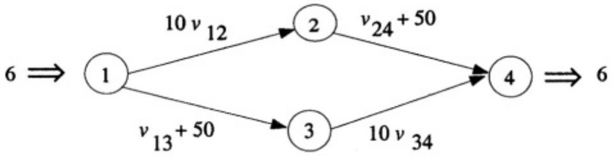
$$m_k(v) = \sum_{a\in A} \delta_{ak} \left\{ s_a(v_a) + v_a s_a'(v_a) \right\} \tag{11.88}$$

$$\text{and let } \tilde{u}_i^* = \min_{k\in K_i} m_k(v^*)\,, \; i \in I \tag{11.89}$$

$$\text{then} \quad m_k(v^*) \begin{cases} = \tilde{u}_i^* & \text{if } h_k > 0 \\ \geq \tilde{u}_i^* & \text{if } h_k = 0 \end{cases}, \quad k \in K. \tag{11.90}$$

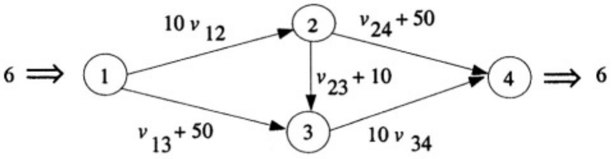
In general, the user optimal flows are different from the system optimal flows for a given network. For large, uncongested networks, the differences may be small, since for near constant travel times there is no difference between the user and system optimal flows.

Braess (1968) provided a classical example of such differences, by using the network shown below, where the user cost functions are indicated on the links



Braess Network

The user optimal flows are three units on the path $1-2-4$ and three units on the path 1–3–4. The total cost is 6 x 83 = 498. If the link (2,3) is now added with the user cost function $v_{23} + 10$,



Braess Network with additional link

the user optimal equilibrium flows are two units on each of the paths of the network with total cost of 6 x 92 = 552. Hence, the addition of an arc increases the total travel cost and the cost to each user. This apparent paradox is understood by computing the system optimal solution which is three units of flow each along the paths 1–2–4 and 1–3–4, even when the arc (2,3) is present. There is no reason to anticipate that the total travel costs will be reduced when a link is added, since these costs are not minimized by the flows which are “user optimal”.

The occurrence of Braess’ paradox is significant in network design and improvement decisions. If the user optimal principle is suitable as the behavioral representation of the route choice, then, given a particular network, it may be advantageous to restrict the use of some of its links and care must be exercised to properly evaluate link additions in order to detect possible occurrences of increase in total costs and increase of service levels for some of the users. Fisk and Pallottino (1981) report the occurrence of the Braess’ paradox in networks used in practice.

Since the system optimal flows lead to a more efficient use of a transportation network, in the sense that the total cost of the trips taken is minimized, the issue of charging tolls that would render user optimal flows to coincide with the system optimal flows has been debated for years (see Arnott and Small, 1994). From a modeling viewpoint, if link tolls equal to $v_a s_a^{'}(v_a)$, $a \in A$, would be charged on each link, then a user optimal route choice results in system optimal flows. The tolls $v_a s_a^{'}(v_a)$, $a \in A$, are referred to as marginal social cost pricing (MSCP) tolls. The notion of marginal cost pricing may be found as early as 1844 with the work of Dupuit.

A difficulty with marginal social cost pricing (MSCP) tolls is that they can be expensive, both to implement and to the users of the network. By the formula, a toll would be imposed for every link with traffic flow. In the Braess example, for instance, the MSCP tolls form the vector

$$\beta_{MSCP}^T = [\beta_{(1,2)}, \beta_{(2,4)}, \beta_{(1,3)}, \beta_{(3,4)}, \beta_{(2,3)}] = [30, 3, 3, 30, 0].$$

Thus there are tolls on four of the five arcs in the network and a collection mechanism (toll booth) would be required for each. Further, the total tolls collected would be 198, which is 40% of the total system travel cost of 498. Note however, that a sufficiently high toll on link (2,3) would prevent its presence in a user optimal flow distribution and no one would "pay" that toll.

The MSCP toll policy and the Braess example raise the question of what other toll vectors β would ensure that the tolled user optimal solution is a solution of the untolled system optimal problem. Recently this question has been addressed by Bergendorff, Hearn and Ramana (1997) and Hearn and Ramana (1998) who have shown that the set of all such tolls is the β part of the following linear system in variables (β, ρ):

$$s(\bar{v}) + \beta \geq A^T \rho^i \quad \forall i \in I \tag{11.91}$$

$$(s(\bar{v}) + \beta)^T \bar{v} = \sum_{i \in I} g_i \, E_i^T \, \rho^i, \tag{11.92}$$

where A is the node-arc incidence matrix of the network, E_i is a column incidence vector for (O-D) pair i with +1 in position O and -1 in position D, and $\overline{v}$ is the vector of system optimal flows. This result requires uniqueness of solutions in both the system and the user problems. Note that $\beta \geq 0$ would be appended to these equations to ensure nonnegative tolls. Otherwise, negative tolls, or subsidies, are possible.

Given this result, a procedure can be devised to obtain tolls which meet some secondary criteria as well as ensuring system optimal flows. The first step is to solve the system problem and obtain $\bar{v}$. Since the problem has the same mathematical structure as the user problem, i.e., it is a convex multicommodity flow problem, it can be solved with the optimization algorithms described earlier in this chapter.

The second step is to define another optimization problem which has constraints defined by (11.91) and (11.92) plus any additional conditions desired. Examples would be requiring that tolls be nonnegative or requiring certain $\beta_a = 0$ to prevent tolls where they are impractical, such as on arcs that represent neighborhood streets. The objective function is a matter of choice for the traffic engineer. Hearn and Ramana (1998) have suggested several:

MINSYS Minimize the total system tolling charges when the tolls are nonnegative, so that users pay as little as possible.

MINMAX Minimize the maximum toll on any individual arc. This can be extended to minimizing the maximum toll relative to the uncongested arc travel time.

MINREV Allow negative as well as positive tolls, thus charging users on some arcs and subsidizing the travel on others, and constrain the total tolls collected to be θ, where $\theta \in [-s(\overline{v})^T \overline{v}, \infty)$. When the net collected is $\theta = 0$, these are known as ROBINHOOD (RH) tolls.

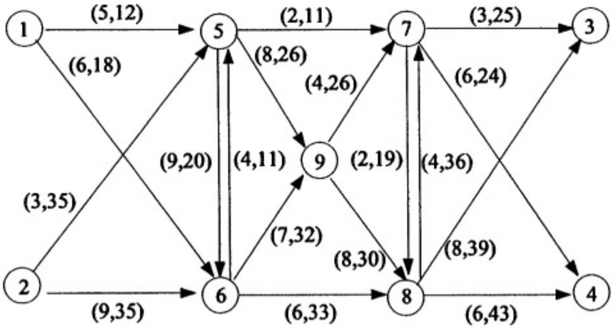


Figure 11-3. The nine node network

MINTB Minimize the number of toll booths required, with tolls nonnegative.

MINTB/RH Combine MINTB with RH tolls: minimize the number of toll booths while requiring the total tolls and subsidies collected to be zero.

To provide a comparison of the formulations listed above as well as comparison with MSCP tolls, Hearn and Ramana (1998) have employed a nine node example which has data similar to large-scale traffic assignment problems. The network is shown in the Figure 11–3. It has 18 links and all of the links have cost functions with the same structure:

$$s_a(v) = s_a(v_a) = T_a(1 + 0.15(v_a/b_a)^4)$$

where T_a and b_a are constants. In the figure, the tuple near link a is (T_a, b_a). There are four OD–pairs: (1–3), (1–4), (2–3) and (2–4), and the demands are 10, 20, 30, 40, respectively.

The various tolls for this example are given in Table 11-1, and the observations below are adapted from the reference.

A comparison between MINSYS and MSCP shows that the tolling pattern and toll amounts are quite different. Further, the total toll system cost (total system cost + total toll cost) in the MSCP case is equal to 3747 (2254 + 1493) and in the MINSYS case equal to 3142 (2254 + 888). So with the MSCP principle the users of the nine-node network pay 68% more in tolls than with the MINSYS pricing principle. The MINSYS solution also happened to coincide with the MINTB solution, so it also gives the minimum number of toll booths, 5 versus the 14 of MSCP.

The maximum toll on any link is 16.88 for MSCP and 11.2 for MINSYS (link (5,7) in both cases). When this maximum is minimized (MINMAX) the largest toll reduces to 8.00. MINMAX also provides another set of nonnegative tolls which significantly reduce the total tolls when compared with MSCP; the total toll is 28% higher in the latter case.

ROBINHOOD and MINTB/RH introduce the idea of negative tolls, but it is clear that this concept requires further examination. For example, it can be shown that $\beta = -s(\bar{v})$, provides a negative toll for every link, implying that the users are totally reimbursed for their time on the network. It is difficult to imagine that such a policy would ever be put in place. However, the selective use of negative tolls luring users to certain links might have some appeal.

The Elastic Demand Case. In the elastic demand system problem, the objective is to maximize net user benefit, the difference between total user benefit and the system cost. From basic economic principles, the total network user benefit from travel is $\sum_{i \in I} \int_0^{g_i} w_i(z)dz$ and, as before, the system cost is defined by $s(v)^T v$. Denoting the solution of this optimization problem as $(\bar{v}, \bar{g})$, and assuming that both it and the solution of the variational problem (11.14) are unique, Hearn and Yildirim (2002) have shown that the toll set for the elastic demand problem is the β part of the following linear inequality system in β and ρ^i variables:

$$(s(\bar{v}) + \beta) \geq A^T \rho^i \quad \forall i \in I \tag{11.93}$$

$$w_i(\bar{g}_i) \leq E_i^T \rho^i \quad \forall i \in I \tag{11.94}$$

$$(s(\bar{v}) + \beta)^T \bar{v} = w(\bar{g})^T \bar{g}. \tag{11.95}$$

Table 11–1. The nine node problem - alternative tolls

Link	System Optimum			Alternative Tolls β for Nine Node Problem				
	v_a	$s_a(v_a)$	$v_a s_a(v_a)$	MSCP	MINSYS & MINTB	MINMAX	RH	MINTB/RH
(1,5)	9.411	5.284	49.728	1.135				
(1,6)	20.589	7.541	155.262	6.162				
(2,5)	38.334	3.648	139.842	2.590	4.000	8.000	4.000	
(2,6)	31.666	9.905	313.652	3.618		4.000		-4.000
(5,6)	.000	9.000						
(5,7)	21.303	6.220	132.505	16.880	11.200	8.000	2.877	8.000
(5,9)	26.442	9.284	245.487	5.135				
(6,5)	0.000	4.000						
(6,8)	39.474	7.843	309.595	7.370	7.200	7.200	-0.816	
(6,9)	12.781	7.027	89.812	0.107				
(7,3)	29.608	3.885	115.027	3.541	4.000	7.200	2.618	
(7,4)	20.757	6.504	135.004	2.014		3.200		-2.126
(7,8)	0.000	2.000				1.079		
(8,3)	10.392	8.006	83.198	0.024			-1.689	
(8,4)	39.243	6.624	259.946	2.497			-0.307	1.874
(8,7)	0.000	4.000						0.121
(9,7)	29.062	4.937	143.479	3.746	3.200		-5.123	
(9,8)	10.162	8.016	81.459	0.063			-8.016	-7.200
Total Time =			2253.918					
Total Tolls = $\beta^T v^S$				1493.458	887.574	1167.572	0.000	0.000
Total Tolls/Total Time (%)				66.38	39.38	51.80	0.00	0.00
Toll Booths				14	5	7	8	6

where A and E_i are defined as before.

One important difference with the fixed demand case is that *all tolls in the elastic demand toll set have the same total toll revenue,* since, as the last equation shows:

$$\beta^T \bar{v} \;=\; w(\bar{g})^T \bar{g} \;-\; s(\bar{v})^T \bar{v}.$$

In Hearn and Yildirim (2002) the framework for determining alternative tolls has been defined similar to the fixed demand case: first, solve the system problem for the flows and demands, $(\bar{v}, \bar{g})$. Then using these values, define and optimize a secondary objective with respect to the toll set (11.93)-(11.95). Even though total toll revenue is constant, it could be appropriate to solve for the minimum number of toll booths (MINTB) or to minimize the maximum toll (MINMAX), and even to look for alternative schemes that include subsidies as well as tolls (MINREV).

As an illustration, demand functions for the nine node network (Figure 11–3) are defined as follows:

$$\begin{aligned}
g_{(1-3)} &= g_{(1-3)}\big(w_{(1-3)}\big) = 10 - 0.5w_{(1-3)} \\
g_{(1-4)} &= g_{(1-4)}\big(w_{(1-4)}\big) = 20 - 0.5w_{(1-4)} \\
g_{(2-3)} &= g_{(2-3)}\big(w_{(2-3)}\big) = 30 - 0.5w_{(2-3)} \\
g_{(2-4)} &= g_{(2-4)}\big(w_{(2-4)}\big) = 40 - 0.5w_{(2-4)}
\end{aligned}$$

Table 11–2 demonstrates the differences in demands and costs for the system (SOPT–ED) and user solutions (UOPT-ED) of the elastic demand nine node problem. For example, the total demand in the system problem is 57.411, but it is 60.753 in the user problem, a 5.5% difference.

Table 11–2. Demands and costs for the nine node elastic demand (ED) problem

OD-PAIR	SOPT-ED SOLUTION	UOPT-ED SOLUTION
(1-3)	(0.000, 20.000)	(0.151, 19.698)
(1-4)	(9.696, 20.607)	(10.698, 18.605)
(2-3)	(19.476, 21.047)	(20.672, 18.656)
(2-4)	(28.239, 23.523)	(29.232, 21.537)

Table 11–3 provides optimal flows for the system and user problems. Although the UOPT-ED solution has higher user benefit than the SOPT-ED solution, it also has a higher system cost due to more traffic on the network. Thus the net user benefit in the system problem is greater. Examining the individual link flow values, notice that they are within 10% of each other on almost all links of the network but not on links (5,7), (5,9) and (9,7). The *total* flow between nodes 5 and 7 for both user and system problems is within 10%, but the link flows differ substantially. Relative to the system solution, in the user problem (5,7) is over utilized while (5,9) and (9,7) are under utilized. Therefore, systems efficiency is increased by diverting traffic on (5,7) to the route 5–9–7. This can be done by making (5,7) less attractive, i.e., increasing the cost by tolling (5,7).

Table 11–4 contains alternative toll vectors for the nine node elastic demand problem and, for comparison, the MSCP tolls are also listed. As expected, tolls on the link (5,7) are high for the tolling schemes with positive tolls, namely, MSCP, MINMAX and MINTB. However, MINREV, which allows negative tolls, rewards travel on the 5–9–7 route with subsidies in order to achieve the SOPT-ED solution. For this example the toll revenue is 268.519 for all tolling schemes and this is 17.44% of net user benefit. MSCP and MINREV tolls require 10 toll booths. MINMAX has an objective value of 8.00 with eight toll booths. It happens that MINTB obtains the same result (i.e., the maximum toll is 8.00) with only five toll booths. Thus it could be argued that the MINTB solution is best in that it achieves the SOPT-ED solution and is cheapest to implement.

To illustrate how tolling affects route costs consider the routes 2–6–8–4 and 2–5–7–4. The delay functions on the first route give a total cost of 21.505, while on the second this total is only 13.504. However, the tolls on the first route are 2.018 and they are 10.018 on the second. Thus the total route cost is 23.523, in agreement with the total cost in Table 11–2. There is no incentive for additional (2–4) trip demand. Any increase would result in lower user benefit and higher costs since the demand and cost functions are strictly monotone.

Table 11–3. The nine node problem - SOPT-ED and UOPT-ED solutions

	SOPT-ED SOLUTION			UOPT-ED SOLUTION		
Link	$\bar{v}_a$	$s_a(\bar{v}_a)$	$\bar{v}_a s_a(\bar{v}_a)$	v_a^*	$s_a(v_a^*)$	$v_a^* s_a(v_a^*)$
(1,5)		5.000			5.000	20.000
(1,6)	9.696	6.076	58.914	10.849	6.119	66.380
(2,5)	31.715	3.303	104.769	34.458	3.423	117.940
(2,6)	15.999	9.059	144.938	15.446	9.051	139.805
(5,6)		9.000			9.000	
(5,7)	17.978	4.140	74.433	26.442	12.016	317.730
(5,9)	13.738	8.094	111.188	8.016	8.011	64.215
(6,5)		4.000			4.000	
(6,8)	25.696	6.331	162.677	26.295	6.363	167.308
(6,9)		7.000			7.000	
(7,3)	19.476	3.166	61.657	20.823	3.217	66.979
(7,4)	12.239	6.061	74.180	13.785	6.098	84.063
(7,8)		2.000			2.000	
(8,3)		8.000			8.000	
(8,4)	25.696	6.115	157.125	26.144	6.123	160.078
(8,7)		4.000		0.151	4.000	0.604
(9,7)	13.738	4.047	55.594	8.016	4.005	32.108
(9,8)		8.000			8.000	
User Benefit			2544.75			2613.50
System Cost			1005.474			1217.21
Net User Benefit			1539.284			1396.285

Table 11–4. The nine node problem (ED) - Alternative tolls

Link	MSCP	MINREV	MINMAX	MINTB
(1,5)			8.000	
(1,6)	0.303	2.085	2.085	0.067
(2,5)	1.214	-7.444		2.018
(2,6)	0.236	2.018	2.018	
(5,6)		6.218		
(5,7)	8.561		8.000	8.000
(5,9)	0.374	-3.953		
(6,5)				
(6,8)	1.323			
(6,9)				
(7,3)	0.663	17.882	2.438	0.420
(7,4)	0.243	17.462	2.018	
(7,8)		15.408		
(8,3)			2.320	
(8,4)	0.459			2.018
(8,7)			0.716	
(9,7)	0.187	-4.047		
(9,8)		9.408		
$\beta^T v^*$	268.519	268.519	268.519	268.519
$\beta^T v^*$/(**NUB**) (%)	17.44	17.44	17.44	17.44
Toll Booths	10	10	8	5

Second Best Toll Pricing. The examples above illustrate recent results in congestion pricing theory showing that traffic planners seeking optimal use of an urban traffic network can also meet secondary objectives such as minimizing the cost of installing tolling stations. It offers an alternative to the traditional theory that congestion tolls should be based on marginal social cost pricing. Any such tolls that obtain a system optimal solution are known as *first best* tolls. By contrast, there is the notion of *second best* toll pricing, where achieving the system optimal flows is not possible because tolling is disallowed on certain links of the network. Current research includes the analysis of small traffic assignment such as the work of McDonald (1995) and the derivation of toll formulas for larger problems as in Verhoef (1999).

Stated mathematically, the second best toll model is a mathematical program with equilibrium constraints, or an (MPEC). Letting V denote the set of all flows and demands that satisfy (11.2)-(11.3) and Y be the set of links that cannot be tolled, the elastic demand second best tolling problem MPEC is

$$\max \quad NUB(v,g,\beta) = \sum_{i\in I} \int_0^{g_i} w_i(z)dz - s(v)^T v$$

$$\text{s.t.}$$

$$\begin{aligned} &(v,g) \in V && \\ &(s(v)+\beta)^T(u-v) - w(g)^T(d-g) \geq 0 && \forall (u,d) \in V \\ &\beta_a = 0 && \forall a \in Y. \end{aligned}$$

where, as before, the objective is to maximize the net user benefit (NUB) function, which here depends not only on v and g, but also on the tolls β.

It is recognized that MPECs can be very difficult to solve numerically. However, a heuristic variation of the toll pricing procedures described above for obtaining first best tolls, has been developed by Yildirim (2001). It can be summarized as follows: First determine whether adding constraints $\beta_a = 0,\ a \in Y$ to (11.93)-(11.95) renders the toll set empty. If not, then first best tolls can be found. If there are no first best tolls, then the problem above can be solved for a *local stationary point* by replacing the equilibrium constraints with the equivalent Karush-Kuhn-Tucker conditions and then employing a standard nonlinear programming code to obtain a *local stationary point,* $\left(\tilde{v},\tilde{g},\tilde{\beta}\right)$. The quality of this solution can be determined by recognizing that

$$NUB\left(\tilde{v},\tilde{g},\tilde{\beta}\right) \in [NUB(v^*,g^*), NUB(\bar{v},\bar{g})]$$

where (v^*,g^*) and $(\bar{v},\bar{g})$ are, respectively, the solutions of the untolled UOPT-ED and SOPT-ED problems, respectively. In some cases, restricting the tolls to only a part of the network may still provide a close to optimal $NUB(\bar{v},\bar{g})$. To illustrate, consider again the elastic demand nine node example above, with user and system solutions given in Table 11–3, and assume that tolls are only allowed on the following arcs: (2,5), (5,7) and (8,7). Note that none of the first best tolls given in Table 11–4 are feasible for this problem, since for example, all have a toll on link (1,6).

A second best solution for this problem is given in Tables 11–5 and 11–6. For this special example, the objective value $NUB\left(\tilde{v},\tilde{g},\tilde{\beta}\right)=1535.13$ is within 0.3% of the SOPT-ED value given in Table 11–3. Further, the total toll revenue, $\tilde{\beta}^T\tilde{v}=147.17$ which is approximately one-half the first best toll revenue of 268.519, which is given in Table 11–3. Thus, the second best tolling solution has the interesting property that it provides not only tolls on the allowed links, but also provides a nearly optimal net user benefit at a significantly reduced toll cost to the users. As theoretical and empirical research in toll pricing continues, a goal will be determining whether this situation occurs often in practice.

Table 11–5 Demands and costs for the nine node elastic demand (ED) - Second best problem

OD-PAIR	SBTP Solution
(1-3)	(0.000, 20.000)
(1-4)	(10.647, 18.706)
(2-3)	(20.591, 18.817)
(2-4)	(29.165, 21.670)

Table 11–6. The nine node elastic demand (ED) - Second best solution

Link	$\tilde{v}_a$	$s_a(\tilde{v}_a)$	$s_a(\tilde{v}_a)\tilde{v}_a$	Tollable	$\tilde{\beta}$
(1,5)		5.000			
(1,6)	10.647	6.110	65.056		
(2,5)	32.777	3.346	109.679	1	0.080
(2,6)	16.978	9.074	154.075		
(5,6)		9.000			
(5,7)	18.069	4.184	75.608	1	8.000
(5,9)	14.708	8.122	119.474		
(6,5)		4.000			
(6,8)	27.625	6.442	177.965		
(6,9)		7.000			
(7,3)	20.591	3.207	66.038		
(7,4)	12.186	6.059	73.847		
(7,8)		2.000			
(8,3)		8.000			
(8,4)	27.625	6.153	169.990		
(8,7)		4.000		1	1.37
(9,7)	14.708	4.061	59.737		
(9,8)		8.000			

11.9 Conclusion

The study of network equilibrium models, and related solution algorithms may be considered to have reached a mature stage. A variety of models may be formulated and solved efficiently on contemporary computing platforms. Applications of network equilibrium models are abundant and relatively common in the practice of transportation planning. However, some of the basic premises of the formulation of these models such as the additivity of link costs to form the cost of a path and the static analysis of "average flows" during a selected time period, open the way to the study of more complex models. The recent interest in temporal on dynamic network equilibrium models is already attracting the attention of many researchers and will probably result in interesting new developments.

11.10 References

We refer the interested reader to two references which contain more complete presentations of the topic of this Chapter. These are

Florian, M. and Hearn, D. (1995) Network Equilibrium Models and Algorithms. Chapter 6 in *Handbooks in OR & MS, Vol.8,* M.O. Ball *et al.,* Eds., 485–550.

Patriksson, P. (1983) *The Traffic Assignment Problem: Models and Methods,* VNU Science Press, 223 pp.

The cited references in the text are the following:

Aashtiani, H,Z. and Magnanti, T.L. (1981) Equilibria on a congested transportation network. *SIAM Journal on Algebraic and Discrete Methods,* **2**, 213–226 .

Arezki, Y. and Van Vliet, D. (1990) A full analytical implementation of the PARTAN/Frank-Wolfe algorithm for equilibrium assignment. *Transportation Science,* **24**, 58–62.

Arnott, R. and Small, K. (1994) The economics of traffic congestion. *American Scientist,* **82**, 446–455.

Barcelo, J., Ferrer, J.L. and Grau, R. (1996) AIMSUN2 and the GETRAM simulation environment. Technical Report, Departamento di Estadistica i Investigaciou Operativa, Universitat Politecnica de Catalunya.

Beckmann, M.J., McGuire, C.B. and Winsten, C.B. (1956) *Studies in the economics of transportation,* Yale University Press, New Haven, CT.

Ben-Akiva, M. and Bierlaire, M. (1999) Discrete choice methods and their applications to short term travel decisions. Handbook of Transportation Science, Chapter 2. Kluwer Academic Publishers, Norwell, Massachusetts.

Ben-Akiva, M., Koutsopoulos, H.N. and Mukundan, A. (1994) A dynamic traffic model system for ATMS-ATIS operations. *IVHS Journal,* **2**, 9–24.

Bergendorff, P., Hearn, D.W. and Ramana, M.V. (1997) Congestion toll pricing of traffic networks. *Nework Optimization,* (Edited by W.W. Hager, D.W. Hearn, and P.M. Pardalos) Springer–Verlag series *Lecture Notes in Economics and Mathematical Systems,* 52–71.

Braess, D. (1968) Über ein Paradox der Verkehrsplannung. *Unternehmenstorchung,* **12**, 258–268.

Bruynooghe, M., Gibert, A. and Sakarovitch, M. (1969) Une médthode d'affectation du trafic in *Proceedings of the 4th International Symposium on the Theory of Road Traffic Flow,* Karlsruhe,(1968), W. Leutzbach and P. Baron, eds., Beiträge zur Theorie des Verkehrsflusses Strassenbau und Strassenverkehrstechnik, Heft 86, Herausgegeben von Bundesminister für Verkehr, Abteilung Strassenbau, Bonn, 198–204.

Cascetta, E. and Cantarella, G.E. (1991) A day-to-day and within-day dynamic stochastic assignment model. *Transportation Research,* **25A**, 277–291.

Dafermos, S. (1980) Traffic equilibrium and variational inequalities. *Transportation Science,* **14**, 42–54.

Dafermos, S. (1972) The traffic assignment problem for multiclass-user transportation networks. *Transportation Science,* **6**, 73–87.

Dafermos, S. and Nagurney, A. (1984) Sensitivity analysis for the asymmetric network equilibrium problem. *Mathematical Programming,* **28**, 174–184.

Dow, P. and Van Vliet, D. (1979) Capacity restrained road assignment. *Traffic Engineering and Control,* 261–273.

Dupuit, Y. (1844) De la mesure de l'utilité des travaux publics. Annales des travaux publics, Annales des ponts et chaussées, **8**, 332–375 .

Evans, S.P. (1976) Derivation and analysis of some models for combining trip distribution and assignment. *Transportation Research,* **10**, 37–57.

Fisk, C. and Pallottino, S. (1981) Empirical evidence for equilibrium paradoxes with implications for optimal planning strategies. *Transportation Research A,* **15A**, 3, 245–248.

Florian, M., Guélat, Y. and Spiess, H. (1987) An efficient implementation of the PARTAN variant of the linear approximation for the network equilibrium problem. *Networks,* **17**, 319–339.

Florian, M., Mahut, M. and Tremblay, N. (2001) A hybrid optimization mesoscopic simulation dynamic traffic assignment model. 2007 *IEES Intelligent Transportation Systems Proceedings,* 118–121.

Florian, M. and Nguyen, S. (1976) An application and validation of equilibrium trip assignment methods. *Transportation Science,* **10**, 379–389.

Florian, M. and Spiess, H. (1983) On binary mode choice/assignment models. *Transportation Science,* **17**, 32–47.

Frank, M. and Wolfe, P. (1956) An algorithm for quadratic programming. *Naval Res. Log. Quart.,* **3**, 95–110.

Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, M.L. and Wie, B.W. (1993) A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research,* **4**, 179–191.

Hall, M. (1978) Properties of the equilibrium state in transportation networks. *Transportation Science,* **12**, 208–216.

Hearn, D.W., Lawphongpanich, S. and Ventura, Y.A. (1985) Finiteness in restricted simplicial decomposition. *Operations Research Letters,* **4**, 125–130.

Hearn, D.W., Lawphongpanich, S. and Ventura, Y.A. (1987) Restricted simplicial decomposition: computation and extensions. *Mathematical Programming Study,* **31**, 99–118.

Hearn, D.W. and Ramana, M.V. (1998) Solving congestion toll pricing models. *Equilibrium and Advanced Transportation Modeling,* (Edited by P. Marcotte and S. Nguyen), Kluwer Academic Publishers, 109–124.

Hearn, D.W. and Yildirim, M.B. (2002) A toll pricing framework for traffic assignment problems with elastic demand. To appear in the edited volume *Current Trends in Transportation and Network Analysis – papers in honour of Michael Florian,* Kluwer Academic Publishers, 10pp.

Jayakrishnan, R., Mahmassany, H.S. and Hu, T.Y. (1994) An evaluation tool for advanced traffic information and management systems in urban networks. *Transportation Research* C, **3**, 129–147.

Knight, F.H. (1981) Some fallacies in the interpretation of social costs. *Quarterly Journal of Economics,* **38**, 306–312.

LeBlanc, L.J., Helgason, R.V. and Boyce, D.E. (1985) Improved efficiency of the Frank-Wolfe algorithm for convex network programs. *Transportation Science,* **19**, 445-462.

LeBlanc, L.J., Morlok E.K. and Pierskalla, W.P. (1975) An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research,* **5**, 309–318.

Leutzbach, W. (1988) Introduction to the theory of traffic flow. Springer-Verlag, Berlin Heidelberg.

Luenberger, D.G. (1965) Introduction to linear and nonlinear programming. Addison-Wesley.

McDonald, J.F. (1995) Urban highway congestion: an analysis of second-best tolls. *Transportation,* **22***,* 353–369.

Merchant, D.K. and Nemhauser, G.L. (1978a) A model and algorithm for the dynamic traffic assignment problem. *Transportation Science,* **12**, 183–199.

Merchant, D.K. and Nemhauser, G.L. (1978b) Optimality conditions for a dynamic traffic assignment model. *Transportation Science,* **12***,* 200–207.

Messmer, A. and Papageorgiou, M. (1995) METANET: A macroscopic simulation program for motorway networks. *Traffic Engineering and Control,* **31**, 466–470.

Nguyen, S. (1976) A unified approach to equilibrium methods for traffic assignment. In *Traffic Equilibrium Methods,* Proceedings 1974, M. Florian, Ed., vol. **118**. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, 148–182 .

Payne, H.J. (1979) FREEFLO: A macroscopic simulation models of freeway traffic. *Transportation Research Record,* **772**, 68–75.

Powell, W.B. and Sheffi, Y. (1982) The convergence of equilibrium algorithms with predetermined step sizes. *Transportation Science,* **16**, 45–55.

Royden, H.L. (1963) Real analysis. Collier MacMillan Ltd.

Sheffi, Y. and Powell, W.B. (1982) An algorithm for the equilibrium assignment problem with random link times. *Networks,* **12**, 191–207.

Smith, M.J. (1993) A new dynamic traffic model and the existence and calculation of dynamic user equilibria on congested capacity constrained road network. *Transportation Research B,* **27B**, 49–63.

Smith, M.J. (1979) Existence, uniqueness and stability of traffic equilibria. *Transportation Research B,* **13B**, 295–304.

Verhoef, E.T. (1999) Second-best congestion pricing in general static transportation networks with elastic demands. *Research Report,* Free University Amsterdam.

von Hohenbalken, B. (1977) Simplicial decomposition in nonlinear programming algorithms. *Mathematical Programming,* **13,** 49–68.

Wardrop, J.G. (1952) Some theoretical aspects of road traffic research. Proc. Inst. Civil Engineers, Part II, 325–378.

Wu, J.H., Chen, Y., and Florian, M. (1997) The continuous dynamic network loading problem: a mathematical formulation and solution method. *Transportation Research,* **32B**, 173–187.

Xu, Y., Wu, J.H., Florian, M., Marcotte, P. and Zhu, D.L. (1999) Advances in the continuous dynamic network loading problem. To appear in *Transportation Science.*

Yang, Q. and Koutsopoulos, H.N. (1996) A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research* C, **4**, 113–129.

Yildirim, M.B. (2001) Congestion toll pricing models for variable demand traffic assignment problems. Ph.D. dissertation, University of Florida.

This page intentionally left blank

12 STREET ROUTING AND SCHEDULING PROBLEMS

Lawrence Bodin, Vittorio Maniezzo and Aristide Mingozzi

12.1 Introduction

The chapter by Teodor Crainic in this book discusses the long haul truck routing problem. Some of the key differences between the long haul truck routing problem and the vehicle routing problems considered in this chapter are as follows:

1. In the long haul truck routing problem, the routes can extend over several days whereas, in the vehicle routing problem described herein, the routes are of one-day duration.

2. In the long haul truck routing problem, the locations can be scattered over a wide region, even the entire United States, whereas in the vehicle routing problems considered herein, the locations requiring service are packed in a small region.

Because of these differences (along with other considerations), the algorithms and data requirements for the vehicle routing problem considered in this paper can be considerably different from the algorithms and requirements of the long haul trucking problem.

Some of the major assumptions of most vehicle routing problems (VRP) discussed in the literature are the following:

a. Routes can be no longer than a specified duration T (such as T= 8 hours).

b. Each route represents a single day of work.

c. Overtime may be permitted at a predefined cost.

d. There is a single depot where all vehicles begin and end their route.

e. The fleet of vehicles is homogeneous; that is to say, all vehicles are identical.

f. Vehicle capacity such as weight or volume is always satisfied.

g. Time windows exist. A time window at a location to be serviced is defined as [L, U] where L is the earliest possible time to begin the service at the location and U is the latest possible time to begin the service at the location. It is normally assumed that the time window is hard; that is to say, the service of the location must begin between L and U. In a route, if the vehicle arrives at the location before L, the hard time window forces the vehicle to wait until L to begin service at the location. The time that the vehicle must wait before beginning the service at L is called the wait time of the vehicle at the location. Vehicle wait time represents nonproductive time and can be considered a cost to the organization. With hard time windows, it is assumed that it is infeasible to begin to service a location after U.

Although the above assumptions occur in most VRP, these assumptions can be modified or extended when solving actual vehicle routing problems as follows:

a. Start and End Times. Each vehicle has its own earliest start time and latest end time. In this case, each vehicle can have a different duration. As above, each route represents a single day of work.

b. Multiple Depots. In cases where there are multiple depots and each depot is autonomous with its own fleet of vehicles and geographical area to serve, then the overall VRP can be decomposed in a number of single-depot VRP. In cases where the geographic area that each depot services is not well defined, there may be an interaction between the depots so that it is impossible to consider any depot in isolation. In some problems, moreover, it may be possible that a vehicle route visits more than one depot in a day. For example, the vehicle can originate at one depot and service some locations. The vehicle then goes to a second depot to reload (perhaps with a product not available at the first depot) and visit another set of locations. The vehicle then goes to another depot, etc. The route concludes with the vehicle returning to the original depot at the end of the day.

c. Heterogeneous Fleet. In a heterogeneous fleet, each vehicle can have different characteristics such as capacity, length of workday, etc. Generally, the vehicle fleet is broken down into vehicle classes where each vehicle in a class is homogeneous. With a heterogeneous fleet, moreover, there may be some restrictions on the vehicle types that can service the stops. These restrictions are called vehicle/stop dependencies. Vehicle/stop dependencies occur in such applications as the scheduling of sanitation vehicles for either residential or containerized pickup. A vehicle/stop dependency can occur at a stop because of access restrictions, height of a bridge covering the road, width of a street segment or the location of the containers needing service. For example, a side-loading sanitation vehicle cannot service a container positioned at the end of a narrow alley even if the side-loading vehicle can traverse the alley. Variants of this type of dependency include vehicle-commodity incompatibilities (some vehicle types may not be able to carry certain types of products) and commodity-commodity incompatibilities (certain pairs of commodities cannot coexist on any vehicle because of possible contamination).

d. Multiple Trips. A trip corresponds to a set of stops which when serviced exhaust the capacity of the vehicle. This situation occurs frequently in sanitation routing and the routing of vehicles that service the collection boxes by a postal service. A sanitation vehicle doing either residential pickup or container pickup may make 2-3 trips to a disposal facility during its route while a postal truck collecting mail at the collection boxes may make 2-3 trips to the sort facility during its route.

e. Break Times and Lunch Times.

f. Soft Time Windows. If the time window [L, U] at a location is soft, the time to begin the service of the location can begin before L or after U. The algorithms designed to form routes in the presence of soft time windows generally associate a penalty with a location that is being assigned to a route if the service of this location causes a violation of the time window of any location on the route. In this way, the assignment of a location to a vehicle that violates a time window becomes less desirable.

g. Mixed Deliveries and Collections. In many situations the same vehicles that deliver goods are also used to collect goods. A typical example is when trucks that deliver goods from a warehouse to a set of stops must then collect raw materials from one or more warehouses and deliver the raw materials to one or more plants. A special case of mixed deliveries and collections is called backhauling. One very popular version of backhauling is when the vehicles perform all the deliveries in the route first and then perform all of the collections.

h. Combined Pickup and Deliveries. In a combined pickup and delivery problem, the service of a stop requires the transportation of goods from a specified pickup location to a specified destination location. Thus, every entity to be serviced has both a pickup location and a destination location specified. An example of a combined pickup and delivery problem is a courier service. Combined pickup and delivery problems can be broken down into full truckload pickup and delivery problems and partial truckload pickup and delivery problems. In the full truckload pickup and delivery problem, a vehicle moves a single dedicated load from the pickup location to the delivery location. In the partial truckload pickup and delivery problem, the vehicle can make several pickups before a delivery and several deliveries preceding the pickup at another location.

i. Multiple Commodities. If more than one commodity is to be delivered by the same vehicle, then the vehicle might have to be broken down into separate vehicle compartments. Examples of these problems include the delivery of gasoline to service stations, delivery of refrigerated and non-refrigerated items to supermarkets, etc.

Two general classes of vehicle routing problems (VRP) are point-to-point routing problems (PTPRP) and neighborhood routing problems (NRP). In PTPRP (also called node routing problems), distinct locations with known demand are to be serviced by a fleet of vehicles with known capacity. Examples of PTPRP include the routing of vehicles that deliver goods from a warehouse to a specified set of locations, the routing of sanitation vehicles for containerized pickup (the large bins at shopping centers, industrial parks, hospitals and schools), the routing and scheduling of a utility's field service operations, and the routing of vehicles for the delivery of newspapers. Most papers in the literature are concerned with variants of PTPRP.

In a NRP, the requirements for service are the street segments in an area of interest (AOI). In a NRP, not all street segments in the AOI require service, A street segment is said to require service if there exists at least one location on the street segment that requires service. Applications of NRP include the scheduling of household refuse collection vehicles, meter readers, initial telephone book delivery and local postal delivery operations.

Street routing and scheduling problems (SRP) can be defined as 1) all PTPRP where the locations to be serviced are assigned to the street segments or intersections of a digital street network database and 2) all NRP. SRP occur primarily in local delivery operations. Examples of SRP include the routing of residential and containerized sanitation vehicles, the scheduling of vehicles for telephone book and newspaper deliveries, the scheduling of meter readers, the routing of field service operations for public utilities and the routing of vehicles for other traditional local pickup and delivery operations. The features and characteristics of SRP and how these problems differ from more traditional VRP are the focus of this paper.

In Section 12.2, special considerations that make SRP different from traditional VRP and the impact of these considerations on the algorithms for solving SRP are presented. In Section 12.3, the goals in developing good solutions to SRP are discussed. In Section 12.4, exact algorithms for solving different types of vehicle routing problems and an outline of a heuristic algorithm for solving SRP and achieving the goals presented in Section 12.3 are presented. In Section 12.5, a discussion of the challenges that exist in solving SRP and some new technologies that may play a role in solving SRP in the future are presented.

12.2 WHY STREET ROUTING AND SCHEDULING PROBLEMS DIFFER FROM TRADITIONAL ROUTING AND SCHEDULING PROBLEMS

In this section, considerations that help to differentiate SRP from more traditional VRP are described. In the remainder of this paper, a stop will refer to either an individual location or a street segment that requires service.

Service Times

In many SRP, the service time at a stop can be extremely small. Examples of service times in actual applications are as follows:

1. In delivering newspapers, the service time at a stop can be a few seconds if the paper is thrown from the automobile to the curb or 1-2 minutes if the paper is delivered to the door.

2. In residential sanitation collection, the service time at a stop is the time it takes to service all of the locations on the street segment and traverse the street segment. Generally, in residential sanitation collection problems, it only takes a few seconds to collect the refuse at each location requiring service so that it only takes a couple of minutes to service a street segment

3. In containerized sanitation pickup problems, the service time at a stop is the time it takes the truck to lift the container and dump the contents of the container into the truck. Generally, this operation generally takes 1-3 minutes.

In many SRP, the number of stops (individual locations or street segments) on a route can be large (over 100) and the addition or deletion of a stop marginally affects the duration of the route. For example, we have found that in residential sanitation collection problems each route services 600 to 1500 individual locations. In these problems, if there are about 10 locations on each street segment, then there are 60 to 150 street segments (or stops) on a route. Because of the density of the stops requiring service, residential sanitation collection is generally considered to be a NRP and the routing and scheduling procedures are carried out over the street segments rather than the individual locations requiring service.

Euclidean Distance Computations and Shortest Paths

In a traditional PTPRP, the time between two stops is usually assumed to be based on the Euclidean distance (or some function of the Euclidean distance) between these stops. These times are used in the algorithm to break the stops down into routes and to sequence the stops on each route.

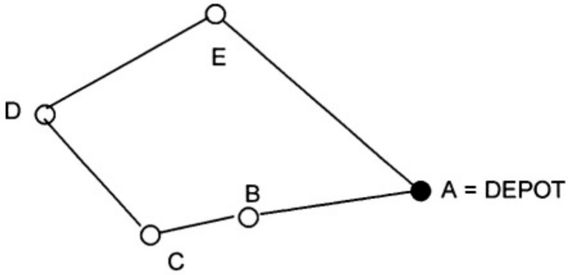


Figure 12.1: Euclidean Distance Route over Five Stops.

Figures 12.1-12.3 illustrate the problem that can arise in using Euclidean distances in a SRP. In Figure 12.1, the solution, A-B-C-D-E-A, to a 5-stop traveling salesman problem is displayed. This solution used the Euclidean distance between each pair of stops in order to determine the order to service the stops. To solve the problem in this way, a digital street network data base is not needed. In some SRP, routes that are developed using Euclidean distances is adequate. For example, if the stop are spatially spread out over the underlying street network (for example, each stop located in a different 5 digit zip code), then the solution found using Euclidean distances may be accurate enough. In other cases, using Euclidean distances can lead to inferior routes.

In Figure 12.2, the stops in Figure 12.1 are superimposed on top of a street network, as would be the case if this problem were solved as an SRP. The solution displayed in Figure 12.2 is found as follows. The sequencing of the stops is found by computing the Euclidean distance between the stops. The travel path is found by determining the actual travel path between the stops knowing the sequence for servicing the stops. In Figure 12.2 (and in Figure 12.3), the travel path is generated under the assumptions that all stops have to be serviced on the side of the street on which the stops are located and no U-turns are allowed.

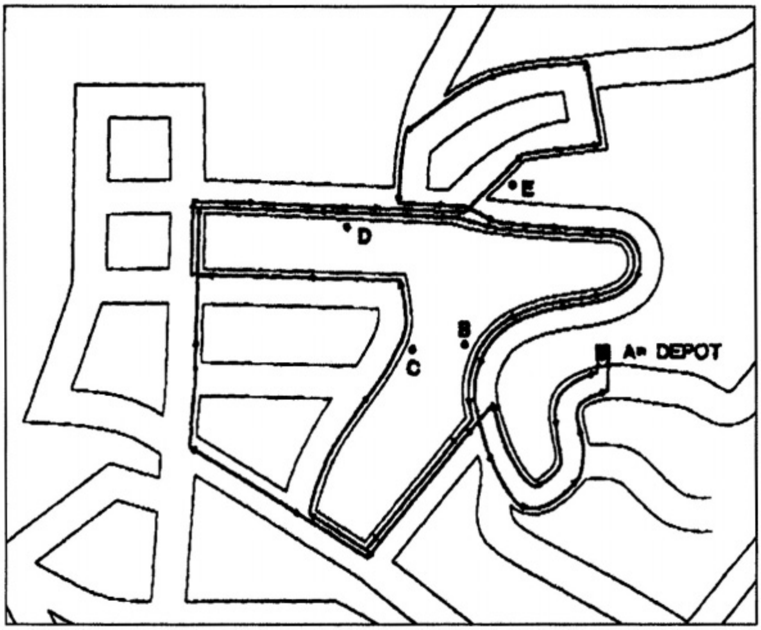


Figure 12.2 - Actual Travel Path for Euclidean Distance Route in Figure 12.1

Order of the route is maintained. No U-turns are allowed.
Service is carried out on the location's side of the street.
To get from A to B, without U-turns and service on the side of the street, the route has to go around the block.

In Figure 12.3, a route (and travel path) is generated under the same assumptions as the route displayed in Figure 12.2 except that, in determining the sequence, the shortest path between the stops is used as the travel time between the stops rather than the Euclidean distance. In this case, the order for servicing the stops in this solution is A-C-D-E-B-A. In terms of route duration, the solution shown in Figure 12.3 is clearly superior to the solution displayed in Figure 12.2.

Thus, in many SRP, using the underlying street network to compute the travel times between stops can lead to the generation of routes and travel paths which are more realistic and more acceptable to the user than the routes generated by using the Euclidean distance between the stops. Furthermore, options can be integrated into the algorithms for travel path generation to take into account issues such as street crossing difficulties and difficult turns (such as U-turns), natural barriers such as lakes and bridges, one-way streets and two-way streets, and ensuring that the stops are serviced on the correct side of the street. It is virtually impossible to take these considerations into account if Euclidean distances are used. In practice, these issues can play an important role in ensuring that the solution is implemented. However, to take these issues into account, the user is faced with an increase in 1) the computer time needed to get a solution, 2) the time and cost required to develop the system and 3) the accuracy of the underlying digital street network data base.

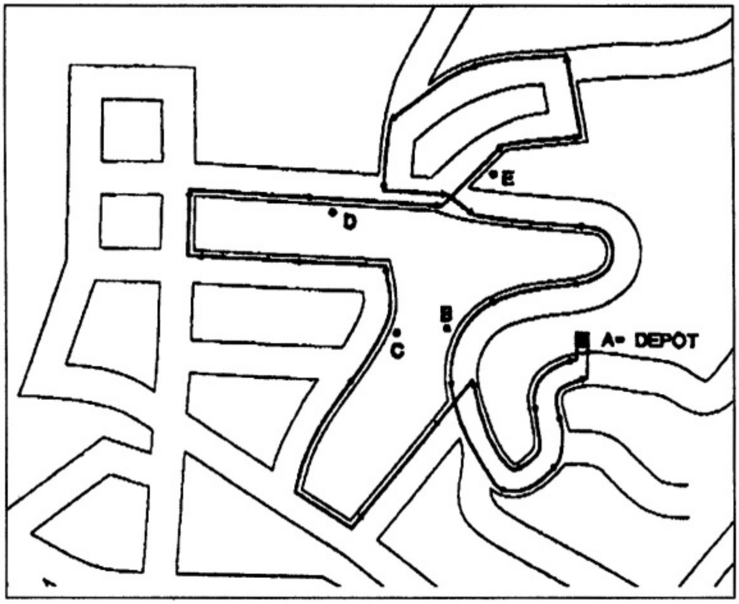


Figure 12.3: Travel Path Over the Street Network without U-turns

No U-turns are allowed.
Service is carried out on the location's side of the street.

Locating the Stops to be Serviced in SRPs

Closely aligned with the discussion in 12.2.2 is whether the stops should be located at their nearest intersection or on the correct street segment and side of street. Geocoding is the process of determining the correct street segment in a digital street network data base to associate with a location that requires service. In some cases, geocoding stops at their nearest intersection can lead to travel paths that give misleading results. This situation is illustrated in Figures 12.4 and 12.5.

In Figure 12.4, stops A, B, C, and D have been geocoded to the same intersection. Assume that these stops are on the same route. The portion of the travel path for the route servicing these four stops is displayed in Figure 12.4. This travel path does not traverse the street segments where stops B and C are located. This inaccuracy can lead to obvious errors such as incorrect estimates in the duration of the routes, travel paths which need severe revision when actually being traversed and errors in deciding upon the stops that can be feasibly placed on a route. On the other hand, in Figure 12.5, stops A, B, C, and D are coded to their correct street segment and side of the street. The travel path for the route displayed in Figure 12.5 is different and longer than the travel path in Figure 12.4 but more realistically represents the actual vehicle travel path and gives a better estimate of the duration of the route.

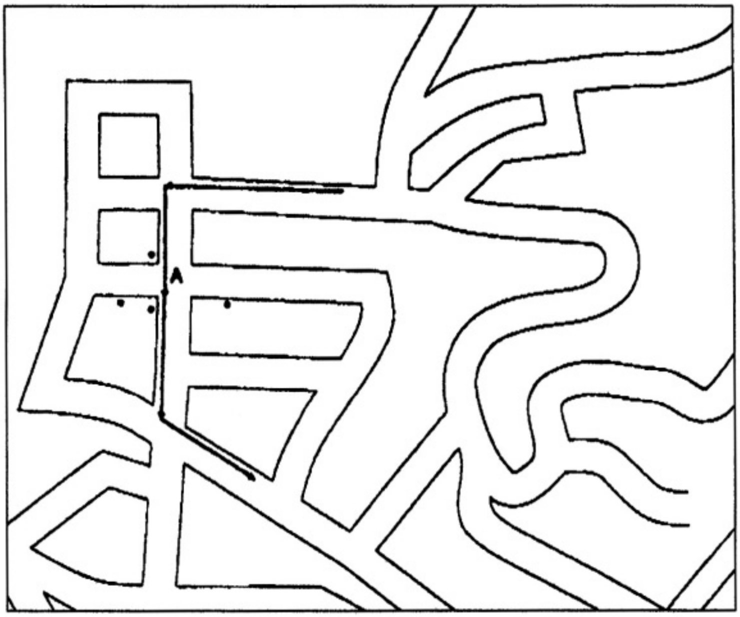


Figure 12.4 - Locations at 4 Street Segments adjacent to an Intersection

All 4 locations associated at A.
One-way street is vertical street denoted by arrow.
Travel path can be assumed to follow on-way street and is just the traversal of the one-way street in the vertical direction.

In applications where the stops are not dense or the driver is allowed to determine the travel path, locating stops at their nearest intersection is adequate. In other applications, the stops must be located on the appropriate street segment and side of street in order to get reasonable and realistic results. If these issues are important to the user, then the algorithms for solving the SRP must be implemented to account for these factors. These considerations lead to increases in system development time, execution time of the algorithms and costs of preparing the data.

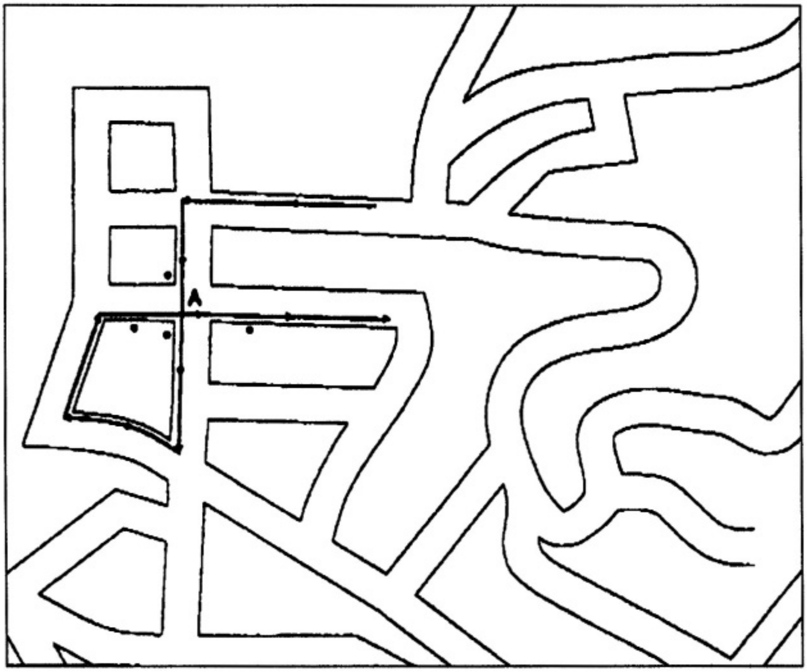


Figure 12.5 - Travel Path over street Network Obeying One-Way Street and Delivery on the correct side of the Street
As in Figure 12.4, one-way street is vertical street.
No U-turns are allowed. Notice how the path has to wander to service all four locations.

Integration of Driving and Walking Components

In some neighborhood routing applications, it is important to generate a travel path where some of the streets are walked and other streets are driven. This situation occurs in the scheduling the postal carriers who deliver mail for the United States Postal Service and the scheduling of meter readers. In these situations, the travel path for a route can be an integration of walking loops and a driving path. Each walking loop begins and ends at the same location (the location where the vehicle is parked) and can contain streets that are traversed but not serviced (deadhead walking streets). The driving path contains streets that require service and streets that are deadheaded (traversed but not serviced). Solving mixed driving and walking problems require details that are difficult to obtain when using standard approaches for solving these problem..

Geographic Data Bases and Geographic Information Systems (GIS)

As noted above, a critical component in solving SRPs is to have an accurate digital street network data base or geographic data base. Existing digital street network data bases for the same area can substantially differ in quality. Navigational quality digital data bases contain virtually all streets in the area, are accurate with respect to address ranges, overpasses and underpasses, streets divided by a median, and one way streets. In these data bases, stops can be accurately geocoded to the correct street segment and reliable shortest paths over the network can be found.

Data bases that do not have this precision can be used for solving some SRPs but generally require considerable manual updating. To correct most of the errors in a digital street network data base that is not accurate can be expensive. If a less accurate digital street network data base, then the computation of shortest paths can be inaccurate and some of the desired reports cannot be generated accurately. An easy way for a system to lose credibility with the crews executing the routes is to give each crew a travel path that is not accurate.

Generally, computer systems for solving SRPs include a Geographic Information Systems (GIS). The routing algorithms are embedded in the GIS and the street and stop data are stored in the GIS. With a GIS, an accurate digital street data base and the correct set of algorithms, the user can 1) geocode most stops automatically, 2) compute the travel time between stops as the shortest time travel path over the street network, 3) edit the digital street data base, 4) execute the algorithms, 5) display the results and 6) manually alter the solution. In the past few years, accurate digital street data bases and highly functional commercial GIS have become more available so that realistic solutions to a wide class of SRP can be found. The discussion of the requirements for accurate geographic data bases and the appropriate functionality of a GIS for use in a SRP can become quite complex and we have just given a brief overview of these topics.

User Intervention

In many street routing systems, the user can change the solution found by the algorithms. Such a solution is called a user generated solution and has a better chance of being implemented than a solution that is formed without manual intervention. The user generated solution is able to consider secondary constraints that the user is aware of but the algorithms are not able to capture. For example, in forming partitions for a SRP, there may be a major street that is difficult to cross. The user wants to form partitions where there are few crossings of this street from secondary streets. Partitioning procedures have great difficulty in capturing this constraint if the area over which the partitions are being formed lies on both sides of this street. The user can intervene in the solution process by splitting the service area into areas of interest where the borders of these areas of interest contain the troublesome streets. Then, the partitioning is carried out over each of these areas of interest, one area at a time. Then, the user can manually swap stops between partitions that lie in adjacent areas of interest without significantly compromising the quality of the partitioning.

12.3 GOALS IN SOLVING STREET ROUTING AND SCHEDULING PROBLEMS

The three most important goals (or criteria) in solving most SRP are the following:

1. Minimize the number of vehicles,

2. Minimize total travel time,

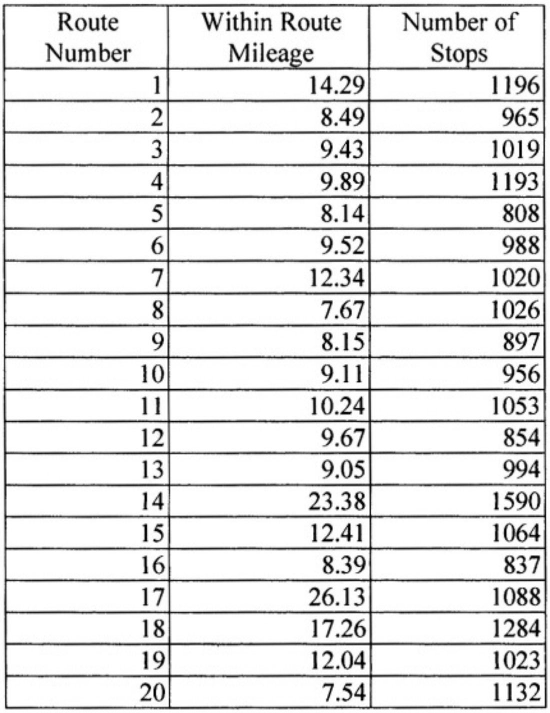
3. Ensure that every route in balanced in terms of time to carry out the routes (this time is called the workload of the route).

The first two goals are the traditional objectives of VRP. The goal of balancing workload is more specific to SRP.

We have observed numerous applications where the workload of the routes range between 4 to 10 hours, the crew that works the 4 hour route was paid for a full 8 hour day and the crew that works the 10 hour day earns 2 hours of overtime. The benefits in eliminating imbalances in the workload are 1) savings can be achieved, 2) overtime can be reduced and 3) the crews believe that they are receiving an equitable and fair deal.

An example on the use of workload balance in a residential sanitation routing problem that we solved is now presented. The Within Route Mileage and Number of Stops for the existing routes are given in Table 12.1. The Within Route Mileage is the total mileage covered by the crew in servicing the route and does not consider the distance between the route and the disposal facility, the route and the depot and the depot and the disposal facility. The number of stops are the number of residences on the route.

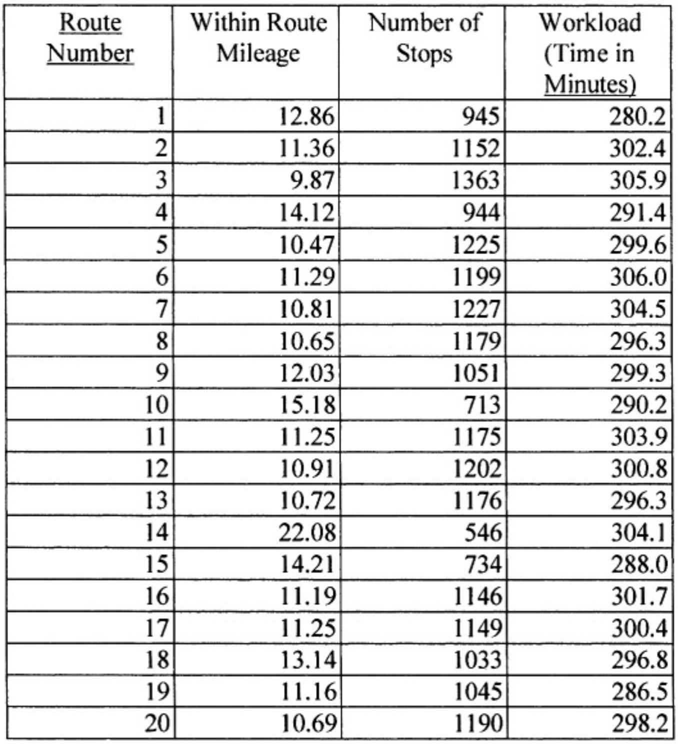
Route imbalance is illustrated in Routes 14, 16 and 17 of the existing routes. The statistics for these routes are denoted in bold in Table 12.1. Route 14's Within Route Mileage is 23.38 miles and contains 1590 stops, route 17's Within Route Mileage is 26.13 miles and contains 1088 stops whereas route 16's Within Route Mileage is 8.39 miles and contains 837 stops. In practice, the crews servicing routes 14 and 17 required over twice the time to complete the route as the crew servicing route 16. Moreover, on heavy volume days, two crews had to be sent out to service routes 14 and 17 - costing additional revenues. These imbalances were so severe that they caused discontent among the crews. The crews requested a route readjustment be carried out in order to achieve better balance. This request and the implementation of a recycling program in the area were the bases of the study that we carried out.



Route Number	Within Route Mileage	Number of Stops
1	14.29	1196
2	8.49	965
3	9.43	1019
4	9.89	1193
5	8.14	808
6	9.52	988
7	12.34	1020
8	7.67	1026
9	8.15	897
10	9.11	956
11	10.24	1053
12	9.67	854
13	9.05	994
14	23.38	1590
15	12.41	1064
16	8.39	837
17	26.13	1088
18	17.26	1284
19	12.04	1023
20	7.54	1132

Table 12.1 - Breakdown of the Existing Routes

In Table 12.2, the Within Route Mileage, Number of Stops and Within Route Travel Time for the computer generated routes in this region are presented. Forming 20 routes was specified in advance and route balance was the primary goal. The Within Route Travel Time or workload was computed as the sum of 1) the total service time on all the streets on the route and 2) the total within route deadhead time for the route. The total service time on a street was determined by linear regression using historic data on the actual time to service a route. The independent variables in the regression were number of stops on an existing route, length of the streets requiring service on an existing route and housing density on an existing route. The total service time on each street was computed from the regression coefficients for the independent variables when each existing route was considered an entity by itself. Times between the route and the depot and disposal facility were not considered in the workload. As can be seen, reasonable workload balance was achieved. The maximum deviation in estimated workload on the 20 routes is 26 minutes and the workload on most of the routes is 300 minutes ± 6 minutes.



Route Number	Within Route Mileage	Number of Stops	Workload (Time in Minutes)
1	12.86	945	280.2
2	11.36	1152	302.4
3	9.87	1363	305.9
4	14.12	944	291.4
5	10.47	1225	299.6
6	11.29	1199	306.0
7	10.81	1227	304.5
8	10.65	1179	296.3
9	12.03	1051	299.3
10	15.18	713	290.2
11	11.25	1175	303.9
12	10.91	1202	300.8
13	10.72	1176	296.3
14	22.08	546	304.1
15	14.21	734	288.0
16	11.19	1146	301.7
17	11.25	1149	300.4
18	13.14	1033	296.8
19	11.16	1045	286.5
20	10.69	1190	298.2

Table 12.2 - *Breakdown of the Computer Generated Routes*

In conclusion, route balance is a powerful goal in solving many SRPs . Route balance is most effective when the network is reasonably dense (urban and suburban areas) and the service times at the stops are small. In this way, the number of stops on each route is large and the deadhead travel time between stops is small.

12.4 ALGORITHMS FOR SOLVING VEHICLE ROUTING PROBLEMS

How to best solve vehicle routing problems has been an area of intense research for many years. There have been hundreds (if not thousands) of papers written on solving variants of vehicle routing problems. Euler in the early 1700s was one of the first to examine a vehicle routing problem. In honor of his analysis, this problem is called the Koenigsberg bridge problem (or one vehicle Chinese Postman problem or the fundamental problem in graph theory). Analysis of these problems has lead to the development of significant new theories in vehicle routing and other areas of combinatorial optimization. Practitioners have used these results to solve many applications, including the applications described in this paper and in other papers and books. The April, 1999 special issue of the Journal, Computers and Operations Research

(Volume 26, Number 4) and the book by Lawler, Lenstra, Rinnooy Kan and Shmoys eds. (1985) are devoted to the Traveling Salesman problem. Ball (1995), Bodin, Golden, Assad, and Ball (1983), Christofides (1985), Christofides and Mingozzi (1990) and Golden and Assad (1986, 1988) are survey articles or special issues of journals or books on vehicle routing. Laporte and Osman (1995) have a bibliography containing over 500 articles on vehicle routing.

In spite of all of this research into vehicle routing, there is no algorithm that optimally solves every VRP. Algorithms and procedures have been developed for optimally solving certain classes of vehicle routing problems but these procedures do not operate very well or do not generate feasible solutions if conditions on the problem change. The procedures for optimally solving the traveling salesman problem may not generate a feasible solution to the multi-vehicle VRP without modification. An algorithm for optimally solving the multi-vehicle VRP can be used to solve a traveling salesman problem (TSP) but this algorithm most likely will be slow computationally.

In this chapter, it is impossible to survey all of the approaches (heuristic, exact and meta-heuristic) for solving vehicle routing problems. Instead, we have decided to highlight some of the new procedures that have been designed for solving certain classes of vehicle routing problems. In particular, we have chosen to describe exact procedures for solving certain classes of vehicle routing problems and a generic heuristic procedure for solving street routing and scheduling problems. Some of these procedures are new and have not been published in the open literature.

Some street routing and scheduling problems can be solved by using the exact procedures described in 12.4.1-12.4.4 if route balance is not a major criterion. In the statement of the problems, it is assumed that a non-negative cost d_{ij} is associated with every edge {i,j if the problem is symmetric or arc {i,j} if the problem is asymmetric. In most papers, the symmetric distances are assumed to be Euclidean. However, in any of these cases, these distances (or travel times or costs) can be computed as shortest path distances. Moreover, if one is willing to solve an asymmetric problem, then turn restrictions, one way streets, and other factors mentioned in Section 12.2 can be considered in the computation of these distances. Thus, the exact procedures can be used to solve SRPs if the data used in the problem is appropriately defined.

Exact Algorithms for the Capacitated Vehicle Routing Problem

The Capacitated Vehicle Routing Problem (CVRP) is a special case of PTPRP. The CVRP is the problem of designing feasible routes for a set of homogeneous vehicles that make up a vehicle fleet. The objective in forming these routes is to minimize the total travel time of all of the routes. Each route begins and ends at the depot and contains a subset of the stops requiring service. A solution to this problem is feasible if the vehicle capacity on each route is not exceeded and all stops are assigned to a route. In the simplest statement of the CVRP, there are no lower and upper bounds on the duration of each route. As such, there are no route balance considerations.

The CVRP has been shown to be NP-hard. The fact that few algorithms have been produced to date which can solve the CVRP optimally reflects the difficulty of this problem. In view of the large number of practical constraints that appear in real-world CVRPs, all exact methods investigate a basic problem which is at the core of all vehicle routing problems. We call this core problem, the basic CVRP.

The basic CVRP ignores a large number and variety of constraints and complications that are often found in real-world problems. Some of these constraints that were discussed in the earlier sections of this chapter and can easily be incorporated in heuristic methods for solving the basic CVRP. However, these constraint and complications represent a small fraction of those found in practice. Also, some of these constraints can be inserted into exact branch and bound algorithms designed for the basic CVRP by rejecting, during the branching process, infeasible solutions. However, if it is not possible to change the lower bound computationally in order to take into account the new constraints, the performance of the resulting exact algorithm becomes very poor and only small CVRPs can be solved to optimality.

The Basic CVRP and its Extensions

The basic CVRP considered in this chapter is as follows. A complete undirected network (or graph) G=(V,E) is given where V=(0, 1, ..., n) is the set of vertices and E is the set of undirected arcs (edges). A non-negative cost d_{ij} is associated with every edge $\{i,j\} \in E$. V'=V\{0} is a set of n vertices, each vertex corresponds to a stop and vertex 0 corresponds to the depot. Henceforth, $i \in V$ will be used interchangeably to refer both to a stop and to its vertex location. Each stop i requires a supply of q_i units from depot 0. A set of M identical vehicles of capacity Q is located at the depot and used to service the stops; these M vehicles comprise the homogeneous vehicle fleet. It is required that every vehicle route start and end at the depot and that the load carried by each vehicle is no greater than Q. The route cost corresponds to the distance traveled on a route and is computed as the sum of the costs of the edges forming the route. The exact algorithms described in this paper find an optimal solution to the CVRP. An optimal solution to the CVRP is a set of M feasible routes, one for each vehicle, in which all stops are visited, the capacity of each vehicle is not exceeded and the sum of the route costs is minimized.

CVRP can be divided into symmetric CVRP and asymmetric CVRP. In the symmetric CVRP, the underlying network G=(V,E) is assumed to be undirected. The network for the symmetric CVRP is the network for the basic CVRP. In the asymmetric CVRP, the underlying network G'=(V,A) is assumed to be directed; i.e., every arc in A is assumed to be directed.

In the case of street routing and scheduling problems, the stops are located on a street network and the travel time between stops is computed as the shortest travel time path between stops. If the travel time matrix is symmetric, then we have the symmetric or basic CVRP. In the symmetric CVRP, it can be assumed that there are no one-way streets and turn and street crossing difficulties are not considered in setting up the travel time matrix. The asymmetric CVRP consider one-way streets and turn and street crossing difficulties since the travel time matrix is not symmetric. In this paper, we consider procedures for solving the symmetrical CVRP since the symmetric CVRP appears to be more difficult to solve to optimality than the asymmetric CVRP..

Exact algorithms for the asymmetric CVRP have been proposed by Laporte, Mercure and Norbert (1986) and by Fischetti, Toth and Vigo (1994). The latter method can solve exactly problems with 300 stops and average vehicle utilization ($\sum_{i \in V} q_i$)/M*Q $\cong 0.80$ where $\sum_{i \in V} q_i$ represents the total volume of all of the stops, M represents the number of vehicles in the fleet and Q represents the capacity of any vehicle. However, the size of problems solved to optimality decreases to 70 stops if the average vehicle utilization is increased to about 0.90. Neither Laporte, Mercure and Norbert (1986) nor Fischetti, Toth and Vigo (1994) report any attempt to solve symmetric CVRP.

Average vehicle utilization is a very important concept. As the average vehicle utilization approaches 1, the vehicles become more tightly packed and finding a feasible solution becomes more difficult. In many CVRP encountered in practice, the average vehicle utilization is at least 0.95. From the results stated in the previous paragraph, it is quite apparent that the current best methods for the asymmetric CVRP cannot be used for solving the symmetric CVRP (or possibly asymmetric CVRP that are almost symmetric). In particular, these approaches cannot be used to solve the set of symmetric test problems proposed in the literature where the vehicle utilization is about 0.95. Thus, in order to solve the symmetric CVRP to optimality, a separate set of formulations and algorithms have to be considered. These formulations and algorithms are the thrust of this section.

We now review some of the more significant exact methods for solving the basic CVRP. These exact methods can be classified into the following categories:

(1) Branch and cut

(2) Branch and Bound

(3) Dynamic programming

(4) Set partitioning based methods

(5) Commodity flow based methods

We then present the exact algorithms for the CVRP with Time Windows constraints (VRPTW) and the CVRP with Backhauls (VRPB) in order to show how exact CVRP methods can be extended to deal with the complexities of real-world routing problems.

This chapter is by no means exhaustive in describing all exact methods and results obtained on the CVRP. Additional results can be found in Christofides et al. (1979), Christofides (1985), Fisher (1995), Desrosiers et al. (1995), Laporte (1992, 1997) and Laporte and Osman (1995).

Branch and Cut Algorithms

Branch and cut methods try to extend to the symmetric CVRP (the graph G is assumed to be undirected) the successful results of polyhedral combinatorics developed for the traveling salesman problem by Chvatal (1973), Grötschel and Padberg (1979), Grötschel and Padberg (1985). These methods are based on the following formulation of the CVRP. Let x_{ij} be an integer variable representing the number of vehicles traversing the undirected arc (edge) {i,j}, and let r(S) be the number of vehicles needed to satisfy the demand of stops in S.

The basic CVRP can be formulated as the following integer program.

$$\text{Min} \quad \sum_{i<j} c_{ij} x_{ij} \tag{12.1}$$

$$\text{s.t.} \quad \sum_{i<j} x_{ij} + \sum_{i>j} x_{ji} = 2, \qquad i \in V' \tag{12.2}$$

$$\sum_{i \in S} \sum_{j \in V' \setminus S} x_{ij} \geq 2\, r(S), \qquad S \subseteq V', S \neq \varnothing \tag{12.3}$$

$$\sum_{j \in V'} x_{0j} = 2M \tag{12.4}$$

$$x_{ij} \in \{0, 1\}, \qquad i \in V', j \in V', i<j \tag{12.5}$$

$$x_{0j} \in \{0, 1, 2\}, \qquad j \in V' \tag{12.6}$$

Constraints (12.2) are the degree constraints for each stop. Constraints (12.3) are the capacity constraints which, for any subset S of stops, that does not include the depot, impose that r(S) vehicles enter and leave S, where r(S) is the minimum number of vehicles of capacity Q required for servicing the stops in S (i.e., $Q \cdot r(S) \geq \sum_{i \in S} q_i$). Constraints (12.3) are also called generalized subtour elimination constraints. It is NP-hard to compute r(S), since it corresponds to solve a bin-packing problem where r(S) is the minimum number of bins of capacity Q that are needed for packing the quantities q_i : $i \in S$. However, inequalities (12.3) remain valid if r(S) is replaced by a lower bound to its value, such as $\left\lceil \sum_{i \in S} q_i / Q \right\rceil$, where $\lceil y \rceil$ denotes the smallest integer not less than y. Constraint (12.4) states that M vehicles must leave and return to the depot while constraints (12.5) and (12.6) are the integrality constraints. Finally, $x_{0j} = 2$ corresponds to a route containing only stop j.

This formulation cannot be solved directly by a general purpose integer programming algorithm because constraints (12.3) are too numerous to be enumerated a priori. The exact algorithms based on this formulation share the following structure. Constraints (12.3) are relaxed. Then, at each iteration, the LP relaxation of the resulting problem that includes only a subset of constraints (12.3) is solved. If the optimal LP solution to this problem is integer and all constraints (12.3) are satisfied, then it is an optimal CVRP solution. Otherwise, constraints of type (12.3) that are violated by the optimal LP solution are added into the LP relaxation and a new iteration is performed.

This cutting plane procedure can be integrated into a branch and cut scheme to solve the CVRP to optimality. Laporte, Norbert and Desrochers (1985), using a procedure similar to the one described above, were able to solve to optimality randomly generated problems with 50 to 60 stops and average vehicle utilization $(\sum_{i \in V} q_i)/(M*Q) = 0.74$.

The LP relaxation of this formulation can be strengthened by adding other valid inequalities that hold for every feasible CVRP solution but might be violated once the integrality constraints (12.5) and (12.6) are relaxed. Laporte and Norbert (1984) describe valid inequalities for the CVRP based on the comb inequalities developed by Chvatal (1973) and Grötschel and Padberg (1979) for the traveling salesman problem. Cornuejols and Harche (1993) used comb inequalities to improve the LP relaxation of formulation (12.1) – (12.6). With the resulting branch and cut method, they were able to solve the 50 stop test problem described in Christofides and Eilon (1969).

A more sophisticated branch and cut algorithm based on formulation (12.1) – (12.6) has been proposed by Augerat et al. (1995). In addition to capacity constraints, they used new classes of valid inequalities, such as comb and extended comb inequalities, generalized capacity constraints and hypotour inequalities. These new inequalities lead to significant improvements in the quality of the bound. The resulting branch and cut algorithm has been able to solve some large CVRP test problems. One of these problems involved 135 stops and represents the largest CVRP problem ever solved to date and reported on in the literature. However, this branch and cut has failed in solving to optimality a well known 75 stop problem described in Christofides and Eilon (1969).

Branch and Bound Algorithms

The effectiveness of branch and bound algorithms is entirely dependent on the quality of the bounds used to limit the tree search. We will, therefore, discuss the derivation of such bounds.

Christofides, Mingozzi and Toth (198la) and Fisher (1994) use different modifications of spanning trees to obtain valid lower bounds for the symmetric basic CVRP. Christofides, Mingozzi and Toth (198la) observe that the removal from a CVRP solution with M routes of any set S_0 containing $y \leq M$ edges adjacent to the depot and of any set S_1 containing M – y edges not adjacent to the depot – one edge from each route – produces a tree with degree K = 2M – y at the depot. This tree is called the Degree Center Tree (K-DCT). If it is known that the maximum number of single stop routes (i.e., the routes made up by the depot and only one stop) is M_1, then y can be fixed a priori to be an integer in the range $M_1 \leq y \leq M$.

Let v(y), for any given y, be the sum of following three cost terms – 1) the cost of the least-cost K-DCT, 2) the cost of the y edges of minimum cost incident to the depot and 3) the cost of the M – y minimum cost edges not incident to the depot. v(y) is a valid lower bound to the CVRP for any y, $M_1 \leq y \leq M$. Hence, a valid lower bound to the CVRP is given by $\mathrm{Max}[v(y) : M_1 \leq y \leq M]$. Details of a polynomial algorithm for computing this lower bound are provided in Christofides, Mingozzi and Toth (198la). The lower bound can be strengthened by introducing Lagrangean penalties on the violated degree constraints (12.2) in the formulation of the basic CVRP given by (12.1) - (12.6). Christofides, Mingozzi and Toth have embedded this lower bound into a branch and bound algorithm and have solved to optimality CVRP problems containing 10 to 25 stops.

Fisher (1994a) describes an exact branch and bound algorithm for the CVRP where the lower bound is computed using a generalization of spanning trees, called M-trees. A M-tree is defined to be a set of n + M edges that span the graph G where n is the number of stops in the problem and M is the number of routes. If routes with a single stop are not allowed, then any CVRP solution is a M-tree with the depot having degree equal to 2M. Fisher shows that the CVRP can be modeled as the problem of finding a minimum cost M-tree with the degree of the depot constrained to be 2M and some side constraints that impose 1) vehicle capacity constraints and 2) the requirement that each stop is visited exactly once. As single stop routes are not allowed, the CVRP can be formulated using (0-1) binary variables as follows.

$$\min_{x \in X} \sum_{i<j} c_{ij} x_{ij} \qquad (12.7)$$

s.t. (12.2), (12.3) and (12.5).

where $X = \{x : x \text{ defines a M-tree satisfying } \sum_{j=1}^{n} x_{0j} = 2M\}$

The lower bound is computed by solving the Lagrangean problem obtained after dualizing the side constraints (12.2) and (12.3). The optimal solution of the Lagrangean problem is provided by the minimum cost M-tree with depot degree equal to 2M. A polynomial algorithm for computing the degree-constrained maximum M-tree is given in Fisher (1994b). Since the number of constraints (12.3) can be enormous, Fisher (1994a) developed an heuristic procedure for choosing an initial subset of these constraints and for dynamically adding and deleting constraints during the subgradient iterations performed to find the best lower bound. This lower bound has been embedded into an exact branch and bound algorithm that has produced the optimal solution for a well-known problem with 100 stops and several real-world problems with 25-71 stops.

Dynamic Programming Algorithms

Dynamic programming (DP) has been applied to solve several types of CVRP or to obtain tight lower bounds. Christofides, Mingozzi and Toth (1981b) present three formulations of the CVRP and introduce the state space relaxation method for relaxing the DP recursions in order to obtain valid lower bounds on the value of the optimal solutions. The computational results show that the ratio "lower bound/optimum" varies between 93.1% and 99.6% when these state space relaxations are used. Christofides (1985) reported that a CVRP involving 50 stops has been solved exactly by this approach. Problems involving up to 125 stops were solved within 2% of the optimum in less than 15 minutes on a CYBER 855.

A generalization of state space relaxation for dynamic programming is described in Mingozzi, Bianco and Ricciardelli (1997). This procedure is used to derive an exact algorithm for solving TSP with time windows and precedence constraints. The derived algorithm outperforms other methods presented in the literature for the same problem, and can be used also on asymmetric TSP. The authors report that TSP problems involving 120 cities and wide time windows can be solved exactly by the proposed method in less that 5 minutes on a Intel 486 (33Mhz) personal computer. Moreover, problems involving tight time windows are directly solved by the bounding procedure at the root node of the tree search.

Algorithms Based on the Set Partitioning Formulation

The set partitioning formulation of the basic CVRP that was initially introduced by Balinski and Quandt (1964) is now described. Let $R = \{1, 2, \ldots, \hat{r}\}$ be the family of all feasible routes. Also, let the index set of the stops in route r be N_r, the optimal cost of route r be d_r and the load of route r be $Q_r = \sum_{i \in N_r} q_i$. Let R_i be the index set of routes visiting stop i and y_r be a (0-1) binary variable whose value is equal to 1 if and only if route r is used in the optimal solution. The basic CVRP can be formulated as follows:

$$\text{(SP)} \qquad \text{Min} \sum_{r \in R} d_r y_r \qquad (12.8)$$

$$\text{s.t.} \quad \sum_{r\in R_i} y_r = 1 \qquad i\in V' \qquad (12.9)$$

$$\sum_{r\in R} y_r = M \qquad (12.10)$$

$$y_r \in \{0, 1\} \qquad r\in R \qquad (12.11)$$

In practice, this formulation cannot be solved directly since the number of variables can run into the millions, even for small size problems, and the computation of the optimal cost d_r for each route r requires solving a TSP on the subgraph defined by vertices $N_r \cup \{0\}$. In this section, we describe three exact algorithms for the basic CVRP that are based on formulation (SP).

In Christofides, Mingozzi and Toth (198la), a lower bound to the SP is developed. This lower bound is obtained by finding a feasible solution of the dual of the linear programming relaxation of the SP. This dual problem to the linear programming relaxation of the SP is denoted by DSP.

$$\text{(DSP)} \qquad z(\text{DSP})=\text{Max} \sum_{i\in V'} u_i + Mu_0$$

$$\text{s.t.} \quad \sum_{i\in N_r} u_i + u_0 \le d_r \qquad r\in R$$

$$u_i \quad \text{unrestricted} \qquad i\in V.$$

where u_i, $i\in V'$, are the dual variables of constraints (12.9) and u_0 is the dual variable of constraint (12.10). In Mingozzi, Christofides, Hadjiconstantinou (1994), it is shown that a feasible solution $u^*=(u_0^*, u_1^*, \ldots, u_n^*)$ to problem DSP is given by

$$u_i^* = q_i \,\text{Min}\left[\frac{\bar{d}_{iq}}{q} : q_i \le q \le Q\right] i\in V' \text{ and } u_0^*=0 \qquad (12.12)$$

where $\bar{d}_{iq}$ is a lower bound to the cost of the least cost route of load q passing thorough stop i; that is to say,

$$\bar{d}_{iq} \le d_r,\ \forall r\in R_i \text{ such that } \sum_{k\in N_r} q_k = q.$$

Therefore, a valid lower bound for the CVRP is given by:

$$z^*(\mathrm{DSP}) = \sum_{i \in V'} u_i^* .$$

A dynamic procedure for computing $\bar{d}_{iq}$, $\forall i \in V'$, and $q_i \leq q \leq Q$ is described in Christofides, Mingozzi and Toth (1981a). The value of the lower bound is improved by placing penalties λ_i, $i \in V'$, on the vertices having degree different from 2 in the solution of the bound. Subgradient optimization is used to maximize the lower bound. This bound is then embedded in a branch and bound algorithm. The resulting algorithm optimally solved CVRP ranging from 10 to 25 stops.

Hadjiconstantinou, Christofides and Mingozzi (1995) describe a new method for computing the values $\bar{d}_{iq}$, $q_i \leq q \leq Q$ $\forall i \in V'$, that is based on the computation of k-shortest paths and q-paths. The resulting CVRP lower bound is superior to $z^*(\mathrm{DSP})$ described above and the branch and bound algorithm is able to optimally solve problems involving up to 50 stops.

Mingozzi, Christofides and Hadjiconstantinou (1994) describe a new method for solving the SP formulation of the CVRP. They propose to solve problem SP using a subset $F \subset R$, (F is a set containing a limited number of routes) so that the resulting problem can be solved by a branch and bound algorithm. The optimal solution obtained for the resulting problem is not guaranteed to be an optimal CVRP solution. However, the method used to generate F permits an estimate of how far the cost of the solution obtained using their method is from the optimal solution (this estimate is called the distance of their solution from the optimal solution).

Let z(UB) be a valid upper bound for the CVRP and let u' be a "good" feasible solution of DSP of cost z'(DSP). The heuristic procedure used by Mingozzi, Christofides and Hadjiconstantinou (1994) for computing u' will be described later. Let $d'_r = d_r - \sum_{i \in N_r} u'_i - u'_0$ be the reduced cost of route $r \in R$ corresponding to the dual solution u' and let $F = F_1 \cup F_2 \cup \ldots \cup F_n$ where F_i is a subset of R_i, $i \in V'$. F is a subset of R that satisfies the following three conditions:

$$\left.\begin{array}{ll} \text{a)} & d'_r \leq z(\mathrm{UB}) - z'(\mathrm{DSP}), \quad r \in F_i \\ \text{b)} & \underset{r \in F_i}{\mathrm{Max}}[d'_r] \leq \underset{r \in R_i \setminus F_i}{Min}[d'_r] \\ \text{c)} & |F_i| \leq \Delta^{\max} \end{array}\right\} \qquad (12.13)$$

where $\Delta^{\max}$ is defined a priori. An efficient dynamic programming procedure, called GENF, for computing the subsets F_i, $i \in V'$, is described in Mingozzi, Christofides and Hadjiconstantinou (1994).

Let SP′ be the set partitioning problem that is obtained from SP by replacing the route set R with the subset F defined above. Let x^* be an optimal integer solution of SP′ of cost $z^*(SP')$ where we assume that $z^*(SP') = \infty$ if the set F does not contain any feasible CVRP solution. If $z^*(SP') < \infty$, then the corresponding solution x^* is a feasible and, possibly, an optimal CVRP solution. It is quite clear that the cost of any feasible CVRP solution containing a route (say *r*) having reduced cost d'_r is greater than or equal to $z'(DSP) + d'_r$. Therefore, if the reduced cost d'_r of every route $r \in R \backslash F$ is greater than z(UB) - z'(DSP), then x^* is an optimal CVRP solution. In fact, any CVRP solution containing a route $r \in R \backslash F$ has a cost greater than $z'(DSP) + d'_r$; but $d'_r \geq z(UB) - z'(DSP)$ and, therefore, $z'(DSP) + d'_r \geq z(UB) \geq z^*$.

In order to verify the optimality of x*, the following two cases have to be considered.

C1. $|F_i| < \Delta^{\max}$, $\forall i \in V'$. In this case, x^* represents an optimal CVRP solution since either $R \backslash F_i = \varnothing$, $\forall i \in V'$, or $d'_r > z(UB) - z'(DSP)$, $\forall r \in R \backslash F$.

C2. $|F_i| = \Delta^{\max}$ for some $i \in V$'. In this case, x^* might not be an optimal CVRP solution.

Let $\mu = \underset{i \in V'}{\text{Min}} \left[\underset{r \in F_i}{\text{Max}} \left[d'_r \right] \right]$. We have the following two subcases under C2:

C2.1: $z^*(SP') \leq z'(DSP) + \mu$ In this subcase, x^* is an optimal CVRP solution since any CVRP solution involving some route of the set R\F will have a cost greater than or equal to $z'(DSP) + \mu$.

C2.2: $z^*(SP') > z'(DSP) + \mu$. In this subcase, x^* may not be an optimal CVRP solution and $z'(DSP) + \mu$ is a valid lower bound for the CVRP.

The optimal SP′solution x^* can be obtained by means of an integer programming solver such as CPLEX.

The core of the method of Mingozzi, Christofides and Hadjiconstantinou is the heuristic procedure, called HDS, used to find a feasible solution u′ of DSP without generating the entire set of routes R. HDS computes a solution u′ of cost z′(DSP) to DSP as the sum of the dual solutions obtained by a sequence of three different relaxations of the CVRP where each relaxation exploits a different substructure of the problem. The value of the lower bound z′(DSP) obtained is greater than the maximum of the values computed by each individual relaxation. The first two relaxations are based on graph theory considerations while the third relaxation requires the generation, by means of procedure GENF, of a limited subset of routes. Mingozzi, Christofides and Hadjiconstantinou report optimal solutions of problems up to 50 stops.

Algorithm GENF can be easily adapted to deal with real-world CVRPs (e.g., involving time windows, delivery and collections, preferences, and so forth) simply by rejecting any infeasible route generated. Computational results concerning CVRP with time windows are reported in a following section. Moreover, this procedure can be easily adapted to deal with the asymmetric CVRP.

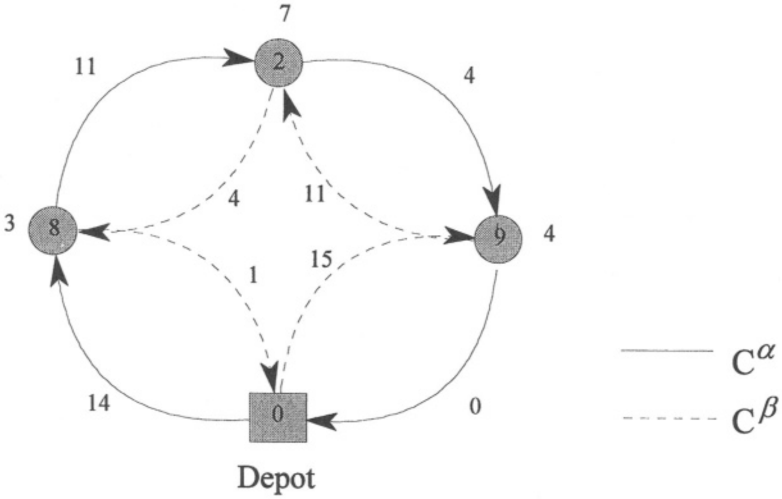


Figure 12.6 Flow Circuits for a 3 Stop Route

A New Algorithm based on a Two-Commodity Network Flow Formulation

Most integer programming formulations of the CVRP use binary variables as vehicle flow variables to indicate if a vehicle travels between two stops in the optimal solution. Fisher and Jaikumar (1981) formulate the CVRP using three index binary variables x_{ijk} as vehicle flow variables to indicate whether vehicle k travels directly from stop i to stop j ($x_{ijk} = 1$ if vehicle k travels directly from stop i to stop j and 0 if not). This formulation is used to derive an algorithm based on Benders decomposition. The master problem in the Bender's decomposition is a generalized assignment problem and the Benders inequalities are derived by solving M independent TSPs, where a TSP is solved over the stops assigned to a vehicle.

The main complication in using this method concerns the derivation of the Benders inequalities. In fact, Benders constraints are derived from the optimal dual variables of the M subproblems. Since the TSP subproblems are integer programs, dual variables cannot be obtained directly. Thus, Fisher and Jaikumar (1981) used this Benders decomposition approach to derive an heuristic algorithm for the CVRP.

In this section, we discuss a new integer programming formulation for the symmetric CVRP, proposed by Baldacci, Mingozzi, Hadjiconstantinou (1998). This formulation, called 2CVRP, is interesting in many different ways. It can be shown that its LP-relaxation satisfies a weak form of the subtour elimination constraints. The formulation can also be modified to accommodate different constraints and, therefore, is capable of being extended to different routing problems. The two-commodity formulation has been used by Lucena (1986) to derive new lower bounds for the VRP and by Langevin et al. (1993) for solving the TSP and the Makespan Problem with time windows. Baldacci, Mingozzi, Hadjiconstantinou (1998) use the two-commodity approach to derive new integer programming formulations for the CVRP, the TSP with mixed deliveries and collections (TSPDC) and the TSP with Backhauls (TSPB).

The idea behind this formulation is to use two flow variables, x_{ij} and x_{ji}, to represent an edge {i,j} of a feasible CVRP solution. If a vehicle travels from i to j, then x_{ij} represents the load of the vehicle and x_{ji} represents the empty space on the vehicle (i.e., $x_{ji} = Q - x_{ij}$), whereas, if the vehicle travels from j to i then x_{ij} and x_{ji} represent the empty space on the vehicle and the load respectively. Thus, the flow variables x_{ij} define two flow circuits for any feasible solution that represents a route. One circuit is defined by the flow variables representing the vehicle load while the second circuit is defined by the flow variables representing the empty space on the vehicle.

In Figure 12.6, a three stop route for a vehicle of capacity Q=15 is shown. Also, in Figure 12.6, the two circuits C^{α} and C^{β} represented by the flow variables x_{ij} defining the route are displayed. Circuit C^{α} is formed by the variables representing the vehicle load. Thus, the flow $x_{08} = 14$. indicates the total demand of the three stops, $x_{82} = 11$ represents the load of the vehicle in traveling from 8 to 2 after having unloaded 3 load units at stop 8, $x_{29} = 4$ represents the load of the vehicle in traveling from 2 to 9 after having unloaded 7 load units at stop 2, finally $x_{90} = 0$ represents the load of the vehicle in returning to the depot after having unloaded the remaining 4 load units at stop 9.

Circuit C^{β} is formed by the variables representing the empty space on the vehicle. Thus, $x_{09} = 15$ indicates that the vehicle arrives empty at the depot, $x_{92} = 11$ represents the empty space of the vehicle in traveling from 2 to 9, $x_{28} = 4$ represents the empty space in traveling from 8 to 2, finally $x_{80} = 1$ represents the empty space of the vehicle leaving the depot. Every edge {i,j} of the route has $x_{ij} + x_{ji} = Q$.

The new CVRP formulation is as follows. Let ξ_{ij} be a 0-1 binary variable equal to 1 if edge {i,j} is in solution and 0 if edge (i,j) is not in the solution. Let x_{ij} be the flow value of arc (i,j), $i,j \in V$, $i \neq j$.

$$\text{(2CVRP)} \quad z(\text{CVRP}) = \text{Min} \sum_{\{i,j\} \in E} c_{ij}\xi_{ij} \qquad (12.14)$$

$$\text{s.t.} \quad \sum_{j \in V} x_{ij} - \sum_{j \in V} x_{ji} = -2q_i, \qquad i \in V' \qquad (12.15)$$

$$\sum_{j \in V} x_{0j} - \sum_{j \in V} x_{j0} = MQ - \sum_{j \in V'} q_j \qquad (12.16)$$

$$x_{ij} + x_{ji} = Q\xi_{ij}, \qquad \{i,j\} \in E \qquad (12.17)$$

$$\sum_{i<j} \xi_{ij} + \sum_{i>j} \xi_{ji} = 2, \qquad i \in V' \qquad (12.18)$$

$$\sum_{j \in V'} \xi_{0j} = 2M \qquad (12.19)$$

$$x_{ij} \geq 0, \qquad i,j \in V \qquad (12.20)$$

$$\xi_{ij} \in \{0,1\}, \qquad \{i,j\} \in E \qquad (12.21)$$

Constraints (12.15), (12.16) and (12.20) define a feasible flow for variables $\{x_{ij}\}$. Constraints (12.17)-(12.19) force the degree of each stop to be 2 and the degree of the depot to be 2M, respectively. Constraints (12.21) are the integrality constraints. The supply-demand pattern involved ensures that there are paths from vertex 0 to any vertex in V' and back from any of the vertices in V' to vertex 0. Since, from (12.17) and (12.20), $x_{ij} + x_{ji} = Q\xi_{ij}$, $\forall \{i,j\} \in E$, the capacity of the vehicle will never be exceeded in the route allocated to it.

Baldacci, Mingozzi and Hadjiconatatinou (1998) describe a valid lower bound for the CVRP that is obtained from the LP-relaxation of formulation 2CVRP by adding valid inequalities that are satisfied by any feasible integer solution but not necessarily verified by the LP-relaxation. They consider inequalities such as flow inequalities and generalized subtour elimination constraints as described by Augerat et al. (1995). The flow inequalities are derived from the observation that in any feasible integer solution, if $\xi_{ij} = 1$ for some edge not incident to the depot, then $x_{ij} \geq q_i$ and $x_{ji} \geq q_i$. Therefore, $x_{ij} \geq q_i \xi_{ij}$ and $x_{ji} \geq q_i \xi_{ij}$, $\forall \{i,j\} \in E$, are valid inequalities and are called flow inequalities. The value of the lower bound is obtained by iteratively adding the violated inequalities to the LP relaxation in the same way as adding cutting planes to the LP relaxation. A branch and cut algorithm based on this bound has been used by Baldacci, Mingozzi and Hadjiconatatinou (1998) to solve well known CVRP test problems with up to 100 stops.

Formulation 2CVRP becomes a valid formulation for the symmetric TSP by setting $q_i = 1$ $\forall i \in V'$, $M=1$ and $Q = n - 1$. Moreover, Baldacci, Mingozzi and Hadjiconatatinou (1998) use the two commodity approach to derive new integer programming formulations for the TSP with mixed deliveries and collections (TSPDC) and for the TSP with backhauls (TSPB). These formulations are used to derive new lower bounds and new branch and cut algorithms. The computational results show that for both TSPDC and TPB the lower bounds are very tight and the exact algorithms can solve problems with up to 150 stops.

Exact Algorithms for the Vehicle Routing Problem with Time Windows

The basic CVRP does not consider several commonly encountered real-world constraints. One very common constraint in real-world problems is to have time windows at the stops. The resulting problem, denoted by VRPTW, inherits the structural definition of the basic CVRP described in Section 12.4.1 and requires the following additional characteristics.

A travel time t_{ij} is associated with each edge $\{i,j\} \in E$. At each stop $i \in V'$ is associated a service time σ_i required by a vehicle to visit the stop and to unload the quantity q_i (we assume $\sigma_0=0$). The start time of the service at stop i must be within a given time window $[a_i, b_i]$. A vehicle is permitted to arrive at stop i before the beginning of the time window and wait at no cost until time a_i. Also vehicles are time-constrained at the depot in that each vehicle must leave the depot and return back within the time window $[a_0, b_0]$.

We now present a mixed-integer programming formulation of the VRPTW, adapted from the one proposed by Kohl and Madsen (1997). This formulation involves two types of decision variables. The 0-1 binary variable x_{ijk} is 1 if and only if vehicle k visits stop j immediately after visiting stop i and 0 if not. The continuous variable s_{ik} denotes the time vehicle k begins service at stop i . It is assumed that s_{0k} denotes the departure time of vehicle k from the depot.

The VRPTW can be formulated as follows:

$$\text{Min} \quad \sum_{k=1}^{M} \sum_{i \in V} \sum_{j \in V} c_{ij} x_{ijk} \tag{12.22}$$

$$\text{s.t.} \quad \sum_{k=1}^{M} \sum_{j \in V} x_{ijk} = 1, \qquad i \in V' \tag{12.23}$$

$$\sum_{i \in V'} q_i \sum_{j \in V} x_{ijk} \le Q, \qquad k = 1, \dots, M \tag{12.24}$$

$$\sum_{j \in V'} x_{0jk} \le 1, \qquad k = 1, \dots, M \tag{12.25}$$

$$\sum_{i \in V} x_{ijk} - \sum_{i \in V} x_{jik} = 0, \qquad j \in V, k = 1, \dots, M \tag{12.26}$$

$$s_{ik} + \sigma_i + t_{ij} - L(1 - x_{ijk}) \le s_{jk}, \qquad i \in V, j \in V', k = 1, \dots, M \tag{12.27}$$

$$s_{ik} + \sigma_i + t_{i0} - L(1 - x_{i0k}) \le b_0, \qquad i \in V', k = 1, \dots, M \tag{12.28}$$

$$a_i \le s_{ik} \le b_i, \qquad i \in V, k = 1, \dots, M \tag{12.29}$$

$$x_{ijk} \in \{0, 1\}, \qquad i,j \in V, k = 1, \dots, M \tag{12.30}$$

Constraints (12.23) state that each stop must be visited exactly once. Constraints (12.24) are the capacity limitation on the vehicles. Constraints (12.25) force each vehicle to be used at most once and constraints (12.26) state that if a vehicle visits a stop, it must also depart from it. Constraints (12.27) impose that vehicle k cannot arrive at stop j before $s_{ik} + \sigma_i + t_{ij}$, if it travels from i to j. Constraints (12.28) force each vehicle k to return to the depot before time b_0. The scalar L can be any large number. Constraints (12.29) ensure that all time windows are respected and these time windows are assumed hard and constraints (12.30) are the integrality constraints.

Being a generalization of the CVRP, the VRPTW is NP-hard. Since the VRPTW occurs in many applications, several research efforts have been devoted to finding solutions to this problem by means of exact and heuristic methods. In this section, we concentrate on the exact approaches. A review of heuristic procedures can be found in Desrosiers et al. (1995).

The most successful exact algorithms to date follow one of three main approaches: column generation (set partitioning formulation), Lagrangean decomposition (variable splitting) or M-tree relaxation (M-trees are discussed in section 12.4.1). The M-tree relaxation approach proposed by Fisher et al. (1997) extends the approach presented by Fisher (1994) for the basic VRP to the VRPTW. The idea for the basic CVRP is to identify subsets S of stops which must be serviced by at least r(S) vehicles. Analogous considerations on the time windows lead to constraints specifying that not all arcs in a route which violates time window constraints can be simultaneously in the solution. These constraints are then relaxed in a Lagrangean fashion, obtaining a problem that is a degree constrained M-tree problem. Subgradient optimization can then be used to tighten the resulting bound. This method has solve to optimality clustered problems up to 100 stops under the assumption that each route contains at least two stops. This approach loses effectiveness when the stops are not clustered or when the time windows are tight. In these cases, the number of violated time-windows constraints increases and the interplay of subgradient optimization and violated constraint identification becomes more computationally demanding.

The other two approaches (column generation and Lagrangean decomposition) are both based on the observation that only assignment constraints, specifying that each stop must be assigned to exactly one vehicle, are formulated with summation over all vehicles. When these constraints are relaxed, or confined in a subproblem following a decomposition, the resulting problem is an Elementary Shortest Path problem with capacity and time window restrictions for each vehicle and is a NP-hard problem (Dror, 1995). However, effective dynamic programming strategies have been proposed for its relaxation. In these procedures, non-elementary paths are accepted. These non-elementary paths permit negative costs that induce cycles. The cycles cause stops to be serviced more than once in a single path.

Desrocher et al. (1992) make use of this possibility of decomposition of the VRPTW in a column generation framework. The problem is decomposed into a master problem and a subproblem. The master problem ensures that all stops are serviced by suitably choosing a subset of paths, where the paths are generated as solutions of the subproblem. The master problem is formulated as a set partitioning problem where each column represents a path. The solution of this problem specifies both how many vehicles are used to service all the stops and the path to be followed by each of the vehicles (all of the vehicles are assumed identical). For implementation reasons, since the paths represented by the columns are not elementary paths but can contain cycles, the master problem is actually formulated as a set covering problem. The set covering solution can be converted to a set partitioning solution by means of a branch and bound algorithm. The set covering matrix is initialized with paths starting from the depot, reaching one stop and returning to the depot.

The optimal solution of the LP relaxation of the set covering problem produces a set of associated dual optimal values. These dual optimal values are used in the subproblem to generate columns with negative reduced costs. Each subproblem is a shortest path problem with capacity and time window constraints. The arc costs in the subproblems are derived from the master dual optimal values. Several negative reduced cost paths can be found when solving each of the subproblems. These paths are included as new columns in the next iteration of the master problem. The master and the subproblems are solved iteratively until either a predefined maximum number of columns has been generated or the reduced cost of the generated columns are strictly positive.

At this point, the value of the objective function of the optimal LP solution of the master problem is a lower bound to the VRPTW. A branch and bound strategy is then employed to find a feasible and possibly optimal solution to the VRPTW. The optimality of a solution is proved when either 1) the initial LP solution is integer, or 2) at each node of the search tree, a sufficient number of additional columns is generated to rule out the possibility that a non-generated column may enter the optimal integer solution. This method has successful problems with up to 100 stops. This procedure performed best on problems where the stops are clustered, worse on problems where the stops are uniformly randomly located over the area, and experienced the most difficulty on random-clustered problems (i.e. problems where some stops are randomly located and the other stops are clustered).

A column generation method was also used by Kohl, Desrosiers, Madsen, Solomon and Soumis (1997), where the master problem is strengthened by means of valid inequalities that are added as extra rows of the set partitioning formulation. The inequalities added are k-path cuts with k = 2 (introduced by Laporte Nobert and Desrochers (1985)). Subsets of stops that require at least 2 vehicles to service but are serviced by less than two vehicles in the optimal (fractional) LP solution of the master are identified. The identification of these subsets cannot be done in polynomial time. However, efficient algorithms for identifying these subsets are proposed by Kohl, Desrosiers, Madsen, Solomon and Soumis (1997).

The cutting plane generation was inserted in the column generation as a means to strengthen the LP at the root node. In fact, columns are generated by means of the capacity and time window constrained shortest path subproblem until no more negative reduced costs paths exist or an a priori specified number of columns is generated (we use 20000 columns). Then, cutting planes (2-paths and subtour elimination constraints) are identified. If the solution proposed by the LP after the addition of these cuts is still fractional, a branching strategy is initiated. At each node of the search tree, new columns are generated (provided their total number is below the specified number (20000) but no new cuts are added. As shown in Table 12.3, this approach has proved very effective. It has solved all clustered problems and most random and random-clustered problems up to 100 stops.

A solution strategy based on Lagrangean decomposition has been proposed by Kohl and Madsen (1997). As with the previous algorithms, the problem is decomposed into a master problem requiring each stop to be assigned to a route and a subproblem for each vehicle consisting of an elementary shortest path problem with capacity and time window restrictions. This problem is NP-hard and cannot be solved exactly for reasonable size problems. For this reason, Kohl and Madsen consider a relaxed problem that allows a path servicing the same stop more than once. This relaxed problem is then solved by means of a dynamic programming algorithm based on the one proposed by Desrochers (1988).

When the assignment constraints are relaxed in a Lagrangean fashion, the purpose of the master problem is to find the optimal Lagrangean multipliers. Thus, the VRPTW decomposes into a number of shortest path subproblems where the costs of the shortest paths are affected by the values of the Lagrangean multipliers. The optimal Lagrangean multipliers are found using a method exploiting the benefits of subgradient methods as well as bundle methods proposed by Lemaréchal, Strodiot and Bihain(1981).

If the penalty optimization does not yield a feasible VRPTW solution, a branch and bound strategy is initiated. Mingozzi, Baldacci and Palumbo (1998) describe a new exact algorithm for the VRPTW that is an extension of the exact algorithm proposed by Mingozzi, Christofides and Hadjiconstantinou (1994) for the basic CVRP (see section 12.4.1). However, the computational performance of this algorithm for the VRPTW is greatly improved if the procedures used to compute the lower bound are customized to deal directly with time windows. Mingozzi, Baldacci and Palumbo (1998) propose a new heuristic procedure that takes into account the time window constraints for solving the dual of the LP-relaxation of the set partitioning formulation of the VRPTW. This method is based on the state space relaxation of the dynamic programming formulation of the TSP with time window introduced by Christofides, Mingozzi and Toth (1981b) described in section 12.4.1. This procedure is combined with other two heuristics to obtain a valid lower bound. Moreover, Mingozzi, Baldacci and Palumbo (1998) describe a new route generation method that use a lower bound to fathom the enumeration of states that cannot lead to feasible routes. The computational results show that the lower bound is tight and the resulting method can solve to optimality difficult VRPTW problems with 100 stops that cannot be solved by other exact algorithms.

Table 12.3 shows the number of problems solved and the number of optimal solution found by different methods on three classes of test problems proposed by Solomon (1987). The approach described in Mingozzi, Baldacci and Palumbo (1998) appears to be the most successful approach for solving these VRPTW exactly. In Table 12.3, R means random problems, C means clustered problems and RC means random-clustered problems.

Vehicle Routing Problem with Backhauls

The Vehicle Routing Problem with Backhauls (VRPB) is an extension of the basic CVRP. In the VRPB, M identical vehicles stationed at a central depot are to be used to supply a set of stops. The stops are broken down into two sets called Linehaul customers and Backhaul customers. The *Linehaul* customers are stops that require deliveries by the vehicles from the depot. The *Backhaul* customers have products that the vehicles collect and these products are unloaded at the depot. In the VRPB, each of the M vehicles must be used (i.e. all M routes are used in the solution), each route is to begin and end at the depot and all Linehaul customers and Backhaul customers are to be serviced. In this version of the VRPB, the routes are constrained so that each Backhaul customer on a route is serviced after every Linehaul customer on the route is serviced. Moreover, it is required that both the total load supplied to the

	R		C		RC		Total	
	Solv.	Opt. found	Solv.	Opt. found	Solv.	Opt. found	Solv.	Opt. found
Desrochers et al. (1992)	21	21	21	21	8	8	50	50
Fisher et al. (1997)	5	4	15	15	1	1	21	20
Kohl, Madsen (1997)	0	0	27	27	0	0	27	27
Kohl et al. (1997)	24	12	27	27	18	4	69	43
Mingozzi et al. (1998)	26	22	27	27	16	13	69	62
Num. of test problems	36		27		24		87	

Table 12.3 - Problems solved and optimal solutions found on Solomon test problems

to the Linehaul customers on a route and the total load collected from the Backhaul customers on the route must not exceed the vehicle capacity Q. The objective is to minimize the total distance travelled by the M vehicles.

In order to ensure feasibility, it is assumed that $M \geq [M_L, M_B]$, where M_L is the minimum number of vehicles need to visit all the Linehaul customers and M_B is the minimum number of vehicles need to visit all the Backhaul customers. In the literature, it is assumed that routes containing only Backhaul customers are not allowed. However, the exact methods for solving the VRPB described below can be easily extended to consider the more general case where this assumption is not present.

Two exact methods have been proposed for the VRPB. Toth and Vigo (1997) describe an exact branch and bound method based on the integer programming formulation. They describe a Lagrangean lower bound based on a relaxation that leads to the determination of Shortest Spanning Arborescences and min-cost flow problems. The lower bound is strengthened in a cutting plane fashion by adding violated capacity-cut constraints.

Mingozzi, Baldacci and Giorgi (1997) propose an exact algorithm for the VRPB. This algorithm is based on an integer formulation requiring the following two set of paths – 1) all feasible paths starting from the depot and visiting delivery stops only, 2) all feasible paths starting from a backhaul stop, visiting backhaul stops only and ending at the depot. The integer programming formulation combines M delivery paths with M backhaul paths in order to form M routes that visits all stops. The lower bound is computed by heuristically solving the dual of the LP-relaxation of the integer formulation. This heuristic method involves two different procedures that exploit two different relaxations and do not require the generation of the sets of all feasible paths. The dual solution obtained and a valid upper bound are used to drastically reduce the number of variables in the integer program so that the resulting problem can be solved to optimality by means of an integer programming code such as CPLEX 3.0.

Extensive computational tests on two classes of problems proposed in the literature show that the lower bound introduced by Mingozzi, Baldacci and Giorgi is greater than the lower bound of Toth and Vigo. Moreover, the Mingozzi, Baldacci and Giorgi approach can optimally solve larger problems than the Toth and Vigo procedure. However, the computational tests on these two procedures give ambiguous results since each approach solves test problems that the other approach fails to solve. Thus, at this point in time, we regard both approaches are competitive for solving the VRPB exactly.

Algorithm For Street Routing And Scheduling Problems

In the previous parts of section 12.4, exact algorithms for solving various vehicle routing problems have been presented. In all of these problems, there is no length of path restriction so that the route balancing does not play a role. In this section, we present an outline of a generic algorithm for solving street routing and scheduling. This algorithm assumes that the service time on each of the stops is small (for example, no more than 3-5 minutes), the number of stops on each of the routes is large (say, greater than 50), there are no time windows and route balancing is an important criterion. This algorithm may not be as effective if the stops have time windows, the number of stops on some of the routes is small, the service time at some of the stops is large and route interlacing or overlap is acceptable.

This algorithm is a 'cluster first, route second' procedure (Bodin, Golden, Assad and Ball (1983)). The algorithm first partitions a region into routes. It then generates the travel paths and determines the amount of deadheading on a route. If the solution is not good enough, then the next iteration through the procedure is carried out with revised estimates of the number of vehicles needed to service the stops. If the routes look reasonable (balance has be attained a prespecified number of iterations has been carried out), then the solution is examined to ensure that there is not a significant amount of overlap or (interlacing) on the routes. If necessary, the systems that we have built that use this procedure then allow the user to manually exchange stops between routes in order to improve balance or remove interlacing. More details of this procedure are given below.

Each of the steps outlined below generally has to be specialized to account for specific applications. In other words, general procedures to solve all street routing and scheduling procedures have not been developed. However, we believe that the approach outlined presented below can serve as a general approach for solving SRP.

Step a: Estimating the Number of Routes to Form

The number of routes to service a region is estimated and this estimate need not be an integer. Suppose that the estimate of the number of routes to form in a region is 9.5. Then the user must decide whether to form i) 9 regular (or full) routes and pay overtime, ii) 9 regular routes and a remnant (or partial) route or iii) 10 regular routes. This initial estimate is derived without an accurate knowledge of the time to drive between the routes and the depot.

Knowing the number of routes to form over the area of interest, a target workload for a route is established. The user then declares the target lower and upper bounds on the workload in any route. In the above example, if 10 regular routes are to be formed, then the target workload can be 440 minutes and the acceptable workload interval is [440 – 15 minutes, 440+15 minutes]. This goal on workload balance can be assumed to be soft so that the acceptable workload interval can be violated. If all routes when formed fall within the acceptable workload interval, then the solution is considered balanced.

Step b: Form an Initial Set of Partitions

The region over which the routes are to be formed is broken down into the number of partitions established in Step a. This partitioning is carried out in two parts - an initial partitioning step and an automatic swapping of stops between partitions. The purpose of this step is to assign every stop to a partition and that the workload in each partition falls within the accepted workload interval.

Step c: Form Travel Paths

Travel paths are then generated over each partition. With these travel paths, accurate estimates of the amount of nonproductive time (deadhead time) on each route and, therefore, accurate estimates of the workload in each partition can be determined. These estimates include the time between the depot and each partition. Moreover, in the case of the routing of vehicles for residential and containerized sanitation problems, further estimates include the time between the disposal facility and each partition and the number of trips to the disposal facility to make.

Step d: Are the Partitions Balanced?

If this solution is better than the “best solution found so far,” then this solution becomes the new “best solution found so far.” If the workload on all of the partitions (including the deadhead travel times) fall within the accepted workload interval or if a specified number of iterations of the algorithm has been carried out, then the algorithm stops. Otherwise, the workload estimate is revised and the algorithm returns to Step c with a new estimate on the number of routes. The workload estimate on this step can differs from the workload estimate in Step a since estimates of deadhead time on each of the routes are now known.

Generally, this algorithm takes between 3-6 iterations to converge and can be used regardless of whether the stops are street segments or actual locations. We have found that this procedure generates partitions that are balanced and contain little overlap (or interlacing). This procedure works best if the stops do not have time windows. If the stops have time windows, then the routes may turn out balanced but have interlacing or be formed with little interlacing but may not be balanced. Also, if the route allows the possibility of making more than one trip to the depot during the route, then volume and weight do not play a role in the partitioning.

Manual intervention between steps b and c and between steps c and d can be very useful in this process by allowing the user to 1) remove interlacing or 2) perform a more complex set of stop exchanges than were considered with the automatic procedures in Step c. If stops are manually exchanged between steps c and d, then step c must be repeated to get travel paths as well as more accurate estimates of deadheading.

The above procedure works regardless of whether a remnant route is formed. In neighborhood routing problems, remnant routes are useful for the following reasons:

i. They allow the above procedure to form partitions that are balanced and the workload in all partitions except for the partition that represents the remnant route are approximately the specified length of the workday.

ii. The remnant route can be strategically located so that the remnant routes from two or more adjacent areas of interest can be pieced together to form a complete route.

12.5 Challenges in Solving SRPs

Street Routing and Scheduling Problems are becoming an important application area. There are some documented successes and sophisticated commercial software systems have been developed. Powerful desktop computers and effective and user-friendly interfaces and GIS software offer the prospect of continued development in this area. Situations are occurring where street routing and scheduling problems are becoming an integral part of the logistics and distribution systems of many organizations.

SRP have been hindered by the lack of accurate geographic data bases. SRP place a demand on the accuracy on the digital street data bases not required by other mapping based applications nor by other classes of vehicle routing problems. The developers of geographic data bases have not played enough attention to form navigational quality digital street data bases (because they are expensive to create). Moreover, some of the GIS do not have enough functionality to allow for the editing of these digital street data bases.

However, the situation is improving. More accurate (and, in some cases, navigational quality) geographic data bases that can support street routing applications being developed commercially in the United States, Canada, Europe and elsewhere. The Windows based GIS contain increased functionality, making them more amenable to building street routing and scheduling systems.

New technologies in allied areas are also becoming commercially available. In particular, organizations are beginning to use Global Positioning Systems (GPS) and for real time vehicle location and tracking and enhanced paging systems for two way communication between the dispatcher and the crew. As such, we believe street routing and dispatch systems using on-board computing and GPS linked to a central processing unit for dispatching service stops to the vehicles will be implemented.

12.6 REFERENCES

Augerat, P., J. M. Belenguer, E. Benavent, A. Corber'an, D. Naddef, and G. Rinaldi. (1995). Computational results with a Branch and Cut Code for the Capacitated Vehicle Routing Problem. *Rapport de recherche 1 RR949-M, ARTEMIS-IMAG,* Grenoble France.

Baldacci R., M.Boschetti and A.Mingozzi. (1998). Heuristic Procedures for the Multi-Depot Vehicle Routing Problem with Pick-up and Delivery. *Working paper. Department of Mathematics, University of Bologna,* Italy.

Baldacci R., A.Mingozzi and E.Hadjiconstantinou (1998). Exact Algorithms for Routing Problems based on the Two-Commodity Flow Formulation. *Working paper. Department of Mathematics, University of Bologna,* Italy.

Balinski, M and R. Quandt (1964). On an Integer Program for a Delivery Problem. *Operations Research,* 12, 300-304.

Ball, M. (1988). *Allocation/routing: Models and Algorithms. In Vehicle Routing: Methods and Studies,* B.L. Golden and A. Assad (eds.), North-Holland, Amsterdam.

Ball, M. O. (1995), *Network Routing,* Elsevier.

Ball, M., T.L.Magnanti and G.L.Nemhauser (eds.) (1995). Network Routing. Volume 8 of *Handbooks in Operations Research and Management Science,* North-Holland.

Bodin, L.D., B. L. Golden, A. A. Assad, and M. O Ball (1983). Routing and Scheduling of Vehicles and Crews. The State of the Art. *Computers & Operations Research,* 10, 69-211.

Christofides, N. (1985). *Vehicle Routing. In The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization,* E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, (eds), 431-448. John Wiley & Sons Ltd., Chichester.

Christofides, N. and S. Eilon (1969). An Algorithm for the Vehicle Dispatching Problem. *Operations Research Quarterly,* 20, 309-318.

Christofides, N. and A. Mingozzi (1990). Vehicle Routing: Practical and Algorithm Aspects. In *Logistics: Where Ends Have to Meet,* C.F.H. van Rijn (editor), Pergamon Press.

Christofides, N., A. Mingozzi, and P. Toth (1979). The Vehicle Routing Problem. In *Combinatorial Optimization,* N. Christofides, A. Mingozzi, P. Toth and C. Sandi, (eds), J. Wiley, 315-338.

Christofides, N., A. Mingozzi, and P. Toth (1981a). Exact Algorithms for the Vehicle Routing Problem based on Spanning Tree and Shortest Path Relaxation. *Mathematical Programming,* 10, 255-280.

Christofides, N., A. Mingozzi, and P. Toth (1981b). State Space Relaxation Procedures for the Computation of Bounds to Routing Problems. *Networks,* 11, 145-164.

Christofides, N., A. Mingozzi, P. Toth, and C. Sandi (1979). *Combinatorial Optimization.* John Wiley & Sons, Chichester.

Chvatal, V. (1973). Edmons polytopes and weakly Hamiltonian graphs. *Mathematical Programming* 5, 29-40.

Cornuejols, G. and F. Harche. (1993). Polyhedral Study of the Capacitated Vehicle Routing. *Mathematical Programming,* 60, 21-52.

CPLEX Optimization Inc. 1993-1996. *Using the CPLEX Callable Library and CPLEX Mixed Integer Library.* 930 Tahoe Blvd #802-297, Incline Village, NV 89451, U. S. A.

Desrochers M. (1988). A Generalized Permanent Labelling Algorithm for the Shortest Path Problem with Time Windows, *INFOR,* 26, 191-211.

Desrochers M., J.Desrosiers and M.Solomon (1992). A New Optimization Algorithm for the Vehicle Routing Problem with Time Windows, *Operations Research,* 40, 342-354.

Desrosiers J., Y.Dumas, M.M.Solomon and F.Soumis (1995). Time Constrained Routing and Scheduling In *Handbooks in Operations Research and Management Science,* Vol.8: Network Routing, M.O.Ball, T.L.Magnanti, G.L.Nemhauser (eds.), North-Holland, 35-139.

Dror M. (1995). Note on the complexity of the shortest path models for column generation in VRPTW, *Operations Research,* 42, 977-978.

Finke, G., A. Claus, and E. Gunn. (1984). A Two-commodity Network Flow Approach to the Traveling Salesman Problem. *Congress. Numerantium,* 41, 167-178.

Fischetti M., P.Toth and D.Vigo (1994). A Branch-and-Bound Algorithm for the Capacitated Vehicle Routing Problem on Directed Graphs. *Operations Research,* 42, 846.

Fisher, M. L. and R. Jaikumar. (1981). A generalized assignment heuristic for vehicle routing. *Networks,* 11, 109-124.

Fisher, M. L. (1994a). Optimal solution of vehicle routing problems using minimum K-Trees. *Operations Research,* 42 , 626-642.

Fisher, M. L. (1994b). A Polynomial Algorithm for the Degree-constrained Minimum K-Tree Problem. *Operations Research,* 42 , 775-779.

Fisher, M. L. (1995). Vehicle routing. In *Handbooks in Operations Research and Management Science,* Vol 8: Network Routing M. O. Ball, T. L. Magnanti, C. L. Monma and G. L. Nemhauser (eds.), North-Holland, Amsterdam, 1-33.

Fisher M.L., K.O.Jörnsten, O.B.G.Madsen (1997), Vehicle Routing with Time Windows: Two Optimization Algorithms, *Operations Research,* 45, 488-492.

Golden, B. L. and A. Assad, eds. (1986). Special Issue on Time Windows. *American Journal of Mathematical and Management Analysis,* 6 (3 and 4), 251-399.

Golden, B. L. and A. Assad (1988). *Vehicle Routing: Methods and Studies.* North-Holland, Amsterdam.

Grötschel M. and M. W. Padberg (1979). On the Symmetric Traveling Salesman Problem: I and II. *Mathematical Programming,* 16, 265-280.

Grötschel, M. and M. W. Padberg (1985). Polyhedral theory, in: E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan and D. B. Shmoys, (eds.), *The Traveling Salesman Problem,* John Wiley & Sons, Chichester, 251-305.

Grötschel M. and M. W. Padberg (1985). Polyhedral Theory. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, editors, *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization,* 251-305. John Wiley & Sons Ltd., Chichester.

Hadjiconstantinou, E., N. Christofides, and A. Mingozzi (1995). A new exact algorithm for the vehicle routing problem based on q-paths and K-shortest paths relaxations. *Annals of Operations Research,* 61, 21-43.

Kohl N., J.Desrosiers, O.B.G.Madsen, M.M.Solomon and F.Soumis (1997) 2-paths cuts for the vehicle routing problem with time windows, *Les Cahiers du GERAD* G-97-19.

Kohl N. and O.B.G.Madsen (1997) An optimization algorithm for the vehicle routing problem with time windows based on Lagrangean relaxation, *Operations Research,* 45, 395-406.

Langevin, A., M. Descrochers, J.Desrosiers, E. Gelinas and F. Soumis (1993). A Two commodity Flow Formulation for the Traveling Salesman and the Makespan Problems with Time Windows. *Networks,* 23, 631-640.

Laporte, G. (1992). The Vehicle Routing Problem: An overview of exact and approximate algorithms. *European Journal of Operational Research,* 59, 345-358.

This page intentionally left blank

13 LONG-HAUL FREIGHT TRANSPORTATION

Teodor Gabriel Crainic

13.1 Introduction

Freight transportation is a vital component of the economy. It supports production, trade, and consumption activities by ensuring the efficient movement and timely availability of raw materials and finished goods. Transportation accounts for a significant part of the final cost of products and represents an important component of the national expenditures of any country (Crainic and Laporte 1997).

The freight transportation industry must achieve high performance levels in terms of economic efficiency and quality of service. The former, because a transportation firm must make a profit while evolving in an increasingly open, competitive, and still mainly cost-driven market. The latter, because transportation services must conform to the high standards imposed by the current paradigms of production and management such as small or no inventory associated with just-in-time procurement, production and distribution, on-time personalized services, and customer-driven quality control of the entire logistics chain. For the transportation firm, these standards concern particularly total delivery time and service reliability, which are often translated into objectives such as "be there fast but within the specified limits" or "offer high quality service and consistent performance".

The political evolution of the world impacts the transportation sector as well. The emergence of free trade zones together with the opening of new markets due to political changes and the resulting globalization of the economy have tremendous consequences for the evolution of transportation systems, not all of which are yet apparent or well understood. For example, open borders generally mean that firms are no longer under obligation to maintain a major distribution center in each country. In consequence, distribution systems are reorganized around fewer but bigger warehouses and transportation services are operated over longer distances. A significant increase in road traffic is a normal consequence of this process, as may be observed in Europe.

Changes to the regulatory environment have an equally powerful impact on the operation and competitive environment of transportation firms. The deregulation drive of the 1980s has seen governments remove numerous rules and restrictions, especially with regard to the entry of new firms in the market and the fixing of tariffs and routes. This resulted in a more competitive industry and in changes to the number and characteristics of transportation firms. At the same time, a number of new policies and regulations resulting from quality-of-life concerns start to significantly impact the operations of the freight transportation-related firms. Two major examples: (i) more stringent safety regulations; (ii) policies aimed towards increasing the volume of inter (and multi) modal freight movements while decreasing the utilization of trucks. The latter result from environmental and energy efficiency concerns and are particularly important in Europe. The evolution of technology is another major factor that modifies how freight transportation is organized and operated. This is not a new trend. Indeed, transportation has followed the industrial innovations and adapted, for example, to advances in traction technologies and fuels. What is new is that, arguably the major technological factor inflecting the evolution of transportation has to do with information and software rather than the traditional hardware. The tremendous expansion of Internet and the electronic-society, eloquently illustrated by the growing importance of electronic market places and business-to-business exchanges, dramatically alters the interactions of carriers and shippers. Intelligent Transportation Systems, on the other hand, both offer means to efficiently operate and raise new challenges, as illustrated by the evolution towards real-time modification to planned routes to account for changes in traffic conditions or new demands. More complex planning and operating procedures are a direct result of these new policies, requirements, technologies, and challenges.

Freight transportation must adapt to and perform within these rapidly changing political, social, and economic conditions and trends. In addition, freight transportation is in itself a complex domain: many different firms, organizations, and institutions, each with their own set of objectives and means, make up the industry; infrastructure and even service modifications are capital-intensive and usually require long implementation delays; important decision processes are often strongly interrelated. It is thus a domain where accurate and efficient methods and tools are required to assist and enhance the analysis, planning, operation, and control processes.

The focus of the chapter is on *long-haul (intercity) transportation,* that is, on transportation operations that are mainly concerned with the movement of goods over relatively long distances, between terminals or cities. Goods may be moved by rail, truck, ship, etc., or any combination of modes. The objective of the chapter is to present the main freight transportation planning and management issues, to briefly review the associated literature, to describe a number of major developments, and to identify trends and challenges. In order to keep the length of the chapter within reasonable limits, optimization-based operations research methodologies are privileged.

The chapter is organized as follows. Section 13.2 presents an overview of freight transportation systems and planning issues. Section 13.3 is dedicated to models which attempt to analyze multi-modal, multicommodity transportation systems at the regional, national or global level. Section 13.4 reviews network design formulations which are often associated with the long-term evolution of transportation infrastructures and services. These formulations also appear prominently when service design issues are considered as described in Section 13.5. Of the many operational issues related to the movement of freight, we focus on one of the most important in Section 13.6: the allocation and repositioning of resources, particularly empty vehicles. To conclude, Section 13.7 attempts to identify a number of interesting problems and methodological challenges.

13.2 Freight Transportation Systems

Demand for freight transportation derives from the interplay between *producers* and *consumers* and the significant distances that usually separate them. Producers of goods require transportation services to move raw materials and intermediate products, and to distribute final goods in order to meet demands. *Carriers* supply transportation services. Railways, shipping lines, trucking companies, and intermodal container and postal services are examples of carriers. Considering the type of service they provide, ports, intermodal platforms, and other such facilities may be described as carriers as well. *Shippers,* which may be producers of goods or some intermediary firm (*brokers*), attribute demand to supply. Governments contribute the infrastructure: roads and highways, as well as a significant portion of ports, internal navigation, and rail facilities. Governments also regulate (e.g., dangerous and toxic goods transportation) and tax the industry.

When examining freight transportation, one often distinguishes between producers that own or operate their own transportation fleet (which then become carriers for their own freight), and "for hire" carriers, which perform transportation services for various shippers. From a planning and operations point of view, a more interesting and practical classification differentiates between: (1) Long-haul transportation (this chapter) and vehicle routing and distribution, *VRP,* problems (Golden and Assad 1988, Ball *et al.* 1995, Dror 2000, Toth and Vigo 2002, Chapter 12, etc.); (2) The multi-modal transportation system of a region, irrespective of its dimensions (Section 13.3), and the transportation services of a particular carrier (Sections 13.5 and 13.6); (3) *Consolidation* transportation where one vehicle or convoy may serve to move freight for different customers with possibly different initial origins and final destinations, and *door-to-door* transportation operations *customized* for a particular customer.

Most freight transportation planning issues exhibit a *multicommodity* nature. In most cases, several distinct commodities must be moved. Even when the transportation system or study is dedicated to one commodity only, the traffic between different origin and destination points must be individually accounted for. Most of the time, both conditions must be satisfied simultaneously.

Customized Transportation

Truckload trucking offers a typical example of door-to-door long distance transportation. In this mode, a vehicle – truck – is usually dedicated to each customer. When the customer calls, a truck with a driver or driving team is assigned to it. The truck is moved to the customer-designated location, and it is loaded. It then moves to the specified destination; this is the long-haul transportation operation. At destination, the truck is unloaded, and the driver calls the dispatcher to give its position and request a new assignment. The dispatcher may indicate a new load, ask the driver to move empty to a new location where demand should appear in the near future, or have the driver wait and call later.

The truckload carrier thus evolves in a highly dynamic environment, where little is known with certainty regarding future demands, travel times, waiting delays at customer locations, precise positions of loaded and empty vehicles at later moments in time, and so on. Service is tailored for each customer and the timely assignment of vehicles to profitable demands is of the outmost importance.

The development of efficient *resource management and allocation* strategies are therefore at the heart of the management process. These strategies attempt to maximize the volume of demand satisfied (loads moved) and the associated profits, while making the best use of the available resources: drivers, tractor and trailer fleets, etc. Navigation services ensured by for-hire ships share some of these dynamic and stochastic characteristics.

Consolidation Transportation

When demands of several customers are served simultaneously by using the same vehicle or convoy, one cannot tailor services individually for each customer.

Carriers must establish regular services (e.g., a container ship from Seattle to Singapore) and adjust their characteristics (route, intermediary stops, frequency, vehicle and convoy type, capacity, speed, etc.) to satisfy the expectations of the largest number of customers possible. Externally, the carrier then proposes a series of *routes,* or *services,* each with its operational characteristics. Services are often grouped in a *schedule* that indicates departure and arrival times at the stops of the route. Internally, the carrier builds a series of rules and policies that affect the whole system and are often collected in an *operational plan* (also referred to as *load* or *transportation* plan). The aim is to ensure that the proposed services are performed as stated (or as closely as possible), while operating in a rational, efficient, and profitable way. The presence of terminals where cargo and vehicles are consolidated, grouped, or simply moved from one service to another characterizes this type of transportation performed by Less-Than-Truckload (LTL) motor carriers, railways, shipping lines, postal and express shipment services, etc. Freight transportation in some countries where a central authority more or less controls a large part of the transportation system also belongs to this category. We include all these systems under *service* or *consolidation* transportation.

The underlying structure of a large consolidation transportation system consists of a rather complex network of terminals connected by physical or conceptual links. Air and sea lines correspond to the latter, while road, highways, and rail tracks are typical examples of the former. The network may belong entirely or partially to the carrier. Rail transportation belongs traditionally to the first category, while LTL motor carriers exemplify the second: LTL carriers generally own the terminals but operate on public roads. It is noteworthy that the current policy of the European Union to separate the infrastructure ownership and the service provider (the carrier) operations is moving rail transportation in Europe towards a more LTL-like mode of operations. Some carriers prefer not to own any infrastructure, however, and only rent space as needed. Intermodal container carriers generally belong to this category, their terminal operations being often organized in ports and railway yards.

Freight demand is defined between given points of this network. Other than its specific origin, destination, and commodity-related physical characteristics, such as weight and volume, each individual shipment may present any number of particular service requirements in terms of delivery conditions, type of vehicle, and so on. A profit or cost also usually accompanies a specific demand. The consolidation carrier moves the freight by services performed by a large number of vehicles: rail cars, trailers, containers, ships, etc. Vehicles move, usually on specified routes and sometimes following a given schedule, either individually or grouped in convoys such as rail or barge trains, or multi-trailer assemblies. Convoys are formed and dismantled in terminals. Other major terminal operations include freight sorting and consolidation, its loading into or unloading from vehicles, as well as vehicle sorting, grouping, and transferring from one convoy to another. Terminals come in several designs and sizes and may be specialized in certain operations or the handling of particular products, or offer a complete set of services. In all cases, terminal operations are vital to the performance of a consolidation transportation system.

Figure 13.1 illustrates the network of a consolidation transportation system. Nodes A, B, and C represent major consolidation centers, also referred to as *hubs,* linked by high frequency and capacity services. Nodes 1–9 stand for the origin and destination terminals where freight and vehicles are consolidated at the beginning and end of the journey, and which are linked to hubs by feeder services. The figure also emphasizes the possibility for a terminal to be linked to more than one hub and illustrates the local pickup and delivery operations usually associated to terminals. Such an organization allows a much higher frequency and quality of service among hubs and a more efficient utilization of resources. The drawback is the increased delays – longer routes and more time spent in terminals – experienced by passengers or goods. This explains partly why there is hardly any "pure" hub-and-spoke

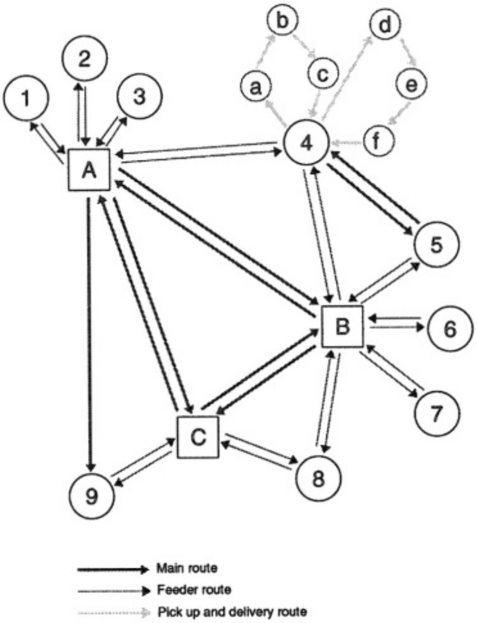


Figure 13.1 Network Representation of Consolidation Transportation

systems in operation, direct transportation being organized for high demand or high priority origin-destination pairs. The links between terminals 4 and 5, and from hub A to terminal 9 in Figure 13.1 illustrate this option. Note that smaller firms may take advantage of consolidation systems and identify profitable niches by offering direct services to markets that large firms serve through hubs.

To further clarify these notions, consider the case of railway transportation that operates networks made up of single or double track lines that link many large and small classification yards, in which rail cars are grouped and trains are formed, pickup and delivery stations, junction points, etc. (Assad 1980, Cordeau, Toth, and Vigo 1998). Here, everything begins when a customer issues an order for a number of empty cars or, alternatively, when freight is brought into the loading facility following a pickup operation. At the appropriate yard, rail cars are selected, inspected, and then delivered to the loading point. Once loaded, cars are moved to the origin yard (possibly the same) where they are sorted, or *classified,* and assembled into *blocks.* A block is a group of cars, with possibly different final destinations, arbitrarily considered as a single unit for handling purposes from the yard where it is made up to its destination yard where its component cars are separated. Rail companies use blocks to take advantage of some of the economies of scale related to full train loads and the handling of longer car strings in yards. The block is eventually put on a train and this signals the beginning of the journey. During the long-haul part of this journey, the train may overtake other trains or be overtaken by trains with different speeds and priorities. When the train travels on single-track lines, it may also meet trains traveling in the opposite direction. Then, the train with the lowest priority has to give way and wait on a side track for the train with the higher priority to pass by. At yards where the train stops, cars and engines are regularly inspected. Also, blocks of cars may be *transferred,* i.e., taken off one train and put on another. When a block finally arrives at destination, it is separated from the train, its cars are sorted, and those having reached their final destination are directed to the unloading station. Once empty, the cars are prepared for a new assignment, which may be either a loaded trip or an empty repositioning movement.

One source of complication in rail freight transportation is the complex nature of the main yard activities: the classification of cars and the composition of trains. The modeling of yard operations as well as that of their interactions with the rest of the system are critical components of any comprehensive rail model. It is interesting to note that, traditionally, in most rail systems cars spend most of their lifetime in yards: being loaded and unloaded, being classified, waiting for an operation to be performed or for a train to come, or simply sitting idle on a side track. Also of interest is the fact that most rail companies have separated the operations and yards dedicated to intermodal services from those used for their regular services in an attempt to cut delays, especially those associated with yard operations, and improve the quality of this time-sensitive and highly competitive service.

Similarly to rail transportation, LTL networks may encompass different types of terminals. Local traffic is picked up by "small" trucks and is delivered to *end-of-line* terminals where it is consolidated into larger shipments before long-haul movements. Symmetrically, loads from other parts of the network arrive at end-of-lines to be unloaded and moved into delivery trucks for final delivery. *Breakbulks* are terminals where traffic from many end-of-line terminals is unloaded, sorted, and consolidated for the next portion of the journey. Breakbulks are the hubs of LTL networks, as major yards are the hubs of rail transportation systems In Figure 13.1, nodes 1–9 represent end-of-lines, while nodes A, B, and C stand for breakbulk terminals.

LTL motor carrier transportation follows the same basic operational structure described for rail but on a simpler scale and with significantly more flexibility due to the fundamental difference in infrastructure: While rail transportation is "captive", trucks may use any of the existing links of the road and highway network as long as they comply with the weight regulations. Furthermore, a truck is only formed of a tractor and one or several trailers (when more than one trailer is used, these are smaller and are called "pups"). Consequently, terminal operations are generally simpler; freight is handled to consolidate outbound movements but there are no significant convoy-related operations. LTL transportation may become rather complex, however, as soon as one considers the option to use rail (the *trailer-on-flat-car – TOFC* – option) for long distances.

It is interesting to note that intermodal container transportation may be viewed as either door-to-door or consolidation transportation. For the customer, it is door-to-door transportation. On request, containers are delivered, loaded, moved through a series of terminals and vehicles (of which the customer has little knowledge even when the exact position of the shipment is available), and are delivered to the final destination where the goods are unloaded. For the shipping company, it is a consolidation transportation system. Containers from many customers must be moved to a port by truck, barge, or rail, or a combination thereof. There, containers are grouped and loaded on a ship that navigates a well-established route, according to a tight schedule, and delivers the containers at the destination port. From there, a land transportation system delivers the containers to the final destination by using a variety of modes and terminals. Container transportation systems that operate exclusively on land may also be encountered. In this case, rail trains and inland terminals usually play the role of ships and ports. The continuous increase in the size of container ships operated on international lines exacerbates the consolidation characteristics of intermodal container transportation systems. Indeed, the huge size of the newest generation of container ships forbids them from entering many ports and makes routes with many stops uneconomical. Consequently, long-course ships stop only at a selected number of important ports – the hubs –, while smaller vessels and land transportation modes ensure delivery of containers to the other ports and final destinations.

A similar argument may be made for express letter and small package services. For customers, it is obviously a door-to-door, high quality and reliable service. For the company, it is a consolidation transportation system that usually makes use of various air, truck, and rail services. The company implements a VRP-type of service to interact with its customers and collect and distribute letters and packages. The collection and distribution centers where mail is sorted and consolidated play a role similar to that of end-of-line terminals in LTL transportation. To reach its destination, a letter or package usually passes through at least one major hub. These terminals do for express mail services what breakbulks do for LTL motor carriers. To link its national hubs and major collection and distribution centers, the company may operate its own planes, as well as use scheduled passenger flights or train services. When distances are moderate, trucks may be used as well.

Empty Flows

A constant characteristic of any freight transportation system is the need to move empty vehicles. This follows from the imbalances that exist in trade flows and that result in discrepancies between vehicle supply and demand in various zones or terminals of the system.

To correct these differences, vehicles must be moved, *repositioned,* in order to have them available to satisfy the demand of the next period. Some repositioning decisions are straightforward. When, for example, unit trains are used to move coal or iron ore from mining fields to the port on the only rail line linking the two, cars, once unloaded, are simply formed into a return train. In most cases, however, the decision of *how many* and *where* to send vehicles appears much more complicated. The alternatives are many, due to the numerous possibilities for movement and the uncertainty of future supply and demand. The search for the most economic *empty repositioning* or *empty balancing* strategy is thus a significant problem in itself, and we will find the preoccupation with these issues in many of the problems and models addressed in the following sections.

Service Schedules

Another notion often encountered in transportation planning has to do with *schedules* and *scheduled services.*

In the general sense, a schedule specifies timing information for each possible occurrence of a service during a given time period: departure time at the origin, arrival/departure time information at intermediary stops, and arrival time at the final destination. The schedule may also include indications on the *cut-off* time: the latest moment freight may be given to the carrier and still meet the scheduled departure of the service. Schedules are omnipresent in passenger transportation by air, rail, bus or ship and are strictly enforced (most of the time). The case is less clear for freight transportation. On the one hand, there are no schedules in door-to-door transportation, except for cut-off times. At the other end of the spectrum, regular navigation lines usually operate according to strict schedules (high port utilization fees constitute an important incentive to follow the schedules). Much air cargo is moved on passenger planes and therefore follows strict schedules. All-cargo air services are also usually operated according to well-established schedules.

LTL trucking follows much less stringent rules. Many carriers operate on a "go when full" policy. Alternatively, earliest and latest departures may be planned, as well as the distribution of departures during the evening, which usually is the busiest period. The goal of this process is to offer customers late cut-off times and to ensure that trucks arrive at destination terminals within certain limits – at the opening of business in the morning, for example. The focus on increased customer service and tighter operations (including crew schedules) is increasing the utilization of scheduled services, however. Actually, schedules are build for part of the traffic only, representing the regular part of operations. Departures may then be added or cancelled to adjust for each day's particular conditions. In all cases, the dispatcher is responsible for orchestrating the operations, as well as for avoiding empty movements.

The tradition in most rail systems around the world was to follow some variant of the "go when full" rule. Even when schedules were prepared, they were mostly indicative of the ideal departure times and served as a basis for various dispatching rules for yard masters (e.g., "a train may leave one hour before planned departure if full and conditions down the line are appropriate"). The high volume of passenger trains already in the system, as well as the desire to decrease total transit time and improve connections, has pushed European rail companies toward more stringent schedules for their freight trains. Some companies operate according to fixed schedules and bookings similar to the ones used for passenger transportation. In recent years, North American companies have also migrated toward scheduled service operations (at least for part of their traffic) with various degrees of rapidity and success. The issues are different for overloaded systems, such as the Indian and Chinese railways, where the demand for passenger and freight transportation significantly exceeds the capacity of the system. In such environments, the emphasis is less on "scheduling" and more on managing the train and line operations to operate freight trains in between the passenger traffic.

Planning Levels

Transportation systems thus appear as rather complex organizations that involve a great deal of human and material resources and that exhibit intricate relationships and tradeoffs among the various decisions and management policies affecting their different components. It is convenient to classify these policies according to the following three *planning levels:*

1. *Strategic* (long-term) planning at the firm level typically involves the highest level of management and requires large capital investments over long-term horizons. Strategic decisions determine general development policies and broadly shape the operating strategies of the system. These include the design of the physical network and its evolution, the location of major facilities (e.g., terminals), the acquisition of major resources such as motive power units, and the definition of broad service and tariff policies.

   Strategic planning also takes place at the international, national and regional levels, where the transportation networks or services of several carriers are simultaneously considered. National or regional transportation departments, consultants, international shippers and forwarders, for example, engage in this type of activity. Sections 13.3 and 13.4 present models aimed at strategic issues at the system and firm levels, respectively.
2. *Tactical* (medium-term) planning aims to determine, over a medium-term horizon, an efficient allocation and utilization of resources to achieve the best possible performance of the whole system. Typical tactical decisions concern the *design of the service network* and may include issues related to the determination of the routes and types of service to operate, service schedules, vehicle and traffic routing, repositioning of the fleet for use in the next planning period. Tactical planning models are the object of Section 13.5.
3. *Operational* (short-term) planning is performed by local management, yard masters and dispatchers, for example, in a highly dynamic environment where the time factor plays an important role and detailed representations of vehicles, facilities and activities are essential. Important operational decisions concern: the implementation and adjustment of schedules for services, crews, and maintenance activities; the routing and dispatching of vehicles and crews; the dynamic allocation of scarce resources. Section 13.6 addresses operational planning issues.

This classification highlights how data flows among decision-making levels and how policy guidelines are set. The strategic level sets the general policies and guidelines for decisions taken at the tactical level, which determines goals, rules and limits for operational and real-time decisions. The data flow follows the reverse route, each level of planning supplying information essential for the decision making process at a higher level. This hierarchical relationship emphasizes the differences in scope, data, and complexity among the various planning issues, prevents the formulation of a unique model for the planning of freight transportation systems, and calls for different model formulations that address specific problems at particular levels of decision making.

13.3 Strategic System Analysis and Planning

The focus of the models and methods presented in this section is broad: strategic planning issues at the international, national and regional level, where the movements of several commodities through the transportation networks and services of several carriers are considered simultaneously. The main questions address the evolution of a given transportation system and its response to various modifications in its environment: changes to existing infrastructure, construction of new facilities, evolution of the "local" or international socio-economic environment resulting in modifications to the patterns and volumes of production, consumption, and trade, variations in energy prices, changes to labor conditions, new environment-motivated policies and legislation, carrier mergers, introduction of new technologies, and so on and so forth. These questions are often part of cost-benefit analyses and comparative studies of investment alternatives – especially when the available monetary resources are limited – and are asked by regional or national planning agencies and regulatory authorities, as well as international financial institutions such as the World Bank. Private firms are also interested in these questions, particularly companies involved in the financing of transportation infrastructures, or firms that plan and operate the distribution of goods using several transportation modes.

The prediction of multicommodity freight flows over a multi-modal network is an important component of transportation science and has attracted significant interest in recent years. One notes, however, that, perhaps due to the inherent difficulties and complexities of such problems, the study of freight flows at the national or regional level has not yet achieved full maturity, in contrast to passenger transportation where the prediction of car and transit flows over multi-modal networks has been studied extensively and several of the research results have been transferred to practice (Florian and Hearn 1995, Cascetta 2001; see also Chapter 11).

A "complete" strategic planning tool identifies and represents the fundamental components of a transportation system – demand, supply, performance measures and decision criteria – and their interactions. It yields product flow volumes and associated performance measures defined on a network representation of the transportation system. It aims to achieve a sufficiently good simulation of the global behaviour of the system to both offer a correct representation of the current situation and serve as an adequate analysis tool for planned or forecast scenarios and policies. It has to be tractable and produce results that are easily accessible. This constitutes an extremely broad scope and it is thus unrealistic to believe that a single formulation, mathematical or otherwise, or a single procedure may encompass all relevant elements, address all important issues, and fulfill all goals. Consequently, a strategic planning tool appears as a set of models and procedures. Other than data manipulation (e.g., collection, fusion, updating, validation, etc.) and result analysis (e.g., cost-benefit, environmental impacts, energy consumption policies, etc.) tools, the main components are: (i) *Supply modeling* representing the transportation modes, infrastructure, carriers, services, and lines; vehicles and convoys; terminals and inter-modal facilities; capacities and congestion; economic, service, and performance measures and criteria, (ii) *Demand modeling* that captures the product definitions, identifies producers, shippers, and intermediaries and represents production, consumption, and zone-to-zone (region-to-region) distribution volumes, as well as mode choices; Relations of demand and mode choice to the performance of economic policies and transportation system performance are also addressed here, (iii) *Assignment* of multi-product flows (from the demand model) to the multi-mode network (the supply representation). This procedure simulates the behaviour of the transportation system and its output forms the basis for the analyses that conduct to the specification of the strategic plan. Therefore, it has to be both precise in reproducing current situation and general to produce robust analyses of future scenarios based on forecast data.

A complete survey of demand and mode choice estimation methodologies is beyond the scope of this chapter. In the following, we only cite some of the most frequently used methodologies and associated references.

The modeling of demand corresponds to an image of the economic activities of a country, production, consumption, import and export of goods. For planning purposes, its output is a series of product (or commodity group) specific demand matrices indicating the volumes to be moved from one region or zone to another. It is often completed by the modeling of mode choice, which specifies for each product and origin-destination combination on what transportation infrastructure or services the demand may be moved.

A number of countries have developed *input/output* models of their economy that serve to determine the basic production and attraction of goods (Isard 1951;

Casceta 2001 and references within). In order to use an input/output model, it is necessary to disaggregate the model inputs and outputs by region and then further disaggregate them by the zonal subdivision of the national planning model. This process is complex and is usually done in an analysis and computing environment which is not necessarily integrated with that used for the supply representation and the computation of flows by product. When an input/output model is not available, the initial determination of origin-destination matrices is carried out by using national statistics on production, consumption, imports and exports combined with sectorial surveys designed to complete missing or unreliable information. This process may be tedious since one has to reconcile data from several sources that may be collected by using different geographical subdivisions or inconsistent product definitions. The results of the disaggregated input/output model or the ad-hoc estimation procedures serve for the initial computation of origin-destination matrices for each product but without a subdivision by mode.

A second class of models that is well studied for the prediction of interregional commodity flows is the *spatial price equilibrium model* and its variants (Friesz, Tobin, and Harker 1983, Harker and Friesz 1986a,b, and Harker 1987; see also Florian and Hearn 1995, or Nagurney 1993). This class of models determines simultaneously the flows between producing and consuming regions, as well as the selling and buying prices that satisfy the spatial equilibrium conditions. Simply stated, a spatial equilibrium is reached provided that for all pairs of supply and demand regions with a positive commodity flow, the unit supply price plus the unit transportation cost is equal to the unit demand price; the sum is larger than this price for all pairs of regions with no exchanges. The transportation network used in these models is usually represented in a simplistic way (bipartite networks). These models rely to a large extent on the supply and demand functions of producers and consumers, respectively, which are rarely available and quite difficult to calibrate. There are relatively few applications of this class of models for the determination of demand by product. The few applications reported in the literature deal with specific products which have a particular importance, such as crude oil, coal or milk products.

The mode choice definition may be rather general, e.g., petroleum moves by ship and pipeline or, alternatively, extremely specific indicating the particular multi-modal path for a given product, shipper, and origin-destination pair, or anywhere in between. The level of detail of modal specification needs not to be the same for all products or inter-zonal trade flows. The specification of the mode choice for a given product may be inferred from historical data and shipper surveys or it may result from a formal description and modeling effort (Winston 1983). *Random utility models,* developed and largely used for the analysis and planning of person transportation systems, have been proposed for freight transportation as well (Cascetta 2001) but their use in actual applications is scarce. The huge number of paths that have to be explicitly generated and stored, coupled to the challenge to perform this task for forecast data, may explain this phenomenon. At aggregated levels, mode choices have been specified for particularly important product flows by explicitly surveying the major logistic chains used between pairs of macro regions.

Once modal origin-destination matrices have been developed by some means, the next step is to assign them to the network (supply) model by using some route choice mechanism. The results of such an assignment model – product flows and performance measures – form part of the input to demand and cost-benefice modeling and analysis. The actual assignment mechanism may be based on further application of random utility models to the choice of pre-defined paths over a multi-modal network or on network optimization models. It is noteworthy that the attributes of pre-defined paths are determined by the state of the network at generation time and are not responsive to assignment results. Thus, for example, congestion conditions are very difficult to represent. Moreover, the utility and choice models have to be calibrated, and all paths have to be generated, for each scenario, which is quite difficult to perform when forecast data is used.

Another class, *network optimization models,* is generally considered to be more appropriate for the type of planning issues considered here. These formulations enable the prediction of multicommodity flows over a multi-modal network that represents the transportation facilities at a level of detail appropriate for a nation or region but with relatively little abstraction. The demand for transportation services is exogenous and may originate from an input-output or spatial equilibrium model, if one is available, or from other sources, such as observed demand or scaling of past observed demand. The choice of mode or subsets of modes used are exogenous and intermodal shipments are permitted. Within the specified mode choice, the optimization (assignment) engine determines the best (with respect to the specified network performance measures) multi-modal paths for each product and origin-destination pair. In this sense, these models may be integrated with econometric demand models as well. The emphasis is on the proper representation of the network and its several transportation modes, the corresponding intermodal transfer operations, the various criteria used to determine the movement of freight, the interactions and competition for limited resources captured through the representation of congestion effects, and the associated estimation of the traffic distribution over the transportation system considered to be used for comparative studies or for discrete time multi-period analyses.

Studies in the 1970's used rather simple network representations. Guélat, Florian, and Crainic (1990) and Crainic *et al.* (1990) review and discuss these efforts. Several studies also attempted to extend spatial equilibrium models to include more refined network representations and to consider congestion effects and shipper-carrier interactions, Friesz, Gottfried, and Morlok (1986) present a sequential model which uses two network representations: detailed separate networks for each carrier, and an aggregate, shipper-perceived network. On each carrier network commodities are transported at the least total cost. On the shipper-perceived network, traffic equilibrium principles are used to determine the carriers that shippers choose to move their traffic. This approach has proven quite successful in the study of logistics of products where a very limited number of shippers and carriers interact and strongly determine the behavior of the system. A typical example is the coal market between the electric utilities in the United States and their suppliers in exporting countries. Friesz and Harker (1985), Harker and Friesz (1986), Harker (1987, 1988), and Hurley and Petersen (1994) present more elaborate formulations. This line of research has not, however, yet yielded practical planning models and tools, mainly because the formulations become too large and complex when applied to realistic situations.

The modeling framework we present is based on the work of Guélat, Florian, and Crainic (1990). The formulation does not consider shippers and carriers as distinct actors in the decisions made in shipping freight. The level of aggregation appropriate for strategic planning of freight flows results in origins and destinations that correspond to relatively large geographical areas and leads to the specification of supply and demand representing, for each of the products considered, the total volumes generated by all the individual shippers. Furthermore, demand for strategic freight studies are often determined from data sources (national freight flow statistics, economic input/output models) which enable the identification of the mode used, but do not contain information on individual shippers. It is thus assumed that the shipper's behavior is reflected in the origin to destination product matrices and in the specification of the corresponding mode choice.

The modeling framework is that of a multi-modal network, made up of modes, nodes, links, and intermodal transfers, on which multiple products are to be moved by specific vehicles and convoys between given origin and destination points. Here, a mode is a means of transportation having its own characteristics, such as vehicle type and capacity, as well as specific cost measures. Depending on the scope and level of detail of the strategic study, a mode may represent a carrier or part of its network representing a particular transportation service, an aggregation of several carrier networks, or specific transportation infrastructures such as highway networks or ports.

The network consists of nodes $\mathcal{N}$, links $\mathcal{A}$, modes $\mathcal{M}$, and transfers $\mathcal{T}$ that represent all possible physical movements on the available infrastructure. To capture the modal characteristics of transportation, a link $a \in \mathcal{A}$ is defined as a triplet (i,m,j), where $i \in \mathcal{N}$is the origin node, $j \in \mathcal{N}$ is the destination node, and $m \in \mathcal{M}$ is the mode allowed on the arc. Parallel links are used to represent situations where more than one mode is available for transporting goods between two adjacent nodes. This network representation is compact and enables easy identification of the flow of goods by mode, as well as various cost functions (e.g., operating cost, time delay, energy consumption, emissions, noise, risk, etc.) by product and mode. Furthermore, the network model resembles the physical network, since, for example, the rail and road infrastructures are physically different. Also, when there are two different types of services on a physical link, such as diesel and electric train services on rail lines, a separate link may be assigned to each service to capture the fact that they have different cost and delay functions. To model intermodal shipments, one must allow for mode transfers at certain nodes of the network and compute the associated costs and delays. Intermodal transfers t at a node of the network are modeled as link to link, hence mode to mode, allowed movements. A path in this network then consists of a sequence of directed links of a mode, a possible transfer to another mode, a sequence of directed links of the second mode, and so on. A transfer thus belongs to path if the two arcs that define it belong to the path. This representation allows for the restriction of flows of certain commodities to subsets of modes (e.g., iron ore may be shipped only by rail and ship) to capture the restrictions that occur in the operation of freight networks and transshipment facilities.

A product is any commodity (or collection of similar products) – goods or passengers – that generates a link flow. Each product $p \in \mathcal{P}$ transported over the multi-modal network is shipped from certain origins $o \in \mathcal{N}$ to certain destinations $d \in \mathcal{N}$ within the network. The demand for each product for all origin-destination pairs is exogenous and is specified by a set of O-D matrices. The mode choice for each product is also exogenous and is indicated by defining for each O-D matrix a subset of modes allowed for transporting the corresponding demand. For example, one may indicate that the traffic out of certain regions must use rail, while in other regions there is a choice between rail and barges. This allows to capture the mode restrictions that occur in the operation of freight networks and transshipment facilities. Let $g^{m(p)}$ be a demand matrix associated with product $p \in \mathcal{P}$, where $m(p) \subseteq \mathcal{M}$ is the subset of modes that may be used to move this particular part of product p.

The flows of product $p \in \mathcal{P}$ on the multi-modal network are the decision variables of the model. Rows on links $a \in \mathcal{A}$ are denoted by v_a^p and flows on transfers $t \in \mathcal{T}$ are denoted by v_t^p; v stands for the vector of all product flows. Cost functions are associated with the links and transfers of the network. For product p, the respective average cost functions $s_a^p(v)$ and $s_t^p(v)$ depend on the transported volume of goods. Then, the total cost of product p on arc a is $s_a^p(v)v_a^p$, and it is $s_t^p(v)v_t^p$ on transfer t. The total cost over the multi-modal network is the function F, which is to be minimized over the set of flow volumes that satisfy the flow conservation and nonnegativity constraints:

$$F = \sum_{p \in P} \left(\sum_{a \in \mathcal{A}} s_a^p(v) v_a^p + \sum_{t \in \mathcal{T}} s_t^p(v) v_t^p \right). \tag{13.1}$$

Let $\mathcal{L}_{od}^{m(p)}$ denote the set of paths that for product p lead from origin o to destination d using only modes in $m(p)$. The path formulation of the flow conservation equations are then:

$$\sum_{l \in \mathcal{L}_{od}^{m(p)}} h_l = g_{od}^{m(p)} \quad o, d \in \mathcal{N}, \ p \in \mathcal{P}, \ m(p) \subseteq \mathcal{M}, \tag{13.2}$$

where h_l is the flow on path $l \in \mathcal{L}_{od}^{m(p)}$. These constraints specify that the total flow moved over all the paths that may be used to transport product p must be equal to the demand for that product. The nonnegativity constraints are:

$$h_l \geq 0, \quad l \in \mathcal{L}_{od}^{m(p)}, \ o, d \in \mathcal{N}, \ p \in \mathcal{P}, \ m(p) \subseteq \mathcal{M}. \tag{13.3}$$

The relation between arc flows and path flows is $v_a^p = \sum_{l \in \mathcal{L}^p} \delta_{al} h_l$, $a \in \mathcal{A}$, $p \in \mathcal{P}$, where $\mathcal{L}^p$ is the set of all paths that may be used by product p, and $\delta_{al} = 1$ if $a \in l$ (and 0, otherwise) is the indicator function which identifies the arcs of a particular path. Similarly, the flows on transfers are $v_t^p = \sum_{l \in \mathcal{L}^p} \delta_{tl} h_l$, $t \in \mathcal{T}, p \in \mathcal{P}$, where $\delta_{al} = 1$ if $t \in l$ (and 0, otherwise). Then, the system optimal multi-product, multi-modal assignment model consists of minimizing (13.1), subject to constraints (13.2) and (13.3). The optimality principle ensures that in the final flow distribution, for each product, demand matrix, and origin-destination pair, all paths with positive flows will have the same marginal cost (lower than on the other paths). The algorithm developed for this problem exploits the natural decomposition by product and results in a Gauss-Seidel-like procedure which allows the solution of large size problems in reasonable computational times (Guélat, Florian, and Crainic 1990).

This network model allows for a detailed representation of the transportation infrastructure, facilities and services, as well as the simultaneous assignment of multiple products on multiple modes. Vehicle and convoy traffic on the links (and transfers) of the network is deduced from the assigned product flows and is used to evaluate congestion conditions and to compute costs. Capacities are considered through congestion or penalty functions. Thus, the model captures the competition of products for the service capacity available, a feature of particular relevance when alternative scenarios of network capacity expansion are considered. It allows for the specification and combination of a wide variety of performance measures and assignment criteria, including user-optimum type of functions when the nature of a particular product requires it. Furthermore, the model is sufficiently flexible to represent the transport infrastructure of one carrier only.

This model and algorithm are embedded in the STAN interactive-graphic system where they are complemented by a large number of tools to input, display, analyze, modify, and output data; specify the network and assignment models; analyze flows, costs, and commodity routings and paths. Matrix-based computing tools may be used to implement a whole gamut of mode choice and demand models. A network calculator can be used to combine network data to implement various performance and analysis models. A path analysis capability allows the visualisation and handling of paths used in assignment and the construction of demand and network performance models based on paths. The data required by the STAN system is organized into a strictly structured data bank. A macro language can be used to program complex operations and procedures. See Larin *et al.* (2000) for a detailed description of the STAN system, components, interfaces, and tools. The STAN system has been applied successfully in practice for scenario analysis and planning, and several agencies and organizations in a number of countries around the world use it. Crainic, Florian, and Léal (1990) present the application of this methodology to the study of freight rail transportation, while several other applications are discussed in Guélat, Florian, and Crainic (1990), Crainic *et al.* (1990), Crainic, Florian, and Larin (1994), Crainic *et al.* (1998, 2002).

13.4 Logistics Network Design

For freight carriers, strategic decisions determine general development policies and broadly shape the operating strategies of the system over relatively long-term horizons. Several such decisions affect the design of the physical infrastructure network: where to locate facilities such as loading and unloading terminals, consolidation centers, rail yards, or intermodal platforms; what type of equipment to install in each facility; on which lines to add capacity; what type of lines or capacity to add; what lines or facilities to abandon; and so on. These issues, which may be collectively identified as *logistics system design,* are the subject of this section.

Logistics system design issues are often addressed by evaluating alternatives using network models for the tactical (Section 13.5) or operational (Section 13.6) planning of transportation activities. When formal models are proposed, these generally appear either as *location* or *network design* formulations. An extensive literature exists on both subjects, addressing the analysis of formulations, the development of algorithms, and the performance of applications for a broad range of problems and issues. Location models are the object of Chapter 10, as well as of Mirchandani and Francis (1990), Daskin (1995), Drezner (1995), and Labbé, Peeters, and Thisse (1995). Labbé and Louveaux (1997) present an annotated bibliography concerning discrete location problems.

In the following, we focus on network design. We give a number of main references and present a general formulation together with a few extensions that may be used in freight transportation planning. For more details, the interested reader should consult the surveys by Magnanti and Wong (1984) and Minoux (1986), the discussions in Ahuja *et al.* (1995), Nemhauser and Wolsey (1988), and Salkin and Mathur (1989), and the annotated bibliography of Balakrishnan, Magnanti, and Mirchandani (1997). Survivability and connectivity issues are particularly important for telecommunication systems and the electronics industry, but may also appear prominently in the transportation industry when service must be ensured to certain regions or between particular zones. Grötschel, Monma, and Stoer (1995) survey the models and solution methods developed for this class of problems. An annotated bibliography may be found in Raghavan and Magnanti (1997).

Network Design

Network design models are extensively used to represent a wide range of planning and operation management issues in transportation, telecommunications, logistics, and production-distribution. These formulations play a particularly important role in decisions concerning the logistics structure, the service network (Section 13.5), and the operations (Section 13.6) of long distance freight transportation systems.

Network design models are defined in terms of a network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ represents the set of *nodes* or *vertices*. Demand for transportation exists at some of these nodes. The set of *arcs* or *links* $\mathcal{A} = \{a = (i,j) | \, i,j \in \mathcal{N}, i \neq j\}$ includes all the possible ways to move directly (no intermediate nodes) between two nodes in $\mathcal{N}$. The set $\mathcal{P}$ includes the *products* or *commodities* that may move on the network. Let i and j be node indices and p the product index.

Other than the usual characteristics – length, capacity, and cost – *fixed costs* may be associated with some or all links of the network. This indicates that as soon as one chooses to use that particular arc, one incurs the fixed cost in excess of the utilization cost, which is in most cases related to the volume of traffic on the link.

The objective of network design formulation thus is to choose links in a network, along with capacities, to enable the demand for transportation to be satisfied at the lowest possible system cost computed as the total fixed cost of the selected links plus the total variable cost of using the network. A *fixed cost network design* formulation may then take the following form:

$$\text{Minimize} \quad \sum_{(ij)\in\mathcal{A}} f_{ij}y_{ij} + \sum_{(ij)\in\mathcal{A}}\sum_{p\in\mathcal{P}} c_{ij}^{p}x_{ij}^{p} \tag{13.4}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{N}} x_{ij}^{p} - \sum_{j\in\mathcal{N}} x_{ij}^{p} = d_{i}^{p} \qquad i\in\mathcal{N},\ p\in\mathcal{P} \tag{13.5}$$

$$\sum_{p\in\mathcal{P}} x_{ij}^{p} \leq u_{ij}y_{ij} \qquad (i,j)\in\mathcal{A} \tag{13.6}$$

$$(y,x)\in\mathcal{S} \qquad (i,j)\in\mathcal{A},\ p\in\mathcal{P} \tag{13.7}$$

$$y\in\mathcal{Y} \qquad (i,j)\in\mathcal{A} \tag{13.8}$$

$$x_{ij}^{p}\geq 0 \qquad (i,j)\in\mathcal{A},\ p\in\mathcal{P} \tag{13.9}$$

where,

y_{ij}: integer variables modeling discrete choice design decisions. When $\mathcal{Y}=\{0,1\}^{|\mathcal{A}|}$ in relation (13.8), $y_{ij}=1$ only if link $(i,j)\in\mathcal{A}$ is *open*, selected for inclusion in the final network or for capacity expansion; $y_{ij}=0$ otherwise, indicating that the link is *closed.* When $\mathcal{Y}=\mathbf{N}_{+}^{|\mathcal{A}|}$, the y_{ij} variables are not restricted to $\{0,1\}$ values and usually represent the number of facilities or units of capacity installed, or the level of service offered (see Section 13.5 for examples in service network design);

x_{ij}^{p}: continuous flow decision variables indicating the amount of flow of commodity p using link (i,j);

f_{ij}: fixed cost of opening link (i,j); when $\mathcal{Y}=\mathbf{N}_{+}^{|\mathcal{A}|}$, the hypothesis is that a f_{ij} cost is incurred for each unit of facility installed or service offered;

c_{ij}^{p}: transportation cost per unit of flow of product p on link (i,j);

u_{ij}: capacity of link (i,j);

d_{i}^{p}: demand of product p at node i.

This is the *linear cost, multicommodity, capacitated* version of the network design formulation; we identify it as *MCND.* Most applications and methodological developments target the formulations where the design variables are restricted to 0 or 1 values. A number of important applications require nonlinear formulations, however, such as the frequency service network design problems presented in

Section 13.5. Some applications also require that flow variables be restricted to integer values, thus increasing the difficulty of these problems. However, since very few methodological developments have been dedicated to such variants of the network design model, the rest of this section focuses on formulations with $\{0, 1\}$ design variables, continuous flow variables, and linear costs.

The objective function (13.4) of the network design formulation (13.4)–(13.9) measures the total cost of the system. An interesting point of view is to consider this objective as also capturing the tradeoffs between the costs of offering the transportation infrastructure or services and those of operating the system to channel the flow of traffic. Equation (13.5) expresses the usual flow conservation and *demand satisfaction* restrictions. Several demand patterns may be defined, resulting in different models. In some cases, a product may be shipped from (one or) several origins to satisfy the demand of (one or) several destinations. These are models where the supply from several origins may be substituted to satisfy a given demand and are often used in the study of the distribution of raw materials. Variants with single product origin (or destination) may also be encountered.

Demand is defined between pairs of origin-destination points in most applications. In this case, and irrespective of the number of true commodities, a product may be associated with each origin-destination pair, by an appropriate modification of the graph that makes multiple copies of the nodes where several commodities originate or terminate their journeys. Let w^p be the total demand of product p. Then,

$$d_i^p = \begin{cases} w^p & \text{if node } i \text{ is the origin of commodity } p \\ -w^p & \text{if } i \text{ is the destination of commodity } p \\ 0 & \text{otherwise.} \end{cases} \qquad (13.10)$$

Constraint (13.6), often identified as a *bundle* or *forcing* constraint, states that the total flow on link (i,j) cannot exceed its capacity u_{ij} if the link is chosen in the design of the network $(y_{ij} = 1)$ and must be 0 if (i,j) is not part of the selected network $(y_{ij} = 0)$. When the capacity is so large that it is never binding (i.e., u_{ij} is at least the largest possible flow on the link), the demand may be normalized to 1 and u_{ij} may be set to $|\mathcal{P}|$. This simplifies the formulation and corresponds to the *uncapacitated* model. Relations (13.8) and (13.9) specify the range of admissible values for each set of decision variables.

Relation (13.7) captures additional constraints related to the design of the network or relationships among the flow variables. Together, they may be used to model a wide variety of practical situations, and this is what makes network design problems so interesting. For example, the set $\mathcal{S}$ may represent topological restrictions imposed on the design of the network, such as precedence constraints (e.g., choose link (*i,j*) only if link (*p,q*) is chosen) or multiple choice constraints (e.g., select at most or exactly a given number of arcs from a specified subset). An important type of additional constraint reflects the usually limited nature of available resources:

$$\sum_{(i,j)\in\mathcal{A}} f_{ij}y_{ij} \le B \tag{13.11}$$

These *budget* constraints illustrate a relatively general class of restrictions imposed upon resources shared by several links. Note that, quite often, budget constraints replace the fixed cost term in the objective function (13.4). *Partial capacity* constraints also belong to this group:

$$x_{ij}^p \le u_{ij}^p \quad (i,j) \in \mathcal{A},\ p \in \mathcal{P} \tag{13.12}$$

They reflect restrictions imposed on the use of some facilities by individual commodities. Such conditions may be used to model, for example, the maximum quantity of some hazardous goods moved by one train or ship.

An equivalent model is the *path*-based multicommodity capacitated network design formulation *PMCND*:

$$\text{Minimize} \quad \sum_{(ij)\in\mathcal{A}} f_{ij}y_{ij} + \sum_{p\in\mathcal{P}}\sum_{l\in\mathcal{L}} k_l^p h_l^p \tag{13.13}$$

$$\text{subject to} \quad \sum_{l\in\mathcal{L}^p} h_l^p = w^p \qquad p \in \mathcal{P} \tag{13.14}$$

$$\sum_{p\in\mathcal{P}}\sum_{l\in\mathcal{L}^p} h_l^p \delta_{ij}^{lp} \le u_{ij}y_{ij} \qquad (i,j) \in \mathcal{A} \tag{13.15}$$

$$y_{ij} \in \mathcal{Y} \qquad (i,j) \in \mathcal{A} \tag{13.16}$$

$$h_l^p \ge 0 \qquad p \in \mathcal{P}, l \in \mathcal{L}^p \tag{13.17}$$

where,

$\mathcal{L}^p$: set of paths for commodity *p*;
h_l^p: flow of commodity *p* on path *l*;
δ_{ij}^{lp}: 1, if arc (*i,j*) belongs to path $l \in \mathcal{L}^p$ for product *p* (0, otherwise);
k_l^p: transportation cost of commodity *p* on path *l*, $k_l^p = \sum_{(ij)\in\mathcal{A}} c_{ij}^p \delta_{ij}^{lp}$;

with $x_{ij}^p = \sum_{l\in\mathcal{L}^p} h_l^p \delta_{ij}^{lp}$.Constraint (13.7) is usually addressed when paths are built. The same mechanisms may also handle some nonlinear route costs. Furthermore, the explicit consideration of path flows may open interesting algorithmic perspectives as illustrated by the tabu search method proposed by Crainic, Gendreau, and Farvolden (2000).

Note that for any setting of the design variables, these models yield capacitated multicommodity minimum cost network flow (*CMCNF*) problems in arc and path formulations, respectively. For uncapacitated design formulations, the subproblem obtained by fixing the design variables becomes an uncapacitated multicommodity flow problem that decomposes into $|P|$ *shortest path* problems (Pallottino and Scutellà 1998). Ahuja (1997) presents an annotated bibliography of these and other network flow problems.

Several problem classes may be derived from these general formulations by an appropriate definition of the network $\mathcal{G}$ and, eventually, of constraints in $\mathcal{S}$ (Magnanti and Wong 1984). Thus, when fixed costs are attributed to nodes, one obtains *location* formulations. Constraints that require the final design to be a Hamiltonian circuit yield the *Traveling Salesman Problem (TSP).* Different sets of constraints on the form of the optimal network design yield the *Steiner* and the *Spanning Tree* problems. The capacitated *Vehicle Routing Problems* may be viewed as a special case of the capacitated spanning tree formulation. This illustrates the richness of the network design models and explains the wide range of their applications.

General Solution Methods

Although relatively simple to state, network design formulations are generally very difficult to solve. From a theoretical point of view, most design formulations are $\mathcal{NP}$-hard. It has also been observed that for capacitated models, linear relaxations yield poor approximations of the mixed-integer polytope resulting in important optimality gaps. In particular, the interplay between link capacities and fixed costs is not adequately represented by these approaches. Moreover, the network flow subproblems are often highly degenerate, increasingly so when the number of commodities becomes larger. Additional algorithmic challenges follow from the very large scale of most applications. Important results have been obtained for some problem classes; for example, uncapacitated and tree-based formulations. However, much work is still needed for more general problem settings. In the following, we point to some of these results and research challenges. The articles mentioned at the beginning of the section and the references they contain offer a more in-depth treatment of the subject.

The previous models are mixed-integer formulations that may be approached by any of the methodologies available for this class of problems (e.g., Nemhauser and Wolsey 1988 or Salkin and Mathur 1989). A widely used methodology is to relax one or several groups of constraints in a Lagrangian fashion to obtain a simpler problem (Geoffrion 1974). A sequence of multiplier adjustments and resolutions of the relaxation subproblem yields a lower bound on the optimal value of the original formulation. As for multipliers, they may be adjusted by using a nondifferentiable optimization technique, subgradient or bundle, for example (Lemaréchal 1989). *Dual ascent* is another often-used approach to obtain this lower bound. In this case, the dual formulation of the linear relaxation of the problem is the starting point. Dual variables are then iteratively and systematically increased, while conforming to the complementary slackness conditions. An upper bound on the optimal value of the design problem is obtained as the objective value of a feasible solution heuristically derived from the solution to the relaxed problem. The lower and upper bounds are then usually integrated into an implicit enumeration scheme such as the *branch-and-bound* algorithm.

The polyhedral structure of the mixed-integer network design formulation may be studied to derive *valid inequalities* (or *cuts*) to be added to the formulation. Briefly, the objective is to construct, or approximate, the convex hull of the mixed-integer programming formulation by adding valid inequalities. If one succeeds and the convex hull is found, the original problem may be solved by linear programming methods. The *cutting plane* method is based on this idea and proceeds via a succession of resolutions of the linear relaxation of the problem and cut generations. If the convex hull can only be approximated, the bounds may be strengthened, yielding more efficient *branch-and-bound* algorithms.

In many cases, the additional complexity introduced to account for the particularities of the application at hand and the large size of the problem instance make the exact resolution of the problem impractical. Heuristics are then used to obtain solutions of, hopefully, good quality. A number of heuristics, e.g., greedly adding or dropping arcs, aim to avoid mathematical programming techniques altogether but are not very successful for capacitated models. The relaxations and dual-ascent methods presented above are also often used as heuristics with interesting results. Modern heuristics, principally *Tabu Search* (Glover and Laguna 1997), *Simulated Annealing* (Laarhoven and Aarts 1987), and *Genetic Algorithms* (Goldberg 1989), are also increasingly being applied.

Much effort has been dedicated to uncapacitated versions of the problem and significant results have been obtained. In particular, Balakrishnan, Magnanti, and Wong (1989) present a dual-ascent procedure that very quickly achieves lower bounds within 1–4 percent of optimality. Used in conjunction with an add-drop heuristic, the method is able to efficiently address realistically sized instances of LTL service network design problems. The attractive performance of the dual-ascent procedure has led to the development of extensions to other design formulations and applications, as illustrated by the work of Barnhart, Jin, and

Vance (2000) on railroad blocking. An exact solution method for the uncapacitated multicommodity fixed charge network design formulation has been recently proposed by Holmberg and Hellstrand (1998). The authors used a Lagrangian relaxation of the demand constraint (13.5) with subgradient optimization to derive lower bounds. Shortest path algorithms on networks derived from the Lagrangian relaxation solutions are used to yield feasible points. The bounds are then used in a branch-and-bound enumeration scheme and the authors discuss various branching and tree search strategies. Experiments were conducted on randomly generated problems and on a number of instances present in the literature (the largest problems solved had 1000 design arcs and 600 commodities) and showed that the branch-and-bound outperformed a state-of-the-art mixed-integer code with respect to problem size and computation time.

Significant results have also been obtained for the *Network Loading* problem. In this particular version of capacitated formulations, the objective is to install, or *load,* on each design arc a number of capacitated facilities, such as different transportation services. The total cost is made up of fixed link costs to install each facility and commodity-specific transportation costs. Total cost must be minimized and the point-to-point transportation demand must be satisfied. Two restrictions characterize this class of models and make their analysis somewhat simpler. First, one may load an integral number of l different capacitated facilities on each arc. Second, the facility capacities are modular, that is, if the capacities are $C_1 < C_2 < \cdots < C_l$, then C_{i+1} is a multiple of C_i. Originating with the work of Magnanti, Mirchandani, and Vachani (1993, 1995), many efforts have been directed toward the polyhedral study of the problem in order to determine valid inequalities and facets to strengthen the formulation (e.g., Epstein 1998). Berger *et al.* (1998) present an efficient tabu search procedure for problems with multiple facilities where the modular restriction is relaxed and flows for each origin-destination pair must follow a single path.

Very few results have been obtained on capacitated problems defined on general networks that are more difficult to solve and pose considerable algorithmic challenges. The capability to compute efficiently good bounds on the optimal value of the design problem is a prerequisite to the development of solution methods that perform on large-scale problem instances with large numbers of commodities. Lagrangian relaxation approaches have been shown appropriate to address this issue (Gendron and Crainic 1994, 1996, Holmberg and Yuan 1996, Gendron, Crainic, and Frangioni 1998). Several Lagrangian relaxations are possible, however, and many offer the same theoretical bound, which is also the bound one obtains from the strong linear relaxation of the formulation (Gendron and Crainic 1994). From an experimental point of view, the computing efficiency and convergence properties of the bounding procedures, as well as the quality of the solution one may actually obtain, are strongly dependent upon the choice of the nondifferentiable optimization technique used to solve the Lagrangian duals, and require careful calibration. Crainic, Frangioni, and Gendron (2001) calibrate and compare subgradient (Camerini, Fratta, and Maffioli 1978, Crowder 1976) and bundle-based methods (Lemaréchal 1989, Hiriart-Urruty and Lemaréchal 1993) for the *shortest path* and *knapsack* relaxations obtained by the dualization of constraints (13.6) and (13.5), respectively. Experiments on a large set of problem instances (largest problem had 700 design arcs and 400 commodities) were used to identify strategies for the efficient design and implementation of each method. The study showed, in particular, that bundle methods converge faster toward the optimal value of the Lagrangian dual, and that they are more robust with respect to parameter calibration.

The lower bounds reported in these studies are within 9 percent of the optimum on average. Feasible solutions were obtained by using resource-based decomposition methods but these yielded poor bounds. Tabu search meta-heuristics offer currently the best procedures for determining high quality feasible solutions. Crainic, Gendreau, and Farvolden (2000) propose a tabu search metaheuristic that identifies good solutions for the path formulation (13.13)–(13.17). The method combines simplex pivot moves and column generation in a tabu search framework where the design objective (13.13) is used to select the next solution from among the possible candidates. Long-term memories record for each design arc the frequency of inclusion in good solutions and guide the diversification of the search. Extensive experiments, on the same set of problems also used by Crainic, Frangioni, and Gendron (2001), have shown that the method dramatically improves the solutions found by the resource decomposition method. The utilization of the cycle-based neighbourhoods proposed by Ghamlouche, Crainic, and Gendreau (2002a) promises to improve further the performances of meta-heuristics for network design. According to this strategy, the search proceeds in the space of the design variables by moving flow of several commodities simultaneously around suitably defined residual networks. Integrating these neighbourhoods in a tabu search-based path relinking method (Glover and Laguna 1992) constitutes the current best methodology for obtaining high quality, feasible solutions to capacitated multicommodity network design problems (Ghamlouche, Crainic, and Gendreau 2002b). The average optimality gap obtained for the same set of test problems was of the order of 2–3 percent, according to the problem type, with a maximum gap of the order of 10 percent. These results correspond to problems for which the optimal solutions are known. Notice that all mentioned meta-heuristics also allowed the resolution of problems too hard for the standard branch-and-bound of a state-of-the-art software in terms of CPU time or memory limitations.

Very few, if any, polyhedral results exist for the general network design formulation (13.4)–(13.9). When actually used, inequalities derived for "simpler" formulations (e.g., location models, uncapacitated network design or network loading problems) are adapted to the more general formulations. See, for example, the work of Kim *et al.* (1999), who use the *cutset* inequalities initially derived for the capacitated loading problem for the design of service networks for express package delivery firms.

These inequalities state that *the total capacity of any cut must support the total demand with endpoints on the two sides of the cut* and they are certainly valid for the general formulation. We do not know, however, if they define facets or how efficient they are. We certainly do not know how to generate these cuts efficiently. Since their number is extremely high, we have little guaranty regarding the efficiency of this procedure. The same questions are also pertinent regarding the other families of cuts proposed in the literature for formulations "similar" to network design. More work is thus required to identify valid inequalities and facets for the MCND and to develop methods to automatically and efficiently generate these new constraints (the *separation* problem). The work of Chouman, Crainic, and Gendron (2001, 2002) contributes towards feeling this gap. The authors adapt and specialize to multicommodity network design a number of important families of valid inequalities. They also introduce a new familly of valid inequalities. Extensive experimentation shows that (1) not all combinations of valid inequalities are equally effective in terms of solution quality, and (2) specialized cuts and procedures yield significant gains in solution quality and, especially, computational efficiency over state-of-the-art general purpose methods.

The situation and needs are similar concerning methods to identify the optimal solution of general MCND formulations. Holmberg and Yuan (1996, 1998) propose a branch-and-bound algorithm based on the Lagrangian relaxation of the flow constraints and subgradient optimization. The results appear promising, but not conclusive, especially when the dimensions of the network and the number of commodities increase. For larger problems, Kim, Barnhart, and Ware (1999) apply a combination of heuristics to reduce the size of the problem and branch-and-bound with column and constraint generation (the so-called *branch-and-price-and-cut*; cuts are added to the root problem only). This constitutes a very interesting overture to a promising algorithmic avenue. See Hoffman and Padberg (1993), Desrosiers *et al.* (1995), Barnhart *et al.* (1998), and Barnhart, Hane, and Vance (2000) for examples of similar algorithmic structures aimed at various complex problems that arise in transportation science and which emphasize the challenges associated with the development of such methods for the MCND.

Parallel computation may help address realistically dimensioned problem instances in reasonable times. In the case of heuristics, parallelism may also enhance the robustness of the method and improve the quality of the solutions (Crainic 2002, Crainic and Gendreau 2002a, Crainic and Toulouse 2002). Applied to branch-and-bound, parallelism may be used to solve the subproblem at each node of the tree (Gendron and Crainic 1994b) or to explore the tree in parallel (Gendron and Crainic 1994a). Many issues still remain to be addressed in this area however. For example, the addition of cuts often destroys the "nice" structure (network, knapsack, etc.) obtained by relaxing some constraints. The relaxation of the cut constraints could them be contemplated. The issue might become even more challenging when constraints are to be generated at nodes other than the root. It is generally believed, however, that the combination of relaxations, polyhedral results, and heuristics within a parallel computation framework constitutes a promising avenue towards a comprehensive solver for capacitated, multicommodity network design.

13.5 Service Network Design

Service network design is particularly relevant to firms and organizations that operate consolidation transportation systems and is typically related to the planning of operations. It is usually part of *tactical planning* activities, although often it is referred to as strategic/tactical or tactical/operational according to the planning traditions and horizons of the firm. The goal is to operate efficiently to answer demand and ensure the profitability of the firm. The "supply" side of this equation implies a system-wide, network view of operations, integrating consolidation activities in terminals, and the selection, routing, and scheduling of services. On the "demand" side, the routing of freight through the network must be planned to ensure timely and reliable delivery according to the customer specifications and the carrier's own targets.

The objectives of the process are complex as well. The customer's expectations have traditionally been expressed in terms of "getting there" at the lowest cost possible. This, combined with the usual cost consciousness of any firm, has implied that the primary objective of a freight carrier was, and still is for many carriers, to operate at the lowest possible cost. Increasingly, however, customers not only expect low tariffs, but also require a high quality service, mostly in terms of speed, flexibility, and reliability. The significant increase in the market share achieved by motor carriers, mainly at the expense of railway transportation, is due to a large extent to this phenomenon. Consequently, one of the major objectives of tactical planning is to achieve the best tradeoff between operating costs and firm profitability, and service performance measured, in most cases, by delays incurred by freight and rolling-stock or by the respect of predefined performance targets.

To illustrate the complexity of decisions and tradeoffs characteristic of tactical planning, consider the routing of a shipment between two terminals of a consolidation transportation system operated, for example, by a railway or LTL motor carrier. Figure 13.2 displays a representation of such a system made up of five terminals and seven services (for simplicity, the actual service routes are not shown). A shipment that originates at terminal A with destination terminal D is sorted (classified) at A and may be routed according to a number of strategies, including:

1. Consolidate it with other shipments going directly to its destination terminal and put it on one of the available direct services, S1 or S2, of possibly different types. If the freight volume is sufficiently high and the customer contract allows it, S1 or S2 might be operated as a dedicated service, such as a full truck moving direct between two end-of-lines or an unit train.
2. Same consolidation, but move the shipment by using a service, such as S3, that stops at one or several other terminals to drop and pickup traffic.
3. Use the same consolidation policy but move the shipment by a direct service S4 to the intermediate terminal C, where it is transferred to another direct service, S5, that moves it to destination. This strategy may outperform the previous one if the service level offerred on the direct routes outweights the transfer costs; in Figure 13.2, it is also the only strategy available to move from terminal B to terminal D.

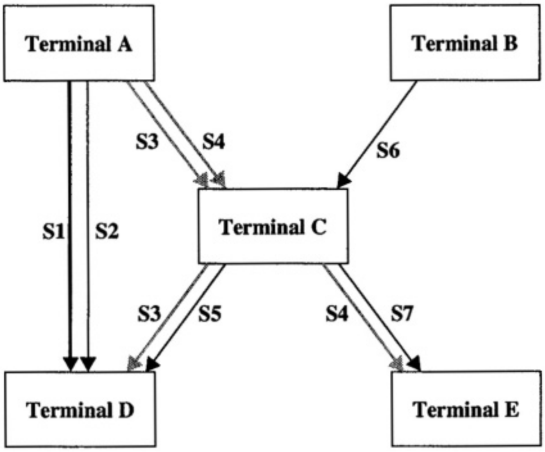


Figure 13.2 Service Network

4. Consolidate the shipment into a load for an intermediate terminal where it will be reclassified and consolidated together with traffic originating at various other terminals into a load for its final destination. The shipment is thus moved by service S3 or S4 from A to C, consolidated together with traffic from B to D and C to D, and then moved by S3 or S5 from terminal C to destination.

Which alternative is "best"? Each has its own cost and delay measures that follow from the service characteristics of each terminal and service. Thus, for example, strategies based on reconsolidation and routing through intermediate terminals may be more efficient when direct services between the origin and destination terminals of the shipment are offered rarely due to generally low level of traffic demand. Such strategies would probably result in higher equipment utilization and lower waiting times at the original terminal; hence, in a more rapid service for the customer. The same decision would also result, however, in additional unloading, consolidation, and loading operations, creating heavier delays and higher congestion levels at intermediary terminals, as well as a decrease in the total reliability of the shipment. On the other hand, to increase the frequency of a direct service between the origin and destination terminals of a shipment would imply a faster and more reliable service for the corresponding traffic, as well as a decrease in the level of congestion at the intermediate terminals at the expense of additional resources, thus increasing the direct costs of the system. Therefore, to select the "best" solution for the customer and the company, one has to simultaneously consider the routing of all traffic, the level of service on each route, and the costs and service characteristics of each terminal. These problems and decisions have network-wide impacts and are strongly and complexly interconnected both in their economic aspects and the space-time dimensions of the associated operations. Therefore decisions should be made globally, network-wide, in an integrated manner (Crainic and Roy 1988). More formally, main decisions made at the tactical level concern the following issues:

1. *Service selection.* The routes – origin and destination terminals, physical route and intermediate stops – on which services will be offered and the characteristics of each service. *Frequency* or *scheduling* decisions are part of this process.
2. *Traffic distribution.* The *itineraries* (routes) used to move the flow of each demand: services used, terminals passed through, operations performed in these terminals, etc.
3. *Terminal policies.* General rules that specify for each terminal the consolidation activities to perform. For rail applications, these rules would specify, for example, the blocks into which cars should be classified (the *blocking* policies), as well as the trains that are to be formed and the blocks that should be

put on each train (the *make up* rules). An efficient allocation of work among terminals is an important policy objective.

4. General *empty balancing* strategies, indicating how to reposition empty vehicles to meet the forecast needs of the next planning period.

Several efforts have been directed toward the formulation of tactical models. See the reviews of Assad (1980), Crainic (1988), Delorme, Roy, and Rousseau (1988), and Cordeau, Toth, and Vigo (1998). Network models, which take advantage of the structure of the system and integrate policies affecting several terminal and line operations, are the most widely developed. Simulation models have been proposed and used by transportation firms to evaluate scenarios and select policies. Network optimization formulations, on the other hand, may efficiently generate, evaluate, and select integrated network-wide operating strategies, transportation plans, and schedules. These models are discussed in this section.

Most service network design and related issues yield *fixed cost, capacitated, multicommodity network design formulations* (Section 13.4). These formulations may be static or dynamic but, up to now, have been generally deterministic. For a clearer view of tactical planning issues and service network design formulations, we distinguish between *frequency* and *dynamic* service network design models.

The former typically addresses strategic/tactical planning issues. The study and representation of interactions and tradeoffs among subsystems and decisions form a central part of this class of approaches. Typical issues addressed by such models concern questions such as: *What* type of service to offer? *How often* over the planning horizon to offer it? Which traffic *itineraries* to operate? What are the appropriate *terminal workloads* and *policies*? Frequency service network design models may be further classified according to the role service levels play in the formulations: *decision* or *output.* In a nutshell, service frequencies are explicit integer decision variables in the first class of models. Formulations that belong to the second class include "operate or not" ({0,1}) decision variables and derive frequencies from traffic flows subject to lower bound restrictions that represent minimum service levels. The output of frequency service network design models, the *transportation* or *load plan,* is used to determine the day-to-day policies that guide the operations of the system and is also a privileged evaluation tool for "what-if" questions raised during scenario analysis in strategic planning. Dynamic formulations are closer to the operational side of things. They usually target the planning of *schedules* and support decisions related to *if* and *when* services leave. Subsections 13.5, 13.5, and 13.5 examine models and methods that belong to each of these three classes. Section 13.5 briefly reviews the literature associated to service network design and tactical planning.

Frequency Service Network Design

The network optimization modeling framework proposed by Crainic and Rousseau (1986) constitutes a prototypical frequency service network design formulation where explicit decision variables are used to determine how often each selected service will be run during the planning period. It is a multi-modal multicommodity model that integrates the service selection and traffic distribution problems with general terminal and blocking policies. Its goal is the generation of global strategies to improve the cost and service performance of the system. It is a modeling framework in the sense that while it may represent a large variety of real situations, it has to be adapted to each application. Rail applications are to be found in Crainic (1982, 1984), Crainic, Ferland, and Rousseau (1984), and Crainic and Nicolle (1986). Roy (1984) and Delorme and Roy (1989) present applications of this framework to LTL trucking. In the following, we present a simplified model in order to emphasize the main modeling issues and challenges.

Let $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ represent the "physical network" over which the carrier operates. Vertices in $\mathcal{T} \subseteq \mathcal{N}$ correspond to nodes where the terminals selected for the particular application are situated. For simplicity, assume that all terminals can perform all operations. The *service network* specifies the transportation services that could be offered to satisfy this demand. Each *service* $s \in \mathcal{S}$ is defined by its route r_s through the physical network; origin, destination, and intermediary terminals where the service stops and work may be performed on its vehicles and cargo; capacity u_{ij}^s on each link of r_s; service class that indicates characteristics such as the mode, preferred traffic or restrictions, speed and priority of the service, etc.

Transportation demand is defined in terms of volume (e.g., number of vehicles) of a certain commodity to be moved between two terminals in $\mathcal{T}$. To simplify, we refer to product p = *(commodity type, origin, destination)* with a positive demand w^p. In the literature, one also finds the terms *market* and *traffic-class* with a similar meaning. Empty vehicles may be included as commodities to be moved between given origin-destination pairs. Traffic moves according to *itineraries.* An itinerary $l \in \mathcal{L}^p$ for product p specifies the service path used to move (part of) the corresponding demand: the origin, destination, and intermediary terminals where operations are to be performed; the sequence of services between each pair of consecutive terminals where work is performed; the commodity class that indicates characteristics such as priority, minimum service level, preferred transportation mode, etc.

Service frequencies y_s, $s \in \mathcal{S}$, define the level of service offered, i.e., how often each service is run during the planning period. To design the service network thus means to decide the frequency of each service contemplated in the planning process such that the demand is satisfied. Many itineraries may be defined for each product and more than one may be actually used, according to the level of congestion in the system and the service and cost criteria of the particular application. Flow distribution decisions are therefore represented by variables h_l^p indicating the volume of product $p \in \mathcal{P}$ moved by using its itinerary $l \in \mathcal{L}^p$. Workloads and general consolidation strategies for each terminal in the system may be derived from these decision variables.

Let $y = \{y_s\}$ and $h = \{h_l^p\}$ be the vectors containing the decision variables. The model states that the total generalized system cost must be minimized, while satisfying the demand for transportation and the service standards:

$$\text{Minimize} \quad \sum_{s \in \mathcal{S}} \Psi_s(y) + \sum_{p \in \mathcal{P}} \sum_{l \in \mathcal{L}^p} \Phi_l^p(y, h) + \Theta(y, h) \tag{13.18}$$

$$\text{subject to} \quad \sum_{l \in \mathcal{L}^p} h_l^p = w^p \qquad p \in \mathcal{P} \tag{13.19}$$

$$y_s \geq 0 \text{ and integer} \qquad s \in \mathcal{S} \tag{13.20}$$

$$h_l^p \geq 0 \qquad l \in \mathcal{L}, p \in \mathcal{P} \tag{13.21}$$

where,

$\Psi_s(y)$: total cost of operating service s;
$\Phi_l^p(y, h)$: total cost of moving the freight of product p by using its itinerary l;
$\Theta(y, h)$: penalty terms capturing various relations and restrictions, such as the limited service capacity.

This model is similar to the path formulation of the capacitated network design model ((13.13)–(13.17)) introduced in Section 13.4, except that the linear cost functions of the latter have been replaced by a notation that indicates more general functional forms. The objective function defines the *total system cost* and includes the total cost of operating a service network at given frequencies, the total cost of moving freight by using the selected itineraries for each product, as well as a number of terms translating operational and service restrictions into monetary vales. $\Psi_s(y)$ and $\Phi_l^p(y, h)$ thus correspond to the *fixed* and *variable* costs, respectively, of the network formulation given the general level of service of the firm and the corresponding traffic pattern. The objective function computes a *generalized* cost, in the sense that it may include various productivity measures related to terminal and transportation operations. Other than the actual costs of performing the operations, one may thus explicitly consider the costs, delays, and other performance measures related to the quality and reliability of the service offered, to evaluate alternatives and determine the most advantageous tradeoffs.

The delays incurred by vehicles, convoys, and freight due to congestion and operational policies in terminals and on the road are generally used as a measure of service quality. Define $T_s(y)$ and $T_l^p(y,h)$ as the total durations of service s and itinerary l for product p, respectively. Equations (13.22)–(13.23) illustrate one approach to use delays to integrate service considerations into the total generalized system cost. On the one hand, unit operating costs C_s^O and C_{lp}^O are computed for each service and product itinerary, respectively. On the other hand, the corresponding total expected service, $E[T_s(y)]$, and itinerary, $E[T_l^p(y,h)]$, times are converted into measures compatible with the operating costs via unit time costs for each traffic (C_{lp}^D) and service (C_s^D) class. These costs are usually based on equipment depreciation values, product inventory costs, and time-related characteristics, such as priority or different degrees of time sensitivity for specific traffic classes.

$$\Psi_s(y) = (C_s^O + C_s^D E[T_s(y)])y_s \tag{13.22}$$

$$\Phi_l^p(y,h) = (C_{lp}^O + C_{lp}^D E[T_l^p(y,h)])h_l^p. \tag{13.23}$$

Although nonlinear functions could be used, unit operation costs C_s^O and C_s^D are usually computed as the sum of the unit costs of all terminal and transportation activities described in the service routes and freight itineraries. For rail applications, these may include hauling costs for trains and cars over the lines of the network, as well as yard handling costs associated with car classification, the transferring of cars and blocks among trains, and the making-up and breaking-down of trains. Similar terms appear in LTL applications: loading, unloading, transdock, and consolidation operations at terminals, energy costs, maintenance, crews, etc.

The expected total delays $E[T_s(y)]$ and $E[T_l^p(y,h)]$ are also computed by summing up the expected delays associated with the terminal and line operations that make up the service and freight routes. No correlation is usually considered. Some durations are simply assumed proportional to the volume of vehicles or traffic handled. It is typically the case for the yard transfer delays for rail applications and intercity transportation time for LTL tracking. In many operations, however, vehicles of different services carrying freight for different products on various itineraries must use the same facilities. It is the case, for example, of most consolidation and classification operations. As a consequence, most time-related functions are built to reflect the increasingly larger delays that result when facilities of limited capacity must serve a growing volume of traffic. Such *congestion* functions are typically derived from engineering procedures and queuing models (see Section 13.6) and are built to represent: average delays due to rail yard operations, particularly car classification and blocking, and train make up; waiting time of trucks at LTL terminals before loading and unloading operations (rail cars and trucks at port loading/unloading facilities experience similar delays); delays incurred by trains when meeting, overtaking, or being overtaken by other trains on the lines of the network; congestion on highways in urban areas; expected departure or connection delays in rail yards, LTL terminals, and maritime ports representing the waiting time for the designated service to be available, and so on.

Average transportation delays do not tell the whole story, however. Often, the goal is not only rapid delivery but also consistent, reliable service. The variance of the total service or itinerary time may then be used to penalize unreliable operations. Equation (13.24) illustrates this approach for the case when service quality targets are announced. Here, each traffic-class has a delivery objective (e.g., 24 hours) and reliability requirements (e.g., target must be achieved for 90 percent of deliveries), noted H_p and n, respectively. A penalty C_{lp}^D is then imposed when the expected itinerary time, adjusted for its standard deviation $\sigma[T_l^p(y,h)]$, does not comply with the service objective. The total itinerary cost then becomes:

$$\Phi_l^p(y,h) = C_{lp}^O h_l^p + C_{lp}^D \,(\min\{0, H_p - E_l^p(y,h) - n\sigma[T_l^p(y,h)]\})^2 \, h_l^p. \tag{13.24}$$

Finally, equation (13.25) illustrates the use of penalty terms to capture various restrictions and conditions. Here, x_{sk} stands for the total volume of freight hauled by service s over its service leg k, $x_{sk} = \sum_{p\in\mathcal{P}} \sum_{l\in\mathcal{L}^p} h_l^p \delta_{sk}^{lp}$, with $\delta_{sk}^{lp} = 1$ if service leg k of service s is used by itinerary l of product p, and 0 otherwise. Thus, in this example, the service capacity restrictions are considered as utilization targets and the over-assignment of traffic is permitted at the expense of additional costs and delays. Tradeoffs between the cost of increasing the level of service and the extra costs of insufficient capacity may then be addressed while the associated mathematical programming problem is solved.

$$\Theta(y,h) = \sum_{s\in S} C_s^P \sum_{(ij)\in r_s} (\min\{0, u_{ij}^s y_s - x_{sk}\})^2. \tag{13.25}$$

The model has the structure of a nonlinear, mixed integer, multi-modal, multicommodity network flow problem. No exact solution method has yet been proposed for this model. The original method described by Crainic and Rousseau (1986) combines a heuristic (based on finite differences in the objective function) that iteratively decreases frequencies from initial high values, with a convex network optimization procedure to distribute the freight. The latter makes use of column generation to create itineraries and descent procedures to optimize the flow distribution. The procedure appeared efficient for the rail and LTL applications considered. Crainic and Roy (1988) and Roy and Crainic (1992) also report on the utilization of this approach to perform scenario and postoptimal analyses, particularly concerning the tradeoffs between the cost of operating the system and the value of time, and the level of demand required to operate direct services over long distances.

Service Frequencies as Derived Output

The load planning model for LTL motor carriers introduced by Powell and Sheffi (1983, 1986, 1989; see also Powell 1986a and Lamar, Sheffi, and Powell 1990) constitutes a major example of frequency service network design formulations that yield service levels as one of their outputs. What follows is a condensed version of this model.

The model is defined on a service network $\mathcal{G} = (\mathcal{T}, \mathcal{S})$ where all nodes are terminals and links represent potential direct services between two terminals. Two types of terminals are considered: *end-of-lines,* where freight originates and terminates; and *breakbulk* consolidation terminals. Although not forbidden, direct movements between end-of-line terminals are extremely rare, especially for very large LTL carriers. Consequently, the design decisions concern only services between end-of-lines and breakbulks, and between breakbulk terminals. This has the benefit of considerably reducing the size of the problem. The main parameters and decision variables that define the model are:

C_{ij}: unit linehaul cost per trailer, loaded or empty, from terminal i to terminal j;

C_i^B: unit trailer handling cost at terminal i, if terminal i is a breakbulk (0, otherwise);

$C_i^E(\cdot)$: a function that computes the trailer handling cost at end-of-line i according to the total number of direct services operated out of i(0, if i is a breakbulk);

w_{od}: number of LTL trailers originating at terminal o and destined for terminal d;

$\mathcal{L}$: set of permissible freight routings, i.e., that respect particular constraints with respect to the association of end-of-line terminals to breakbulks (the so-called *clustering constraints*);

y_{ij}: service design decisions; $y_{ij} = 1$ if the carrier is offering direct service from terminal i to terminal j, and 0 otherwise;

x_{ij}^d: volume of LTL traffic on link (i,j) with destination terminal d; $x_{ij} = \sum_d x_{ij}^d$;

r_{ij}^d: auxiliary flow routing variable (its use simplifies the representation of the clustering constraints);

v_{ij}: flow of empty trailers moving from i to j;

x_i^B: volume of total LTL traffic handled at breakbulk i, that is, the traffic that originates at i plus the traffic that is transferred at the terminal;

M_{ij}: minimum frequency if a direct service is offered from terminal i to terminal j;

$F_{ij}(x_{ij})$: service frequency – the number of trailers dispatched over the planning period, from terminal i to terminal j, where,

$$F_{ij}(x_{ij}) = \begin{cases} \max\{M_{ij}, x_{ij}\} & \text{if } x_{ij} \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{13.26}$$

The model may be written as:

$$\text{Minimize} \quad \sum_{(ij)\in\mathcal{S}} C_{ij}\left[F_{ij}(x_{ij})y_{ij} + v_{ij}\right] + \sum_{i\in\mathcal{T}} \left[C_i^B x_i^B + C_i^E(y)w_i\right] \tag{13.27}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{T}} r_{ij}^d = 1 \ \text{ and } \ \{r_{ij}^d\} \in \mathcal{L}, \qquad i,j,d \in \mathcal{T} \tag{13.28}$$

$$x_{ij}^d = \left[w_{id} + \sum_{k\in\mathcal{T}} x_{ki}^d\right] r_{ij}^d, \qquad i,j,d \in \mathcal{T} \tag{13.29}$$

$$r_{ij}^d \leq y_{ij}, \qquad i,j,d \in \mathcal{T} \tag{13.30}$$

$$\sum_{j\in\mathcal{T}} v_{ij} - \sum_{k\in\mathcal{T}} v_{ki} = w_i, \qquad i \in \mathcal{T} \tag{13.31}$$

$$w_i = \sum_{k\in\mathcal{T}} F_{ki}(x_{ki}) - \sum_{j\in\mathcal{T}} F_{ij}(x_{ij}), \qquad i \in \mathcal{T} \tag{13.32}$$

$$y_{ij} \in \{0,1\}, \qquad (i,j) \in \mathcal{S} \tag{13.33}$$

$$r_{ij}^d \in \{0,1\}, \qquad (i,j) \in \mathcal{S} \tag{13.34}$$

$$x_{ij}^d, v_{ij}^d, \hat{x}_{ij} \geq 0, \qquad i,j,d \in \mathcal{T} \tag{13.35}$$

The objective function (13.27) computes the total cost of dispatching trailers according to the determined service level, moving the loaded and empty trailers, and handling freight in terminals. Constraints (13.28) and (13.29) ensure that freight itineraries obey routing restrictions and that demand is satisfied. Relation (13.30) is the usual linking constraint that ensures that only operated services are used. Equations (13.31) and (13.32) balance the empty flows.

The modeling framework is strongly influenced by the LTL context and the considerable challenges associated with the large size of the LTL carriers operating at the national level in the United States. It may be viewed as an extension of the arc-based multicommodity network design formulation ((13.4)–(13.9)) in Section 13.4, with no explicit capacities and a number of complicating constraints. The authors implemented a heuristic procedure based on the hierarchical decomposition of the problem into a master problem and several subproblems. The master problem is a simple network design problem where the total system cost (13.27) is computed for each given configuration of selected services. The design is modified by adding or dropping one arc at a time (Powell 1986a). Each time the design is modified, the subproblems must be solved and the objective function must be evaluated. The first subproblem concerns the routing of loaded LTL trailers and it is solved by shortest-path-type procedures with tree constraints (Powell and Koskosidis 1992). The empty balancing subproblem is solved as a minimum cost transshipment formulation with adjusted supply and demand to account for timing conditions not included in the original formulation (Roy and Delorme 1989, use a similar approach).

The model and solution method are at the core of an interactive decision support system, dubbed APOLLO (Advanced Planner Of LtL Operations), and has been implemented at a major U.S. LTL carrier. Impressive results are reported with respect to the impact of the system both on load planning operations and strategic studies of potential terminal location. Powell and Scheffi (1986, 1989) present in more details the functionalities of APOLLO and discuss its performances on actual problems. They also emphasize the importance of allowing planners to interact with the software to explore alternatives and to select among various options. In this way, planners are better positioned to understand how the system works and, ultimately, to accept its suggestions. The same modeling framework was also used as the basis for the development of a more comprehensive load planning system called SYSNET (Braklow *et al.* 1992), implemented at one of the largest LTL carriers in the United States. In this version, the issue of running direct services, bypassing breakbulk terminals, was explicitly addressed by including such services into the service network. The routing of the freight also acknowledged the geographic and labor structure of the company and considered the relay points where trailers are passed from one driver to the next. The resulting network representation is huge. Heuristics based on company operating rules are used to prune it before the optimization routines are called upon. Other than the optimization model and procedures, the planning system includes demand forecasting, database management, user monitoring and control functionalities. The system has been used with great success to build the load plan, to study the location and dimension of breakbulks, to determine the routing of loaded and empty trailers, and to study which directs should be added or dropped.

Deterministic Dynamic Service Network Design

When schedules are contemplated, a *time* dimension must be introduced into the formulation. This is usually achieved by representing the operations of the system over a certain number of *time periods* by using a *space-time* network.

The representation of the physical network is replicated in each period. Starting from its origin in a given period, a service arrives (and leaves, in the case of intermediary stops) later at other terminals. Services thus generate temporal service links, between different terminals at different time periods. Temporal links that connect two representations of the same terminal at two different time periods may represent the time required by terminal activities or the freight waiting for the next departure. The costs associated with the arcs of this network are similar to those used in the static formulations of the previous subsections. Additional arcs may be used to capture penalties for arriving too early or too late.

There are again two types of decision variables. Integer design variables are associated with each service. Restricted to $\{0, 1\}$ values, these variables indicate whether or not the service leaves at the specified time. When several departures may take place in the same time period, general (nonnegative) integer variables must be used. (Note that one can always use $\{0, 1\}$ variables only by making the time periods appropriately small.) Continuous variables are used to represent the distribution of the freight flows through this service network.

The resulting formulations are network design models similar to those presented in Section 13.4, but on a significantly larger network due to the time dimension. Actually, any of the two previous modeling frameworks, service network design with frequency variables or derived output, may be used as the basis for a dynamic scheduling model. The sheer size of the dynamic network, as well as the additional constraints usually required by the time dimension, makes this class of problems even harder to solve than the static ones. Thus, the pioneering effort of Morlok and Peterson (1970), which integrated blocking, train formation, and train scheduling into a very large mixed integer formulation, never did yield a solution method or an application. Heuristic methods have been used so far.

Farvolden and Powell (1991, 1994) present a dynamic service network design model for LTL transportation. The formulation allows for several departures in the same period, but the simpler $\{0, 1\}$ version is solved. An efficient primal-partitioning with column generation algorithm is used to solve the freight routing problem for any given service configuration (Farvolden, Powell, and Lustig 1992). This was also used to determine the dual variables for service links used to develop an add-drop heuristic for the design problem. The methodology appeared interesting, especially concerning the quality of the evaluation of the add and drop moves. No comprehensive experimental analysis is available, however. Equi *et al.* (1997) determine which shipments of a given good are to be performed and the schedules of the vehicles that will undertake them. The model is a mixed-integer formulation. The proposed heuristic decomposes the problem according to a Lagrangian-type decomposition and proceeds in two steps: a metaheuristic implements tabu search ideas to approximate the design subproblem, and a transportation problem addresses the scheduling part. The methodology has been successfully applied to the problem of transporting wood from cutting areas to ports.

Haghani (1989) attempts to combine the empty car distribution with the train make-up and routing problems. The dynamic network includes normal and express modes for each service route for each time period, but traffic on each link is prespecified and access to express links is restricted to given markets. Travel times are fixed. Linear functions are used for costs and delays, except for classification, which makes use of a convex congestion function. The dynamic service network design has continuous empty and loaded car flows and integer engine flows. A heuristic decomposition approach is used to solve somewhat simpler problems and appears efficient for small rail systems. The study also points to better performances, in terms of operating costs, of an integrated formulation as compared to the "traditional" hierarchical approach.

Gorman (1998a) also attempts to integrate the various service network design aspects into a scheduled operating plan that minimizes operating costs, meets the customer's service requirements, and obeys the operation rules of a particular railroad. Model simplifications must be introduced in order to achieve a comprehensive mathematical network design formulation. The solution method is innovative. A hybrid metaheuristic, a tabu-enhanced genetic search, is used to generate candidate train schedules, which are evaluated on their economic, service, and operational performances. On relatively small but realistic problems, the metaheuristic performed very well. A major U.S. railroad has successfully used this model for strategic scenario analysis of their operations (Gorman 1998b). This work emphasizes the interesting perspectives offered by modern heuristics in addressing complex service network design problems.

More Service Network Design Models

Several other service network design modeling efforts make use of $\{0,1\}$ mixed integer network flow formulations similar to the network design models in Section 13.4. Keaton (1989, 1991, and 1992) proposes a model to develop operating plans for railroads. The model aims to determine which pairs of terminals to connect by direct service, and whether to offer more than one train a day, as well as the routing of freight and the blocking of rail cars. The service network is made up of one network for each origin-destination pair of terminals in the system with positive demand. Links represent trains and connections in yards, as well as *a priori* determined blocking alternatives. Continuous car flows and integer train connections represent the decision variables. All cost functions are linear – there is no congestion and fixed average yard delays are used. The model minimizes the total cost computed as the sum of fixed train costs, car time-related costs, and classification costs. The maximum number of blocks that may be built in a yard yields the linking constraints. Feasibility constraints limit the maximum number of connections and the minimum number of trains for a given pair of terminals. Solutions were obtained by using a Lagrangian relaxation of the linking constraints combined with various heuristics based on operation rules. Results were mixed. While the model was used to perform a number of analyses on relatively small systems, convergence difficulties were also reported.

Newton (1996), Newton, Barnhart, and Vance (1998), and Barnhart, Jin, and Vance (2000) also address the rail blocking problem. They formulate it as a network design problem, where nodes correspond to classification yards, and candidate blocks correspond to arcs. No fixed costs are associated with blocks, but several capacity restrictions are introduced to limit the number of blocks and the total volume of freight processed at each yard. The first two references present a path-formulation and a branch-and-price solution approach (Barnhart *et al.* 1998). In the third paper, a dual-based Lagrangian relaxation is used to decompose the problem into easier-to-address subproblems: a continuous multicommodity flow problem and an integer block formulation that selects blocks that satisfy the yard capacity constraints. Subgradient optimization is used to solve the Lagrangian dual, while column generation is applied to the flow subproblem. To solve the block subproblem, a branch-and-cut approach is used, where constraints that force the connectivity of at least one path for each commodity are added to the nodes of the enumeration tree. With these constraints, the Lagrangian relaxation identifies a better bound than previously. By incorporating significant data preprocessing to reduce the number of potential blocks and paths, the method could address the problem of a major American railroad and propose blocking plans that represent significant cost improvements.

Kuby and Gray (1993) developed an early model for the design of the network of an express package delivery firm. It is a path-based $\{0, 1\}$ network design model, similar to formulation (13.13)–(13.17), where multistop aircraft routes must be selected in and out of a given hub. Paths were generated *a priori,* and the model was solved with a standard mixed-integer package. Analyses illustrated the cost effectiveness of a design with multiple stops over a pure hub-and-spoke network.

Kim, Barnhart, and Ware (1999) propose more comprehensive models for the design of the multi-modal version of the problem (Barnhart and Schneur 1996, address a simplified version of the problem). Here, several hubs and aircraft types are considered, while trucks may perform pickup and delivery activities, as well as transportation over limited distances. The problem is further complicated by *time window* restrictions on pickup and delivery times at major collection centers, as well as on the sorting periods at hubs. One product is considered in the application. The authors examine arc, path, and tree-based formulations, and select the latter since it significantly reduces the size of the problem. To solve the linear relaxation of the resulting formulation, the authors combine heuristics to further reduce the size of the problem, cut-set inequalities, and column generation. Branch-and-bound is then used to obtain an integer solution. The paper by Kim and Barnhart (1997) presents a good summary of the authors' experience with these difficult problems and the branch-and-price-and-cut methodology.

The design of postal networks and services forms a class of problems very close to those just mentioned. The LTL frequency service network design by Roy (1984) has already been applied to the design of express letter services for Canada Post. The reorganization of the German postal services belongs to the same problem class, albeit on a more comprehensive scale. To bring the problem down to manageable proportions, Grünert and Sebastian (2000; see also Grünert, Sebastian, and Thärigen 1999 and Buedenbender, Grünert, and Sebastian 2000) decompose it into several subproblems: the optimization of the night airmail network, the design of the ground feeding and delivery transportation system, the scheduling of operations. Vehicle routing models and techniques are used for the routing and scheduling tasks. A discrete dynamic network design formulation, similar to those discussed in Section 13.5, is also proposed. The air network design formulation is further decomposed into a direct flight problem and a hub system problem; both yield fixed cost, multicommodity, capacitated network design formulations with side constraints. To optimize these formulations, the authors propose combinations of classical heuristics, tabu search and evolutionary metaheurstics, and exact mathematical programming methods (e.g., branch-and-bound). A decision support system integrates the models and associated solution methods, as well as the tools required to handle the data, models, and methods, and to assist the decision process.

Armacost, Barnhart, and Ware (2002) also address next-day express (air) delivery service design but through a different methodological approach. The authors transform the problem formulation by defining variables that represent combinations of service routes. The new variables implicitly account for the flow distribution and thus yield a pure design formulation for which stronger bounds and thus more efficient solution methods may be derived. The results obtained on data from a major U.S. express shipment firm is very encouraging and emphasizes the need to continue to explore the network design formulations for new insights and more efficient solution methods.

13.6 Operational Planning and Management

The ultimate goal of any transportation firm is to make profits and improve, or at least maintain, its competitive position. To this end, strategic and tactical plans can be drawn up to guide operations, but the operational capabilities of the firm will ultimately determine its performance.

There are many different issues that must be addressed at the operational level in order to ensure that demand is satisfied within the required service criteria and that the resources of the carrier are efficiently used. Most of these issues must consider the *time* factor. For example, an empty truck must be assigned and moved following a customer request; empty rail cars have to be repositioned, otherwise, soon, idle equipment will be observed at some terminals while others will not be able to satisfy demands; a container must arrive in time to be loaded on the departing ship; a truck has to pick up a load within a specified time window; and so on and so forth. For other types of operations, the very notion of a planned solution does not make sense and the whole operation must continuously adapt and react in real time. Consider, for example, truckload motor carrier services where drivers learn their next assignment only after the current task is concluded. Thus, the need to answer customer requests in real time, to conform to time restrictions on operations, and to integrate in today's decisions their possible impact on future decisions and performances, emphasize the *dynamic* aspect of operational planning and management issues for a freight carrier.

Many models traditionally used in transportation planning use known static data as their input. Tactical planning formulations, for example, consider aggregated forecast demand data as "known". However, the real world in which these models are used is in a constant state of change and solutions cannot always be implemented as planned. If traffic is slower than predicted, vehicles may arrive late at customers' locations or at the terminal. Forecasted customer requests for empty containers or trailers may not materialize while unexpected demands may have to be satisfied. The planned supplies of empty vehicles at depots may thus be unsettled and additional empty movements may have to be performed. Consequently, the dynamic aspect of operations is further compounded by the *stochasticity* inherent to the system, that is, by the set of uncertainties that are characteristic of real-life management and operations. Increasingly, these characteristics are reflected in the models and methods aimed at operational planning and management issues, as illustrated in this section.

Crew Scheduling

Crews are assigned to vehicles and convoys in order to support the planned operations. There are also numerous other issues related to manpower management such as the scheduling of reserve crews, terminal employees (e.g., Nobert and Roy 1998), maintenance crews, etc. A significant body of methodological and technological knowledge has been developed to deal with these issues, especially in the context of transit (bus and passenger rail) and airline transportation. Some form of *set covering* model is generally used. The resulting mixed-integer formulation is usually very large and it is addressed by column generation and branch-and-price techniques. See, for example, Barnhart and Talluri (1997), Desrosiers *et al.* (1995), Desaulniers *et al.* (1998a,b).

These methodologies were developed for applications where detailed schedules are known and adhered to. Consequently, although a few similar developments have targeted crew scheduling issues in the freight transportation industry (e.g., Crainic and Roy 1990, 1992), currently it appears that better results can be achieved by applying the class of methodologies used to dynamically allocate resources to tasks described in Section 13.6. Crew scheduling issues and models form the subject of Chapter 14 of this book.

Terminal and Linehaul Operations

Terminal and line managers, operators, and dispatchers face a host of control and dispatch issues that form the subject of an extensive literature. The corresponding models and methods aim either to analyze and plan operations or to assist the real-time dispatch of resources and control of operations. A brief enumeration of a number of important issues and references follows.

Terminal models mainly address issues related to the estimation of delays associated with the various operations: load or unload freight, classify vehicles, form blocks and trains, transfer freight between vehicles or convoys, etc. The restricted number of available resources and the large volumes of freight and vehicles that require service result in congestion conditions usually evaluated through the average (and, sometimes, the variance) of the associated waiting time. Delays may also result from the need to wait for the planned connection. Queuing formulations are generally used to derive models for these phenomena (e.g., Crainic 1988 and Chapter 5). Petersen (1977b,c; see also Petersen 1971a,b and Petersen and Fullerton 1975) presented what is probably the first comprehensive analysis of yard delays. Bulk service queues, where service is performed for groups of customers (cars) emerge as the main methodological approach. They are difficult to solve, however, in all but the simplest limiting cases. Turnquist and Daskin (1982) use similar formulations for their rail yard model but relax a number of restrictions in order to obtain a more tractable model. Daskin and Walton (1983) propose a set of queuing models to represent the lightering operations (transfers from large ocean tankers to smaller vessels) in crude transportation.

Powell (1981, 1986b; see also Powell and Humblet 1986) undertook a significant study of bulk queues and their applications to modeling delays in transportation terminals. He proposed efficient numerical methods (Powell 1986d) and closed-form approximations (Powell 1986c) to compute the moments of the distributions. Closed-form approximation formulas have also been proposed by Crainic and Gendreau (1986). Such closed-form approximations of delays in freight terminals (as well as on the lines of the systems) are equally important as generators of functions and measures for the service network design and the strategic planning models presented in Sections 13.5 and 13.3, respectively. A different perspective on yard-blocking performance is offered by Daganzo (1987a,b). Based on direct analyses of the departure schedules, policies, and operational rules, formulae are determined for a number of performance measures – number of tracks in the yard and number of switches per car, for example – for various blocking strategies.

Many rail line models aim to represent the delays that result when trains meet (on single-track lines) or when one train overtakes another. When traffic volumes are low, analytical formulae may be obtained directly from the corresponding operating rules (Petersen 1974, 1975a,b). Queuing models are again the methodology of choice, the approach being similar to that used to analyze rail yard operations when congestion conditions occur (Petersen 1977c). More recently, Chen and Harker (1990) and Harker and Hong (1990) consider the case when services are scheduled and evaluate the mean and variance of delays on double and single-track lines, respectively. Hallowell and Harker (1996) evaluate and predict performance on partial double-track rail line with scheduled traffic.

The preceding models may be combined and, eventually, approximated, to yield formulations that may be used in more comprehensive planning systems. Petersen and Taylor (1982), Petersen (1984), Crainic, Ferland, and Rousseau (1984), and Crainic, Florian, and Léal (1990) integrate queuing submodels or functions into the planning systems they propose.

A different line of research relative to rail line models addresses issues related to the scheduling and pacing of trains on a line. Jovanović and Harker (1991) propose a mixed integer formulation, solved by branch-and-bound, to assist the tactical (weekly or monthly) scheduling of trains on a line. The model is embedded in the SCAN1 software. The issue of optimally pacing trains over a line is addressed by Kraay, Harker, and Chen (1991), Higgins, Ferreira, and Kozan (1995) and Higgins, Kozan, and Ferreira (1996, 1997). Network-based mixed integer formulations also appear here. The application of genetic, tabu search, and hybrid metaheuristics to the same problem is explored by Higgins, Ferreira, and Kozan (1997).

Empty Vehicle Distribution and Repositioning

A particularly important and challenging issue for freight carriers is the need to move empty vehicles. Indeed, the geographic differences in demand and supply for each commodity type often result in an accumulation of empty vehicles in regions where they are not needed and in deficits of vehicles in other regions that require them. Then, vehicles must be moved empty, or additional loads must be found, in order to bring them where they will be needed to satisfy known and forecasted demand in the following planning periods. This operation is known as *repositioning* and is a major component of what is known as *fleet management.* In its most general form, fleet management covers the whole range of planning and management issues from procurement of power units and vehicles to vehicle dispatch and scheduling of crews and maintenance operations. Often, however, the term designates a somewhat restricted set of activities: allocation of vehicles to customer requests and repositioning of empty vehicles.

Moving vehicles empty does not directly contribute to the profit of the firm but it is essential to its continuing operations. Consequently, one attempts to minimize empty movements within the limits imposed by the demand and service requirements. *Empty balancing,* the distribution of empty vehicles to balance the supply and demand in future periods, is a major objective of dispatchers and a central component of planning and operations of most transportation firms. This issue must also be considered at the tactical level. In rail transportation, for example, empty rail cars are put on the same trains as loaded ones and thus contribute to an increase in the number of trains, in the volume of vehicles handled in terminals and, ultimately, in system costs and delays. For planning purposes, the demand for empty cars may be approximated and introduced in tactical model by viewing empties as another commodity to be transported (e.g., Crainic, Ferland, and Rousseau 1984). A similar approach may also be used for the planning of multimodal regional or national systems (e.g., Crainic, Florian, and Léal 1990). The issue is also relevant in LTL trucking where empty balancing is an integral part of a transportation plan. In this case, a load plan is first obtained for the actual traffic demands, and an empty balancing model is then solved to reposition the empties (see Delorme and Roy 1989 and Braklow *et al.* 1992, for example).

Numerous studies reflect the significant research and development effort that has been dedicated to empty vehicle distribution issues. Interested readers may start exploring this field with the review of Dejax and Crainic (1987). It includes contributions going back to the '60s and spans the whole spectrum of modeling approaches from simple static transport models to formulations that integrate the dynamic and stochastic characteristics of the problem. In the following, we recall some of the main articles and models in this field.

The first empty vehicle allocation models used straightforward transportation formulations (e.g., Leddon and Wrathall 1967, Misra 1972, Baker 1977). Given estimations of future supply and demand of empty cars of a homogeneous fleet at the yards of the network, and the cost in car-hours, for each pair of yards, the distribution of empty cars is optimized to minimize the total cost.

A significant step forward in modeling capabilities was achieved with the explicit consideration of the time perspective. A space-time diagram represents the various paths that vehicles may travel to reach their proper destination at a specified time (Figure 13.3 illustrates such a network). The resulting formulation takes the form of a deterministic dynamic transshipment network model, where flows are optimized such that either the total cost is minimized, or the profitability of the system is maximized. Starting with the pioneering contributions of White

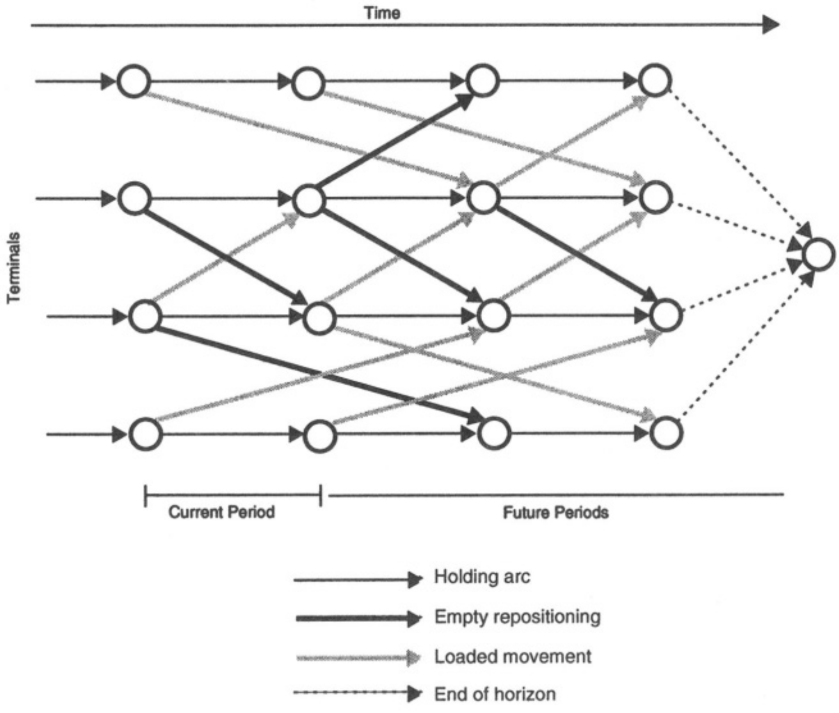


Figure 13.3 Dynamic Network Representation of Operations

(1968) and White and Bomberault (1969) for rail car distribution, and of White (1972) for container allocation, many models that aimed for the distribution of empty vehicles, took the form of a dynamic transshipment network optimization problem (e.g., Herren 1973, 1977, McGaughey, Gohring, and McBrayer 1973). Linear programming and network flow algorithms were usually applied. This line of research is still very active today. The formulations are more complex, though. Multiple commodities, substitutions, integer flows, are some of the characteristics that add realism to these formulations (Shan 1985, Chih 1986, Turnquist and Markowicz 1989, Markowicz and Turnquist 1990, Turnquist 1994; and others). Alternatively, the strict schedules and booking policies enforced by many European railways impose additional conditions on empty vehicle distribution, such as limited hauling capacity for empties, and pre-defined itineraries (Joborn 1995, Holmberg, Joborn, and Lundgren 1998, Joborn *et al.* 2001).

The explicit consideration of uncertainties in empty vehicle distribution models constitutes another significant methodological contribution. The first comprehensive effort in this direction was made by Jordan and Turnquist (1983) for rail. The formulation aims to maximize the profits of the firm, and integrates revenues from performing the service as well as various costs from moving cars between yards, holding them at yards, or from not filling orders due to stockouts. The model structure is again a multicommodity, dynamic network. Stochasticity of supply, demand, and travel times is explicitly considered. The resulting model is a nonlinear optimization formulation, solved by using the Frank-Wolfe algorithm (1956). A similar approach is proposed by Beaujon and Turnquist (1991) for a model that simultaneously considers vehicle inventories at terminals and their allocation in order to answer fleet-sizing issues. The whole research area addressing the dynamic allocation of limited resources in uncertain environments naturally continues these important developments.

Dynamic Allocation of Resources

Many operational problems, fleet management in particular, dynamically allocate limited resources to requests and tasks. For example, empty vehicles, trailers and rail cars are allocated to the appropriate terminals; motive power tractors and locomotives to services; crews to vehicles or services; loads to driver-truck combinations; empty containers from depots to customers and returning containers from customers to depots; and so on. All these problems have several common characteristics:

1. Some future demands are known, but most can only be forecasted, and unpredictable requests may happen.

2. Many requests materialize in real or quasi-real time and must be acted upon in relatively short time.
3. Once a resource is allocated to an activity, it is no longer available for a certain duration (whose length may be subject to variations as well).
4. Once a resource becomes available again, it is often in a different location than its initial one.
5. The value of an additional unit of a given resource at a location greatly depends on the total quantity of resources available (which are determined from previous decisions at potentially all terminals in previous periods) and the current demand.

This is an extremely rich field both for research and development and for applications. In a sense, it extends and complements the empty vehicle distribution problems described previously. The latest developments in the field also allow to plan and control the activities of several resources simultaneously (Powell 1996b, 1998, Powell and Carvalho 1998b, Powell and Shapiro 2001). Dynamic and stochastic network formulations have been, and continue to be, extensively studied for these problems. This has resulted in important modeling and algorithmic results. A number of these results have been transferred to industry (Powell *et al.* 1992, for example). The interested reader should consult the excellent synthesis and review by Powell, Jaillet, and Odoni (1995) and the numerous references quoted in this work. In the following, we briefly illustrate two main modeling approaches.

One may represent dynamic allocation issues by an activity graph similar to the one displayed in Figure 13.3. Here, the operations of a simple four-terminal system are schematically drawn for a certain length of time, which is arbitrarily divided into three periods. At each terminal, there are a number of vehicles that are available to satisfy customer requests during the current period and in future ones. Customer demands have precise characteristics, such as the origin and destination of movement, and pickup and delivery dates (with time window restrictions, eventually). At any period, a vehicle may be assigned to a customer demand in the current period and at the current location, moved to another location to satisfy a known future request, held at the current location, or moved empty to another location in preparation for future, forecasted demands.

Accepting requests and performing the corresponding movements implies expenses and generates revenues. Crainic, Gendreau, and Dejax (1993) developed a model for the assignment and management of a heterogenous fleet of containers where loaded movements are exogenously accepted. Here, the objective is to minimize the total operating cost, including substitutions and stockouts. Several other models also address the issue of whether a request is profitable with respect to the operation of the system and should therefore be accepted. Indeed, repositioning empty vehicles does not generate any immediate revenues. One may be ready to incur these expenses, however, in the hope that, as a consequence, vehicles will be adequately posted to take advantage of future (known, forecasted or estimated) opportunities. Refused requests represent lost business opportunities, while accepted but unsatisfied ones generally result in penalties. Powell *et al.* (1992) and Powell (1996a) present such applications to truckload motor carrier transportation.

A classical modeling approach for this class of problems is to consider the entire planning horizon with the objective of maximizing the *total system profit* computed as the sum of the profit resulting from decisions taken for the current period, plus the expected profit over future periods. The usual constraints apply: satisfy the demand; do not use more than the number of available vehicles; adhere to specific operations rules; etc. When the state of the system and its environment in future periods is known, or assumed to be known, the resulting formulation is deterministic and is often written as a network flow optimization model with additional constraints.

The major difficulty with this approach becomes apparent when the uncertainties in future demands, as well as, eventually, uncertainties related to performing the operations, are explicitly considered. In this case, decisions taken "now" for future periods cannot be based on sure data, but only on estimations of how the system will evolve, which demand will materialize, and so on. From a mathematical programming point of view, random variables are used to represent the stochastic elements and decisions in future periods. Consequently, the expectation of future profits that appears in the objective function of the model becomes a very complex, recursive stochastic equation where the statistical expectation of the total profit must be computed over all possible realizations of all random variables.

To address this complex issue, the model generally takes the form of a recourse formulation. Such formulations are based on the idea that today's decisions are taken within today's deterministic context but using an estimation of the variability of the random factors, and that their consequences are reflected in later decisions. The recourse represents these later decisions which must be taken to adjust the initial policies once the actual realization of the random variables is observed. In the simplest possible recourse formulation, called *simple recourse,* it is assumed that one does not attempt to change the decisions but pays a penalty when the observed value of a random variable is different from the estimation. More complex formulations, such as nodal, tree, and network recourse attempt to evaluate the possible modifications to the initial decisions, and the impact on the total expected profit. Refer to Powell (1988), Frantziskakis (1990), Powell and Frantzeskakis (1994), and Powell and Cheung (1994a,b) for details. Powell (1987), Frantziskakis and Powell (1990), Cheung and Powell (1996a), Chen and Powell (1999), Powell and

Cheung (2000) extend the recourse methodology and present increasingly more complex and precise methods to approximate the recourse function of multistage, dynamic, stochastic networks. An excellent analysis of the application of these approaches to the dynamic fleet management problems for truckload motor carriers, as well as a discussion of the merits and difficulties of stochastic formulations, may be found in Powell (1996a). Cheung and Powell (1996b) further compare these approaches in the context of dynamic distribution problems. Cheung and Chen (1998) apply the same type of methodology to the problem of distributing empty containers in an international maritime system.

These formulations, which are generally difficult to solve, also make use of various criteria to discretize, aggregate, and end time. For example, in Figure 13.3, the theoretically infinite future planning and operation horizon has been reduced to three periods. When the recourse formulation is solved, the periods could be further aggregated, all future periods being considered as one; this corresponds to a two-period formulation, as opposed to n-period, otherwise. Then, in actual applications, the models are used in a rolling horizon environment where, as time advances, a new period is added at the end of the horizon. An important issue is then how to approximate what happens in all the periods beyond the artificially fixed end of the horizon, and how to integrate this approximation into the recourse function. Powell, Jaillet, and Odoni (1995) present an excellent review of this class of formulations.

A different approach recently championed by Powell (1995), Powell *et al.* (1995), and Powell and Carvalho (1998a; see also Carvalho 1996, Carvalho and Powell 2000) addresses resource allocation problems as *Logistic Queueing Networks, LQN.* In this case, at each node of the time-space diagram there are two queues: one of resources and one of tasks requesting resources. Figure 13.4 illustrates a possible configuration for two terminals over two periods. Two "resources" are managed, vehicles and loads, and their levels currently known or approximated at each terminal are displayed. Available vehicles may be allocated to loads already at terminals. Arrows illustrate other possible actions: move loaded vehicles from one terminal to another, where they will increase the inventory of empty vehicles; hold empty vehicles for use in subsequent periods; move empty vehicles to reposition them at a different terminal; determine where to send a vehicle that becomes empty and what vehicles from which terminal to assign to new loads. The objective is to maximize the total profit generated by operating the system to satisfy demand.

The basic idea of the *LQN* methodology is to cast the formulation as a recursive dynamic model and to decompose the resulting optimization problem for each period into "easy-to-solve" local subproblems. In the applications described, each subproblem corresponds to the assignment of vehicles to tasks (loaded or empty movements, for example) at a given terminal. But, in order to evaluate the worth

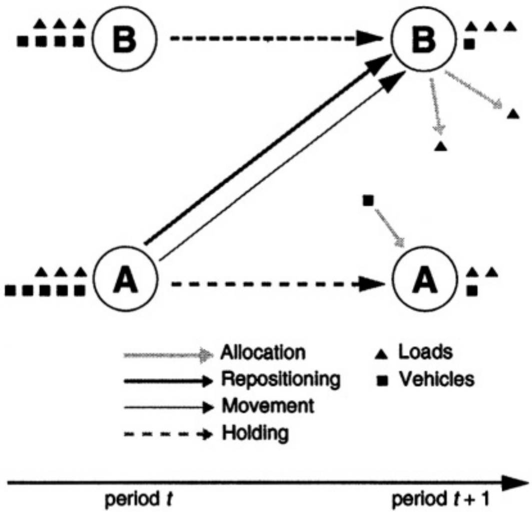


Figure 13.4 Logistics Queuing Network

of allocating a vehicle to a loaded or repositioning movement, one has to know, or evaluate, not only the operating costs and the profit of the load, but also the value of the empty vehicle at the destination terminal. Furthermore, the dynamics of the system make this value depend on future decisions at all terminals. At each period and for each vehicle type, these values are approximated by measuring how desirable it is to have one more vehicle at each terminal. The resulting *potentials* are then used to build a linear approximation of the part of the recursive objective function that corresponds to future periods. Gradients of this approximated objective functions with respect to the supply of vehicles at terminals are used to adjust the potentials, as well as the upper limits on empty movements.

The general solution approach proceeds iteratively in a series of forward and backward passes along the time axis. At each iteration, the forward pass assigns vehicles to tasks, the backward pass computes gradients, and a control adjustment phase modifies the potentials and the bounds on empty movements. The process continues until "convergence" is ensured. The latest developments (Godfrey and Powell 2002a,b) use nonlinear approximations and present truly impressive results for realistic fleet management applications.

The LQN approach appears to offer a very interesting framework for a wide variety of real situations that may be efficiently represented and solved. It offers, in particular, a rather straightforward way to explicitly take into account various considerations, such as time windows, labor restrictions and substitutions, by addressing them at the level of the local subproblem. The application of LQN methodology to the real-time management of fleets of containers and flatcars for intermodal operations presented by Powell and Charvalho (1998b) offers very encouraging results both in terms of actual results (significant savings in operation costs are forecasted) and of further applications to resource allocation problems and service network design models.

13.7 Perspectives

We have presented a number of major issues, models, and methodologies in long distance freight transportation planning and management. Many significant methodological advances have been achieved and many have been successfully transferred to actual practice. However, many research opportunities and challenges still exist.

The advent of *Intelligent Transportation Systems (ITS)* will have a tremendous impact on the planning and operations of freight transportation. ITS technologies increase the flow of available data, improve the timeliness and quality of information, and offer the possibility to control and coordinate operations in real-time. Significant research efforts are required to adequately model the various planning and management problems under ITS and real-time information, and to develop efficient solution methods. Some efforts have already been undertaken relative to the real-time dispatching, assignment, routing, and re-routing of vehicles (Regan 1997, Regan, Mahmassani, and Jaillet 1995, 1996a,b 1998, Yang, Mahmassani, and Jaillet 1998, Yang, Jaillet, and Mahmassani 2002, Gendreau *et al.* 1996, 1998, 1999, Gendreau and Potvin 1998) and the study of the impact of new technologies on the planning and performance of intermodal classification yards (Bostel and Dejax 1998).

The rapid and sustained development of the *electronic business* way of interacting with customers and partners is already modifying how transportation firms plan and operate. In many respects, e-business and ITS are related and the challenges associated to the real-time response mentioned above are also encountered here. E-business also brings (or should bring) easier access to loads through various e-marketplaces. Many of these offer increasingly sophisticated auction mechanisms to determine the allocation of loads and the associated prices. The fleet management models and tools have to integrate these possibilities. A major challenge is related to determining on what loads to bid and the bidding strategy, in particular when loads that would combine in interesting routes must be negotiated separately (e.g., Abrache, Crainic, and Gendreau 2001, Chang, Crainic, and Gendreau 2002, Crainic and Gendreau 2002). It is difficult at this time to adequately predict the whole extend of impact of ITS and e-business on transportation science theory, methods, and practice, but we are convinced that it will be major and comprehensive.

The study of network design formulations and solution methods still offers considerable challenges; from a theoretical point of view, of course, but also when contemplating applications to huge problem instances with very large number of commodities. The same may be said of dynamic and stochastic formulations. In fact, one observes that more and more formulations explicitly consider the dynamic and stochastic characteristics of the problems under study. The trend may be observed not only for issues traditionally associated with actual operations, but also for problems considered "tactical", such as load planning and service network design. Generally speaking, however, the literature does not offer trusted solution methods capable of addressing scheduled (dynamic) service network design problems of realistic dimensions and complexity. The study of the formulations and their properties (e.g., reformulations, bounds, cuts) should be continued. A number of decomposition ideas (according to the time period or node, for example) have also been advanced and are worth investigating. Such approaches will also present "natural" parallelization characteristics that should facilitate the implementation of efficient solution methods.

Metaheuristics play an increasingly important role in obtaining good solutions to difficult problems within reasonable computing times. Work is still needed, however, to develop more efficient and more robust procedures and to better understand the conditions under which each approach performs best. Hybrids, combining characteristics of two or more metaheuristics, offer interesting, but challenging perspectives.

Parallel and distributed computation offers another challenging perspective with potentially great rewards: to solve realistically modeled and dimensioned problem instances within reasonable times. Each class of problems and algorithms presents its own challenges. The parallel exploration of branch-and-bound trees, and the collaborative search undertaken by several metaheuristics or by metaheuristics and exact methods, are only two exciting research areas. Parallel computing also offers the possibility of designing an architecture to efficiently answer complex requests in real, or quasi-real time. These ideas that have just begun to be considered (e.g., Séguin *et al.* 1997), have a great potential for the development of intelligent and efficient decision support tools for ITS and other real-time transportation systems.

Acknowledgments

I wish to express my gratitude toward my colleagues and friends, Professors Bernard Gendron and Gilbert Laporte, who read earlier versions of this paper.

Their comments have helped improve the presentation significantly. I also wish to thank Ms. Lucie Larin for the professional drawing of the figures.

Financial support for this project was provided by N.S.E.R.C. (Canada) and the Fonds F.C.A.R. (Québec).

References

Abrache, J., Crainic, T.G., and Gendreau, M. (2001). Design Issues for Multi-object Combinatorial Auctions, In *Proceedings of The Fourth International Conference on Electronic Commerce Research (ICECR-4),* volume 2, pages 412–423. Cox School of Business, Southern Methodist University, Dallas TX.

Ahuja, R.K. (1997). Flows and Paths. In Dell'Amico, M., Maffioli, F., and Martello, S., editors, *Annotated Bibliographies in Combinatorial Optimization,* pages 283–309. John Wiley & Sons, New York, NY.

Ahuja, R.K., Magnanti, T.L., Orlin, J.B., and Reddy, M.R. (1995). Applications of Networks Optimization. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Network Models,* volume 7 of *Handbooks in Operations Research and Management Science,* pages 1–83. North-Holland, Amsterdam.

Armacost, A.P., Barnhart, C., and Ware, K.A. (2002). Composite Variable Formulations for Express Shipment Service Network Design. *Transportation Science,* 36(1).

Assad, A.A. (1980). Models for Rail Transportation. *Transportation Research A: Policy and Practice,* 14A:205–220.

Baker, L. (1977). Overview of Computer Based Models Applicable to Freight Car Utilization. Report prepared for U.S. Department of Transportation, NTIS, Sprinfield VA, U.S.A.

Balakrishnan, A., Magnanti, T.L., and Mirchandani, P. (1997). Network Design. In Dell'Amico, M., Maffioli, F., and Martello, S., editors, *Annotated Bibliographies in Combinatorial Optimization,* pages 311–334. John Wiley & Sons, New York, NY.

Balakrishnan, A., Magnanti, T.L., and Wong, R.T. (1989). A Dual-Ascent Procedure for Large-Scale Uncapacitated Network Design. *Operations Research,* 37(5):716–740.

Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors (1995). *Network Routing,* volume 8 of *Handbooks in Operations Research and Management Science.* North-Holland, Amsterdam.

Barnhart, C., Hane, C.A., and Vance, P.H. (2000a). Using Branch-and-Price-and-Cut to Solve Origin-Destination Integer Multicommodity Flow Problems. *Operations Research,* 48(2):318–326.

Barnhart, C., Jin, H., and Vance, P.H. (2000b). Railroad Blocking: A Network Design Application. *Operations Research,* 48(4):603–614.

Barnhart, C., Johnson, E.L., Nemhauser, G.L., Savelsbergh, M.W.F., and Vance, P.H. (1998). Branch-and-Price: Column Generation for Solving Huge Integer Programs. *Operations Research,* 46(3):316–329.

Barnhart, C. and Schneur, R.R. (1996). Network Design for Express Freight Service. *Operations Research,* 44(6):852–863.

Barnhart, C. and Talluri, K. (1997). Airlines Operations Research. In A.E. McGarity and C. ReVelle, editors, *Civil and Environmental Engineering Systems: An Advanced Applications Text.* John Wiley, New York.

Beaujon, G.J. and Turnquist, M.A. (1991). A Model for Fleet Sizing and Vehicle Allocation. *Transportation Science,* 25(1):19–45.

Berger, D., Gendron, B., Potvin, J.-Y, Raghavan, S., and Soriano, P. (2000). Tabu Search for a Network Loading Problem with Multiple Facilities. *Journal of Heuristics,* 6(2):253–267.

Bostel, N. and Dejax, P. (1998). Models and Algorithms for Container Allocation Problems on Trains in a Rapid Transshipment Shunting Yard. *Transportation Science,* 32(4):370–379.

Braklow, J.W., Graham, W.W., Hassler, S.M., Peck, K.E., and Powell, W.B. (1992). Interactive Optimization Improves Service and Performance for Yellow Freight System. *Interfaces,* 22(1):147–172.

Buedenbender, K., Grünert, T., and Sebastian, H.-J. (2000). A Hybrid Tabu Search/Branch and Bound Algorithm for the Direct Flight Network Design Problem. *Transportation Science,* 34(4):364–380.

Camerini, P.M., Fratta, L., and Maffioli, F. (1978). On Improving Relaxation Methods by Modified Gradient Techniques. *Mathematical Programming Study,* 3:26–34.

Carvalho, T.A. (1996). *A New Approach to Solving Dynamic Resource Allocation Problems.* PhD thesis, Department of Civil Engineering and Operations Research, Princeton University, Princeton, NJ, U.S.A.

Carvalho, T.A. and Powell, W.B. (2000). A Multiplier Adjustment Method for Dynamic Resource Allocation Problems. *Transportation Science,* 34(2):150–164.

Cascetta, E. (2001). *Transportation Systems Engineering: Theory and Methods.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Chang, T.S., Crainic, T.G., and Gendreau, M. (2001). Designing Advisors for Fleet Management and E-Auctions. Technical report, Centre de recherche sur les transports, Université de Montréal, Canada.

Chen, B. and Harker, P.T. (1990). Two Moments Estimation of the Delay on a Single-Track Rail Line with Scheduled Traffic. *Transportation Science,* 24(4):261–275.

Chen, Z.-L. and Powell, W.B. (1999). A Convergent Cutting-planne and Partial-sampling Algorithm for Multistage Linear Programs with Recourse. *Journal of Optimization Theory and Applications,* 103(3):497–524.

Cheung, R.K. and Chen, C.-Y. (1998). A Two-Stage Stochastic Network Model and Solution Methods for the Dynamic Empty Container Allocation Problem. *Transportation Science,* 32(2):142–162.

Cheung, R.K.-M. and Powell, W.B. (1996a). An Algorithm for Multistage Dynamic Networks with Random Arc Capacities, with an Application to Dynamic Fleet Management. *Operations Research,* 44(6):951–963.

Cheung, R.K.-M. and Powell, W.B. (1996b). Models and Algorithms for Distribution Problems with Uncertain Demands. *Transportation Science,* 30(1):43–59.

Chih, K. C.-K. (1986). *A Real Time Dynamic Optimal Freight Car Management Simulation Model of the Multiple Railroad, Multicommodity, Temporal Spatial Network Flow Problem.* PhD thesis, Princeton University, Princeton, NJ, U.S.A.

Chouman, M., Crainic, T.G., and Gendron, B. (2001). Revue des inégalités valides pertinentes aux problèmes de conception de réseaux. Publication CRT-2001-38, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Chouman, M., Crainic, T.G., and Gendron, B. (2002). Valid Inequalities for Multicommodity Caparitated Fixed Charge Network Design. Technical report, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Cordeau, J.-F, Toth, P., and Vigo, D. (1998). A Survey of Optimization Models for Train Routing and Scheduling. *Transportation Science,* 32(4):380–404.

Crainic, T.G. (1982). *Un modèle de plannification tactique pour le transport ferroviaire des marchandises,* PhD thesis, Université de Montréal, Montréal, QC, Canada.

Crainic, T.G. (1984). A Comparison of Two Methods for Tactical Planning in Rail Freight Transportation. In J.P. Brans, editor, *Operational Research'84,* pages 707–720. North-Holland, Amsterdam.

Crainic, T.G. (1988). Rail Tactical Planning: Issues, Models and Tools. In Bianco, L. and Bella, A.L., editors, *Freight Transport Planning and Logistics,* pages 463–509. Springer-Verlag, Berlin.

Crainic, T.G. (2002). Parallel Computation, Co-operation, Tabu Search. In Rego, C. and Alidaee, B., editors, *Adaptive Memory and Evolution: Tabu Search and Scatter Search.* Kluwer Academic Publishers, Norwell, MA.

Crainic, T.G., Dufour, G., Florian, M., and Larin, D. (1999). Path Analysis in STAN. In C. Zopounidis and D. Despotis, editors, *Proceedings 5th International Conference of the Decision Sciences Institute,* pages 2060–2064. New Technologies Publications, Athens, Greece,

Crainic, T.G., Dufour, G., Florian, M., and Larin, D. (2002). Path Recovery/Reconstruction and Applications in Nonlinear Multimodal Multicommodity Networks. In Gendreau, M. and Marcotte, P., editors, *Transportation and Network Analysis – Current Trends.* Kluwer Academic Publishers, Norwell MA.

Crainic, T.G., Ferland, J.-A., and Rousseau, J.-M. (1984). A Tactical Planning Model for Rail Freight Transportation. *Transportation Science,* 18(2): 165–184.

Crainic, T.G., Florian, M., Guélat, J., and Spiess, H. (1990a). Strategic Planning of Freight Transportation: STAN, An Interactive-Graphic System. *Transportation Research Record,* 1283:97–124.

Crainic, T.G., Florian, M., and Larin, D. (1994). STAN: New Developments. In A1 S. Khade and R. Brown, editors, *Proceedings of the 23rd Annual Meeting of the Western Decision Sciences Institute,* pages 493–498. School of Business Administration, California State University, Stanislaus CA.

Crainic, T.G., Florian, M., and Léal, J.-E. (1990b). A Model for the Strategic Planning of National Freight Transportation by Rail. *Transportation Science,* 24(1):1–24.

Crainic, T.G., Frangioni, A., and Gendron, B. (2001). Bundle-Based Relaxation Methods for Multicommodity Capacitated Network Design. *Discrete Applied Mathematics,* 112:73–99.

Crainic, T.G. and Gendreau, M. (1986). Approximate Formulas for the Computation of Connection Delays under Capacity Restrictions in Rail Freight Transportation, In *Research for Tomorrow's Transport Requirements,* volume 2, pages 1142–1155. Fourth World Conference on Transport Research, Vancouver, Canada.

Crainic, T.G. and Gendreau, M. (2002a). Cooperative Parallel Tabu Search for Capacitated Network Design. *Journal of Heuristics.*

Crainic, T.G. and Gendreau, M. (2002b). Freight Exchanges and Carrier Operations: Issues, Models, and Tools. Technical report, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada,

Crainic, T.G., Gendreau, M., and Dejax, P.J. (1993). Dynamic Stochastic Models for the Allocation of Empty Containers. *Operations Research,* 43:102–302.

Crainic, T.G., Gendreau, M., and Farvolden, J.M. (2000). A Simplex-Based Tabu Search Method for Capacitated Network Design. *INFORMS Journal on Computing,* 12(3):223–236.

Crainic, T.G. and Laporte, G. (1997). Planning Models for Freight Transportation. *European Journal of Operational Research,* 97(3):409–438.

Crainic, T.G. and Nicolle, M.-C. (1986). Planification tactique du transport ferroviaire des marchandises: quelques aspects de modélisation. In *Actes du Premier Congrès International en France de Génie Industriel,* pages 161–174. CEFI-AFCET-GGI, Paris. Tome 1.

Crainic, T.G. and Rousseau, J.-M. (1986). Multicommodity, Multimode Freight Transportation: A General Modeling and Algorithmic Framework for the Service Network Design Problem. *Transportation Research B: Methodology,* 208:225–242.

Crainic, T.G. and Roy, J. (1988). O.R. Tools for Tactical Freight Transportation Planning. *European Journal of Operational Research,* 33(3):290–297.

Crainic, T.G. and Roy, J. (1990). Une approche de recouvrement d'ensembles pour 1'établissement d'horaires des chauffeurs dans le transport routier de charges partielles. *R.A.I.R.O. – Recherche opérationnelle,* 24(2):123–158.

Crainic, T.G. and Roy, J. (1992). Design of Regular Intercity Driver Routes for the LTL Motor Carrier Industry. *Transportation Science,* 26(4):280–295.

Crainic, T.G. and Toulouse, M. (2002). Parallel Strategies for Meta-heuristics. In F. Glover and G. Kochenberger, editors, *State-of-the-Art Handbook in Metaheuristics.* Kluwer Academic Publishers, Norwell, MA.

Crowder, H. (1976). Computational Improvements for Subgradient Optimization. In *Symposia Mathematica,* volume XIX. Academic Press, London.

Daganzo, C.F. (1987a). Dynamic Blocking for Railyards: Part I. Homogeneous Traffic. *Transportation Research B: Methodological,* 21B(1):1–27.

Daganzo, C.F. (1987b). Dynamic Blocking for Railyards: Part II. Heterogeneous Traffic. *Transportation Research B: Methodological,* 21B(1):29–40.

Daskin, M.S. (1995). *Network and Discrete Location. Models, Algorithms, and Applications.* John Wiley & Sons, New York, NY.

Daskin, M.S. and Walton, C.M. (1983). An Approximate Analytic Model of Supertanker Lightering Operations. *Transportation Research B: Methodological,* 17B(3):201–219.

Dejax, P.J. and Crainic, T.G. (1987). A Review of Empty Flows and Fleet Management Models in Freight Transportation. *Transportation Science,* 21(4):227–247.

Delorme, L., Roy, J., and Rousseau, J.-M. (1988). Motor-Carrier Operation Planning Models: A State of the Art. In Bianco, L. and Bella, A.L,, editors, *Freight Transport Planning and Logistics,* pages 510–545. Springer-Verlag, Berlin.

Desaulniers, G., Desrosiers, J., Gamache, M., and Soumis, F. (1998a). Crew Scheduling in Air Transportation. In T.G. Crainic and G. Laporte, editors, *Fleet Management and Logistics,* pages 169–185. Kluwer Academic Publishers, Norwell, MA.

Desaulniers, G., Desrosiers, J., Ioachim, I., Solomon, M.M., Soumis, F., and Villeneuve, D. (1998b). A Unified Framework for Deterministic Time Constrained Vehicle Routing and Crew Scheduling Problems. In T.G. Crainic and G. Laporte, editors, *Fleet Management and Logistics,* pages 57–93. Kluwer Academic Publishers, Norwell, MA.

Desrosiers, J., Dumas, Y., Solomon, M.M., and Soumis, F. (1995). Time Constrained Routing and Scheduling. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Network Routing,* volume 8 of *Handbooks in Operations Research and Management Science,* pages 35–139. North-Holland, Amsterdam.

Drezner, Z., editor (1995). *Facility Location. A Survey of Applications and Methods.* Springer-Verlag, New York, NY.

Dror, M., editor (2000). *Arc Routing: Theory, Solutions and Applications.* Kluwer Academic Publishers, Norwell, MA.

Epstein, R. (1998). *Linear Programming and Capacitated Network Loading.* PhD thesis, Sloan School of Management, Massachusetts Institute of Technology.

Equi, L., Gallo, G., Marziale, S., and Weintraub, A. (1997). A Combined Transportation and Scheduling Problem. *European Journal of Operational Research,* 97(1):94–104.

Farvolden, J.M. and Powell, W.B. (1991). A Dynamic Network Model for Less-Than-Truckload Motor Carrier Operations. Working Paper 90-05, Department of Industrial Engineering, University of Toronto, Toronto, ON, Canada.

Farvolden, J.M. and Powell, W.B. (1994). Subgradient Methods for the Service Network Design Problem. *Transportation Science,* 28(3):256–272,

Farvolden, J.M., Powell, W.B., and Lustig, I.J. (1992). A Primal Partitioning Solution for the Arc–Chain Formulation of a Multicommodity Network Flow Problem. *Operations Research,* 41(4):669–694.

Florian, M. and Hearn, D. (1995). Networks Equilibrium Models and Algorithms. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Network Routing,* volume 8 of *Handbooks in Operations Research and Management Science,* pages 485–550. North-Holland, Amsterdam.

Franck, M. and Wolfe, P. (1956). An Algorithm for Quadratic Programming. *Naval Research Logistics Quaterly,* 3:95–110.

Frantzeskakis, L. (1990). *Dynamic Networks with Random Arc Capacities, with Application to the Stochastic Dynamic Vehicle Allocation Problem.* PhD thesis, Department of Civil Engineering and Operation Research, Princeton University, Princeton, NJ, U.S.A.

Frantzeskakis, L. and Powell, W.B. (1990). A Successive Linear Approximation Procedure for Stochastic Dynamic Vehicle Allocation Problems. *Transportation Science,* 24(1):40–57.

Friesz, T.L., Gottfried, J.A., and Morlok, E.K. (1986). A Sequential Shipper-Carrier Network Model for Predicting Freight Flows. *Transportation Science,* 20:80–91.

Friesz, T.L., Tobin, R.L., and Harker, P.T. (1983). Predictive Intercity Freight Network Models. *Transportation Research A: Policy and Practice,* 17A:409–417.

Friesz, T.L. and Harker, P.T. (1985). Freight Network Equilibrium: A Review of the State of the Art. In A.F. Daughety, editor, *Analytical Studies in Transport Economics,* chapter 7. Cambridge University Press, Cambridge.

Gendreau, M., Badeau, P., Guertin, F., Potvin, J.-Y, and Taillard, É.D. (1996). A Solution Procedure for Real-Time Routing and Dispatching of Commercial Vehicles. In *Third Annual World Congress on Intelligent Transportation Systems.* distributed on CD.

Gendreau, M., Guertin, F., Potvin, J.-Y, and Séguin, R. (1998). Neighborhood Search Heuristics for a Dynamic Vehicle Dispatching Problem with Pick-Ups and Deliveries. Publication CRT-98-10, Centre de recherche sur les transports, Université de Montréal.

Gendreau, M., Guertin, F., Potvin, J.-Y, and Taillard, É.D. (1999). Tabu Search for Real-Time Vehicle Routing and Dispatching. *Transportation Science,* 33(4):381–390.

Gendreau, M. and Potvin, J.-Y. (1998). Dynamic Vehicle Routing and Dispatching. In T.G. Crainic and G. Laporte, editors, *Fleet Management and Logistics,* pages 115–126. Kluwer Academic Publishers, Norwell, MA.

Gendron, B. and Crainic, T.G. (1994). Relaxations for Multicommodity Network Design Problems. Publication CRT-965, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Gendron, B. and Crainic, T.G. (1994a). Parallel Branch-and-Bound Algorithms: Survey and Synthesis. *Operations Research,* 42(6):1042–1066.

Gendron, B. and Crainic, T.G. (1994b). Parallel Implementations of Bounding Procedures for Multicommodity Capacitated Network Design Problems. Publication CRT-94-45, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Gendron, B. and Crainic, T.G. (1996). Bounding Procedures for Multicommodity Capacitated Network Design Problems. Publication CRT-96-06, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Gendron, B., Crainic, T.G., and Frangioni, A. (1998). Multicommodity Capacitated Network Design. In Sansó, B. and Soriano, P., editors, *Telecommunications Network Planning,* pages 1–19. Kluwer Academic Publishers, Norwell, MA.

Geoffrion, A.M. (1974). Lagrangean Relaxation for Integer Programming. *Mathematical Programming Study,* 2:82–114.

Ghamlouche, I., Crainic, T.G., and Gendreau, M. (2002a). Cycle-based Neighbourhoods for Fixed-Charge Capacitated Multicommodity Network Design. *Operations Research.*

Ghamlouche, I., Crainic, T.G., and Gendreau, M. (2002b). Path Relinking, Cycle-based Neighbourhoods and Capacitated Multicommodity Network Design. *Annals of Operations Research.*

Glover, F. and Laguna, M. (1993). Tabu search. In Reeves, C., editor, *Modern Heuristic Techniques for Combinatorial Problems,* pages 70–150. Blackwell Scientific Publications, Oxford.

Glover, F. and Laguna, M. (1997). *Tabu Search.* Kluwer Academic Publishers, Norwell, MA.

Godfrey, G.A. and Powell, W.B. (2002a). An Adaptive Dynamic Programming Algorithm for Dynamic Fleet Management I: Single Period Travel Times. *Transportation Science,* 36(1).

Godfrey, G.A. and Powell, W.B. (2002b). An Adaptive Dynamic Programming Algorithm for Dynamic Fleet Management II: Multiperiod Travel Times. *Transportation Science,* 36(1).

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, MA.

Golden, B.L. and Assad, A.A., editors (1988). *Vehicle Routing: Methods and Studies.* North-Holland, Amsterdam.

Gorman, M.F. (1998a). An Application of Genetic and Tabu Searches to the Freight Railroad Operating Plan Problem. *Annals of Operations Research,* 78:51–69.

Gorman, M.F. (1998b). Santa Fe Railway Uses an Operating-Plan Model to Improve Its Service Design. *Interfaces,* 28(4):1–12.

Grötschel, M., Monma, C.L., and Stoer, M. (1995). Design of Survivable Networks. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Network Models,* volume 7 of *Handbooks in Operations Research and Management Science,* pages 617–672. North-Holland, Amsterdam.

Grünert, T. and Sebastian, H.-J. (2000). Planning Models for Long-haul Operations of Postal and Express Shipment Companies. *European Journal of Operational Research,* 122:289–309.

Grünert, T., Sebastian, H.-J., and Thäerigen, M. (1999). The Design of a Letter-Mail Transportation Network by Intelligent Techniques. In Sprague, R., editor, *Proceedings Hawaii International Conference on System Sciences 32.*

Guélat, J., Florian, M., and Crainic, T.G. (1990). A Multimode Multiproduct Network Assignment Model for Strategic Planning of Freight Flows. *Transportation Science,* 24(1):25–39.

Haghani, A.E. (1989). Formulation and Solution of Combined Train Routing and Makeup, and Empty Car Distribution Model. *Transportation Research B: Methodological,* 23B(6):433–431.

Hallowell, S.F. and Harker, P.T. (1996). Predicting On-Time Line-Haul performance in Scheduled Railroad Operations. *Transportation Science,* 30(4):364–378.

Harker, P.T. (1987). *Predicting Intercity Freight Flows.* VNU Science Press, Utrech, The Netherlands.

Harker, P.T. (1988). Issues and Models for Planning and Regulating Freight Transportation Systems. In Bianco, L. and Bella, A. L., editors, *Freight Transport Planning and Logistics,* pages 374–408. Springer-Verlag, Berlin.

Harker, P.T. and Friesz, XL. (1986a). Prediction of Intercity Freight Flows I: Theory. *Transportation Research B: Methodological,* 20B(2):139–153.

Harker, P.T. and Friesz, T.L. (1986b). Prediction of Intercity Freight Flows II: Mathematical Formulations. *Transportation Research B: Methodological,* 20B(2):155–174.

Harker, P.T. and Hong, S. (1990). Two Moments Estimation of the Delay on a Partially Double-Track Rail Line with Scheduled Traffic. *Journal of the Transportation Research Forum,* 31(l):38–49.

Herren, H. (1973). The Distribution of Empty Wagons by Means of Computer. An Analytical Model for the Swiss Federal Railways (SSB). *Rail International,* 4(1):1005–1010.

Herren, H. (1977). Computer Controlled Empty Wagon Distribution on the SSB. *Rail International,* 8(l):25–32.

Higgins, A., Ferreira, L., and Kozan, E. (1995). Modeling Single-Line Train Operations. *Transportation Research Record,* 1489:9–16.

Higgins, A., Ferreira, L., and Kozan, E. (1997a). Heuristic Techniques for Single Line Train Scheduling. *Journal of Heuristics,* 3(1):43–63.

Higgins, A., Kozan, E., and Ferreira, L. (1996). Optimal Scheduling of Trains on a Single Line Track. *Transportation Research B: Methodological,* 30B:147–161.

Higgins, A., Kozan, E., and Ferreira, L. (1997b). Modelling the Number and Location of Sidings on a Single Line Track Railway. *Computers & Operations Research,* 3:209–220.

Hiriart-Urruty, J.B. and Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms II.* Springer-Verlag, Berlin.

Hoffman, K.L. and Padberg, M. (1993). Solving Airline Crew-Scheduling Problems by Branch-and-Cut. *Management Science,* 39:657–682.

Holmberg, K., Joborn, M., and Lundgren, J.T. (1998). Improved Empty Freight Car Distribution. *Transportation Science,* 32(2):163–173.

Holmberg, K. and Hellstrand, J. (1998). Solving the Uncapacitated Network Design Problem by a Lagrangian Heuristic and Branch-and-Bound. *Operations Research,* 46(2):247–259.

Holmberg, K. and Yuan, D. (1998). A Lagrangean Approach to Network Design Problems. Report LiTH-MAT-R-1998-01, Department of Mathematics, Linköping Institute of Technology.

Holmberg, K. and Yuan, D. (2000). A Lagrangean Heuristic Based Branch-and-Bound Approach for the Capacitated Network Design Problem. *Operations Research,* 48(3):461–481.

Hurley, W.J. and Petersen, E.R. (1994). Nonlinear Tariffs and Freight Network Equilibrium. *Transportation Science,* 28(3):236–245.

Isard, W. (1951). Interregional and Regional Input-Output Analysis: A Model of a Space-Economy. *The review of Economics and Statistics,* 33:318–328.

Joborn, M. (1995). *Empty Freight Car Distribution at Swedish Railways – Analysis and Optimization Modeling.* PhD thesis, Division of Optimization, Department of Mathematics, University of Linköping, Linköping, Sweden.

Joborn, M., Crainic, T.G., Gendreau, M., Holmberg, K., and Lundgren, J.T. (2001). Economies of Scale in Empty Freight Car Distribution in Scheduled Railways. Publication CRT-2001-12, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Jordan, W.C. and Turnquist, M.A. (1983). A Stochastic Dynamic Network Model for Railroad Car Distribution. *Transportation Science,* 17:123–145.

Jovanović, D. and Harker, P.T. (1991). Tactical Scheduling of Rail Operations: the SCAN I Decision Support System. *Transportation Science,* 25(l):46–64.

Keaton, M.H. (1989). Designing Optimal Railroad Operating Plans: Lagrangian Relaxation and Heuristic Approaches. *Transportation Research B: Methodological,* 23B(6):415–431.

Keaton, M.H. (1991). Service–Cost Tradeoffs for Carload Freight Traffic in the U.S. Rail Industry. *Transportation Research A: Policy and Practice,* 25A(6):363–374.

Keaton, M.H. (1992). Designing Optimal Railroad Operating Plans: A Dual Adjustment Method for Implementing Lagrangian Relaxation. *Transportation Science,* 26:263–279.

Kim, D. and Barnhart, C. (1997). Transportation Service Network Design: Models and Algorithms. Report, Center for Transportation Studies, Massachusetts Institute of Technology.

Kim, D., Barnhart, C., Ware, K., and Reinhardt, G. (1999). Multimodal Express Package Delivery: A Service Network Design Application. *Transportation Science,* 33(4):391–407.

Kraay, D., Harker, P.T., and Chen, B. (1991). Optimal Pacing of Trains in Freight Railroads: Model Formulation and Solution. *Operations Research,* 39(l):82–99.

Kuby, M.J. and Gray, R.G. (1993). The Hub Network Design Problem with Stopovers and Feeders: The Case of Federal Express. *Transportation Research A: Policy and Practice,* 27A(1):1–12.

Laarhoven, P. and Aarts, E.H.L. (1987). *Simulated Annealing: Theory and Applications.* Reidel, Dordrecht.

Labbé, M., Peeters, D., and Thisse, J.-F. (1995). Location on Networks. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Network Routing,* volume 8 of *Handbooks in Operations Research and Management Science,* pages 551–624. North-Holland, Amsterdam.

Labbé, M. and Louveaux, F.V. (1997). Location Problems. In Dell'Amico, M., Maffioli, F., and Martello, S., editors, *Annotated Bibliographies in Combinatorial Optimization,* pages 261–281. John Wiley & Sons, New York, NY.

Lamar, B.W., Sheffi, Y., and Powell, W.B. (1990). A Capacity Improvement Lower Bound for Fixed Charge Network Design Problems. *Operations Research,* 38(4):704–710.

Larin, D., Crainic, T.G., Simonka, G., James-Lefebvre, L., Dufour, G., and Florian, M. (2000). *STAN User's Manual, Release 6.* INRO Consultants, Inc., Montréal, QC, Canada.

Leddon, C.D. and Wrathall, E. (1967). Scheduling Empty Freight Car Fleets on the Louisville and Nashville Railroad. In *Second International Symposium on the Use of Cybernetics on the Railways,* pages 1–6.

Lemaréchal, C. (1989). Nondifferentiable Optimization. In Nemhauser, G.L., Rinnoy Kan, A.H.G., and Todd, M.J., editors, *Optmization,* volume 1 of *Handbooks in Operations Research and Management Science,* pages 529–572. North-Holland, Amsterdam.

Magnanti, T.L., Mirchandani, P., and Vachani, R. (1993). The Convex Hull of Two Core Capacitated Network Design Problems. *Mathematical Programming,* 60:233–250.

Magnanti, T.L., Mirchandani, P., and Vachani, R. (1995). Modeling and Solving the Two-Facility Capacitated Network Loading Problem. *Operations Research,* 43:142–157.

Magnanti, T.L. and Wong, R.T. (1986). Network Design and Transportation Planning: Models and Algorithms. *Transportation Science,* 18(l):l–55.

Markowicz, B.P. and Turnquist, M.A. (1990). Applying the LP Solution to the Daily Distribution of Freight Cars. Technical report, Cornell University.

McGaughey, K., Gohring, W., and McBrayer, R.N. (1973). Planning Locomotive and Caboose Distribution. *Rail International,* 3:151–158.

Minoux, M. (1986). Network Synthesis and Optimum Network Design Problems: Models, Solution Methods and Applications. *Networks,* 19:313–360.

Mirchandani, P.S. and Francis, R.L., editors (1990). *Discrete Location Theory.* John Wiley & Sons, New York, NY.

Misra, S. (1972). Linear Programming of Empty Wagon Disposition. *Rail International,* 3:151–158.

Morlok, E.K. and Peterson, R.B. (1970). A Final Report on a Development of a Geographic Transportation Network Generation and Evaluation Model. In *Proceedings of the Eleventh Annual Meeting,* pages 99–103. Transportation Research Forum.

Nagurney, A. (1993). *Network Economics – A Variational Inequality Approach.* Kluwer Academic Publishers, Norwell, MA.

Nemhauser, G.L. and Wolsey, L.A. (1988). *Integer and Combinatorial Optimization.* Wiley, New York, NY.

Newton, H.N. (1996). *Network Design Under Budget Constraints with Application to the Railroad Blocking Problem.* PhD thesis, Industrial and Systems Engineering, Auburn University, Auburn, Alabama, U.S.A.

Newton, H.N., Barnhart, C., and Vance, P.H. (1998). Constructing Railroad Blocking Plans to Minimize Handling Costs. *Transportation Science,* 32(4):330–345.

Nobert, Y. and Roy, J. (1998). Freight Handling Personnel Scheduling at Air Cargo Terminals. *Transportation Science,* 32(3):295–301.

Pallottino, S. and Scutellà, M.G. (1998). Shortest path Algorithms in Transportation Models: Classical and Innovative Aspects. In Marcotte, P. and Nguyen, S., editors, *Equilibrium and Advanced Transportation Modelling,* pages 245–281. Kluwer Academic Publishers, Norwell, MA.

Petersen, E.R. (1971a). Bulk Service Queues: With Applications to Train Assembly Times. Working Paper 71-2, CIGGT, Queen's University, Kingston, Canada.

Petersen, E.R. (1971b). Queues with Random Batch Size with Applications to Railroad Modeling. Working Paper 71-77, CIGGT, Queen's University, Kingston, Canada.

Petersen, E.R. (1974). Over-the-Road Transit Time for a Single Track Railway. *Transportation Science,* 8(l):65–74.

Petersen, E.R. (1975a). A Primal-Dual Traffic Assignment Algorithm. *Management Science,* 22:87–95.

Petersen, E.R. (1975b). Interference Delays on a Partially Double-Tracked Railway with Intermediate Signalling. In *Proceedings of the Sixteenth Annual Meeting.* Transportation Research Forum.

Petersen, E.R. (1977a). Capacity of a Single Track Rail Line. Working Paper 77–38, CIGGT, Queen's University, Kingston, Canada.

Petersen, E.R. (1977b). Railyard Modeling: Part I. Prediction of Put-through Time. *Transportation Science,* 11(1):37–49.

Petersen, E.R. (1977c). Railyard Modeling: Part II. The Effect of Yard Facilities on Congestion. *Transportation Science,* 11(1):50–59.

Petersen, E.R. (1984). Rail Analysis Interactive Language (RAIL): A Decision Support System. In Florian, M., editor, *Transportation Planning Models,* pages 363–380. North-Holland, Amsterdam.

Petersen, E.R. and Fullerton, H.V., editors (1975). The Railcar Network Model. Working Paper 75–11, CIGGT, Queen's University, Kingston, Canada.

Petersen, E.R. and Taylor, A.J. (1993). A Structured Model for Rail Line Simulation and Optimization. *Transportation Science,* 16(2): 192–206.

Powell, W.B. (1981). *Stochastic Delays in Transportation Terminals: New Results in the Theory and Application of Bulk Queues.* PhD thesis, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, U.S.A.

Powell, W.B. (1986a). A Local Improvement Heuristic for the Design of Less-than-Truckload Motor Carrier Networks. *Transportation Science,* 20(4):246–357.

Powell, W.B. (1986b). Analysis of Vehicle Holding and Cancellation Strategies in Bulk Arrival, Bulk Service Queues. *Transportation Science,* 19:352–377.

Powell, W.B. (1986c). Approximate, Closed Form Moment Formulas for Bulk Arrival, Bulk Service Queues. *Transportation Science,* 20(l):13–23.

Powell, W.B. (1986d). Iterative Algorithms for Bulk Arrival, Bulk Service Queues with Poisson and non-Poisson Arrivals. *Transportation Science,* 20(2):65–79.

Powell, W.B. (1987). An Operational Planning Model for the Dynamic Vehicle Allocation Problem with Uncertain Demands. *Transportation Research B: Methodological,* 21B:217–232.

Powell, W.B. (1988). A Comparative Review of Alternative Algorithms for the Dynamic Vehicle Allocation Problem. In B.L. Golden and A.A. Assad, editor, *Vehicle Routing: Methods and Studies,* pages 249–292. North-Holland, Amsterdam.

Powell, W.B. (1996a). A Stochastic Formulation of the Dynamic Assignment Problem, with an Application to Truckload Motor Carriers. *Transportation Science,* 30(3):195–219.

Powell, W.B. (1996b). Toward a Unified Modeling Framework for Real-Time Logistics Control. *Military Journal of Operations Research,* I(4):69–79.

Powell; W.B. (1998). On Languages for Dynamic Resource Scheduling Problems. In T.G. Crainic and G. Laporte, editor, *Fleet Management and Logistics,* pages 127–157. Kluwer Academic Publishers, Norwell, MA.

Powell, W.B. and Carvalho, T.A. (1998a). Dynamic Control of Logistics Queueing Networks for Large-Scale Fleet Management. *Transportation Science,* 32(2):90–109.

Powell, W.B. and Carvalho, T.A. (1998b). Real-Time Optimization of Containers and Flatcars for Intermodal Operations. *Transportation Science,* 32(2):110–126.

Powell, W.B., Carvalho, T.A., Godfrey, G.A., and Simaõ, H.P. (1995a). Dynamic Fleet Management as a Logistics Queueing Network. *Annals of Operations Research,* 61:165–188.

Powell, W.B. and Cheung, R.K.-M. (1994a). A Network Recourse Decomposition Method for Dynamic Networks with Random Arc Capacities. *Networks,* 24:369–384.

Powell, W.B. and Cheung, R.K.-M. (1994b). Stochastic Programs over Trees with Random Arc Capacities. *Networks,* 24:161–175.

Powell, W.B. and Cheung, R.K.-M. (2000). SHAPE – A Stochastic Hybrid Approximation Procedure for Two-Stage Stochastic Programs. *Operations Research,* 48(l):73–79.

Powell, W.B. and Frantzeskakis, L. (1994). Restricted Recourse Strategies for Dynamic Networks with Random Arc Capacities. *Transportation Science,* 28(l):3–23.

Powell, W.B. and Humblet, P. (1986). Queue Length and Waiting Time Transforms for Bulk Arrival, Bulk Service Queues with a General Control Strategy. *Operations Research,* 34:267–275.

Powell, W.B., Jaillet, P., and Odoni, A. (1995b). Stochastic and Dynamic Networks and Routing. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Network Routing,* volume 8 of *Handbooks in Operations Research and Management Science,* pages 141–295. North-Holland, Amsterdam.

Powell, W.B. and Koskosidis, Y.A. (1992). Shipment Routing Algorithms with Tree Constraints. *Transportation Science,* 26(3):230–245.

Powell, W.B. and Shapiro, J.A. (2001). A Representational Paradigm for Stochastic, Dynamic Resource Transformation Problems. *Annals of Operations Research.*

Powell, W.B. and Sheffi, Y. (1983). The Load-Planning Problem of Motor Carriers: Problem Description and a Proposed Solution Approach. *Transportation Research A: Policy and Practice,* 17(6):471–480.

Powell, W.B. and Sheffi, Y. (1986). Interactive Optimization for Motor Carrier Load Planning. *Journal of Business Logistics,* 7(2):64–90.

Powell, W.B. and Sheffi, Y. (1989). Design and Implementation of an Interactive Optimization System for the Network Design in the Motor Carrier Industry. *Operations Research,* 37(l):12–29.

Powell, W.B., Sheffi, Y., Nickerson, K.S., Butterbaugh, K., and Atherton, S. (1992). Maximizing Profits for North American Van Lines' Truckload Division: A New Framework for Pricing and Operations. *Interfaces,* 18(1):21–41.

Raghavan, S. and Magnanti, T.L. (1997). Network Connectivity, In Dell'Amico, M., Maffioli, F., and Martello, S., editors, *Annotated Bibliographies in Combinatorial Optimization,* pages 335–354. John Wiley & Sons, New York, NY.

Regan, A.C. (1997). *Real-Time Information for Improved Efficiency of Commercial Vehicle Operations.* PhD thesis, University of Texas at Austin, Austin TX, U.S.A.

Regan, A.C., Mahmassani, H.S., and Jaillet, P. (1995). Improving the Efficiency of Commercial Vehicle Operations Using Real-Time Information: Potential Uses and Assignment Strategies. *Transportation Research Record,* 1493:188–198.

Regan, A.C., Mahmassani, H.S., and Jaillet, P. (1996a). Dynamic Decision Making for Commercial Vehicle Operations Using Real-Time Information. *Transportation Research Record,* 1537:91–97.

Regan, A.C., Mahmassani, H.S., and Jaillet, P. (1996b). Dynamic Dispatching Strategies under Real-Time Information for Carrier Fleet Management. In Lesort, J.B., editor, *Transportation and Traffic Theory,* pages 737–756. Pergamon.

Regan, A.C., Mahmassani, H.S., and Jaillet, P. (1998). Evaluation of Dynamic Fleet Management Systems: A Simulation Framework. *Transportation Research Record,* 1645:176–184.

Roy, J. (1984). *Un modèle de planification globale pour le transport routier des marchandises.* PhD thesis, Ecole des Hautes Etudes Commerciales, Université de Montréal, Montréal, Canada.

Roy, J. and Crainic, T.G. (1992). Improving Intercity Freight Routing with a Tactical Planning Model. *Interfaces,* 22(3):31–44.

Roy, J. and Delorme, L. (1989). NETPLAN: A Network Optimization Model for Tactical Planning in the Less-than-Truckload Motor-Carrier Industry. *INFOR,* 27(1):22–35.

Salkin, H.M. and Mathur, K. (1989). *Foundations of Integer Programming.* North-Holland, Amsterdam.

Séguin, R., Potvin, J.YJ., Gendreau, M., Crainic, T.G., and Marcotte, P. (1997). Real-time Decision Problems: An Operational Research Perspective. *Journal of the Operational Research Society,* 48:162–174.

Shan, Y. (1985). *A Dynamic Multicommodity Network Flow Model for Real-Time Optimal Rail Freight Car Management.* PhD thesis, Princeton University, Princeton, NJ, U.S.A.

Toth, P. and Vigo, D., editors (2002). *The Vehicle Routing Problem,* volume 9 of *SIAM Monographs on Discrete Mathematics and Applications.* SIAM.

Turnquist, M.A. (1994). Economies of Scale and Network Optimization for Empty Car Distribution. Technical report, Cornell University.

Turnquist, M.A. and Daskin, M.S. (1982). Queuing Models of Classification and Connection Delay in Railyards. *Transportation Science,* 16(2):207–230.

Turnquist, M.A. and Markowicz, B.P. (1989). An Interactive Microcomputer-Based Planning Model for Railroad Car Distribution. Technical report, Cornell University.

White, W. (1972). Dynamic Transshipment Networks: An Algorithm and its Application to the Distribution of Empty Containers. *Networks,* 2(3):211–236.

White, W.W. (1968). A Program for Empty Freight Car Allocation. Technical Report 360D.29.002, IBM Contributed Program Library, IBM Corporation, Program Information Department, Hawthorne, N.Y.

White, W.W. and Bomberault, A.M. (1969). A Network Algorithm for Empty Freight Car Allocation. *IBM Systems Journal,* 8(2):147–171.

Winston, C. (1983). The Demand for Freight Transportation: Models and Applications. *Transportation Research A: Policy and Practice,* 17A:419–427.

Yang, J., Jaillet, P., and Mahmassani, H.S. (2002). Study of a Real-Time Multi-Vehicle Truckload Pick-Up and Delivery Problem. *Transportation Science,* forthcoming.

Yang, J., Mahmassani, H.S., and Jaillet, P. (1999). On-Line Algorithms for Truck Fleet Assignment and Scheduling under Real-Time Information. *Transportation Research Record,* 1667:107–113.

14 AIRLINE CREW SCHEDULING

Cynthia Barnhart, Amy M. Cohn,
Ellis L. Johnson, Diego Klabjan,
George L. Nemhauser, Pamela H. Vance

14.1 Introduction

Crew Scheduling

Crew scheduling can be defined as the problem of assigning a group of workers (a *crew*) to a set of tasks. The crews are typically interchangeable, although in some cases different crews possess different characteristics that affect which subsets of tasks they can complete.

Crew scheduling problems appear in a number of transportation contexts. Examples include bus and rail transit, truck and rail freight transport, and freight and passenger air transportation. There are many common elements to all of these problems, including the need to cover all tasks while seeking to minimize labor costs, and a wide variety of constraints imposed by safety regulations and labor negotiations. Nonetheless, each application also has its own unique characteristics and its own research challenges. In fact, most crew scheduling research focuses on a particular application, rather than the general case.

In this chapter, we focus on the *airline crew scheduling problem.* [For additional details on crew scheduling in the railway industry we refer the reader to Caprara et al., 1998 and for crew scheduling in mass transit systems to Wilson, 1999.] There are a number of reasons for focusing on airlines. First, they provide a context for examining many of the elements common to all crew scheduling problems. Second, the airline problem is truly a planning problem in the sense that airlines typically have a fixed schedule that changes at most monthly. Therefore, substantial time and resources can be (and are) allocated to solving it. Third, airline crews receive substantially higher salaries than equivalent personnel in other modes of transportation; the savings associated with an improved airline crew schedule can be quite significant. Finally, a large number of restrictive rules, mandated both by the FAA (or equivalent governing agencies for non-U.S. carriers) and strong labor unions, greatly restrict the set of feasible solutions, making airline crew scheduling one of the hardest crew scheduling problems. For all of these reasons, the airline crew scheduling problem has received the greatest level of attention, both from industry and from the academic community.

Airline Planning

Crew scheduling is just one of a number of challenging planning problems faced by airlines, see Figure 14.1. Although these problems are closely interrelated, they are typically solved sequentially, due to their size and complexity. Airlines usually begin by solving a *schedule design problem,* in which they determine the flights to be flown during a given time period. In the next step, the *fleet assignment problem,* they decide what type of aircraft (such as Boeing 767, 727, etc.) to assign to each flight, as a function of the forecasted demand for that flight. The *maintenance routing problem* follows, in which individual aircraft are assigned to flights so as to ensure that each aircraft spends adequate time at specific airports in order to undergo routine maintenance checks. Having completed these three tasks, the airlines then address the problem of scheduling crews.

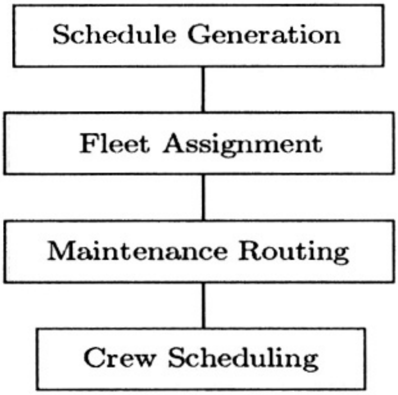


Figure 14.1 Schedule Planning

Within airline crew scheduling, there are significant differences between how international and domestic operations are scheduled. In the U.S., for example, international flight networks tend to be relatively sparse, with a limited number of flights into and out of an airport. U.S. domestic operations, in contrast, are characterized by *hub-and-spoke networks* with large numbers of arrivals followed by departures (called *banks* or *complexes*) occurring at hub airports in relatively short periods of time. International flight networks, however, are characterized by *point-to-point networks* with operations spread throughout the network. Another distinction is that international networks typically operate on a *weekly schedule,* while *daily schedules* are usually assumed for domestic operations. Moreover, unlike domestic operations, it is not uncommon for international operations to *deadhead* crews, that is fly them as passengers on some of the flights within their schedule in order to re-position them for future assignments. Barnhart et al., 1995 study the deadheading problem. All of these differences affect how crews are scheduled.

There are also significant differences between how *cockpit* and *cabin crews* are scheduled. For example, crews of pilots and other cockpit personnel usually remain together for much of their schedule. Cabin crews tend to vary more frequently, with flight attendants scheduled as individuals, rather than as part of a prescribed crew. Another key difference is that cockpit crews are heavily restricted in the number of fleet types that they are qualified to fly; cabin crews have greater latitude in the range of aircraft types that they can staff.

We will focus our attention on the problem of scheduling cockpit crews. For additional information on other forms of airline crew scheduling, we refer the reader to Day and Ryan, 1997 and Kwok and Wu, 1996.

14.2 The Crew Scheduling Problem

Each cockpit crew is qualified to fly a specific fleet type or set of closely related fleet types, known as a *fleet family.* Therefore, we solve a separate crew scheduling problem for each crew type, which includes only those flights that have been assigned to the corresponding fleet types.

The input to a crew scheduling problem is the set of flights to be covered. Flights are grouped together to form *duty periods,* which are series of sequential flight legs comprising a day's work for a crew. Duties are then strung together to form *pairings,* crew trips spanning one or more work days separated by periods of rest. Finally, monthly *schedules* are made up of multiple pairings with time off in between. These four components, i.e., flights, duties, pairings, and monthly schedules, are the building blocks of crew scheduling.

Associated with each of these building blocks is a distinct set of constraints. These typically come from three sources. First, governing agencies such as the FAA in the U.S. restrict crew scheduling, primarily for safety purposes. Second, labor organizations often enter into collective bargaining agreements concerning the crews' work conditions. Third, the airlines themselves pose added constraints, for example, to make the schedule more robust. In addition to these constraints, each building block is associated with a distinct cost structure. These constraints and cost structures are described in greater detail in the sections that follow.

Work Rules and Pay Structures

The most elemental decision in the crew scheduling problem is to decide which crew to assign to a given flight. The cost of such an assignment is a complex computation. Crews are not salaried, but rather are paid for the time that they spend flying, plus some added compensation for excess time spent on the ground between flights and during rest periods. Given this, we can think of the 'cost' associated with an individual flight simply as the duration of that flight. Because individual crews for a given fleet family cannot be distinguished in the crew pairing problem, crew costs are usually expressed in terms of time rather than cost. The total flying time in the system is clearly fixed and provides a lower bound on the optimal crew cost. The objective in crew scheduling is therefore to minimize *pay-and-credit,* the payments made above and beyond the cost of the actual flying time.

Duty Periods A number of rules restrict what combinations of flights can be flown by the same crew. A sequence of flights that can be flown by a single crew over the course of a work day is a *duty period.* Note that the same crew members typically stay together throughout the duration of a duty period.

Duties are constrained by a number of restrictions. The most obvious of these is that flights must be sequential in space and time. Furthermore, there is a restriction on the *minimum idle time* between two sequential flights, sometimes referred to as *connect time* or *sit time.* There is also a restriction on the *maximum idle time* allowed between two sequential flights. Additionally, there is a *maximum elapsed time* for a duty period. Finally, strict regulations govern the *total number of flying hours,* known as *block time,* that a crew can incur during the course of a single duty period.

The crew cost associated with a duty period is usually expressed as the maximum of three quantities. The first quantity is the *flying time.* The second quantity is a fraction (for example, $\frac{5}{8}$) of the *total elapsed time* of the duty period. The third quantity is a *minimum guaranteed* number of hours. This pay structure primarily compensates crews for flying time, but also provides additional pay for those crews assigned to very short duties or to duties with extensive idle time between the flights. Formally the cost of a duty period d can be expressed as

$$b_d = \max\{f_d \cdot elapse, fly, min\_guar\},$$

where b_d is the cost in minutes, $f_d \cdot elapse$ is a fraction of the elapsed time *elapse, fly* is the number of minutes of flying in the duty period, and *min_guar* is the minimum guarantee expressed in minutes.

Pairings Often a duty period starts and ends at different airports. Therefore, the crew cannot always return home at the end of a duty period but instead must often *layover* until the next day's duty period begins. Typically, crews spend anywhere from one to five days in a row away from home. A sequence of duties and layovers is known as a *pairing*. In general, a crew will stay together for all of the duties within a pairing.

There are a number of logical constraints on what constitutes a feasible pairing. Clearly, a pairing's first duty period must begin at the crew's domicile, called also the *crewbase,* and the last duty period must end there as well. In addition, each duty period must begin at the same airport where the previous duty period ended.

Pairings are further constrained by a complex array of rest requirements, flying time restrictions, and other constraints. These include the *maximum number of duties* in a pairing, the *minimum* and the *maximum amount of rest* between duties, and the *maximum elapsed time* of a pairing, also known as *time-away-from-base* (TAFB). One particularly complicated constraint is the *8-in-24* rule, which is imposed by the FAA in the U.S. This rule requires extra rest if a pairing contains more than 8 hours of flying in any 24 hour period. Generally, this occurs when the 24 hour period in question spans two consecutive duty periods. It is allowable to have more than 8 hours of flying in a 24 hour period only so long as the *included rest,* i.e., the layover between the two duty periods involved, and the rest following the second duty period, also known as the *compensatory rest,* are of sufficient length.

In the U.S., the cost of a pairing has two components. The first component, similar to the cost of a duty period, is the maximum of three quantities. The first of these quantities is the *sum of the costs of the duties* contained in the pairing. The second quantity is some fraction of the *total elapsed time* of the pairing. The third quantity is a *minimum guaranteed* number of minutes per pairing, which is typically the number of duty periods in the pairing times a fixed minimum guaranteed number of minutes per duty period. In addition to this, we include a second component, which represents the extra costs associated with the rest period between two duties, such as meals and lodging. Formally, the cost of a pairing p is

$$c_p = \max\left\{f_p \cdot \mathrm{TAFB}, ndp \cdot mg, \sum_{d \in p} b_d\right\} + \sum_{\substack{\hat{d}\in p, \bar{d}\in p \\ \hat{d}\to\bar{d}}} e(\hat{d},\bar{d}),$$

where $d, \hat{d}, \bar{d}$ represent the duty periods in p. Here, $\hat{d} \rightarrow \bar{d}$ indicates that duty period $\bar{d}$ immediately follows duty period $\hat{d}$ in p. In addition, mg and f_p are constants, ndp is the number of duty periods in p, and $e(\hat{d}, \bar{d})$ is the extra cost associated with the rest between duty periods $\hat{d}$ and $\bar{d}$.

European carriers tend to have a fixed salary for each crew. In this case the cost of a pairing is either ndp or 1.

Schedules Just as a duty period is a sequence of flights with sit times in between, and a pairing is a sequence of duties with layovers in between, a schedule is simply a sequence of pairings with periods of time off in between. However, a key difference between schedules and the other building blocks is that schedules are associated with *individual crew members,* rather than complete crews. The reason is that each crew member has different needs for time-off throughout the schedule period, which is typically a month. These include vacation time, training time, etc. Thus, in assigning crew schedules, we must take into account the needs and preferences of individual crew members.

In addition to individual crew member needs, we also have constraints similar to those seen for duties and pairings, for example, limits on the *maximum monthly flying time,* the *maximum duty time* in a month, the *minimum number of consecutive days off,* the *minimum total number of days off,* the *minimum rest* between pairings, and so forth.

Given this key difference, it is not surprising that the cost of a schedule is quite different from the other components. Whereas the focus within duties and pairings is on actual labor costs, the cost of a schedule is considered to be more a function of crew satisfaction and of workload balance.

The Crew Pairing and Crew Assignment Problems

The crew scheduling problem is typically divided into two subproblems. First, the *crew pairing problem* is solved. In this problem, we select a set of pairings such that each flight is included in exactly one pairing and pay-and-credit is minimized. Then, the *crew assignment problem* is solved. In this problem, the chosen pairings are combined with rest periods, vacations, training time, and other breaks to create extended individual work schedules, typically spanning a period of about one month.

The Crew Pairing Problem The domestic U.S. crew pairing problem is typically solved in three stages: *daily, weekly exceptions,* and *transition.*

The first stage, the daily problem, considers the set of flights which are flown at least four days per week. In this first stage, we treat these flights as though they all operate daily. We therefore want to find a minimum cost set of feasible pairings such that every flight in this set is covered exactly once. The pairings in this solution are then assumed to be repeated daily. The daily problem forms a good approximation since in the U.S. most of the flights operate every day, with a few exceptions on weekends.

For pairings that span multiple days, we assume that one crew will be assigned to each of the different duties within that pairing on any given day. For example, suppose the solution includes a three-day pairing made up of duties *A, B,* and *C.* On any given day, there will be one crew starting their trip with duty period *A,* another crew that began the trip the day before and is now covering duty period *B*, and a third crew on the final day of their trip, covering duty period *C*. This, in conjunction with the constraint that pairings cannot cover the same flight more than once, ensures that on any given day, every flight will be covered by exactly one crew.

Note that a solution to the daily problem will not be completely feasible in practice, because it assumes that all flights are flown every day of the week. Pairings that cannot be flown on certain days of the week because one or more of the flights do not operate on that particular day are referred to as *broken pairings.* The second crew pairing stage, the weekly exceptions problem, constructs new pairings to correct these broken pairings and also to cover those flights that are flown three or fewer days per week. Thus, in the weekly exceptions problem, we must associate flights with a specific day of the week. Accordingly, pairings become specific to days-of-week as well. Typically, deadhead flights are also needed in order to find good solutions to the weekly exceptions problem. Combined, the daily and weekly exceptions pairing solutions cover each flight in the weekly schedule exactly once.

Finally, note that airlines change their flight offerings on a regular basis, often quarterly and, to some degree, even monthly. Therefore, multi-day pairings can be problematic at the end of a monthly flight schedule. For example, on the last day of the month, a new crew must begin each pairing to cover that day's flights. However, the remaining days of the pairing may not be valid given that different flights might be offered in the next month's schedule. We therefore must solve a third stage of the crew pairing problem, the transition problem. This problem constructs pairings to cover flights for a small number of days spanning the changeover from one monthly flight schedule to another.

In all three of these problems, we emphasize that the object is to minimize pay-and-credit, the labor costs above and beyond the minimum required flying time.

The three stages described above are typical of U.S. domestic problems. More generally, we can think of two types of crew pairing problems, *weekly problems* and *dated problems.* Weekly problems yield sets of pairings that are repeated weekly; pairings that start at the end of the week wrap back around to the beginning of the week. Dated problems, on the other hand, correspond to specific days of the month. In the U.S. case, the daily pairing problem and weekly exceptions problem collectively solve the weekly problem, while the transition problem is a dated problem.

The Crew Assignment Problem Given the solution to the crew pairing problem, i.e., a minimum cost set of pairings that cover all flights throughout a monthly period, we must then assign specific individuals to these pairings. This occurs in the *crew assignment problem.*

Just as the crew pairing problem selects a minimum cost set of pairings (strings of sequential flights that satisfy a variety of rules) such that every flight is covered, the crew assignment problem selects a set of schedules (strings of sequential pairings that satisfy a variety of rules) such that every pairing is covered. In this context, a pairing corresponds to specific days in the schedule.

In spite of their similarities, these two problems are addressed separately, both in industry and in the academic literature. There are two primary reasons for this. First, in the crew pairing problem we assign complete crews to flights, while in the crew assignment problem crew members are scheduled individually, with each pairing being covered by multiple crew members. Second, the crew pairing problem focuses on minimizing labor costs, while in the crew assignment problem greater emphasis is placed on satisfying crew requests and seeking a balanced distribution of work.

In the U.S., the crew assignment problem is solved in two stages. In the first stage, a set of schedules is constructed such that each pairing is included in exactly as many schedules as are needed to fully staff the flight. Then, in the second stage, these schedules are assigned to individual crew members using a *bidline approach,* where schedules are allocated to crew members through a system in which crew members bid on their preferred work schedules. The schedules are then awarded by the airline based on crew priority, often related to seniority.

In Europe, on the other hand, individualized schedules, called *rosters,* are often constructed directly, taking into consideration the particular needs or requests of each crew member and, in some cases, crew seniority as well.

14.3 Formulations

The Crew Pairing Problem

Crew pairing models are typically formulated as *set partitioning problems,* in which we want to find a minimum cost subset of the feasible pairings such that every flight segment is included in exactly one chosen pairing.

Let F be the set of flight segments to be covered and let P be the set of all feasible pairings. Decision variable y_p is equal to 1 if pairing p is included in the solution, and 0 otherwise. Column p has a 1 in row i of the constraint matrix if flight i is included in pairing p and a 0 otherwise.

The crew pairing problem is

$$\begin{aligned} &\min \sum_{p \in P} c_p y_p \\ &\sum_{p: i \in p} y_p = 1 \qquad i \in F \qquad\qquad (14.1) \\ &y_p \in \{0, 1\} \qquad p \in P\,. \end{aligned}$$

Note that this formulation requires the explicit enumeration of all pairings. Enumerating pairings can be difficult both because of the numerous work rules that must be checked to ensure legality and, more importantly, because of the huge number of potential pairings. In fact, for most real instances, explicit enumeration of the constraint matrix is not possible. For example, a domestic problem on a hub-and-spoke network with several hundred flights typically has billions of pairings. Thus, heuristic local optimization approaches or column generation methods (described in Section 14.4) are used to solve all but the smallest ofproblem instances.

This basic set partitioning model is used for all three phases of crew pairing optimization. The models differ in the set of flights F that define the constraints of the problem. For the daily problem, there is one constraint for each flight that is repeated four or more times per week. The underlying assumption in solving this problem is that each pairing in the solution will be flown starting each day of the week. Recall that this presupposes that pairings are constrained to cover a flight leg at most once.

In the weekly and dated problems, F contains all of the flights in the flight schedule. In the weekly problem the flights are associated with a specific day of the week whereas in the dated problem they are associated with a specific date. Note that in the dated problems, we can relax the restriction that a pairing does not cover the same flight more than once since flights are associated with specific dates in these problems. For weekly problems, the same relaxation is valid for all of the pairings with time away from base shorter than a week.

Balancing Constraints Many airlines also add *crewbase balancing constraints* to the basic crew pairing model. These constraints ensure that the distribution of work over the set of crewbases is matched to the crew resources. They require that the number of hours of work contained in the chosen pairings which originate at a given crewbase must be between specified lower and upper bounds, which are a function of the number of crews stationed at that crewbase. Constraints of this form are known as *two-sided knapsack constraints.*

Example. We illustrate the crew pairing formulation using an example composed of the following seven flights.

Flight #	Orig	Dest	Start	End	Frequency
1	A	B	08:00	09:00	not 67
2	B	C	10:00	11:00	
3	C	D	13:00	14:00	not 7
4	C	A	15:00	16:00	
5	D	A	15:00	16:00	not 6
6	A	B	17:00	18:00	
7	B	C	11:00	12:00	not 67

The last column indicates the flight schedule. For example, flight 1 is operated every week day, while flight 5 is operated every day except Saturday. We assume, for simplicity, that all of the airports are in the same time zone.

We first consider the daily problem. Suppose that the valid duty periods are

$$D_1 = \{1\} \qquad D_2 = \{2\} \qquad D_3 = \{3\} \qquad D_4 = \{4\}$$
$$D_5 = \{5\} \qquad D_6 = \{6\} \qquad D_7 = \{7\} \qquad D_8 = \{1,2\}$$
$$D_9 = \{1,7,3\} \qquad D_{10} = \{2,3\}\,.$$

Assuming that airports A, C, and D are crewbases, we have six pairings, which can be expressed in terms of the duty periods as

$$P_1 = \{D_4, D_8\} \qquad P_2 = \{D_9, D_5\} \qquad P_3 = \{D_5, D_6, D_{10}\}$$
$$P_4 = \{D_4, D_6, D_7\} \qquad P_5 = \{D_1, D_7, D_4\} \qquad P_6 = \{D_5, D_7, D_9\}\,.$$

Pairing P_6 covers flight 7 twice and therefore it is not considered in the daily problem. Notice that an additional pairing could have been defined by the set of duties $\{D_4, D_1, D_2\}$. However, this pairing covers the same flights as pairing P_1. Given that both pairings originate at the same crewbase, only one of the two pairings (the less costly) need to appear in the model. In this example, we assume that $\{D_4, D_8\}$ has lower cost than $\{D_4, D_1, D_2\}$.

Assuming pairing costs $c_1 = c_2 = c_3 = c_4 = 4$ and $c_5 = 5$, from (14.1) we obtain the following formulation.

$$\begin{array}{llllllcll}
\min & 4y_1 & +4y_2 & +4y_3 & +4y_4 & +5y_5 & & & \\
 & y_1 & +y_2 & & & +y_5 & = & 1 & \text{(flight 1)} \\
 & y_1 & & +y_3 & & & = & 1 & \text{(flight 2)} \\
 & & y_2 & +y_3 & & & = & 1 & \text{(flight 3)} \\
 & y_1 & & & +y_4 & +y_5 & = & 1 & \text{(flight 4)} \\
 & & y_2 & +y_3 & & & = & 1 & \text{(flight 5)} \\
 & & & y_3 & +y_4 & & = & 1 & \text{(flight 6)} \\
 & & y_2 & & +y_4 & +y_5 & = & 1 & \text{(flight 7)} \\
 & y_1, & y_2, & y_3, & y_4, & y_5 & \in & \{0,1\} &
\end{array}$$

If we require that at least 3 hours and at most 6 hours of pay be assigned to crewbases A and D, and at most 5 hours of pay to crewbase C, then the crew balance constraints are

$$\begin{array}{ll}
3 \leq 4y_2 + 3y_5 \leq 6 & \text{(Crewbase A)} \\
0 \leq 3y_1 + 3y_4 \leq 5 & \text{(Crewbase C)} \\
3 \leq \quad 4y_3 \quad \leq 6 & \text{(Crewbase D).}
\end{array}$$

An optimal solution to this problem uses pairings 3 and 5, for a total cost of 9.

To obtain a solution to the weekly problem, we need, in addition to solving the daily problem, to solve a weekly exceptions problem to repair the broken pairings. The weekly exceptions problem consists of the following flights.

	P_3	P_5
Friday	5	1
Saturday	6	
Sunday	6,2	4
Monday	2,3	7,4
Tuesday		4

Alternatively, we could have solved the problem as a single weekly problem. Such a problem would have 43 flight segments and thus 43 cover constraints. Each pairing would appear multiple times, associated with the appropriate days of the week. For example, pairing P_1 would have five copies, one starting on each day of the week except Friday and Saturday. The copy that starts on Sunday wraps around in time since the next duty period is on Monday. In addition, this weekly problem also includes the pairing P_6. ■

The Crew Assignment Problem

In this section, we explain the rostering problem, which has been the focus of much of the crew assignment literature.

Separate rostering problems are solved for each *crew type,* where a crew type is specified both by the crew member rank (such as Captain, First Officer, Flight Engineer, etc.) and the fleet family (such as Boeing 767, Airbus 320, etc.) the crew members are qualified to fly. For a given crew type, the model input includes the set of pairings that must commence each day, and the number of crew members of the specified type that must be assigned to each of these pairings. The constraints of the rostering model require that:

1. Each pairing in the crew pairing solution is contained in the appropriate number of selected schedules. Note that the rostering model contains one constraint for each pairing commencing on a given day, for each day in the rostering period.
2. Each crew member is assigned to exactly one work schedule. If the airline is not required to use all crew members, a crew member might be assigned to an *empty* or *null* schedule – that is, a schedule containing no work.

Let K be the set of crew members of a given type, let S^k be the set of work schedules that are feasible for employee $k \in K$, and let P be the set of dated pairings to be covered. [A *dated pairing* is a pairing together with the starting date of the pairing.] n_p represents the minimum number of crew members that must be assigned to pairing $p \in P$ and γ_p^s is 1 if pairing $p \in P$ is included in schedule s and 0 otherwise. Decision variable x_s^k equals 1 if schedule $s \in S^k$ is assigned to employee $k \in K$ and 0 otherwise. c_s^k, the cost of schedule $s \in S^k$ for employee $k \in K$, represents the schedule cost, which might represent how close the schedule is to the crew member's stated preferences, or be set so as to minimize the number of crew members used. The latter is done by assigning very low costs to null assignments.

Given this notation, we write the crew rostering formulation, Gamache and Soumis, 1998, for a given crew member type as

$$\begin{aligned}
&\min \sum_{k \in K} \sum_{s \in S^k} c_s^k x_s^k \\
&\sum_{k \in K} \sum_{s \in S^k} \gamma_p^s x_s^k \geq n_p && \text{for all } p \in P && (14.2) \\
&\sum_{s \in S^k} x_s^k = 1 && \text{for all } k \in K \\
&x_s^k \in \{0,1\} && \text{for all } s \in S^k, \quad \text{for all } k \in K\,.
\end{aligned}$$

14.4 Solution Algorithms

At their core, the crew pairing and crew assignment models are set partitioning and set covering models with one constraint for each task to be performed (i.e. a flight or pairing to be covered) and one variable for each feasible combination of the tasks.

These problems are difficult for three reasons. First, even determining whether a combination of tasks is feasible can be difficult, given the wide array of rules and regulations that must be enforced. Second, these problems often have an enormous number of variables – often in the hundreds of millions or more. Third, these variables are all integer, further complicating the solution process.

In this section, we discuss solution approaches to address these difficulties. For the purpose of illustration, we focus our attention on the crew pairing problem. However, these ideas are applicable to the crew assignment problem as well.

Historical Solution Approaches

One of the major challenges in solving the crew pairing problem arises from the shear size of the problem. It is not uncommon for a problem containing 300 flights to have billions of pairings. Consequently, in early work on the pairing problem, only a subset of the pairings were constructed, using heuristic rules of thumb to guide and limit the construction process. In fact, until the 1990's, local improvement heuristics represented the state-of-the-art in crew pairing optimization.

A local improvement heuristic for the crew pairing problem starts with a feasible solution to the set partitioning problem. Because most airlines only make minor changes to their flight schedules from one planning period to the next, feasible solutions can usually be constructed manually by modifying the solution used in the previous planning period. Then, to find improved schedules, the heuristic randomly selects a small number of pairings in the current solution and searches for a better solution for the flights covered by that subset of the pairings. The search is usually performed by enumerating all possible pairings for the subset of flights and solving the small set partitioning IP to optimality using branch-and-bound. Often, these small set partitioning problems can be solved quickly since the LP relaxations of set partitioning problems with a small number of rows frequently have integral or near-integral solutions. This process is repeated until no further improvements are found or until some preset time limit is reached. This approach is taken in Anbil et al., 1991 and Gershkoff, 1989.

Pairing Generation

In each iteration of a local search heuristic the current incumbent solution is improved by considering only a small subset of the flights and the pairings covering only these flights. Therefore these heuristics lack the ability to consider the whole flight network in a single step and they need a large number of iterations before finding a good solution. An additional drawback of the local search heuristics is that they do not provide a lower bound on the best possible solution value. Thus, it is hard to estimate how far the current solution is from the optimum. To circumvent these two obstacles more global approaches are needed where at each iteration pairings covering all of the flights are generated.

Network Structure for Pairing Generation There are two main types of networks that have been developed in the literature for generating pairings. The first, called a *flight network,* has an arc for each flight in the schedule and arcs representing possible connections between flights. The second type of network, a *duty period network,* has an arc for each possible duty period and arcs representing possible overnight connections between the duties.

The network used to model international problems is typically *duty period,* rather than *flight,* based. That is, nodes represent the start or end of a duty period and an arc is included in the network for each possible duty period. *Connection arcs* between duties are included if two duties can be flown consecutively by the same crew. Domestic operations, on the other hand, typically use flight networks, because of the large number of feasible duties.

Each crew pairing is represented by a network path, but only the subset of paths satisfying certain requirements represent pairings. For example, in a flight network, a sequence of flights may give a path through the network, but that does not guarantee that the resulting duty periods will contain no more than the allowable hours of flying or that the resulting pairing will contain fewer than the maximum number of duty periods allowed.

Flight Network A typical flight network, Minoux, 1984, Desrosiers et al., 1991, has nodes representing the departure and arrival of each flight as well as a source *s* and a sink *t*. There is an arc representing each flight in the schedule. If there is a sparsity of flights arriving and departing an airport, it is often necessary to include potential deadhead flights in order to achieve a good, or even feasible, solution. For daily problems each flight arc is replicated as many times as the maximum number of calendar days allowed in a pairing but pairings are generated only from flights operating on the first day. For weekly problems, flight arcs are replicated as many times as the maximum number of weeks allowed in a pairing. Because pairings are often not allowed to exceed one week in duration, flight arcs need to be replicated only once so that pairings that cross over from the end of the week to the beginning of the next week, i.e. from Sunday to Monday, can be generated. Of course we could allow pairings to cross over without replicating flights by introducing arcs from the last day of the week back to the first. However, maintaining an acyclic network by replicating flights simplifies the shortest path algorithm for finding attractive pairings.

The source node is connected to the departure node of each flight that originates at a specified crewbase. The arrival node of every flight that ends at that crewbase is connected to the sink. There are also arcs representing legal connections between flights. A pair of flights will have a connection arc between them if the arrival airport of the first is the same as the departure airport of the second and the time between the two flights is either a legal connection within a duty period or a legal overnight rest. Note, however, that the required duration of an overnight rest might be a function of the attributes of the duty period that precedes it and plus perhaps other attributes of the pairing.

Figure 14.2 shows a partial flight network for the following flight schedule.

Flight 1: AIRPORT A – AIRPORT B 08:00 – 09:00
Flight 2: AIRPORT B – AIRPORT C 10:00 – 11:00
Flight 3: AIRPORT C – AIRPORT D 13:00 – 14:00
Flight 4: AIRPORT D – AIRPORT A 15:00 – 16:00

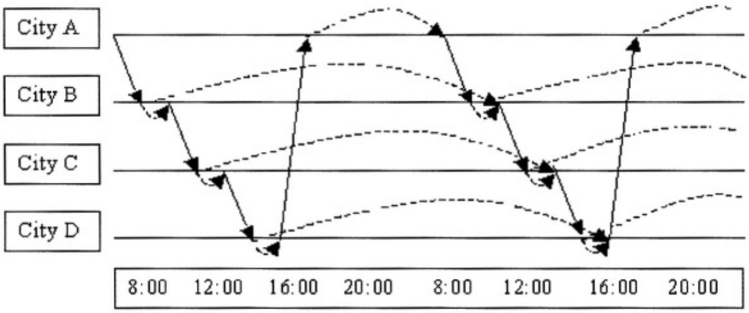


Figure 14.2 Flight Network

The network spans a two-day time horizon and contains two copies of each flight. The solid arcs represent flights. With each flight we have a transition through both time and space from the departure airport and departure time to the arrival airport and arrival time. The dotted arcs represent possible connections between flights. Of course, connections are only allowed between pairs of flights that arrive and depart from the same airport. Because of minimum and maximum connection time limits, the set of connection arcs in the network will generally be a proper subset of the set of all possible connections. Note that in the figure each arrival node has two connections emanating from it, one to the next departure and the other to the same departing flight one day later. In order to generate pairings originating at a crewbase, for example, Airport A, we would add a source node *s* and sink node *t* to this network. We would then connect *s* to the departure node of every flight arc originating at Airport A and connect the arrival node of every flight arriving at A to node *t*.

It is easy to see that every legal pairing is represented by some *s* – *t* path in this network. However, there are many *s* – *t* paths that do not represent legal pairings. The network structure guarantees that we will not connect two flights that do not have their respective arrival and departure at the same airport, but it does not prevent us from violating other rules like the maximum number of hours of flying allowed in a duty period or the maximum TAFB in a pairing.

Using a duty period network it is possible to build the duty period rules into the network, resulting in a much larger arc set.

Duty Period Network This network, Lavoie et al., 1988, Anbil et al., 1994, Vance et al., 1997b, has nodes representing the departure and arrival of each duty period as well as a source and sink. There are arcs representing each possible duty period in the flight schedule as well as arcs representing legal connections between duties. For daily problems, each duty period is replicated as many times as the maximum number of calendar days allowed in a pairing. Similarly, for weekly problems, each duty period is replicated as many times as the maximum number of weeks allowed in a pairing.

A pair of duties will have a connection arc between them if the arrival airport of the first is the same as the departure airport of the second and the time between them is a legal overnight rest. Remember that the required duration of an overnight rest might be a function of the attributes of the duty period that precedes it and possibly other attributes of the pairing. With the duty period network, unlike the flight network, it is possible to build explicitly into the network the requirements involving the preceding duty period.

For daily problems, the no repeated flight rule can be *partially* enforced by allowing connection arcs only between duties that do not share a common flight. That is, we can ensure that no two consecutive duties share a common flight, but for nonconsecutive duties, e.g. the first and third duties, we cannot prevent flight legs from being repeated in this manner.

Klabjan et al., 2001b propose an approach to store the network compactly. They assume that the duty periods are sorted based on the departure airport and the duty periods originating at the same airport are sorted in increasing order of the departure times. For each node representing a duty period arrival they store two pointers. The first pointer points to the earliest connecting duty period and the second pointer to the last connecting duty period. Due to the imposed order on duty periods, all possible connections are obtained by scanning all duty periods between the two stored duty periods. While compact, this network representation cannot embed into the network rules that involve two duty periods, for example, sharing common flights between two duty periods.

Figure 14.3 shows a two-day duty period network for the schedule shown in Figure 14.2. The solid arcs represent duty periods and the dotted ones represent connections between duties. The lighter solid arcs are the single-flight duty periods corresponding to each of the four flights in the schedule while the darker solid lines correspond to two additional duties, one composed of flights 1 and 2 and the other composed of flights 3 and 4. It is possible to build more duty periods from this set of four flights, but we have chosen to add only two to maintain the simplicity of the example. Note that the single flight duty period arcs arrive much later than the corresponding flight arcs in the flight network. This is because we include the time of the overnight rest in the duration of the duty arc.

To generate pairings in the flight network starting and ending at a crewbase, we add a source and sink node. The source node is connected to the departure node

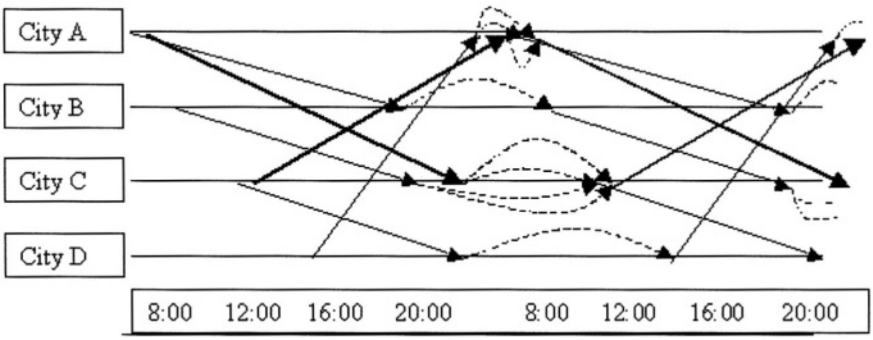


Figure 14.3 Duty Period Network

of each duty period that originates at the specified crewbase. The arrival node of every duty period that ends at that crewbase is connected to the sink.

Many more rules are satisfied by all paths from source to sink in the duty period network than the flight network. However, there are still some rules, such as the 8-in-24 rule and the no repeating flight rule for nonconsecutive duties, that cannot be enforced through the network structure.

Pairing Enumeration Duty period enumeration can be accomplished by a depth-first search approach on the flight network. For each flight arc we construct all duty periods that start with this flight. We attempt to extend the duty period with a flight if there is a corresponding sit connection arc in the flight network and all of the other duty feasibility rules are satisfied. In order to enumerate all duty periods we have to backtrack whenever we have exploited all sit connection arcs originating at a node corresponding to a flight arrival.

Pairings can be enumerated in a similar way either from the flight network or from the duty period network. In pairing enumeration the generation is started from every flight or duty that originates at a crew base. Depth-first search is then used to extend partial pairings or backtrack.

Partial Generation of Pairings Several methodologies for the crew pairing problem require a generation of only a subset of pairings since all of them cannot be handled explicitly. An easy way to achieve this is by generating pairings only on a subset of flights. This approach is taken in Anbil et al., 1991 and Gershkoff, 1989. It is substantially more difficult to generate a subset of pairings that cover all of the flights in the schedule. Andersson et al., 1998 give some details on how this operation is carried out at Carmen Crew Pairing. To generate a subset of pairings, for each flight they limit the number of possible connections. A certain number of short connections is selected and possibly some historically useful connections. Their idea is to solve a flight matching problem for each airport before pairing enumeration. Knowledge of 'good' connections is essential. Moreover, an experienced user might also prune the generation by recognizing useless connections.

Klabjan et al., 2001b propose the generation of random pairings. When extending a branch during enumeration, they choose connections at random. They use the connection times as greedy estimates; that is, the probability of selecting a connection depends on the connection time. Given that short connections are more likely to yield pairings with low cost, the smaller the connection time, the larger the probability of selecting the connection. Because in hub-and-spoke flight networks there are many connections, the connection selection strategy has to be implemented carefully. They propose a similar approach for generating random duties. Based on this random pairing generation they develop an algorithm for the crew pairing problem.

Solving the LP Relaxation

Early work on approximately solving the LP relaxation of (14.1) involves considering a large number of pairings and solving the LP relaxation over these pairings. Anbil et al., 1992 found an optimal solution to the LP relaxation of (14.1) over a large subset of the pairings for an 800 flight instance of a U.S. domestic daily problem. Five and a half million feasible pairings were enumerated and the optimal LP solution was found over this set using a specialized approach referred to as *SPRINT* in which several thousand columns are loaded into the LP solver and the LP is optimized over those columns. Then, most of the nonbasic columns are discarded, and several thousand more columns are added. This process is continued until all columns have been considered. At the end, however, it is necessary to price out all nonbasic columns to prove optimality. Bixby et al., 1992 used a combination of an interior point method and the simplex method to find the optimal LP solution to a very large crew pairing model. Hu and Johnson, 1999 propose a primal-dual algorithm for solving the LP relaxation over a given number of pairings. Their algorithm maintains a dual feasible vector and in every iteration increases the objective value by considering convex combinations of dual solutions.

As optimization solvers and computers became more sophisticated, there was a shift to dynamic column generation techniques that implicitly consider all possible pairings in solving the LP relaxation, Anbil et al., 1994, Desaulniers et al., 1998. In column generation the set partitioning problem with all possible pairings is referred to as the *masterproblem.* Thus (14.1) is the master problem. A *restricted* master problem is one that contains only a subset of the possible pairing columns. The *column generation algorithm* to solve the crew pairing LP involves the following steps:

- *Step 1: Solve the Restricted Master Problem* – Find the optimal solution to the current restricted master problem containing only a subset of all columns.
- *Step 2: Solve the Pricing Subproblem* – Generate one or more columns that may improve the solution. If no columns are found, STOP: the LP relaxation is solved.
- *Step 3: Construct a New Restricted Master Problem* – Add to the restricted master problem the columns generated in solving the subproblem and return to Step 1.

The solution and construction of the restricted master problem (steps 1 and 3) can be achieved using optimization software such as CPLEX or OSL. The solution of the pricing subproblem (step 2), however, should be tailored to exploit the network structure of the problem. The idea is to represent every pairing as a path in a network so that the huge number of pairings can be represented efficiently. This network is then used to identify variables that may improve the solution, without examining all variables. This often can be achieved either by solving *multi-label shortest path problems* on the specially structured network, Desrochers and Soumis, 1989, or by enumeration, Marsten, 1994 and Makri and Klabjan, 2001.

Considerations in Solving the Restricted Master Subproblem Until recently it was believed that the primal simplex algorithm is the most efficient procedure for solving the restricted master subproblem. Given that a primal solution is available from the previous iteration, primal simplex can be warm started. However primal simplex has two drawbacks. First, the LP relaxations of (14.1) tend to be extremely degenerate and therefore primal simplex tend to perform many degenerate steps, and second, the extreme point optimal dual solutions give misleading reduced costs and several iterations of column generation are needed. A standard trick for removing degeneracy is to randomly perturb the right hand sides. After the perturbed LP is solved, the perturbation is removed and the LP is solved to optimality. A different approach is presented in du Merle et al., 1999 by adding surplus and slack variables with penalties. This corresponds to requiring soft lower and upper bounds on the dual variables and penalizing the dual variables if they lie outside the bounds. For crew pairing problems this technique substantially reduces the number of iterations in column generation. Interior point algorithms yield an interior point dual solution, which is a much better indicator of the "usefulness" of a column, but lack the benefit of warm starting.

Barahona and Anbil, 1998 and Barahona and Anbil, 1999 propose a variant of the subgradient algorithm, which they call the *volume algorithm.* At every iteration the dual vector is improved in the direction of the subgradient and a primal feasible solution is obtained by taking a convex combination of previously obtained primal feasible vectors. The volume algorithm is fast, does not have large memory requirements, and it produces excellent dual vectors for use in column generation. Barahona and Anbil, 1999 claim significant performance improvements over both interior point and simplex algorithms.

Pricing Pricing is the problem of selecting pairings that are then added to the restricted master problem in step 2. There are two main questions in pricing: what is the criteria to select the pairings and how to find pairings that meet the criteria.

Traditionally, pairings are selected by the reduced cost criterion. Recently alternative strategies have been proposed. Bixby et al., 1992 use the pairing cost divided by the sum of the dual values of the legs in the pairing as the selection criteria. They report a significant decrease in the number of iterations. Hu and Johnson, 1999 present a primal-dual algorithm to select the columns with the lowest reduced cost based on a dual feasible vector, which is updated in every iteration. They also report a significantly lower number of iterations. Another, yet unexplored, strategy in the context of column generation would be to use the steepest edge pricing rule, Forrest and Goldfarb, 1992.

There are two approaches for finding pairings that best meet the selection criteria. One is combinatorial by using a shortest path algorithm, and the second is the brute force approach of enumerating the pairings. In the following sections we describe both approaches.

Pricing with Shortest Path Algorithms Until recently shortest path approaches have been designed only for the reduced cost criterion. Many algorithms solve the pricing problem to find attractive pairings using multilabel or constrained shortest path methods on specially structured networks, Desrochers and Soumis, 1988. In either the flight or the duty period network, only basic requirements can be built into the network structure. Requirements that cannot be built into the network structure are enforced through the use of labels. For example, we can maintain a label to track the number of hours of flying in the current duty period, the number of duties in the pairing, and the 8-in-24 rule. In addition to the labels that control the pairing feasibility rules, we need labels to capture the nonlinear components of the pairing cost structure.

Multilabel shortest path approaches differ from single-label approaches in that it might be necessary to keep many paths to each intermediate node in the network. For example, in solving the crew pairing problem, often it is not known which of the cost factors will dominate or which rules might prevent a path from resulting in a pairing until the complete pairing is specified. Consequently, it is necessary to keep track of all *nondominated paths* to each node between *s* and *t.* A path is nondominated if there does not exist another single path which is 'better' with respect to all the costs and rules. By better, we mean that either it is cheaper with respect to one of the cost criteria, or it is less restricted with respect to one of the rules. For example, if two paths to the same node have all labels identical except one has more time-away-from-base than the other, by dominance the one with the larger time-away-from-base can be eliminated.

Figure 14.4 illustrates a label update in a multilabel shortest path procedure. Each path carries four labels: the first gives the flying time in the current duty period, the second the elapsed time in the current duty period, the third the number of segments in the current duty period, and the final one, the number of duties in the pairing. At the arrival node of arc A the label values are (3.0, 6.0, 2, 1). For the arrival node of arc B the labels are (4.0, 5.0, 4, 1). Now consider the departure node of arc C. Two connection arcs both terminate at that node. The connection arc from flight A has a duration of two hours and the connection from flight B has a duration of one hour. Thus the two possible paths will have labels (3.0, 8.0, 2, 1) and (4.0, 7.0, 4, 1) respectively. Neither path can be eliminated by dominance because one contains less flying time and the other less elapsed time. Thus, we must now maintain two sets of labels at the departure node of flight C. Note that in this simple example we have only a small number of labels. As the number of labels grows, it is generally more difficult to eliminate paths by dominance so that a large number of potential paths to each node must be stored.

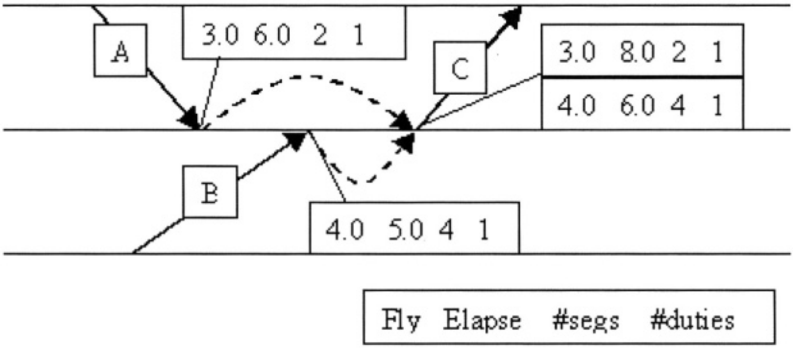


Figure 14.4 Constrained Shortest Path Example

Lavoie et al., 1988 and others were successful in using the multi-label shortest path procedure to solve the pricing subproblem over duty-based networks. This approach works especially well when the number of duty periods is not excessive, as demonstrated by Anbil et al., 1994, who used a duty-based network to solve international crew problems containing about two to three times as many duties as flights. Vance et al., 1997a were also successful in using a duty period network to solve a relatively small domestic daily problem.

Pricing By Enumeration The approach of generating all the pairings is an alternative solution to pricing. Note that the pairing feasibility rules and the cost structure are very complex and therefore a shortest path approach typically requires many labels, which makes dominated paths a rare occurrence. In addition, it might not even be possible to capture some of the feasibility rules with labels. If a change or an update of a feasibility rule is required, major changes in the shortest path code might be necessary. Therefore crew scheduling software vendors prefer to use pairing enumeration in pricing because they have many customers, each with its own feasibility rules.

For medium and large crew pairing instances enumerating all the pairings can be prohibitive and therefore strategies have to be designed to avoid this. Marsten, 1994 and Anbil et al., 1998 describe crew pairing optimizers that use partial enumeration in pricing. Both of these approaches use the reduced cost criterion.

Makri and Klabjan, 2001 use the selection criterion, introduced by Bixby et al., 1992,

$$\min_{p \in P} \left\{ \frac{c_p}{\sum_{i \in p} y_i^*} \,\middle|\, \sum_{i \in p} y_i^* > 0 \right\},$$

where y^* is the optimal dual vector to the restricted master subproblem. Pairings are enumerated, however they prune the enumeration by providing upper bounds on the score of a pairing p, which is defined as $\frac{c_p}{\sum_{i \in p} y_i^*}$.

Finding Good Solutions to the IP

Most state-of-the-art approaches combine column generation for solving the LP relaxation of the set partitioning problem with a branch-and-bound algorithm to find good integer solutions. Because of the large number of possible pairings, for all but the smallest problems, these approaches are heuristic in nature. They fall into one of three general classes. In the first class are algorithms where column generation is performed 'off-line'. That is, a subset of pairings is enumerated up front and the integer program is solved to optimality over this subset. An example of this type of approach can be found in Hoffman and Padberg, 1993. Because even moderate sized problems can have billions of variables, these approaches must work on a very small subset.

The second class of approaches uses dynamic column generation to solve the LP relaxation of the set partitioning problem to optimality or near optimality. Then, branch-and-bound is applied to obtain the optimal IP solution over the subset of columns generated to solve the LP relaxation. Among these approaches is work by Anbil et al., 1994 on the international crew pairing problem and Ryan, 1992 on the rostering problem. Recently Klabjan et al., 2001b proposed an algorithm that solves the LP relaxation of (14.1) and then selects several million pairings with low reduced cost to find an integer solution.

The drawback to these approaches is that there is no guarantee that a good solution, or even a feasible solution, exists among a subset of columns that give a good LP solution.

A third class of algorithms allow dynamic column generation throughout the branch-and-bound tree. We refer to algorithms of this type as *branch-and-price* approaches. Like branch-and-bound, a branch-and-price procedure is a smart enumeration strategy in which an LP relaxation is solved at each node of a branch-and-bound tree. The difference is that the huge constraint matrix requires the use of column generation. Branch-and-price methodology has been applied to a number of problems in transportation, scheduling, and combinatorial optimization. For a survey, see Barnhart et al., 1998.

Recently, a number of groups have developed crew pairing and rostering algorithms using a branch-and-bound framework with column generation, including Desaulniers et al., 1998, Desrosiers et al., 1991, Gamache et al., 1999, Gamache et al., 1998, Gamache and Soumis, 1998, Ryan, 1992, Vance et al., 1997a, and Anbil et al., 1998.

Marsten, 1994 combines dynamic pairing generation with variable fixing to obtain good integer solutions. To find integer solutions, the variables associated with fractional pairings with value close to one are fixed to one sequentially. To limit column generation, new pairings are generated only when the bound from the LP relaxation increases above a pre-set target.

Andersson et al., 1998 decouple pairing generation from the optimization engine. The algorithm generates pairings several times and solves (14.1) over the generated columns. They use the Lagrangian algorithm presented in Wedelin, 1995 to solve the integer programs.

Branching Rules for the Crew Pairing Problem To be able to generate pairings at any node in the branch-and-bound tree, a branching rule that is compatible with the pairing generation procedure is needed. The standard rule of branching on variable dichotomy is difficult to implement. Using such a branching decision, we would either fix a pairing into the solution ($x_j = 1$) or forbid the use of a pairing ($x_j = 0$) with each decision. It is easy to fix a pairing *j* into the solution. There is no need to generate any more pairings containing any of the flights covered by pairing *j,* so these flight arcs may be deleted from the pairing generation network. However, forbidding the use of a specific pairing is difficult since we must forbid specific paths from being returned by the pairing generation procedure. This could require finding the (k + 1)st shortest path if k pairings have been forbidden by branching.

The first branching rule presented here is motivated by a general rule for set partitioning problems developed by Ryan and Foster, 1981. Their rule is based on the simple observation that given a fractional solution to the LP relaxation of a set partitioning problem there must exist two columns whose associated variables are fractional such that they both contain coefficients of one in a common row r and there exists another row s where one column has a coefficient of one and the other has a coefficient of zero. This fact leads to a general branching rule where pairs of rows r and s are required to be covered by the same column on one branch and by different columns on the other.

Essentially the same logic can be used for crew pairing optimization, but the rule is modified to maintain tractability. In general, it is difficult to force two specific flights to either appear only in pairings that contain them both (the first branch) or to never appear together in the same pairing (the second branch). However, it is an easy matter to force two flights to appear consecutively in a pairing or not. If flights r and s satisfy the conditions of the Ryan and Foster rule and they appear consecutively in at least one of the fractional pairings that contains them both, branching can be performed by requiring that they appear consecutively in the pairing that covers them at one node and by requiring that they cannot appear consecutively in any pairing in the solution at the other node. This strategy is sometimes referred to as branching on *follow-ons* because it places restrictions on which flights can follow flight r in the solution. The flight pair r, s is often termed a *follow-on.*

Klabjan et al., 2001b present a different branching rule called *timeline branching.* In timeline branching the decision is based on a flight r and a time t. In one branch we only allow pairings with time of the connection immediately following r less than or equal to t. The other branch considers only pairings with time greater than t of the connection immediately following r. They show that this is a valid branching rule if the departure times are all different, which can be achieved by slightly perturbing them. They also combine follow-on branching or timeline branching with strong branching. Strong branching is a branching rule that chooses the branching variable (follow-on in the context of crew pairing) by carrying out several dual simplex iterations for each branching candidate to estimate the change in the lower bound, Bixby et al., 1995, Linderoth and Savelsbergh, 1999.

Master Problem Modification We discuss the modifications to the master problem for follow-on branching. For timeline branching the modifications are similar. To implement follow-on branching, both the master problem with its existing set of columns and the column generation subproblem must be modified. On the branch where flight *r* must be followed by flight *s* in the same pairing, any pairing in the restricted master problem that contains *r* and/or *s* but does not have the two flights appearing consecutively is eliminated. On the branch where the flights cannot appear consecutively, any pairing with *r* and *s* consecutive is eliminated from the restricted master problem.

Flight Network Modification If a flight network is used, flight *s* can be required to follow flight *r* by eliminating all the connection arcs out of *r* except the one to *s*. Connections into *s* from any flight other than *r* are also deleted. Note that this second modification is not absolutely necessary since requiring *r* to be followed by *s* is sufficient to ensure that *s* will not be preceded by a flight other than *r* in any pairing in the basis. However, for the subproblem, it is computationally advantageous to eliminate as many arcs as possible when a branching decision is fixed. Forbidding the connection is implemented by eliminating the arc connecting *r* to *s*. These network modifications can be accomplished by removing the arcs or by giving them a very high cost so that they will not be used in attractive pairings. The second approach is generally preferable since it simplifies many of the bookkeeping issues associated with storing the network structure and enables the same network (with modified costs) to be used to generate pairings at any node in the branch-and-bound tree.

Duty Period Network Modification If a duty period network is used, the implementation will depend on whether the connection between flights *r* and *s* is an overnight rest. A connection within a duty period (not an overnight) can be required by eliminating all duties that contain either flight but do not have them appearing consecutively. The connection can be forbidden by eliminating all duty periods containing *r* and *s* consecutively. A connection that is an overnight rest can be required by eliminating all duties containing flight *r* that do not end in *r* and all duties containing *s* that do not begin with *s*. Then arcs connecting any duty period ending in *r* to a duty period that does not begin with *s* are deleted, as well as arcs connecting a duty period ending in a flight other than *r* to a duty period beginning with *s*. To forbid the connection, we delete connection arcs from any duty period ending in *r* to a duty period beginning in *s*.

From the above discussion, we see that the branch on follow-on rule can be implemented with either type of pairing generation network simply by eliminating arcs from the network.

Parallel Approaches to Crew Pairing

Crew pairing problems with as few as 300 legs for hub-and-spoke networks or 2000 legs for point-to-point networks can take as much as 10 to 20 hours of CPU time to solve and do not necessarily produce an optimal solution. This is particularly problematic when conducting ‘what-if’ analysis. One way to decrease computation time is to employ parallel algorithms for crew pairing. However, the crew pairing model (14.1) is an integer program and parallel algorithms for integer programs typically do not scale well. Thus, designing parallel algorithms for the crew pairing problem is a challenging problem that has only recently begun to be studied.

One of the most time intensive parts of most crew pairing algorithms is pairing generation. The basic idea of a parallel algorithm for pairing generation is to distribute the legs originating at crewbases (called starting legs) among the processors, with each processor enumerating all the pairings starting with the assigned starting legs. Since the computational times to generate all the pairings starting with a given leg can vary substantially, load balancing algorithms are needed. Goumopoulos et al., 1997 propose a pulling algorithm based on the master/workers paradigm. The master distributes the legs one by one to the workers. Whenever a worker becomes idle, it queries the master for a new starting leg. Klabjan and Schwan, 2000 eliminate the master with the processors exchanging the workload among themselves.

In other research, Alefragis et al., 1998, Sanders et al., 1999, and Alefragis et al., 2000 focus on parallelizing the pairing enumeration and the Lagrangian decomposition algorithm of Andersson et al., 1998. Since the latter algorithm is inherently sequential, they describe the steps for parallelizing an iteration of the algorithm, i.e. updating the Lagrangian multipliers and computing the subgradient. They compute the constraints in parallel and distribute the variables among the processors. Note that due to the fine grain parallelism, this algorithm does not perform well on architectures with high latency such as a cluster of workstations.

An entirely different approach is given in Klabjan et al., 2001b. The LP relaxation is solved in parallel by generating the pairings in parallel and in each iteration the LP is solved with the parallel primal-dual algorithm, Klabjan et al., 2000. The IP over a small subset of pairings is solved with a branch-and-bound algorithm that executes the strong branching rule in parallel. The pairing enumeration algorithm is scalable but the parallel primal-dual algorithm scales only to 20 processors.

The most promising algorithms for parallelization are branch-and-price since coarse granularity is easily achievable by evaluating the branch-and-bound nodes in parallel. Gedron and Crainic, 1994 give a survey on parallel branch-and-bound algorithms. Klabjan, 2001 describes a parallel branch-and-price algorithm. The algorithm evaluates the branch-and-bound nodes in parallel and in addition, each node is evaluated in parallel. An LP relaxation is solved in parallel by embedding parallel column generation in the parallel primal-dual algorithm.

Open Issues

There are still a number of open questions regarding the best method for crew pairing optimization. Whether to use dynamic network-based pairing generation or a fast pricing procedure like SPRINT is not clear. If pairings can be enumerated quickly and accessed off-line in an efficient manner, the SPRINT approach may be preferable to network-based generation. If network-based generation is used, the type of network that will perform most efficiently might be highly schedule-dependent. For point-to-point crew pairing problems, duty period based networks have proven to be efficient because the number of possible duty periods grows relatively slowly with the number of possible flights. However, for hub-and-spoke networks, the results have been mixed. Another open issue is how to manage effectively the number of pairings in the constraint matrix; that is, how many pairings to add at each iteration and whether or not to delete pairings with high reduced cost.

Crew Rostering Solution Approaches

The solution approaches described above apply to both the crew pairing and crew assignment problems. We conclude by briefly highlighting some of the research conducted specifically for the crew rostering version of the crew assignment problem. We also introduce an alternative approach to generating schedules that has been used in the crew rostering literature but could also be applied to crew pairing generation.

Computational Results Ryan, 1992 solves crew rostering problems with 55 crew members and 120 pairings. This results in problems containing as many as 300,000 variables, with solution times ranging from less than 10 minutes to 2–3 hours.

Gamache et al., 1998 construct individualized monthly work schedules for pilots and officers. Schedules are selected for assignment to crew members based on considerations of individual preferences and seniority restrictions. They solve 24 instances of problems at Air Canada, containing up to 108 pilots and 568 pairings. Using cutting planes, solution times range from 1 to 8 hours.

A Constraint Programming Approach In some cases, there exist schedule rules and regulations that cannot easily be captured in a constrained shortest path approach. Moreover, pricing by enumeration may be intractable. Recently *constraint programming* (CP) approaches to rostering have been proposed. Thorough discussions of CP can be found in Lustig and Puget, 2001 and Brailsford et al., 1999.

An example of how constraint programming has been used in crew assignment can be found in the work of Fahle et al., 1999 and Junker et al., 1999. They use CP to solve the pricing problem for generating columns in the crew assignment problem of a major European airline. For each crew, they define a variable that represents the set of duties to be assigned to that crew. The domain contains all feasible subsets of the set of duties to be covered. They are able to specify constraints that capture all of the crew rules and regulations. In addition, they incorporate a shortest path component that leverages dual information from the restricted master. This allows them to reduce the search space significantly. They are able to solve real-world problems successfully, incorporating constraints that could not be captured in a constrained shortest path approach.

14.5 Integrating Crew Pairing with Maintenance Routing and Schedule Design

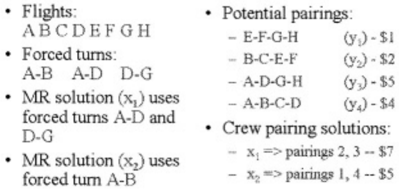
In airline planning, the schedule design, fleet assignment and maintenance routing problems are all solved before the crew scheduling problem. Their solutions then impact the input to crew pairing. For example, by assigning aircraft types to flights in the fleet assignment problem, we partition the flights into smaller sets and solve a separate crew scheduling problem for each of these sets, since individual crews can only fly certain aircraft types. Thus, solving these planning problems sequentially can lead to sub-optimalities, because decisions are made in the earlier problems without taking into account their impact on crew scheduling. A fully integrated approach to the airline planning process is quite difficult, due to its enormous size and complexity. Nonetheless, benefits can be gained by partially integrating elements of the planning process. We provide examples of this in the sections that follow.

Crew Pairing and Maintenance Routing

Although crew scheduling assigns *crews* and maintenance routing assigns *aircraft,* there is a connection between these two problems. Specifically, this connection deals with the amount of time required between two flights for a connection to be crew-feasible. Recall that there is a minimum idle time required between two consecutive flights in a duty period. This time is needed, in part, so that the crew can move through the terminal from the arrival gate of the first flight to the departure gate of the second flight. However, if both of these flights have been assigned to the same aircraft in the maintenance routing problem, then the crew will remain with the aircraft and therefore this time restriction can be relaxed. Such a crew connection, which is feasible only if both flights share a common aircraft in the maintenance routing solution, is known as a *forced turn.*

When constructing a network for the crew pairing problem, we begin with the set of connections that have adequate sit time and then add to this set those forced turns implied by the maintenance routing solution. These additional connections permit new pairing opportunities and thus can improve the quality of the crew pairing solution. However, because the forced turns are determined in the maintenance routing problem without taking into account the crew pairing objective, solving these two problems sequentially can lead to sub-optimal results.

This potential for sub-optimality is demonstrated in the following example, taken from Cohn and Barnhart, 2002. Consider a network of eight flights, denoted *A* through *H.* As shown in Figure 14.5, this network has three potential forced turns, two different feasible solutions to the maintenance routing problem (denoted by MR), and four potential pairings. For each of the maintenance solutions, only a subset of the pairings are feasible, depending on the forced turns implied by the maintenance solution. Thus, given maintenance solution 1, the cost of an optimal crew pairing solution is $7, while maintenance solution 2 yields a crew pairing problem whose optimal cost is $5.



- Flights:
  A B C D E F G H
- Forced turns:
  A-B   A-D   D-G
- MR solution (x_1) uses forced turns A-D and D-G
- MR solution (x_2) uses forced turn A-B

- Potential pairings:
  - E-F-G-H   (y_1) - $1
  - B-C-E-F   (y_2) - $2
  - A-D-G-H   (y_3) - $5
  - A-B-C-D   (y_4) - $4
- Crew pairing solutions:
  - x_1 => pairings 2, 3 -- $7
  - x_2 => pairings 1, 4 -- $5

Figure 14.5 Integrated Example

This small example highlights how a sequential approach to the maintenance routing and crew pairing problems can lead to sub-optimal results. Three different approaches have appeared in the literature to address this.

In the first approach, Klabjan et al., 2002 solve the crew pairing problem *before* they solve the maintenance routing problem. They include *all* potential forced turns in the crew pairing network. Those forced turns contained in the crew pairing solution then become required aircraft turns when solving the maintenance routing problem. Although this approach can potentially lead to maintenance infeasibility, in practice they found feasible solutions for many hub-and-spoke flight networks. Additionally, note that whenever a feasible crew pairing solution is found to be maintenance feasible, this solution is in fact *optimal* for the integrated problems, when the maintenance routing problem is a feasibility rather than an optimization problem. Furthermore, their approach requires no more computational effort than the original sequential approach.

In the second approach, Cordeau et al., 2001 present an integrated model that guarantees maintenance feasibility. They maintain the original string-based maintenance routing and crew pairing formulations. Like Klabjan et al., 2002, they include all potential forced turns in the crew pairing network. They then link the two models by adding one constraint for each potential forced turn. The constraint for forced turn t states that the number of chosen crew pairings containing forced turn t cannot exceed the number of chosen maintenance routes that contain it. This results in a large-scale integer program which they solve by branch-and-bound, where the LP relaxations are solved by using a Benders decomposition approach and column generation.

A third approach is found in Cohn and Barnhart, 2002. Their approach is similar to that of Cordeau et al., 2001, but in place of maintenance string variables, they use variables representing *complete solutions* to the maintenance routing problem. This dramatically reduces the number of constraints, because all of the original maintenance routing constraints are replaced by a single convexity constraint. This reduction in constraints comes at the cost of a potential explosion in the number of variables. However, they prove that only a small subset of the feasible maintenance solutions need to be considered to ensure an optimal result, thereby greatly reducing the size of the problem. Furthermore, they prove that the integrality of the maintenance solution variables can be relaxed. Thus, their integrated maintenance routing and crew pairing model has no more binary decision variables than the basic crew pairing model alone.

Crew Pairing and Schedule Planning

Klabjan et al., 2002 study the impact of flight departure times on the crew pairing problem. If we are allowed to change the flight departure times, then some paths in the duty period network that do not satisfy all of the pairing feasibility rules might correspond to a pairing in a retimed flight schedule. Consider two flights *i* and *j* depicted in Figure 14.6. In the original schedule, leg *j* cannot follow leg *i* in a pairing because it violates the minimum sit connection time. However if leg *j* departs 5 minutes later, then the connection becomes feasible. In a retimed flight schedule, additional paths in the duty period network become pairings not only due to the minimum sit and rest connection times, but also with respect to the maximum duty elapsed time and the 8-in-24 rule.

The model in Klabjan et al., 2002 is identical to (14.1) except that more columns are considered. They develop an algorithm that simultaneously generates paths in the duty period network and new departure times such that the generated paths correspond to pairings in the retimed flight schedule. Each path defines its own flight departure times, but given that in (14.1) every flight segment is covered by exactly one pairing, the solution implies a single departure time for each leg.

They report computational experiments on large fleets of a U.S. domestic carrier. A time window *w* of either 5 or 10 minutes is imposed, i.e. every departure time can be changed by at most ±*w*. New flight departure times should not diverge by much from the original times since otherwise it would affect the fleeting cost and would substantially disrupt passenger connections. On average, the improvement of pay-and-credit for w = 5 minutes is 25% and for w = 10, pay-and-credit decreases by 35%. These data clearly show that the crew cost can be substantially reduced by slightly retiming departures.

Crew Pairing with Regularity

The crew pairing model (14.1) for the weekly problem minimizes the weekly crew cost. However it is unlikely that the resulting pairings could be repeated many

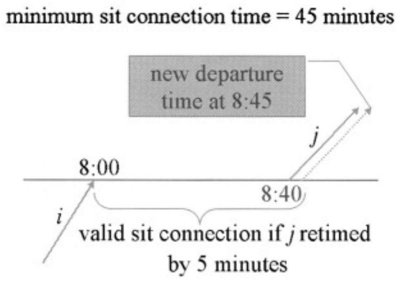

minimum sit connection time = 45 minutes
new departure time at 8:45
j
8:00
8:40
i
valid sit connection if j retimed by 5 minutes

Figure 14.6 Leg Retiming

times in the weekly horizon unless such repetition constraints were specifically imposed. Thus the solution would lack regularity. Regularity is important with respect to crew (and aircraft) schedules, since regular solutions are much easier to implement and manage, and, if possible, crews prefer to repeat itineraries.

In Section 14.2 we described the daily/weekly exceptions methodology for solving the weekly problem. This methodology, first, does not necessarily find the minimum cost crew schedule even if the daily and the weekly exception problems are solved to optimality and second it does not directly take into account regularity. Klabjan et al., 2001a present a new model, called the weekly crew pairing model with regularity, that captures both the crew cost and regularity in a weekly schedule. They solve the model in several stages, where in the first stage they obtain pairings with the highest regularity, i.e. those that can repeat seven days a week, in the second stage the algorithm yields pairings that can be repeated six times in a week, and so forth.

By using approximations and integer programming as a heuristic, they obtain solutions that improve on current practice with respect to both regularity and cost. They report computational results on small and large fleets of a major U.S. domestic carrier. The improvements on crew cost range from 10% to 40% and their solutions have 40-60% higher regularity.

14.6 The Crew Recovery Problem

An airline schedule rarely operates as planned. Maintenance problems, weather conditions, and other unplanned events cause frequent disruptions – on a typical day, several flights will be delayed or canceled. Each disruption can propagate through the system, because it impacts resources such as crews and aircraft that are also needed for subsequent flights. The *crew recovery problem* considers how to modify a crew schedule that has been affected by disruptions.

The recovery problem differs from the planning problem in several ways. One of the most fundamental differences is in the time horizon for solving the problem. Unlike the planning problem, which is solved as part of a multi-week process, the recovery problem must be solved very quickly – often, in minutes. Thus, the goal of the crew recovery problem is to find a good solution quickly.

The second difference between recovery and planning is that the crew recovery problem must take into account the recent flying history of the active crews. Each crew's options are limited as a function of the work that has already been performed during the current pairing.

Third, *reserve crews* can be considered when solving the crew recovery problem. These crews have a minimum guaranteed pay, measured in flying hours, and cannot fly more than a designated monthly maximum. This pay structure adds a further level of complexity to the problem.

Another difference is in the constraints that determine what constitutes a feasible pairing. Most airlines use tighter restrictions in their scheduling problems than are legally mandated by, for example, tightening the minimum rest connection time and the maximum duty elapsed time. This is precisely so that they will have some added flexibility in recovering from disruptions. In the recovery problem, on the other hand, rules on flying hours are often pushed to their legal limits.

Perhaps the most significant difference between crew pairing and crew recovery is in how the objective function is defined. When a schedule is modified to address disruptions, active crews are usually paid *at least* the cost of their originally scheduled pairings; if a crew is assigned to a modified pairing that has higher cost, they receive the higher amount. It is also desirable to keep down the additional costs incurred by reserve crews. Furthermore, there are other objectives that are important as well, such as returning to the original plan quickly and minimizing passenger disruptions. In addition, crew decisions are not made in isolation. They must be made in conjunction with decisions about delaying or canceling future flights, swapping aircraft, and further issues related to the other resources that have been affected by the disruptions. The recovery problem also faces a host of safety and labor constraints that restrict what changes can be made. Therefore, even deciding what objective to use when recovering from disruption can be a difficult question.

Thus, while crew recovery has much in common with crew pairing, it also poses its own unique set of challenges. Although this problem has been addressed in limited fashion in the literature, much work remains to be done. In the following section, we present one approach to modeling the crew recovery problem, which leverages its similarities with crew pairing. In the subsequent section, we highlight some of the research on solving the crew recovery problem.

A Crew Recovery Model

We present a crew recovery model from Lettovský, 1997 and Lettovský et al., 2000 that is similar to the crew pairing model. However, in this recovery model, each pairing is specific to a particular crew. For a given crew, all potential pairings begin at the next time and location that the crew becomes available, i.e. at the end of their current flight or rest period. Each potential pairing for a crew must not only take into account work already completed in the current pairing when ensuring that all rules and regulations are satisfied, but must also ensure the legality of the crew's remaining schedule. That is, there must be enough time at the completion of this modified pairing for a sufficient rest period before the next scheduled pairing is to begin, and restrictions that span multiple pairings in a schedule (such as monthly flying time limits) must not be violated.

The objective of this model is to minimize the cost of adjusted pairings, reserve crews, and deadheaded crews, as well as the cost of canceling flights. The cancelation cost is the cost of re-assigning passengers to other flights as well as hotel and meal costs for affected passengers and some estimate of the loss of good will.

We define the following parameters:

e equipment type experiencing disruption (this may represent several aircraft types if they are crew compatible),
L_e set of flight segments to be covered by crews of equipment type e,
K_e set of crews available for equipment type e (including reserve crews),
P_k set of pairings that can be flown by crew $k \in K_e$,
c_p cost of assigning pairing p,
d_l cost of using flight segment l for deadheading,
q_k cost estimate of returning the crew to its domicile,
f_l cost estimate of canceling flight segment l,
β_{pl} 1 if flight segment l is included in pairing p, 0 otherwise.

The variables are:

$$x_p = \begin{cases} 1 & \text{if pairing } p \text{ is assigned to a crew,} \\ 0 & \text{otherwise,} \end{cases}$$

$$v_p = \begin{cases} 1 & \text{if crew } k \text{ has no pairing assigned,} \\ 0 & \text{otherwise,} \end{cases}$$

$$y_l = \begin{cases} 1 & \text{if flight segment } l \text{ is canceled,} \\ 0 & \text{otherwise,} \end{cases}$$

$$w_l = \text{the number of crews deadheading on flight segment } l.$$

The airline crew recovery problem for a given equipment type e is

$$\min \sum_{k \in K_e} \sum_{p \in P_k} c_p x_p + \sum_{l \in L_e} f_l y_l + \sum_{l \in L_e} d_l w_l + \sum_{k \in K_e} q_k v_k$$

$$
\begin{aligned}
\sum_{k \in K_e} \sum_{p \in P_k} \beta_{pl} x_p + y_l - w_l &= 1 && \text{for all } l \in L_e, \\
\sum_{p \in P_k} x_p + v_k &= 1 && \text{for all } k \in K_e, \\
w_l + \max{}_l \cdot y_l &\leq \max{}_l && \text{for all } l \in L_e, \\
0 \leq w_l &\leq \max{}_l && \text{for all } l \in L_e, \\
x \text{ binary}, y \text{ binary}, w \geq 0, v \geq 0 \, .
\end{aligned}
\tag{14.3}
$$

The first set of constraints guarantees that all flight segments are either canceled or covered at least once. The slack variable on these constraints, w_l, has an upper bound $\max_l$ defined as the maximum number of crews that can deadhead on flight segment l. The second set of constraints ensures that crew k is either assigned to a pairing or is deadheaded to its crewbase. The third set of constraints forces w_l to be zero if flight l is canceled. Note that the integrality of the variables w_l and v_k are implied and hence need not be imposed.

When solving the crew recovery problem, it is important to find a good solution quickly. Furthermore, it makes sense to take advantage of the fact that the original scheduled pairings were optimal in the undisrupted problem environment. Therefore, when solving the crew recovery problem, we do not want to re-assign all crews. Instead, we want to consider only those crews whose pairings were disrupted, as well as a small number of additional crews deemed likely to introduce good "swapping" opportunities. Limiting the scope of the problem in this way can significantly reduce the size of the model and thus improve its tractability. Heuristics for selecting the set of crews to be considered can be found in Lettovský et al., 2000 and Lettovský, 1997.

Crew Recovery Solution Approaches

Very little has been published in the open literature on solving the crew recovery problem. Teodorovic and Stojkovic, 1990 developed a sequential approach based on a dynamic programming algorithm, using the first-in-first-out principle to minimize the crews' ground time. Wei and Yu, 1997 presented a heuristic-based framework for real-time crew re-scheduling. Song et al., 1998 presented a multicommodity integer network flow model and a heuristic search algorithm to solve it. Stojkovic et al., 1998 presented a column generation approach similar to that used for crew pairing problems. Solutions approaches to the crew recovery model presented in the previous section can be found in Lettovský et al., 2000 and Lettovský, 1997.

Models such as (14.3) can be solved to obtain optimal solutions for small disruptions and good feasible solutions for medium sized disruptions. However, for major disruptions such as those caused by snowstorms, and particularly those disruptions that affect multiple airports, further refinement and perhaps even a new approach altogether is needed.

Crew Rostering Recovery

We conclude this section by noting that one potential for inefficiency in the crew recovery approach discussed above is that it limits changes to the pairings currently underway. By requiring the modified pairings to be feasible in conjunction with the remainder of the crew's monthly schedule, opportunities may be missed. On the other hand, considering the entire schedule, rather than just the current pairings, yields an enormous problem. Stojkovic et al., 1998 present an initial approach to this challenging task.

14.7 Robustness in Crew Pairing

The crew pairing problem is solved well before the flight schedule becomes operational. In this planning stage, all flights are assumed to have departure times that are both fixed and known. This assumption is often proven wrong when the crew schedule is actually implemented. For example, the U.S. Department of Transportation reported that the total number of *delay minutes* in the system (based on flight delays of 15 minutes or more) had increased by 11% from 1995 to 1999, Bond, 2000. In the summer of 2000, airline delays received national attention in the U.S., when the airline with the *best* performance record had 25% of its flights delayed by 15 minutes or more.

When crew members' schedules are disrupted in operations, they are nonetheless guaranteed to be paid for their original scheduled workload. In addition, if delays increase their flying or sit time, they may be entitled to added compensation. Furthermore, disruptions may require the use of reserve crews to get back on schedule. Clearly, then, the cost associated with implementing a crew pairing solution may vary significantly from the planned cost. Typically, the planned ratio of pay-and-credit to flying time is below 1% for large fleets, but increases on average to 4% when the schedule is implemented. For smaller fleets, the ratio tends to increase from about 3% to 8%. Solutions of large fleets are much more sensitive to disruptions since they have many tight connections. A disrupted short connection can have a significant impact on the entire flight and crew schedule due to the snowball effects. Such increases in planned cost can translate to millions of dollars in unplanned crew costs. There are two ways that carriers can try to minimize these unplanned costs. The first, discussed in Section 6, is to improve the quality of their recovery procedures. The second is to focus in the planning stage on developing more robust schedules – that is, to minimize the expected operational cost of a schedule rather than its planned cost.

Evaluating Crew Schedules

Robustness is not well-defined, in fact, comparing two different schedules to determine which one is more robust can be quite difficult. In general, comparisons are done by using a simulation to approximate the operating cost of a given schedule for a particular time period (typically, one month). Clearly, such a simulation should reflect the airline operations as closely as possible.

Simulations of partial airline operations, for example, aircraft ground movement and passenger flow, have been developed, see Yu, 1998. Only recently have simulations of integrated airline operations been designed. Kornecki and Vargas, 2000, for example, developed a simulation designed for employee training. Rosenberger et al., 2000 created `SimAir`, a simulation that takes into account most airline operations and has built-in recovery modules. It keeps track of several types of resources, including aircraft, crews, and passengers, and produces a number of statistics such as crew costs and block times. Finally, Schaefer et al., 2000 use a simulation-based approach to design more robust crew schedules.

Models for Robust Airline Crew Pairing

Here we present three approaches for finding robust crew pairings.

Expected Pairing Cost Approach Schaefer et al., 2000 solve a problem very similar to the crew pairing problem (14.1). However, they replace the objective coefficients in this model with $\bar{c}_p$, which they define to be the *expected cost of pairing p.* They then solve this model with the same methodologies as presented in Section 14.4.

Of course, the difficult aspect of this problem is in computing the cost coefficients $\bar{c}_p$, given that the expected cost of a pairing depends in part on the other pairings in the crew schedule. For every pairing they compute the expected cost by running `Simair` under the assumption that the expected cost is independent of the other pairings in a crew schedule. They show that this assumption holds under the push-back recovery procedure. Push-back recovery delays the flights until all of the resources are available.

Once they have computed cost coefficients they solve this modified version of the crew pairing problem and then use `SimAir` to evaluate the quality of their solutions. They report some interesting findings. For example, they observe that it may be preferable to have some pairings in which costs are determined by TAFB or the minimum guarantee pay, rather than flying cost. In addition, they find crew schedules to be more robust when the pay-and-credit of the pairings has low variance and there are not many pairings with zero pay-and-credit. This is intuitive because zero pay-and-credit pairings have minimal connection time and are therefore vulnerable to disruptions.

Schaefer et al., 2000 compared this expected cost approach with a penalty approach that includes penalties in the cost function for such factors as tight connections and elapsed times, and tight 8-in-24 constraints. Better results were obtained by the expected cost model.

Maximizing the Connection Time Ehrgott and Ryan, 2001 and Yen and Birge, 2000 measure robustness as the excess sit connection time above the minimum sit connection time. If k is a sit connection and t is the connection time, they define a penalty by $w_k(minSit - t)$, where w_k is the penalty factor and $minSit$ is the minimum required sit connection time. They define the *robustness cost of a pairing* as the sum of the penalties over all sit connections in the pairing, excluding the sit connections corresponding to the aircraft turns. Their models find a crew schedule that minimizes the robustness cost.

Yen and Birge, 2000 solve the resulting model as a stochastic integer programming model by assuming that t is a random variable. Given a crew schedule, the recourse problem is a large-scale LP. They develop a heuristic based on follow-on branching for solving the model. They sample 100 disruption scenarios and they show the computational results on a problem with 3000 pairings and 50 legs. Their crew schedules tend to have more sit connections corresponding to the aircraft turns and longer connection times. Ehrgott and Ryan, 2001 assume that the connection time t is deterministic and it is taken with respect to the planned flight schedule. They give computational result on fleets from Air New Zealand.

The Crew Pairing Model with Move-up Crews When a crew is delayed or has reached a limit on its flying time for a duty or pairing, it would be highly desirable to have an alternative crew available with which it could swap one or more flights. The *crew pairing model with move-up crews,* presented in Klabjan et al., 2001c and Chebalov and Klabjan, 2002, relies on a recovery procedure that uses crew swaps. In addition to the traditional objective of minimizing pairing costs, they introduce a new objective of maximizing the number of opportunities for crew swapping. Thus, their model is a bicriteria optimization model.

A *move-up crew* for a given flight i is a crew that is on the ground for at least the minimum required connection time, originates at the same crew base as the crew covering *i*, and the two involved crews finish their respective pairings on the same day. If two crews can be swapped in operations, then one crew is a move-up crew. The crew pairing model with move-up crews maximizes the overall number of move-up crews and is solved by a Lagrangian decomposition approach. Computational results show that there are crew schedules with only a slightly higher crew cost but 5 to 10 times more move-up crews than the crew schedules obtained by solving (14.1). Moreover this approach, which attempts to provide protection against uncertainty rather than modeling uncertainty, can be combined with stochastic models that minimize expected cost and/or incorporate penalties.

14.8 Future Directions

Airline crew scheduling has been one of the great successes of operations research, with decision support software installed at all major airlines. Whereas a decade ago solutions to daily problems were typically 10–15% above the lower bound of flying cost, solutions are now typically within at most 1–2% of the lower bound. This improvement in solution quality translates to savings on the order of $50 million annually for a large airline.

Nonetheless, airline crew scheduling is still an active research area with many unsolved problems. We have discussed some recent work on recovery and robust planning in Sections 14.6 and 14.7, but this is clearly just the ‘tip of the iceberg’.

Benefits can be gained by developing more efficient schedules for cabin crews. This problem has received less attention than cockpit crew scheduling, both because cabin crews are significantly less costly and also because it is a much larger problem.

Finally, and perhaps most challenging, is the integration of the crew pairing, fleet assignment, and schedule planning problems, especially since these problems are difficult to solve individually.

References

Alefragis, P., Goumopoulos, C., Housos, E., Sanders, P., Takkula, T., and Wedelin, D. (1998). Parallel crew scheduling in PAROS. In *Proceedings of 1998 Europar/8,* pages 1104–1113.

Alefragis, P., Sanders, P., Takkula, T., and Wedelin, D. (2000). Parallel integer optimization for crew scheduling. *Annals of Operations Research,* 99:141–166.

Anbil, R., Barnhart, C., Johnson, E., and Hatay, L. (1994). A column generation technique for the long-haul crew assignment problem. In Ciriani, T. and Leachman, R., editors, *Optimization in Industry II,* pages 7–24. John Wiley & Sons.

Andersson, E., Housos, E., Kohl, N., and Wedelin, D. (1998). Crew pairing optimization. In Yu, G., editor, *Operations Research in the Airline Industry,* pages 228–258. Kluwer Academic Publishers.

Anbil, R., Forrest, J., and Pulleyblank, W. (1998). Column generation and the airline crew pairing problem. In *Extra Volume Proceedings ICM.* Available from `http://www.math.uiuc.edu/documenta/xvol-icm/17/17.html`.

Anbil, R., Gelman, E., Patty, B., and Tanga, R. (1991). Recent advances in crew pairing optimization at American Airlines. *Interfaces,* 21:62–74.

Anbil, R., Johnson, E., and Tanga, R. (1992). A global approach to crew pairing optimization. *IBM Systems Journal,* 31:71–78.

Barahona, F. and Anbil, R. (1998). The volume algorithm: Producing primal solutions with a subgradient method. Technical Report RC-21103, T. J. Watson Research Center.

Barahona, F. and Anbil, R. (1999). On some difficult linear programs coming from set partitioning. Technical Report RC-21410, T. J. Watson Research Center.

Barnhart, C., Hatay, L., and Johnson, E. (1995). Deadhead selection for the long-haul crew pairing problem. *Operations Research,* 43:491–499.

Barnhart, C., Johnson, E., Nemhauser, G., Savelsbergh, M., and Vance, P. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research,* 46:316–329.

Bixby, R., Cook, W., Cox, A., and Lee, E. (1995). Parallel mixed integer programming. Technical Report CRPC-TR95554, Rice University. Available from `ftp://softlib.rice.edu/pub/CRPC-TRs/reports`.

Bixby, R., Gregory, J., Lustig, I., Marsten, R., and Shanno, D. (1992). Very large-scale linear programming: A case study in combining interior point and simplex methods. *Operations Research,* 40:885–897.

Bond, D. (2000). Commercial aviation on the ropes. *Aviation Week & Space Technology.* September issue.

Brailsford, S., Potts, C., and Smith, B. (1999). Constraint satisfaction problems: Algorithms and applications. *European Journal of Operational Research,* 119:557–581.

Caprara, A., Toth, P., Vigo, D., and Fischetti, M. (1998). Modeling and solving the crew rostering problem. *Operations Research,* 46:820–830.

Chebalov, S. and Klabjan, D. (2002). Robust airline crew scheduling: Move-up crews. In *Proceedings of the 2002 NSF Design, Service, and Manufacturing Grantees Research Conference.*

Conn, A. and Barnhart, C. (2002). Improving crew scheduling by incorporating key maintenance routing decisions. Technical report, Massachusetts Institute of Technology. To appear in *Operations Research.*

Cordeau, J., Stojković, G., Soumis, F., and Desrosiers, J. (2001). Benders decomposition for simultaneous aircraft routing and crew scheduling. *Transportation Science,* 35:375–388.

Day, P. and Ryan, D. (1997). Flight attendant rostering for short-haul airline operations. *Operations Research,* 45:649–661.

Desaulniers, G., Desrosiers, J., Ioachim, I., Solomon, M., and Soumis, F. (1998). A unified framework for deterministic time constrained vehicle routing and crew scheduling problems. In Crainic, T. and Laporte, G., editors, *Fleet Management and Logistics,* pages 57–93. Kluwer Publishing Company.

Desrochers, M. and Soumis, F. (1988). A generalized permanent labeling algorithm for the shortest path problem with time windows. *INFOR,* 26:191–212.

Desrochers, M. and Soumis, F. (1989). A column generation approach to the urban transit crew scheduling problem. *Transportation Science,* 23:1–13.

Desrosiers, J., Dumas, Y., Desrochers, M., Soumis, F., Sanso, B., and Trudeau, P. (1991). A breakthrough in airline crew scheduling. Technical Report G-91-11, Cahiers du GERAD.

du Merle, O., Villeneuve, D., Desrosiers, J., and Hanses, P. (1999). Stabilized column generation. *Discrete Mathematics,* 194:229–237.

Ehrgott, M. and Ryan, D. (2001). Bicriteria robustness versus cost optimization in tour of duty planning at Air New Zealand. Technical report, Univeristy of Auckland.

Fahle, T., Junker, V., Karish, S., Kohl, N., and Vaaben, B. (1999). Constraint programming based column generation for crew assignment. *Journal of Heuristics.* To appear.

Forrest, J. and Goldfarb, D. (1992). Steepest-edge simplex algorithms for linear programming. *Mathematical Programming,* 57:341–374.

Gamache, M. and Soumis, F. (1998). A method for optimally solving the rostering problem. In Yu, G., editor, *Operations Research in the Airline Industry,* pages 124–157. Kluwer Academic Publishers.

Gamache, M., Soumis, F., Marquis, G., and Desrosiers, J. (1999). A column generation approach for large scale aircrew rostering problems. *Operations Research,* 47:247–262.

Gamache, M., Soumis, F., Villeneuve, D., Desrosiers, J., and Gelinas, E. (1998). The preferential bidding system at Air Canada. *Transportation Science,* 32:246–255.

Gedron, B. and Crainic, T. (1994). Parallel branch-and-bound algorithms: Survey and synthesis. *Operations Research,* 42:1042–1066.

Gershkoff, I. (1989). Optimizing flight crew schedules. *Interfaces,* 19:29–43.

Goumopoulos, C., Housos, E., and Liljenzin, O. (1997). Parallel crew scheduling on workstation networks using PVM. In *Proceedings of 1997 EuroPVM-MPI,* volume 1332.

Hoffman, K. and Padberg, M. (1993). Solving airline crew scheduling problems by branch-and-cut. *Management Science,* 39:657–682.

Hu, J. and Johnson, E. (1999). Computational results with a primal-dual subproblem simplex method. *Operations Research Letters,* 25:149–158.

Junker, U., Karisch, S., Kohl, N., Vaaben, B., Fahle, T., and Sellmann, M. (1999). A framework for constraint programming based column generation. In *Proceedings of CP 1999,* pages 261–274.

Klabjan, D. (2001). Next generation airline crew scheduling. Technical report, University of Illinois at Urbana-Champaign. Available from `http://www.staff.uiuc.edu/~klabjan/reports/ngCS.pdf`.

Klabjan, D., Johnson, E., and Nemhauser, G. (2000). A parallel primal-dual algorithm. *Operations Research Letters,* 27:47–55.

Klabjan, D., Johnson, E., Nemhauser, G., Gelman, E., and Ramaswamy, S. (2002). Airline crew scheduling with time windows and plane count constraints. *Transportation Science,* 36:337–348.

Klabjan, D., Johnson, E., Nemhauser, G., Gelman, E., and Ramaswamy, S. (2001a). Airline crew scheduling with regularity. *Transportation Science,* 35:359–374.

Klabjan, D., Johnson, E., Nemhauser, G., Gelman, E., and Ramaswamy, S. (2001b). Solving large airline crew scheduling problems: Random pairing generation and strong branching. *Computational Optimization and Applications,* 20:73–91.

Klabjan, D., Schaefer, A., Johnson, E., Kleywegt, A., and Nemhauser, G. (2001c). Robust airline crew scheduling. In *Proceedings of TRISTAN IV,* pages 275–280.

Klabjan, D. and Schwan, K. (2000). Airline crew pairing generation in parallel. In *Proceedings of the Tenth SIAM Conference on Parallel Processing for Scientific Computing.*

Kornecki, A. and Vargas, D. (2000). Simulation-based training for airline controller operations. In *Proceedings of Society of Computer Simulation 2000 Advanced Simulation Technologies Conference, pages* 162–171.

Kwok, L. and Wu, L. (1996). Development of an expert system in cabin crew pattern generation. *International Journal of Expert Systems,* 9:445–464.

Lavoie, S., Minoux, M., and Odier, E. (1988). A new approach for crew pairing problems by column generation with an application to air transportation. *European Journal of Operational Research,* 35:45–58.

Lettovský, L. (1997). *Airline Operations Recovery: An Optimization Approach.* PhD thesis, Georgia Institute of Technology.

Lettovský, L., Johnson, E., and Nemhauser, G. (2000), Airline crew recovery. *Transportation Science,* 34:337–348.

Linderoth, J. and Savelsbergh, M. (1999). A computational study of search strategies for mixed integer programming. *INFORMS Journal on Computing,* 11:173–187.

Lustig, I. and Puget, J. (2001). Program ≠ program: constraint programming and its relationship to mathematical programming. *Interfaces.* To appear.

Makri, A. and Klabjan, D. (2001). Efficient column generation techniques for airline crew scheduling. Technical report, University of Illinois at Urbana-Champaign. Available from `http://www.staff.uiuc.edu/~klabjan/professional.html`.

Marsten, R. (1994). Crew planning at Delta Airlines. XV Mathematical Programming Symposium. Presentation.

Minoux, M. (1984). Column generation techniques in combinatorial optimization: a new application to crew pairing problems. In *Proceedings XXIVth AGIFORS Symposium.*

Rosenberger, J., Schaefer, A., Goldsman, D., Johnson, E., Kleywegt, A., and Nemhauser, G. (2000). A stochastic model of airline operations. *Transportation Science.* To appear. Available from `http://tli.isye.gatech.edu`.

Ryan, D. (1992). The solution of massive generalized set partitioning problems in air crew rostering. *Journal of the Operational Research Society,* 43:459–467.

Ryan, D. and Foster, B. (1981). An integer programming approach to scheduling. In Wren, A., editor, *Computer Scheduling of Public Transport Urban Passenger Vehicle and Crew Scheduling,* pages 269–280. Elsevier Science B.V.

Sanders, P., Takkula, T., and Wedelin, D. (1999). High performance integer optimization for crew scheduling. In *Proceedings of the HPCS '99.* Available from `http://www.cs.Chalmers.se/~tuomo`.

Schaefer, A., Johnson, E., Kleywegt, A., and Nemhauser, G. (2000). Airline crew scheduling under uncertainty. Technical Report TLI-01-01, Georgia Institute of Technology.

Song, M., Wei, G., and Yu, G. (1998). A decision support framework for crew management during airline irregular operations. In Yu, G., editor, *Operations Research in the Airline Industry,* pages 260–286. Kluwer Academic Publishers.

Stojkovic, G., Soumis, M., and Desrosiers, J. (1998). The operational airline crew scheduling problem. *Transportation Science,* 32:232–245.

Teodorovic, D. and Stojkovic, G. (1990). Model for operational daily airline scheduling. *Transportation Planning Technology,* 14:273–285.

Vance, P., Atamtürk, A., Barnhart, C., Gelman, E., Johnson, E., Krishna, A., Mahidhara, D., Nemhauser, G., and Rebello, R. (1997a). A heuristic branch-and-price approach for the airline crew pairing problem. Technical Report LEC-97-06, Georgia Institute of Technology.

Vance, P., Barnhart, C., Johnson, E., and Nemhauser, G. (1997b). Airline crew scheduling: A new formulation and decomposition algorithm. *Operations Research,* 45:188–200.

Wedelin, D. (1995). An algorithm for large scale 0-1 integer programming with applications to airline crew scheduling. *Annals of Operations Research,* 57:283–301.

Wei, G. and Yu, G. (1997). Optimization model and algorithm for crew management during airline irregular operations. *Journal of Combinatorial Optimization,* 1:305–321.

Wilson, N., editor (1999). *Computer-Aided Transit Scheduling*. Springer Verlag.

Yen, J. and Birge, J. (2000). A stochastic programming approach to the airline crew scheduling problem. Technical report, University of Washington.

Yu, G., editor (1998). *Operations Research in the Airline Industry,* Kluwer Academic Publishers.

15 SUPPLY CHAINS

Randolph W. Hall

15.1 Introduction

In modern economies, the production of goods entails the coordinated effort of manufacturers, transporters and distributors spread across the globe. This symphony of actions, which delivers products to consumers, is called the "supply chain". For complex products, such as automobiles, the supply chain comprises a deep and intricate web, beginning with acquisition of raw materials, continuing with fabrication of parts and components, then assembly of these parts and components into finished vehicles, and finally distribution of vehicles through networks of dealers.

By contrast, supply chains in primitive societies are simple and local. Manufacturing is limited to production of simple tools and implements, relying on materials that are locally acquired or traded, and largely depending on the effort of individuals to fabricate an entire product from start to finish. In primitive societies, items are not transported over great distances, and there is far less inter-dependency among individuals or groups in the creation of goods.

The comparison of automobiles to primitive products (stone tools, baskets, clay pots, etc.) illustrates that supply chains have been a fundamental ingredient in the advancement of technology. It would be impossible to create modern automobiles, for instance, if we were unable to:

- Acquire and transport raw materials from their natural sources, which are spread across the globe.

- Create industries capable of designing and building specialized components, such as braking systems, air bags, audio systems, upholstery, tires, glass, sheet metal, etc.

- Assemble and fabricate vehicles and components in large facilities, which permit the division of labor into specialized tasks and the use of automation technologies.

- Efficiently transport vehicles and their parts over long distances.

Because each of these steps also exhibits natural economies-of-scale, it has become increasingly common for entities in the supply chain to serve larger, and more global, markets. Hence, management and operation of supply chains has become more important and more challenging.

Though the differences between primitive and modern products are obvious, one need not look back more than 250 years to see huge contrasts in the formation of supply chains. Prior to the industrial revolution of the 1800s, mass production and mass distribution were virtually non-existent, due to the absence of four technological ingredients: (1) efficient ground transportation, (2) reliable sources for power generation, (3) ability to communicate instantaneously over long distances, and (4) ability to manufacture products with sufficient accuracy to permit interchangeable parts. Without these ingredients, production was more like the primitive clan than modern society. Products were manufactured by craftsmen, largely from start to finish, serving local markets. Though people did trade materials and some products, it would have been impossible to manage a large and distributed organization, and it would have been impossible to attain a high degree of specialization.

The industrial revolution brought the needed ingredients to construct true supply chains. Locomotives and steam-powered boats permitted inland transportation. Coal engines provided a more reliable source of power for manufacturing plants. The telegraph gave the power to coordinate activities among distributed sites. And machinery reached sufficient accuracy to permit the specialized fabrication of components, for later assembly into finished products.

Over the last 200 years supply chains have continued to evolve and expand. Transportation and communication have become much less expensive. The scale economies of production have become even greater. And specialized knowledge has grown more important in the design of products. Nevertheless, the fundamental questions in supply chain design and management have changed little in this time:

- When and where should value be added to a product, through fabrication, processing, assembly, etc.?

- How should goods be transported among production and consumption locations?

- How should information be used to coordinate and enable supply chain activities?

These questions are addressed in this chapter, with emphasis on the transportation aspects of supply chains. The approach here is more conceptual, and less mathematical, than in other chapters. For further background on the topics, refer to Daganzo (1999) and Simchi-Levi *et al* (2000).

15.2 Production and Distribution Over Time and Space

The most fundamental question in supply chain management is: When and where to "add value" to a product? Value can be added through any of several processes:

Extraction: separating substances from a composite material, as in mining or refining.

Processing: exposing a material to a process that changes its composition, such as heating in a furnace.

Fabrication: changing the shape or form of a material, such as stamping automobile parts from coils of steel.

Assembly: combining materials or parts to create a new product, as in printed circuit board assembly.

Transportation: movement of materials, parts or products from one place to another.

Inventory: storage of materials, parts or products for future use.

The "supply chain" represents that entire sequence of steps that produce a finished product. Prior to the industrial revolution, most products required a small number of steps, with a large portion of these steps completed by individual skilled craftsmen. After the industrial revolution, production steps were increasingly divided among many individuals, each of whom specialized in a particular task. However, it was still relatively common for a single vertically integrated firm to oversee a large portion of these steps. Ford's River Rouge plant, which was used to manufacture the Model T, fits this model. Modern production has continued the trend toward specialization, with greater specialization by firms and production plants. Thus, it has become more important to develop systems for coordinating production among different organizations, located far from each other, possibly on different continents.

Supply Chain Elements

Depending on the complexity of the product, supply chains show substantial variation in depth and breadth. However, most supply chains contain the following elements (Figure 15.1):

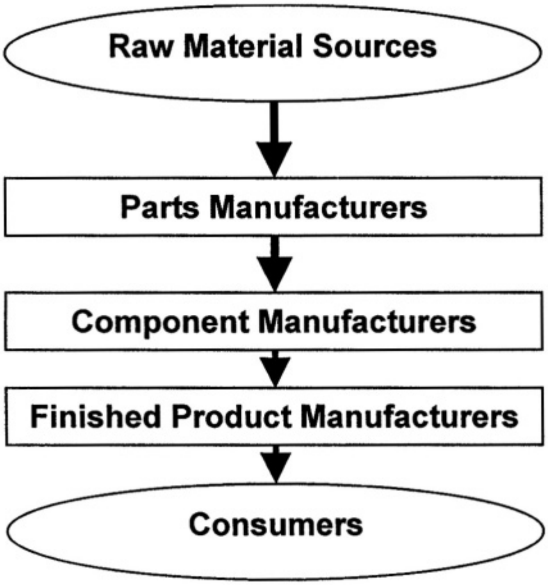


Figure 15.1 Tiers in the Supply Chain

Raw Material Extraction and Refinement: Materials are extracted from places where there are known deposits. While some materials are widely available (e.g., gravel or water), others can only be found in a few locations (e.g., diamonds). Refinement industries (e.g., oil refining in the Middle East) are frequently centered around sources, as processed materials are often more easily transported than the raw material. The availability of raw materials can also affect the positioning of manufacturing industries – the steel industry in Pennsylvania, for example.

Parts Manufacture: Parts represent the most basic and decomposable unit of a product. A part could be a screw, wire, piece of plastic, or possibly a processed substance, such as paint. Parts manufacture is relatively generic, meaning that the same processes and possibly the same parts can be used in the production of many different finished products. Parts manufacturers also tend to serve many different industrial customers, and benefit from the scale economies of serving large markets. Parts manufacturers focus on the processing and fabrication steps of production.

Component Manufacturer: A component represents a collection of parts that have been built to perform a particular function. Components can contain sub-components, which in turn contain their own sub-components. In automobile assembly, a radio is a vehicle component, a circuit board is a radio component, and an integrated circuit is a circuit board component. Thus, component manufacturing alone can entail a long supply chain. Assembly is the dominant process in component manufacture, though processing and fabrication steps may also occur.

Finished Product Manufacture: This is an extension of component manufacture – representing the final step before a product is distributed and sold. Because a finished product may in turn be a component for something else, the distinction is somewhat arbitrary. Like component manufacture, the dominant process is assembly.

Distribution: To complete the supply chain, materials must be transported to part manufacturers; parts must be transported to component manufacturers; components must be transported to finished product manufacturers; and finished products must be transported to consumers. Transportation also occurs – in the name of material handling – within each production facility. Transportation is an inevitable result of specialization in manufacture. So long as a single individual is not producing the entire product, parts and components must be moved from individual to individual, or firm to firm. Distribution also entails storage and warehousing of materials, parts, components and finished products, along with the use of independent retailers and distributors to reach the consumer. Distribution may also encompass some level of assembly tasks. An automobile dealer may customize a vehicle by installing audio systems or wheels, or a computer dealer may customize a PC by installing or swapping storage devices, memory or communication boards.

Though Figure 15.1 indicates that manufacturing is compartmentalized from distribution (the arrow in the figure), there are instances when they occur simultaneously. For instance, fish processing sometimes occur on boats, and railroad trains have been used at times as rolling production lines.

Distribution Systems

The discipline of supply chain management is highly centered on distribution, and not as much on product manufacture, though they are naturally inter-related. The economics of manufacturing strongly influence the scope and scale of distribution. For instance, products that demand a high labor content (e.g., circuit board assembly) may be assembled overseas because of lower wage rates; highly sophisticated products (e.g., integrated circuits) may only be produced in regions that offer highly skilled employees and highly skilled equipment providers. Moreover, schedules for production naturally influence schedules for transportation, as discussed in Chapter 5.

Distribution typically occurs in steps, as products move from point of manufacture to point of consumption. The number of steps depends on the degree to which the manufacturers serve local, national or international markets. The size of the market depends on the technological sophistication of the product along with the costs of transportation. Concrete production is almost always local, because the product is both expensive to transport and relatively unsophisticated. Integrated circuit manufacture is global, because the product is both technologically sophisticated and inexpensive to transport. In between, there are products like automobiles, which are relatively sophisticated but also expensive to transport (partly due to tariffs). These may be manufactured on a national scale.

Global and national products are typically distributed through a *multi-echelon* (or multi-tier) system (Figure 15.2). Manufacturing plants constitute the top-echelon, which may feed into multiple warehouses, each of which serves a large region. These warehouses could then serve multiple distributors (each covering a sub-region), which in turn serve multiple dealers, serving end consumers. In some cases distributors and dealers are independent entities (sometimes called 3rd party logistics providers), who purchase the products from the manufacturer and sell them to their own customers, or work under contract to fulfill orders on behalf of the manufacturer. In other cases dealers and distributors are solely owned franchises of the manufacturer, operating under defined terms. And in other cases all levels in the supply chain fall under the same organization.

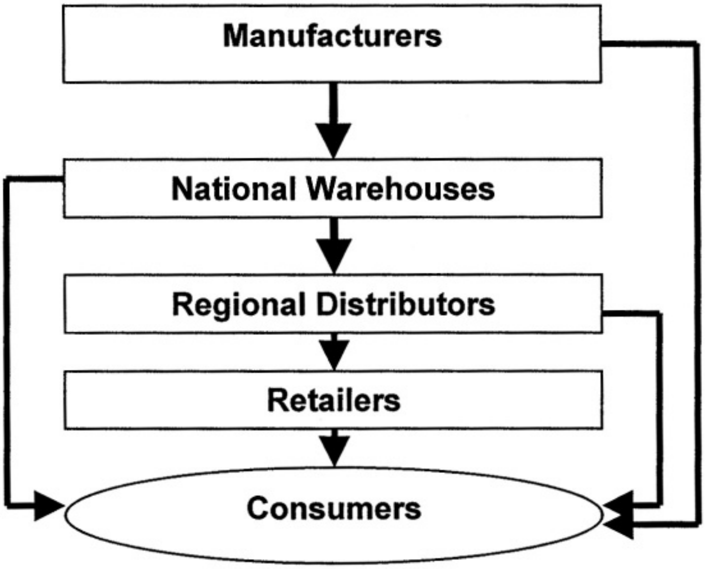


Figure 15.2 Echelons in a Distribution System

Key issues in the design of the distribution system include:

Echelon structure: number of echelons, regions served by each center specialization of warehouses in carrying particular products.

Inventory transfers: number of products stocked in each center, methods and places for acquiring products in the event of an inventory shortage, stock replacement rules.

Ownership: degree to which the manufacturing organization is involved in owning or managing the distribution entities; management of retailer relationships; ability of the manufacturer to bypass dealers and sell direct to consumers.

Integration: degree to which the distribution of replacement parts and components (called the after-market) is integrated with the distribution of original equipment.

Consolidation: methods for combining different types of product, coming from multiple sources and traveling to multiple destinations, into vehicle loads.

With respect to ownership, smaller manufacturers rely more on independent dealers and distributors than large manufacturers. Manufacturers also depend more on independent dealers and distributors when products are relatively inexpensive and simple to maintain. Vehicle and equipment manufacturers, on the other hand, are more involved in retailing, though often through franchises. This is because: the dealer must be capable of both selling and maintaining the product; because the sale frequently requires manufacturer financing; because dealers can more effectively market a single line of products than a wide variety; and because sales volumes are sufficient for a dealer to devote his or her entire effort to a single product line.

With respect to transportation in particular, sometimes the manufacturer, retailer or distributor operates its own private network, and sometimes they rely on an external for-hire network. Private networks are most common where large volumes of freight are sent between origin/destination pairs, which it makes it economical to bypass the terminals operated by for-hire networks. In some instances, private networks provide services that integrate terminal/warehouse operations with transportation. Whereas for-hire networks are designed to rapidly move products through each network node, an integrated private network may allow for the strategic management of inventories in the vicinity of consumer demand.

Private networks and for-hire networks differ in several important respects. First, private networks tend to be asymmetrical, typically with many more supply points than consumption points. Second, flows on private networks tend to be unidirectional. And third, network nodes tend to be multi-functional, providing both a transportation service and a production or inventory related service. For-hire networks share none of these characteristics, yet they still share many properties with respect to transportation operations, as will be discussed later.

Another important factor in ownership is the ability to share and utilize information. Prior to the mid 19th century, sales transactions typically accompanied the transport of goods between locations. During this period, the merchant industry thrived under a model of buying goods in one location, transporting them to another, and selling at a profit upon arrival. New means of communication (telegraph and telephone) enabled a new business model, in which different locations in the supply chain could be managed and coordinated, making it possible for manufacturers to manage the distribution system. More recently, the Internet is changing business models again, by offering a decentralized tool for pulling information from multiple sources, thus reducing the manufacturer advantage in controlling all steps in distribution. Information policies can sometimes produce counter-intuitive results, as in the bullwhip effect (Chen *et al,* 2000; Lee *et al,* 1997). Random perturbations in consumer demand can lead to larger distortions in demand further up the supply chain, as a consequence of ordering and production policies, conflicting objectives and incomplete information. Research in this area has focused on methods for stabilizing the supply chain in the presence of variability in consumer demand.

The issues mentioned so far are hardly new. In some sense, supply chain management has been the dominant ingredient in business strategy since the birth of civilization. Nevertheless, supply chain management is a new area for academic study. Though the field borrows heavily from prior research in inventory management, transportation and location theory, modeling supply chains as an integrated whole is a recent idea.

Role of Technology in Supply Chain Research

The attention given to supply chains is in large part due to the economics of manufacturing computers and related "high-tech" products (Arntzen *et al,* 1995; Lee and Billington, 1995; Brown *et al,* 2000). Unlike traditional products, which have historically encountered inflationary price increases, the cost trends in computers have been predominantly deflationary. As a consequence, the cost of holding inventories – largely due to obsolescence – have become enormous. It is uneconomical for a high-tech company to produce large quantities of a product in anticipation of future demand, because the inventories could become obsolete before the demand is realized. The pressures of price deflation not only exist for the finished product, but also for each step of the supply chain along the way. Thus, manufacturers and distributors are motivated to postpone production as long as feasible – ideally until the customer places his or her order (Garg and Tang, 1997; Lee and Tang, 1997). Gateway and Dell have prospered in this environment by establishing a centralized build-to-order production system, minimizing distribution through retailers, and carefully managing the acquisition of each computer component.

Another factor motivating supply chain research has been the advent of electronic commerce, the Internet and on-line retailing (Bollo and Stumm, 1998; Brynjolfsson and Smith, 2000; Dewan *et al,* 2000; Huang, 2001; Morrison and Wise, 2000; Partyka and Hall, 2000). These technologies have changed the economics of distribution, enabling consumers to bypass "middlemen" in some cases. In traditional shopping, a purchase follows from physical inspection of available products at stores. In electronic shopping, products are not physically inspected, but are instead electronically inspected, through acquisition of on-line product information. Furthermore, the product is transported to the consumer by commercial vehicles (possibly over long distances), rather than by the consumer in a personal vehicle. These steps eliminate the retailer tier from the supply chain, permitting direct purchase from the distributor.

On-line purchases have encountered varying success. Whereas completely transparent products – like airline tickets – are successfully sold on-line; and moderately transparent products – like books and CDs – have achieved some on-line success; other products have not. Various on-line grocers have failed, and highly personalized products, such as clothing, have not moved far beyond the traditional catalogue market.

A last motivator for supply-chain research has been product substitution. Technology already exists for on-line acquisition of music, video and reading material, though the medium has not fully matured. Though products like these are customarily delivered as physical items, there is nothing inherently physical about them. All are ultimately presented to the consumer through the senses of sight and sound, which can be created in virtual environments. Going a step further, one can imagine that future technology will enable localized production of a range of products through the electronic transmission and storage of designs, along with rapid prototyping equipment. Such technology is currently being researched.

Customization and Variety

The interplay between temporal and spatial decisions in supply chain management is illustrated through customization and variety (Boynton *et al,* 1993; Feitzinger and Lee, 1997; Van Hoek, 2000). Unlike the days of Ford's Model T (available in any color, so long as it is black), modern supply chains offer consumers a tremendous range of options within any product line. As discovered by Ford's competitor, Alfred Sloan, variety enables manufactures to create new markets and expand their sales, furthering their scale economies in production. The range of items (or "stock-keeping-units", SKUs) sold in any major grocery or department store illustrates how variety is engrained in our markets.

Depending on cost structure and competitive pressures, different industries follow different strategies toward customization. Here are two basic options.

Customization in Manufacture: A range of product types is created at manufacture through any of these means: (1) establishing a unique production line for each type, (2) using "time-share" for a single line, periodically alternating between different product types, or (3) creating a flexible production line that allows different types to be randomly interspersed. It is also possible for production to take place in a non-linear fashion, as in cellular manufacturing, but we focus on lines here. Option one provides efficiency through specialization, but may demand a greater investment in equipment and labor. Option two provides more efficient utilization of resources, but requires periodic product change-overs, which may be costly. Option three also provides more efficient utilization of resources, but demands highly flexible employees and equipment, as well as a capability for storing a wide range of part and component types in the vicinity of the production line.

A challenge with manufacturer customization is that a great variety of products must be distributed through the supply chain to consumers. With more variety, each warehouse, distributor and retailer must stock more products, at greater cost. They will also encounter relatively more variability in sales, increasing the likelihood of shortages and unsold products. Because of this variability, some manufacturers choose to "build-to-order", meaning that the product is only manufactured after an order is placed. While saving on inventory costs, build-to-order can add to manufacturing and transportation costs, and create lead-time delays in fulfilling demand.

Customization in Retailing In some instances, most of the customization takes place at the point of sale. Instead of stocking finished products, the retailer stocks parts and components, which are assembled into consumer products. Fewer items need to be stocked in such a system, because a small number of components can provide a huge number of combinations for finished products. The disadvantage is that scale economies are lost in the assembly process. As a consequence, retailer customization is only truly viable in relatively simple assembly operations that pose minimal requirements in terms of equipment or labor skills.

Customization in retailing and customization in manufacturing are two ends of a spectrum. Automobile manufacture provides both, with things like body colors and styles determined at the factory, and easy to install options, such as luggage racks, determined at the dealer. Customization can also occur at distributors, either for direct sale to the customer, or to fulfill orders placed through dealers. Lastly, customization is an important factor in product design. When variety is added at point of sale, the product must be designed in a manner that permits easy substitution of parts (e.g., parts easily snap in to a basic chassis), whereas there are fewer restrictions when the product is customized at point of manufacture.

15.3 Transportation Component of Supply Chains

Supply chains could not exist without a capability for moving products over long distances. Though one can imagine a day when products are moved more electronically than physically, supply chains today depend on efficient transportation -- in the conventional sense.

The goods movement portion of the transportation industry can be segmented by mode of travel, commodity carried and scale of operation. Ocean shipping is the predominant mode for transportation over very long distances; rail is most efficient when water is not an option and distances are long; trucking is predominant for local distribution, up to medium distances. Pipelines are common for products like fuel and water. And aircraft are used for urgent shipments over long distances.

For raw materials, transportation is highly specialized by commodity. Special vehicles are needed to carry products like lumber, milk, gasoline, and vegetables, for instance. Finished products are more likely to be containerized or boxed, which allows them to be transported in standardized vehicles.

The scale of operation can be defined by geographic coverage, number of terminals, miles of infrastructure and number of vehicles. In trucking, the largest carriers concentrate on transporting small items over long distances, in great volumes. They require substantial investments in terminals and handling equipment, which creates a barrier to entry for smaller firms. Due to the nature of their equipment and their focus on long distances, ocean shipping and railroad companies are inherently large in scale.

Virtually all forms of transportation exhibit strong scale economies. By attracting larger volumes of freight, companies are better able to: fill their vehicles to capacity, or increase frequency and density; operate larger and more efficient vehicles, and more efficiently spread the costs of investments in terminals and infrastructure. To achieve these scale-economies, freight carriers utilize the process of consolidation: the act of combining shipments with different origins, destinations and times of travel into vehicle loads (Hall, 1987c). There are just three basic approaches to consolidation:

1. **Temporal:** accumulation of shipments over time until a desired load size is attained.
2. **Terminal:** transshipping from incoming to outgoing vehicles at a terminal to attain desired loads. (Figure 15.3)
3. **Vehicular:** routing a vehicle among multiple stops, picking up and dropping off shipments, to attain a desired load size. (Figure 15.3)

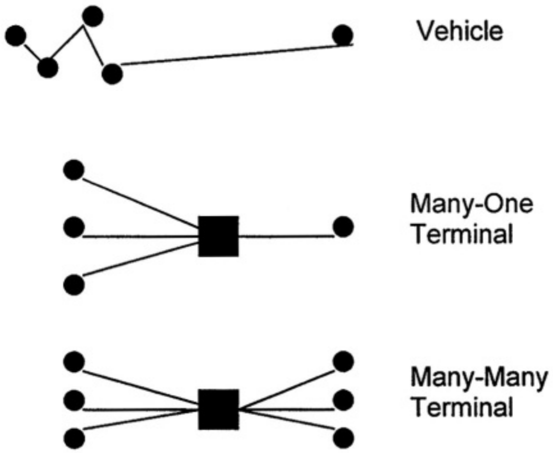


Figure 15.3 Vehicle and Terminal Consolidation Strategies

Whereas consolidation adds to transportation efficiency, it also imposes costs on the carrier and the customer. Temporal consolidation creates delays and inventory costs. Terminal consolidation demands additional investment in facilities, and causes shipments to travel by more circuitous routes. Vehicular consolidation causes both vehicles and shipments to travel by more circuitous routes. Taking all factors into consideration, the optimal design of a transportation/distribution system must balance numerous competing trade-offs between cost and service, usually blending all three consolidation approaches.

The value of consolidation can be illustrated with a simple example. Suppose that a distribution network serves 15 million pounds of freight per day – a rather large quantity -- along with 5000 origins and 5000 destinations. And suppose that the freight is evenly distributed among origin-destination pairs. Then the network serves a total of 25 million (5000 x 5000) origin-destination pairs, which results in just .6 pounds/day per pair. It would certainly be uneconomical to provide a daily direct route between each pair, when the freight is comparable to a large manila envelope. And even if the consolidation cycle were extended to an annual period, the total accumulation would be just 219 pounds per route, still likely not enough to justify direct routes.

As an alternative, suppose that shipments are transported through a system like the one shown in Figure 15.4. Suppose that one hub terminal is created, through which all shipments are processed, along with 50 regional terminals, each of which serves 50 origins and 50 destinations. And suppose that the origins and destinations are served by multi-stop routes, with 10 stops each. Then 500 local pickup and delivery routes would average 30,000 pounds per day in volume, and 50 terminal-to-terminal routes would average 300,000 pounds per day in volume. In this scenario, it becomes economical to provide daily service between all origin destination pairs.

The example illustrates the concept underlying the original Federal Express network (Hall, 1989a; Kuby and Gray, 1993). By focusing its traffic on a single national terminal, Federal Express amassed sufficient freight flows to make daily overnight service economical. Over time, the Federal Express network has grown and become more complex. Nevertheless, it still relies on a terminal network to consolidate its traffic and make transportation efficient.

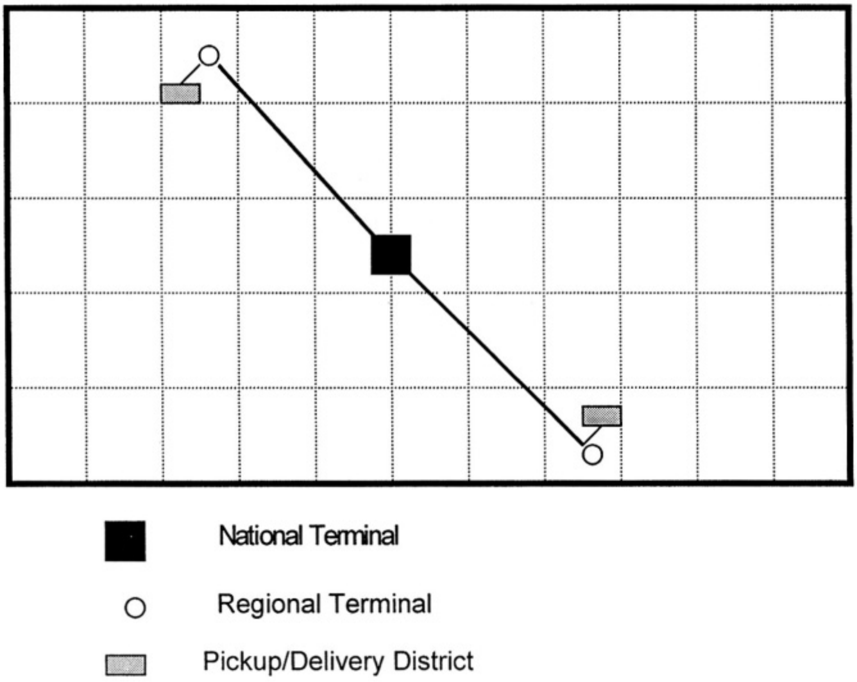


Figure 15.4 Consolidation Through a National Terminal

Terminal Consolidation Strategies

Terminals present many consolidation options and decisions. Adding terminals to a network is desirable from the standpoint of reducing travel distances for local pickup and delivery routes (Hall, 1985b). However, with more terminals, more routes are needed to connect terminals with each other, reducing the advantage of consolidation because each route serves less volume. Adding terminals also increases the investment in terminal infrastructure. Beyond determining how many terminals to operate and where they are placed (Campbell, 1993; Klincewicz, 1998; O'Kelly, 1986; and O'Kelly and Bryan, 1998), various strategies exist for routing freight. Terminals can be classified as follows:

One-One: Though not truly a consolidation strategy, terminals sometimes exist for the sole purpose of transferring vehicles, trailers or containers from one driver to another. These terminals permit drivers to travel part way to a destination, swap loads with another driver, then return to the origin, avoiding an overnight stay and saving on driver costs.

One-Many/Many-One: Some terminals serve the rather simple function of consolidating many incoming shipments into a single longhaul load, or the opposite (de-consolidating from one longhaul to many outgoing; sometimes called a feeder terminal). This is not the predominant use of terminals. A single long route is shorter, and generally more cost efficient, than many local routes (Figure 15.5). Nevertheless, one-many/many-one terminals are useful when it is impossible to complete a long route within a reasonable time period (e.g., a one-day workshift), or when it is impossible for the longhaul vehicle to navigate the local infrastructure (e.g., an ocean ship, or even some large trucks, cannot travel along local streets; Daganzo, 1987a).

Many/Many: A more common use of terminals is for transshipment from many incoming vehicles to many outgoing vehicles. Many-many consolidation is extremely difficult to achieve in vehicles alone, as it is difficult to organize and sort shipments with simultaneous pickup and delivery. Because of the multiplying effect of serving many origins and many destinations, many-many consolidation greatly increases the potential to achieve large loads, even when the flow of freight between origin/destination pairs is small.

The three basic terminal types are the building blocks for the construction of terminal networks, which further define the consolidation strategy in the following ways.

Hierarchies: Some transportation networks utilize a multi-echelon structure, similar to those in inventory systems. An example is provided in Figure 15.6. A service region is broken into terminal areas, and local districts. Shipments traveling from one area to another are processed through the single national terminal. Shipments traveling from one district to another, within an area, are processed through a regional terminal. Shipments traveling within a single district are only processed at the local terminal. Hierarchical strategies, like these, balance the need for achieving large load sizes against the cost of long and circuitous routes.

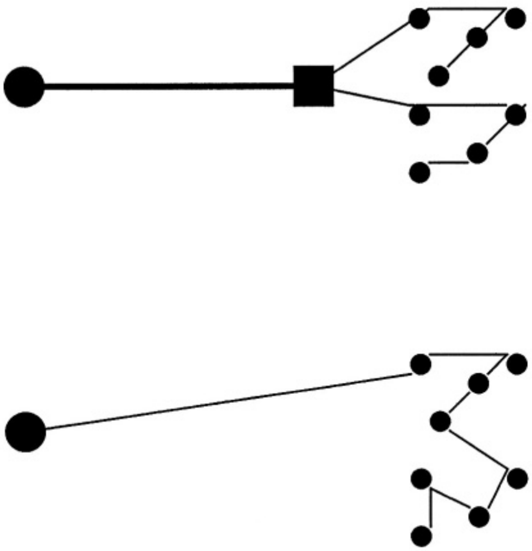


Figure 15.5 Single long-route produces shorter route length than one-many terminal

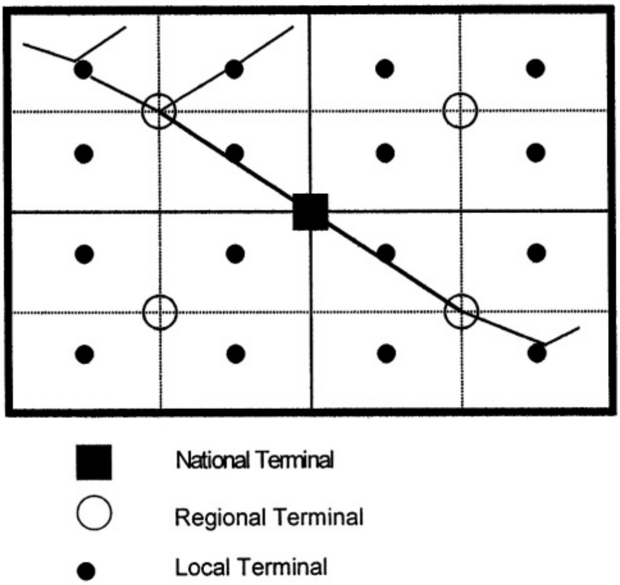


Figure 15.6 Hierarchical Network

Wide Area and Local Networks In most supply chains, origins and destinations tend to be clustered around major cities located in metropolitan areas. As a consequence, distribution networks tend to fall into either of two classes: wide-area, connecting metropolitan areas, and local area, connecting locations within a metropolitan area (Figure 15.7). A local area network (LAN) may contain multiple terminals, each serving a district on the order of 100 to 500 square miles, along with one or more gateway terminals, which provide the interface to the wide area network (WAN) (Hall, 1993). In some sense, an individual factory or building material handling system may be viewed as a sub-local network. Thus, a supply chain can be viewed as the composition of three network layers: the WAN, which connects metropolitan areas, the LAN, which connects sites within a metropolitan area, and the sub-LAN, which connects stations within a building or complex.

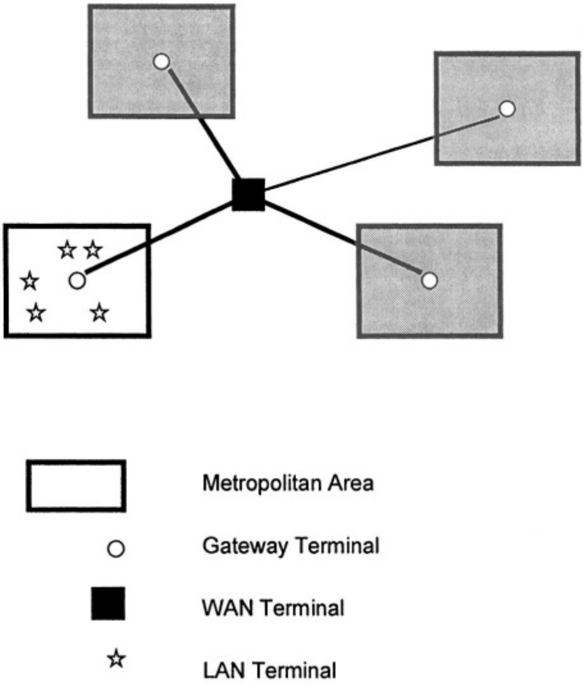


Figure 15.7 Wide Area and Local Area Terminals

For example, companies like UPS and FedEx operate multiple stations (or local terminals) within each metropolitan area. These are the places where pickup and delivery trucks are based. The stations are connected to major terminals, frequently in the vicinity of airports, which provide gateways to the WAN. At the national scale, FedEx operates multiple regional hubs along with a single large national hub, for consolidating shipments sent between metropolitan areas. Large metropolitan areas, such as Los Angeles, provide multiple gateways, though a single large gateway provides a richer set of routing options. This leads to the set of routing choices illustrated in Figure 15.8).

- Shipment can be routed through either a minor gateway or a more distant, but larger, gateway.
- If the shipment travels via the minor gateway, it can either be routed through a national hub or possibly a regional hub.
- If the shipment travels via the major gateway, a direct route may be provided to a major gateway in the vicinity of the destination, reducing longhaul travel distance, eliminating a handling step, and compensating for longer local distances.

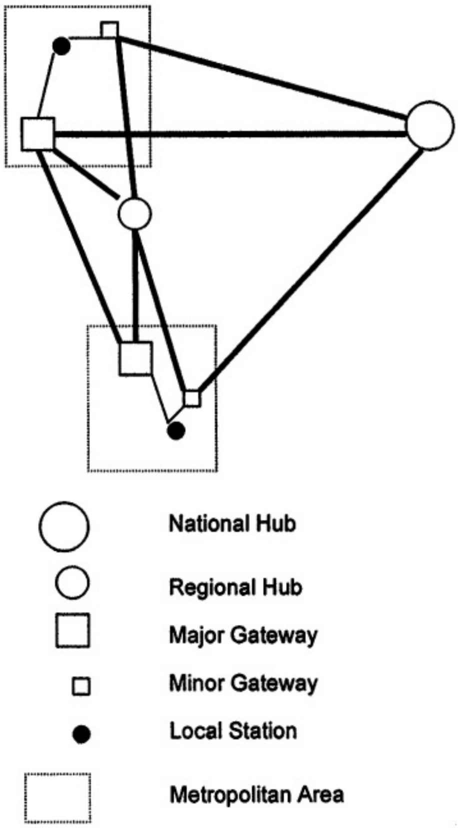


Figure 15.8 Multiple Routing Options from a Large Metropolitan Region

It should be evident that the best route for any given origin destination pair depends on how other freight is routed, creating a highly interdependent optimization problem.

Timing is an important factor in the design of both wide and local area networks. Deliveries typically occur in the morning, pickups in the afternoon, and terminal-to-terminal operations in the evening, when businesses are closed. Whereas local stations are positioned to facilitate direct access to customers, gateways and hubs are less restricted, and in some instances can be situated outside of metropolitan areas. If next day deliveries are desired, aircraft may be used for some terminal routes, and the entire system will be choreographed to enable quick transfer between vehicles and tightly scheduled routes.

Peer Structure A terminal is considered a peer if it falls at the same level in a hierarchy as another terminal. In Figure 15.9, regional terminals are peers. Local terminals, within the same region, are also peers. In a pure hierarchy, direct routes must strictly follow the diagram, having the structure of a tree. However, it is often advantageous to permit routes between peer terminals, or permit routes that cross regional boundaries. In addition, the top echelon of the hierarchy may have multiple terminals, in which case some peer-to-peer routing may be needed to complete deliveries.

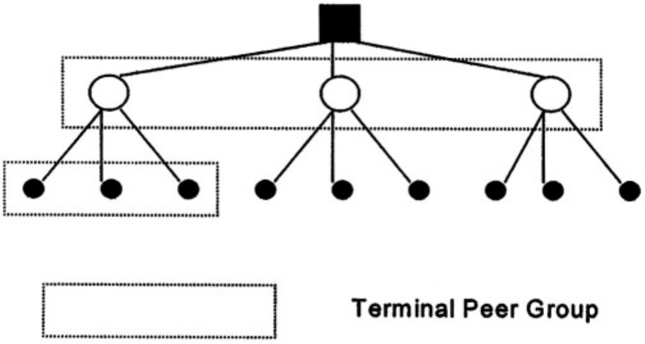


Figure 15.9 Terminal Peer Structure

Figure 15.10 provides an example of a "two-terminal" structure, meaning that shipments visit two peer terminals at the top echelon (Hall, 1987b). Figure 15.11 provides an example of a "one-terminal" structure, meaning that shipments visit one terminal at the top echelon, but cross regional boundaries. The two terminal structure is advantageous for widely distributed origins and destinations with symmetrical flow patterns (i.e., the number of origins is comparable to the number of destinations, as in for-hire networks). The one terminal structure can be advantageous with asymmetrical flow patterns, as can occur in manufacturer supply and distribution networks. The decision of one pattern or another often rests on the number of routes required to connect the network, which depends on the number of origins and destinations served, along with the importance of minimizing travel distances through provision of extra terminals and more direct routes.

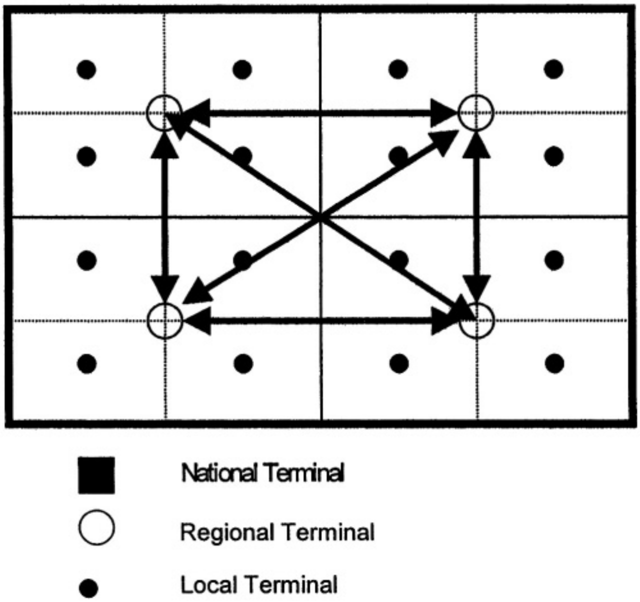


Figure 15.10 Peer-to-Peer Routing in 2-Terminal Pattern

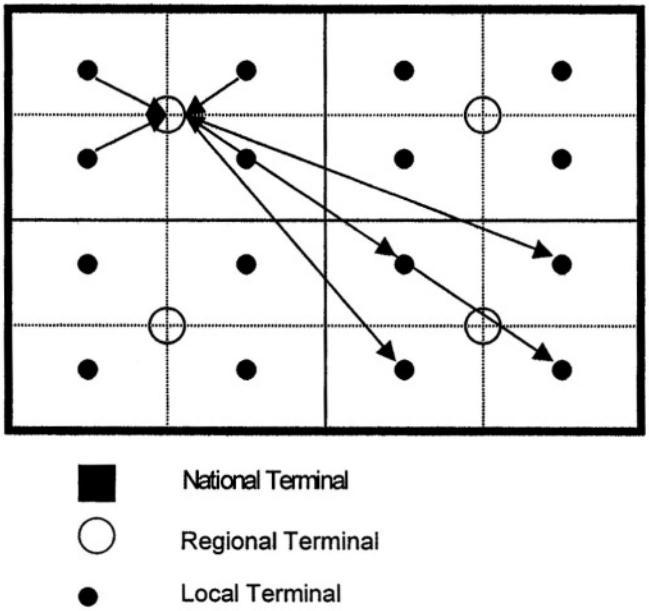


Figure 15.11 Cross-regional Routing in 1-terminal Pattern

Multi-hop Network: In a multi-hop network, shipments are transported between terminals in steps (or hops), with some level of sorting occurring at each step along the way. Because handling and processing occurs multiple times, multi-hop is most viable when shipments are transported in large containers, which are not opened at intermediate stops. An example would be a rail network, where trains travel from classification yard to classification yard and only rail cars are sorted (not their contents). Multi-hop networks have also been used in regional trucking networks, where the majority of shipments are traveling short distances (hence, most shipments are not handled many times). Multi-hop is advantageous from the perspective of keeping driver route lengths shorts, so that they return home in a reasonable time, and from the perspective of minimizing the number of routes emanating from each terminal.

In a multi-hop network, each terminal is connected to its adjacent peers by a direct route. Each of these peers can than act as a consolidator for shipments traveling to more distant destinations, in the general direction of the adjacent peer (Figure 15.12). Because only a small number of routes is needed from each terminal, consolidation levels are high. The disadvantage is that shipments must be re-handled at each terminal, which is impractical if shipments travel very long distances.

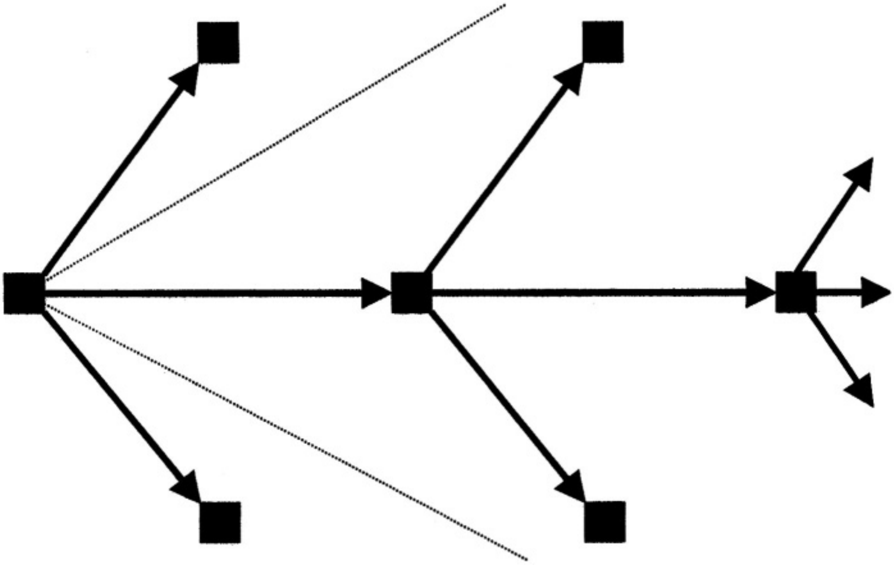


Figure 15.12 Multi-hop Network

Vehicular Consolidation Strategies

Like terminals, vehicles can consolidate shipments traveling between multiple locations, but their capabilities are limited (Daganzo, 1987a, 1988). Whereas a terminal can quite readily group shipments among widely dispersed origins and destinations, vehicles are limited by the distance that they can travel in a reasonable tour – often no more than a driver can cover in a day. Also, vehicles are not designed to facilitate shipment sorting, which makes it difficult to mix pick-ups and deliveries on the same tour. For these reasons, vehicular consolidation is typically limited to tours that begin an end at a terminal, distributor, warehouse or factory (Hall, 1995a).

It should be kept in mind that a stop on a vehicle tour may either be a shipment origin or destination, or it may be a terminal. Thus, it is possible to connect terminals to each other with multi-stop tours. Nevertheless, the primary application of vehicular consolidation is in pickups from shipment origins and delivery to shipment destinations. This is because terminal networks are designed to provide sufficient consolidation to make it economical for direct connections among terminals. If flows are insufficient to permit direct connections, it is often advantageous to eliminate some of the terminals, or change the terminal hierarchy or peer structure in a manner that makes direct connections feasible.

Pickup or Delivery: A common form of tour – especially among private carriers -- is the pure pick-up or pure delivery route. The pure pick-up route (many-to-one), also called a collection route, may be used to bring supplies into a major production facility. Somewhat more common, the pure delivery route (one-to-many) may be used to distribute products to retailers or consumers (e.g., food deliveries to grocers, appliance deliveries to consumers). The latter is more prominent because supply chains have a tendency to expand at lower levels. This is most obviously true at the bottom level of the chain, as there are many more consumers than there are retailers, distributors, manufacturers, or raw material sources. Thus, with more points for delivery than pick-up, one-to many routes are more prominent than many-to-one.

In the absence of sequencing and timing restrictions (e.g., time windows), pure pickup or delivery routes are structured like Figure 15.13. The service region is partitioned into districts, each with sufficient work to fill a vehicle or workshift, and each district is served by a loop. Vehicles travel from a single node (e.g., a terminal, warehouse or factory) to multiple stops located in a district. The route consists of a line-haul portion (to and from the district) and a local portion (within the district; Daganzo 1984a, 1984b).

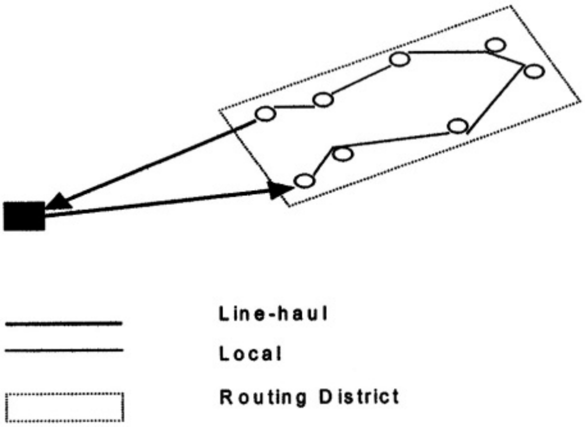


Figure 15.13 Simple Pick-up or Delivery Tour

Pickup and Delivery/Common Node: Another common route type provides both pickup and delivery out of a single node, or location (Daganzo and Hall, 1993). For-hire carriers routinely operate in this manner out of their local terminals. Most commonly, deliveries and pick-ups occur in different parts or the day, the former in the morning and the latter in the afternoon. This approach coincides with natural work patterns, as deliveries are desired as early in the day as possible, whereas pick-ups are desired as late as possible, to provide as much work time as possible before fulfilling orders. A partition of pickups and deliveries is also desirable from the standpoint of shipment handling, as vehicles can be totally unloaded of deliveries prior to filling the space with pick-ups. In reality, some amount of intermixing of pick-ups and deliveries may occur, but only to a limited degree. Because supply chains have a tendency to expand at lower levels, mixed pick-up/delivery routes tend to make more deliveries than pick-ups, and also devote a greater portion of their day to deliveries than to pick-ups.

In the absence of sequencing and timing restrictions, mixed pickup and delivery routes are structured like Figure 15.14. The service region is partitioned into pickup districts and delivery districts. A loop is formed by pairing a delivery district with a pickup district, which are connected with a "deadhead" segment (empty travel). In some instances (e.g., when pickups and deliveries naturally occur in the same area) the pickup and delivery districts coincide, in which case the deadhead segment is eliminated.

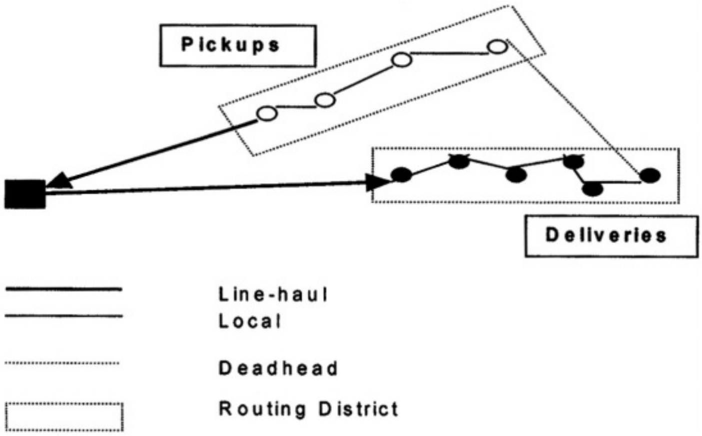


Figure 15.14 Combined Pickup and Delivery Route

Multiple Node Pickup or Delivery: In a multiple node route, a route may begin in the vicinity of one node (or terminal) and end in the vicinity of another. To complete a tour, the route may be paired with a "backhaul" route, which ends up in the vicinity of the original node. Or, in more complicated networks, routes may be combined into longer tours rooted from three or more nodes. The process reduces the empty mileage returning to a facility. A fundamental characteristic of such a network is that each node serves large regions, which encompass other nodes. This may occur, for instance, in manufacturer networks, where each node represents a manufacturing plant specializing in a particular product line. It is unlikely to occur in a for-hire transportation network, as each node (a terminal), would be assigned a unique, and non-intersecting, service territory.

Route structure depends on region, as illustrated in Figure 15.15 (Hall, 1991a). Loads delivered in the vicinity of a home terminal have a two-way, out-and-back, structure, like the pure pickup or pure delivery route in Figure 15.13. Loads delivered in the vicinity of another node (falling outside the "backhaul boundary" of a terminal) can either have a one-way orientation or a two-way orientation, according to the illustrated boundaries (one-way in the center, two-way in the periphery). In Euclidean space, a two-way territory is aligned along the boundary of an ellipse, having the two terminals as foci. A one-way territory is aligned perpendicular to the ellipse. The backhaul boundary is determined by solving the classic transportation problem. In Euclidean space, the boundary has a hyperbolic shape, with foci corresponding to the terminals (Hall, 1989b). In the case of multiple node networks (3 or more), boundaries are created from hyperbolic segments, again defined by solutions to transportation problems.

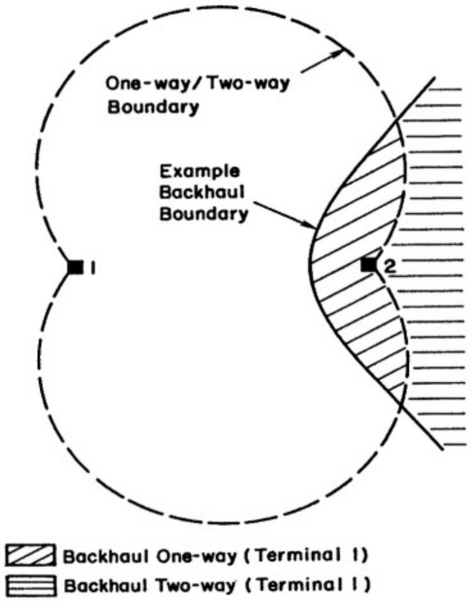

One-way/Two-way Boundary
Example Backhaul Boundary
1
2
▨ Backhaul One-way (Terminal 1)
☰ Backhaul Two-way (Terminal 1)

Figure 15.15 Multi-node Pickup or Delivery Route

General Pickup and Delivery: General intermixing of pickups and deliveries is rare for several reasons. First, shipment travel distances are usually too long to make it practical for a single route to provide the entire transportation service – both pickup and delivery (Daganzo, 1987a, 1988). Second, it is difficult to fully utilize a vehicle's capacity, because new pick-ups can block access to upcoming deliveries. Third, intermixing forces the driver to sort shipments while in route. Without access to specialized sorting equipment, as can be constructed in terminals, the process is highly labor intensive. Last, shipments often need to be brought to a facility for processing and billing purposes.

Mixed pick-up and delivery is used by courier services, where shipments travel relatively short distances (up to about 50 miles), and shipments must be delivered within hours of a request. In such instances, the courier rarely carries more than a few shipments at a time, and operates something like a taxi service. These systems tend to be highly dynamic and unstructured. In some cases a small number of origins in close proximity may be paired with a small number of destinations -- in close proximity to each other, yet far from the origins -- to create a general pickup and delivery tour. This has the natural advantage of circumventing terminal handling, but practical applications are rare, as location patterns tend not to follow this structure.

scheduled tour, and is fed by multiple pickup and delivery routes (Kahmoun and Hall, 1996b). The backbone then operates as a sort of virtual terminal, connecting many locations. However, its practical application is limited to metropolitan regions; otherwise travel time delays become excessive. Mass transit systems frequently adopt this structure, with rail lines serving as the backbone and bus lines serving as the feeder system.

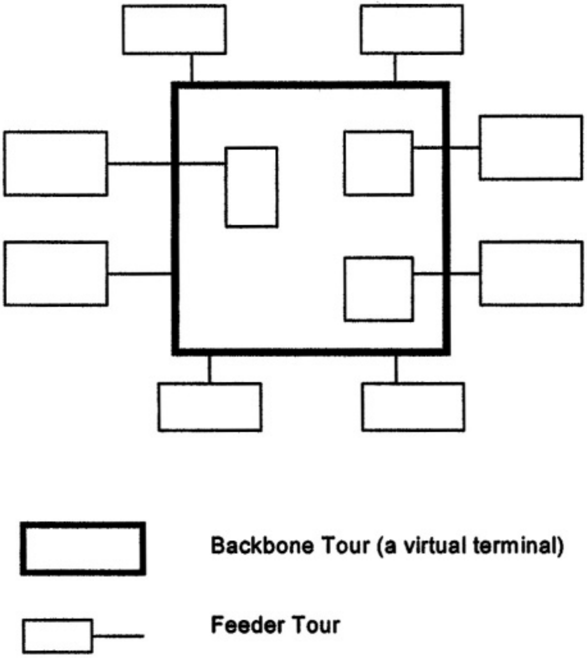


Figure 15.16. Backbone tour acts as a virtual many-to-many terminal

This discussion has assumed that vehicle routes can be characterized by geographic partitions of service regions into route territories. In reality, routes may be overlapping, for several reasons:

- Deliveries or pick-ups must occur within designated time windows, which may be better served by a vehicle from outside a normal territory, or by establishing time-dependent territories (Chapter 12; Daganzo, 1987b).
- Requests for service occur by a dynamic or random process, forcing areas to be repeatedly served throughout the day (Hall, 1996).
- Vehicles must cross into other territories to obtain shipments that fit within available vehicle space, thus maximizing capacity utilization (Hall and Daganzo, 1985; Hall, 1989d; Hall, 1994).

- To maximize utilization of driver time, vehicles cross into other territories to complete a stop with a desired duration, fitting within available workshift time.
- Inventory management at stops necessitates serving stops at particular times. (Ukovich *et al*, 1998).
- Driver learning makes it desirable to permanently assign some locations to a fixed route, while other locations are reassigned daily to balance loads (Zhong, 2002).

There are naturally many more variations on vehicle routing; interested readers should refer to Chapters 12 to 14.

15.4 Optimizing Trade-offs in Logistics Costs

In Chapter 5, the models presented for dispatching optimize one type on consolidation trade-off: the balance between waiting cost and transportation cost when shipping on a single link of a network. At a strategic level, it is important to consider more than one transportation link, and to optimize the total logistics costs of a system. By total logistics costs, we mean those costs that are related to the movement of goods between locations in the supply chain, including the following:

Inventory and Delay: These are functions of the frequency of service on transportation routes (see Chapter 5) and the time duration of transportation routes.

Pickup and Delivery Routes: Initial and final routes that serve shipment origins and destinations, depending on the number of stops per route and the length of these routes.

Terminal Routes: Routes that connect terminals to each other, depending on the length of number of routes needed to connect the network.

Terminals: The number and size of terminals that need to be established and operated.

As discussed in the prior sections, there are many ways to configure a distribution network. The best network for any given situation is the one that optimizes the balance between these costs while providing the desired level of service.

Distance Modeling

The length of transportation routes is one of the primary factors affecting these costs. For instance, the cost of local routes depends on the size of the region served by each local terminal. With a larger region, vehicles travel longer line-haul distance, adding to route length and reducing the number of stops that can be visited in a workshift.

The placement of terminals can also affect the length of terminals routes (Langevin and Campbell, 1996, survey distance approximation methods).

Transportation route lengths depend on the quality and topology of the network infrastructure. Within major cities, streets tend to follow a rectangular grid pattern, which leads to an average route length that is approximately 27% longer than a straight-line Euclidean path (representing the ratio $4/\pi$, or the average value of $\sin(\theta)$ + $\cos(\theta)$ for uniformly distributed θ). Routes connecting major cities tend to be shorter than rectilinear paths, but still exceed straight-line distances by about 10-20% in the United States.

For a pure pickup or pure delivery route (without sequencing or timing constraints, and with uniformly and independently distributed stops), route length, L, may be approximated by a function of the type:

$$L = N\left[\frac{d}{c} + k\sqrt{\frac{1}{\rho}}\right] \qquad (15.1)$$

where N is the number of stops served, d is the average distance from stops to terminal, c is the vehicle capacity measured in stops, ρ is the stop density per unit area and k is a network specific parameter, in the range from .6 to .8 (Daganzo, 1984b). The model has two components, representing line-haul distance and local distance. The average line-haul distance per stop declines as capacity increases, as more stops are served per trip. The local distance per stop does not depend on the capacity, but does depend on the distance separating stops on an optimal tour. This is a function of the stop density: when stops are closer together, local distance declines.

The average stop-terminal distance, d, depends on the size of the region served by the terminal, the placement of the terminal within the region and the network infrastructure. For a centrally located terminal within a circular region, uniformly distributed trip ends, and Euclidean distance, $d = (2/3)\sqrt{A}$, where A is the region's area, d becomes larger for non-circular regions, and for other distance metrics. For instance, for a centrally located terminal within a square region, rectilinear paths and uniformly distributed trip ends, $d = \sqrt{A}$. d also increases when the terminal is displaced from the central point. On the other hand, clustering of trip ends tends to reduce average trip length, especially if it is possible to place terminals in the centers of clusters. There is a fairly large literature on distance calculations; readers may refer to Vaughan (1984), as well as Chapter 9.

A Combined Vehicular and Temporal Model

Consider now a model that combines vehicle dispatch with vehicle routing (consolidation approaches 1 and 3). A region is partitioned into territories, each served by one vehicle. We wish to optimize two decisions: the size of territories (measured in stops), and the time interval between service. One possibility might be to create one long route that serves the entire region, with very frequent vehicle departures. At the other extreme, each individual stop might be served by its own route, but only very infrequently. Both options may result in the same average load sizes, and hence the same level of consolidation. However, the first option produces highly circuitous routes, while the second option produces long delays (due to low frequency). As might be expected, the optimum falls somewhere in between.

In examining a simple case, where stops are uniformly and independently distributed across a region, demand occurs at a constant rate, inventory costs occur at both origin and destination, and all stops are served from a common source, Burns *et al* (1985) obtained the following results:

- The optimal territory size corresponds to vehicle capacity, no matter what time interval is set for the tours. Routes should be made as long as possible, without exceeding capacity, because transportation cost per stop is a declining function of route length. Cost is then:

$$\begin{aligned} C &= \text{cost per unit time and area} \\ &= \frac{d\rho\lambda}{V} + (\beta + \alpha k\sqrt{\frac{1}{\rho}})(\frac{\rho}{T}) + h\rho\lambda T \end{aligned} \tag{15.2}$$

where:

d = average line-haul cost from terminal to stop
T = interval between dispatches
β = a fixed cost of visiting each stop (independent of distance)
h = inventory holding cost, per unit time and unit demand
V = vehicle capacity, measured in units of demand
k = parameter representing distance metric, on the order of .6 to .8
α = average local cost per unit distance
λ = demand rate, per stop
ρ = stop density

- The optimal time interval, *T*, equals:

$$T = \sqrt{\frac{\beta + \alpha k\sqrt{\frac{1}{\rho}}}{h\lambda}} \tag{15.3}$$

Eq. 15.3 is analogous to the basic economic order quantity model in Chapter 5. However, the optimal interval is a function of the local cost of serving a stop (the value in the parentheses of the numerator), and not the line-haul cost. Because the local cost tends to be much smaller than line-haul cost, *T* tends to be much smaller for multiple stop tours than for single stop tours, and stops can be visited more frequently.

Terminal Consolidation

By adding terminals, a distribution network can improve its capability to serve widely dispersed origins and destinations. For illustration, we consider a 2-terminal routing strategy, as shown in Figure 15.10. We assume that terminals are equally spaced on a square lattice within a square region, that shipments travel by rectilinear paths, and that shipment origins and destinations are independently and uniformly distributed. Then the following performance measures apply to the network (Hall, 1984):

d_1 = average linehaul trip length, origin/destination to terminal

$$= x/(2\sqrt{t}) \tag{15.4a}$$

d_2 = average trip length, terminal to terminal

$$= (2x/3)(1-1/t) \tag{15.4b}$$

r_1 = total number of routes connecting origins/destinations to terminals

$$= M \tag{15.5a}$$

r_2 = total number of routes connecting terminals to each other

$$= t(t-1) \tag{15.5b}$$

l_1 = total linehaul length of routes connecting origins/destinations to terminals

$$= xM/2\sqrt{t} \tag{15.6a}$$

l_2 = total length of all routes connecting terminals to each other

$$= (2x/3)(t^2-t) \tag{15.6b}$$

where

x = length of side of square region
t = number of terminals
M = number of origins and destinations

As indicated in Eqs. 15.4a and 15.6a, adding terminals to a network (increasing t) is advantageous from the perspective of reducing the average origin/destination to terminal trip length and route length. And although, the average terminal-to-terminal trip length is an increasing function of t, the entire trip length (origin to destination) always declines as t increases.

As more terminals are added to a network, costs do eventually rise for two reasons. First, the cost of building and operating terminals will increase faster than the transportation savings due to reduced trip length. Second, it will become impossible to fill vehicles to capacity without resorting to multiple stop tours on terminal-to-terminal routes. This is clearly seen in the equations for r_2 and l_2, which increase quadratically as *t* increases. For large values of *t,* freight flows are diluted among too many routes, leading to higher costs per unit freight and distance.

It should be noted that these trade-offs depend little on local pickup and delivery, as local territories are sized according to vehicle capacity, not linehaul trip length. There is, however, a natural dependency between the number of terminals, *t,* and the shipping interval, *T.* If *T* is enlarged, the demand per vehicle route will increase, which would make it economical to add terminals to the network. Nevertheless, because r_2 and l_2 are quadratic functions, the percentage increase in *t* would be substantially smaller than the percentage increase in *T.*

As already indicated, there are numerous variations in network design. While most networks encounter similar trade-offs – with respect to number of terminals, size of pickup/delivery territories, and shipping interval – the cost models are highly dependent on the particular situation. Daganzo (1999) provides numerous examples.

Accounting For Different Requirements and Characteristics

Real supply chains exhibit considerable heterogeneity, in terms of spatial and temporal distribution of freight flows. Whereas average flows may point to a particular design, the specific flows from a particular origin or destination may demand something different. As a consequence, most networks exhibit a mixed structure, combining elements from different fundamental designs. For example:

Bypass: When freight flow is sufficient, direct routes may be created between locations that would not normally be connected.

Multiple Networks: Separate networks may be created to serve different classes of shipments.

Multiple Stops: Locations with less than normal demand may be consolidated into a multi-stop route, when most other locations have direct routes.

Alternate Frequency: Stops with less than normal, or greater than normal, demand may be served with different frequency, to achieve a better balance between waiting/inventory costs and transportation costs.

These are all complicated issues, as the decisions made for one location, or one origin-destination pair, affect the costs for others. Because distribution costs are inherently concave, the marginal cost for serving an additional unit of freight flow is less than the average cost. Therefore, not only is it impossible to determine the optimal route for one unit of flow without simultaneously knowing the flow patterns for all other shipments, but there is tremendous incentive to group different types of shipments and serve them in the same manner. Nevertheless, origin-destination pairs that generate large freight volumes ought to be treated differently than those that do not; it is uneconomical to route a large load through a series of terminals when it could be loaded on a truck and sent direct to the destination. In a broad sense, system design might be viewed as a balancing act between exploiting the economies of consolidation and the needs to serve individual shipments or origin-destination pairs in a manner that is customized to their specific characteristics.

One of the key decisions is when to provide direct shipments between two locations, bypassing an intermediate terminal. If the network exhibits a concave cost structure, and arc flows are uncapacitated (as is often the case), then an optimal flow pattern will provide "all-or-nothing" routing for each origin-destination pair. This means that 100% of the flow for each pair will be assigned to a single path. In addition, in an optimal solution, it must be impossible to reduce the cost by reassigning the flow for any origin-destination pair to an alternate path. Thus, for Figure 15.17, the cost of a direct route must be less than or equal to the incremental cost of serving the freight flow on the terminal route, in order for it to be used. Likewise, if flow is transported on the terminal route, the cost of the direct route must be greater than or equal to the incremental savings due to removing the flow from the terminal route.

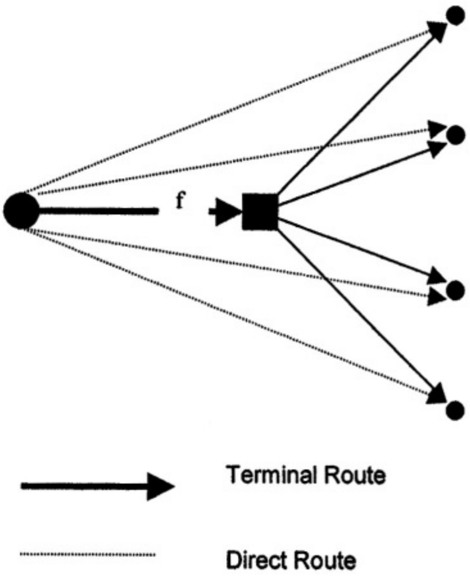


Figure 15.17 Terminal Bypass

These are necessary, but not sufficient, conditions for optimality. In some instances, a locally optimal solution may place all flow on the terminal route, and another locally optimal solution may place all flow on direct routes. This is because the terminal route only becomes attractive once a "critical mass" of flow follows the route. The search for optimality depends on finding all of the locally optimal solutions, and comparing them to determine which is best. In simple networks, like

Figure 15.17, the local optima can be found with a one dimensional search (Hall, 1987a, 1989c); for any value of flow on the terminal arc, f, the incremental cost of adding or reducing flow can be calculated to assign origin-destination pairs to routes. Local optima occur when the flow assigned to the terminal route exactly matches f. By searching across different values of f, local optima can be identified.

A characteristic of distribution networks is that terminals may be used most in intermediate demand ranges. With very high demand, direct routes become economical. With very small demand – especially if the number of origin-destination pairs is small – there may be insufficient demand to reach a critical mass to justify terminal operation. This does not imply that terminals are abandoned completely, however. It may instead mean that shipments are sent via a for-hire network rather than via a private network (Hall, 1995b). Because the for-hire network has the capability of combining shipments from different companies, they have greater potential for achieving the critical mass needed to establish a terminal.

It is obvious that the scale economies in transportation have not driven the industry toward a single monopoly that manages to carry all goods through one network. Instead, we have many specialized networks, and many competing networks. Large truck-load or container load shipments bypass trucking terminals, and are often sent direct from origin-to-destination (or possibly via rail or ship terminals). Intermediate sized shipments are sent via less-than-truckload networks, which provide simple terminals for transferring pallets of boxes from one truck to another. Packages are sent via parcel networks, which operate more elaborate terminals for sorting and processing shipments. And mail is sent through postal networks, which provide highly automated systems for sorting and processing shipments. In each case, shipments flow in different quantities and rates, which necessitates a different solution. There is sufficient uniqueness to justify creating specialized networks. In the case of large freight carriers, like UPS and Federal Express, there may be incentives to establish two or more networks to serve different types of shipments. For instance, smaller shipments may be separated from large shipments, or overnight shipments may be separated from lower priority shipments. Smilowitz(2001).

Another important issue is determining optimal route and shipping frequencies. The complication here is maintaining synchronization among routes and among stops. For instance, suppose that a region is served by multi-stop tours that operate in a constant interval, but the shipping interval for individual stops is permitted to differ from the tour interval. As with network routing, concave costs and interdependency are important in this question. As indicated in Eq. 15.1, the cost of serving an individual stop depends on the stop density – the number of stops *actually served* per unit area each time the route is covered (Hall, 1985a, 1991b; Daganzo, 1985). This number may be less than the total number of stops available, as all stops are not served on all trips. If the shipping frequency is reduced for some stops, then the per-stop cost for serving the remaining stops will increase, which will in turn make it desirable to reduce their shipping frequency. However, if the stop density is known, then frequencies can be optimized for individual stops. Thus, to optimize the entire tour, a search can be completed for the optimal stop density, from which shipping and tour frequencies can be derived.

To make such a route feasible, it should also be recognized that the shipping interval should be an integral multiple of the tour interval. In addition, efficient terminal operation demands some level of synchronization of incoming and outgoing routes. These issues are discussed in Chapter 5.

All of these factors are indicative of the need for optimization tools that can simultaneously optimize a huge number of decision variables in a non-linear cost environment. There are enormous challenges in achieving this goal, but progress is being made through faster computers, meta-heuristics and more efficient optimization codes.

15.5 Final Comments

Supply chain networks are the mechanism for transforming raw materials into finished products, and delivering these products to consumers. Nodes in the supply chain network serve manufacturing related functions and consolidation related functions, often simultaneously. Whereas in transportation alone, goods are intended to be quickly sorted and processed in terminals, in the broader context of supply chains, goods may remain at a node for a longer period to serve additional purposes. Instead of transferring boxes from incoming vehicles to outgoing vehicles, the node may assemble the items into finished products, or simply store them for future use, while also consolidating them into loads.

Whether an organization relies on for-hire networks for distribution, or its own private network, depends on its needs to create a network of supply chain nodes for production and inventory. Grocers, for instance, not only operate large networks of retail stores but also large networks of distribution warehouses. The warehouses are needed because of food perishability, because many food products come from local sources, and because grocers handle sufficient volumes to justify the investments. Grocer warehouses serve multiple functions, while providing the same consolidation function as terminals in for-hire networks. Therefore, they do not use for-hire networks to get products from warehouses to consumers.

Book publishing, on the other hand, has none of the incentives seen in the grocery industry for establishing large networks. Nevertheless, publishers still have the ability to reach widely dispersed markets through for-hire transportation networks. In fact, it is the richness of the transportation infrastructure that enables our highly specialized, and technology focused, economy to function so well. By consolidating different types of products, produced by different firms at different times, relatively small companies can reach wide markets.

In some instances, a blend of for-hire and private networks is common. A company may depend more on for-hire networks for replacement parts than for original equipment, because demands for the latter are larger and more predictable. Or, in a multi-echelon structure, the distribution strategy may depend on the stocking level of the item. For instance, fast moving items may be stocked at all echelons, and distributed through a private network. Slow-moving items may only be stocked at the highest echelon (taking advantage of the “risk pooling” effect, due to averaging of demands across many locations), and distributed through for-hire networks (Blumenfeld *et al,* 1985). By exploiting the for-hire system, the supply network can be responsive for all types of items, even though inventories are kept small. Going one step further, the private network may be specifically designed to exploit the capabilities of the for-hire network by locating warehouses in the vicinity of major hub terminals (e.g., near Federal Express’ Memphis hub or UPS’ Louisville hub). In doing so, next-day deliveries can be guaranteed for orders that are placed late in the day.

As a final comment, availability and usage of information have had profound effects on supply chain management. Historically, retailers have provided the dual function of physically delivering products to customers and offering information to assist in product selection (sometimes, just by inspecting products). Retailers are now squeezed in both directions. First, small package companies (UPS and Federal Express in particular), with their extensive networks, have enabled distributors to rapidly deliver products direct to consumers from centralized locations. Second, the Internet is providing product information in a manner that sometimes replicates and even enhances information available at the store. In many product lines (media, replacement parts, etc.), the potential exists for eliminating retailers and perhaps local distributors. Enhanced multi-media tools may one day enable a consumer to visualize wearing a new outfit, eliminating the need to try it on at a store. But even this innovation would not mean the elimination of local entities for consolidating shipments. Instead of consolidating through inventory-stocking private networks, consolidation would occur through the terminals operated by for-hire networks. Thus, the technologies of the Internet and communication do not alter the fundamental economics of transportation for physical products.

15.6 References

Arntzen, B.C., G.G. Brown, T.P. Harrison and L.L. Trafton (1995). Global supply chain management at Digital Equipment Corporation. *Interfaces,* **25**:1, 69-93.

Blumenfeld, D.E. R.W.Hall and W. Jordan (1985). Trade-off between freight expediting cost and safety stock inventory costs, *Journal of Business Logistics,* **6**, 79-100.

Bollo, D. and M. Stumm (1998). Possible changes in logistic chain relationships due to Internet developments. *ITOR,* **5**, 427-445.

Brown, A.O., H.L. Lee, and R. Petrakian (2000). Xilinx improves its semiconductor supply chain using product and process postponement. *Interfaces,* **30**:4, 65-80.

Brynjolfsson, E. and M.D. Smith (2000). Frictionless commerce? A Comparison of Internet and conventional retailers. *Management Science,* **46**, 563-585.

Burns, L.D., R.W.Hall, D.E. Blumenfeld and C.F. Daganzo (1985). Distribution strategies that minimize transportation and inventory costs, *Operations Research,* **33**, 469-483.

Campbell, J.F. (1993). Continuous and discrete demand hub location problems, *Transportation Research,* **27B** 473-482.

Chen, F., Z. Drezner, J.K. Ryan and D. Simchi-Levi (2000). Quantifying the bullwhip effect in a simple supply chain: the impact of forecast, lead times, and information. *Management Science,* **46**, 436-443.

Daganzo, C.F. (1984a). The length of tours in zones of different shapes, *Transportation Research,* **18B**, 135-146.

Daganzo, C.F. (1984b). The distance traveled to visit N points with a maximum of C stops per vehicle: an analytic model and an application. *Transportation Science,* **18**, 331 -350.

Daganzo, C.F. (1985). Supplying a single location from heterogeneous sources. *Transportation Research,* **19B**, 409-420.

Daganzo, C.F. (1987a). The break-bulk role of terminals in many-to-many logistic networks. *Operations Research,* **35**, 543-555.

Daganzo, C.F. (1987b). Modeling distribution problems with time windows. *Transportation Science,* **21**, 171-179.

Daganzo, C.F. (1988). A comparison of in-vehicle and out-of-vehicle freight consolidation strategies. *Transportation Research,* **22B**, 173-180.

Daganzo, C.F. (1999). *Logistics Systems Analysis,* 3rd Edition. Springer Verlag: Heidelberg, Germany.

Daganzo, C.F. and R.W. Hall (1993). A routing model for pickups and deliveries: no capacity restrictions on the secondary items, *Transportation Science,* **27**, 315-329.

Dewan, R., M. Freimer, and A. Seidmann (2000). Organizing distribution channels for information goods on the Internet. *Management Science,* **46**, 483-495.

Feitzinger, E. and H. Lee (1997). Mass customization at Hewlett Packard. *Harvard Business Review.*

Garg, A. and C.S. Tang (1997). On postponement strategies for product families with multiple points of differentiation. *IIE Transactions,* **29**, 641-650.

Hall, R.W. (1984). Travel distance through transportation terminals on a rectangular grid, *Journal of the Operational Research Society,* **35**, 1067-1078.

Hall, R.W. (1985a). Determining vehicle dispatch frequency when shipping frequency differs among suppliers, *Transportation Research,* **19B**, 421-431.

Hall, R.W. (1985b). Heuristics for selecting facility locations, *Transportation and Logistics Review,* **21**, 353-373.

Hall, R.W. (1987). Direct versus terminal freight routing on a network with concave costs, *Transportation Research,* **21B**, 287-298.

Hall, R.W. (1987). Comparison of strategies for routing shipments through transportation terminals, *Transportation Research,* **21A**, 421-429.

Hall, R.W. (1987). Consolidation strategy: inventory, vehicles and terminals, *Journal of Business Logistic,* **8**, 57-73.

Hall, R.W. (1989a). Configuration of an overnight package air network, *Transportation Research,* **23A**, 139-149.

Hall, R.W. (1989b). Graphical interpretation of the transportation problem, *Transportation Science,* **23**, 37-45.

Hall, R.W. (1989c). Route choice on freight networks with concave costs and exclusive arcs, *Transportation Research,* **23B**, 177-194.

Hall, R.W. (1989d). Vehicle packing, *Transportation Research,* **23B**, 103-121.

Hall, R.W. (1991a). Characteristics of multi-stop/multi-terminal delivery routes, (R.W.Hall) *Transportation Research,* **25B**, 391-403.

Hall, R.W. (1991b). Comments on one warehouse multiple retailer systems with vehicle routing costs, *Management Science,* **37**, 1496-1497.

Hall, R.W. (1991c). Route selection on freight networks with weight and volume constraints, *Transportation Research,* **25B**, 175-189.

Hall, R.W. (1993). Design of local area freight networks, *Transportation Research,* **27B**, 79-95.

Hall, R.W. (1994). Use of continuous approximations within discrete algorithms for routing vehicles. experimental results and interpretation, *Networks,* **24**, 43-56.

Hall, R.W. (1995a). The architecture of transportation systems, *Transportation Research,* **3C**, 129-142.

Hall, R.W. (1995b). Transportation with common carrier and private fleets: system assignment and shipment frequency optimization, *IIE Transactions,* **27**, 217-225.

Hall, R.W. (1996). Pickup and delivery strategies for overnight carriers, *Transportation Research,* **30**, 173-187.

Hall, R.W. and C.F. Daganzo (1985). Vehicle miles for a freight carrier with two capacity constraints. *Transportation Research Record,* No. 1038, 1985.

Huang, J. (2001). Future space: a new blueprint for business architecture. *Harvard Business Review.*

Kahmoun, M. and R.W. Hall (1996b), Design of express mail services for metropolitan regions, *Journal of Business Logistics* **17**, 265-302.

Klincewicz, J.G. (1998). Hub location in backbone/tributary network design: a review. *Location Science:* **6**, 307-335.

Kuby, M.J. and G.R. Gray (1993). The hub network design problem with stopovers and feeders: the case of Federal Express. *Transportation Research,* **27A**, 1-12.

Langevin, A., P. Mbaraga, and J.F. Campbell (1996). Continuous approximation models in freight distribution: an overview. *Transportation Research,* **30B**, 163-188.

Lee, H.L. and C. Billington (1995). The evolution of supply-chain-management models and practices at Hewlett-Packard. *Interfaces,* **25**:5, 42-63.

Lee, H.L., V. Padmanabhan and S.J. Whang (1997). Information distortion in a suppy chain: the bullwhip effect. *Management Science,* **43**, 546-558.

Lee, H.L. and C.S. Tang (1997). Modelling the costs and benefits of delayed product differentiation. *Management Science,* **43**, 40-53.

Newell, G.F. and C.F. Daganzo (1986). Design of multiple vehicle delivery tours—I. A ring-radial network. *Transportation Research,* **20B**, 345-364.

Newell, G.F. and C.F. Daganzo (1986). Design of multiple vehicle delivery tours—II: Other metrics. *Transportation Research,* **20B**, 365-376.

O'Kelly, M.E. (1986). The location of interacting hub facilities. *Transportation Science,* **20**, 92-106

O'Kelly, M.E., and D.L. Bryan (1998). Hub location with flow economies of scale. *Transportation Research,* **32B**, 605-616.

Morrison, D. and R. Wise (2000). Beyond the exchange: the future of B2B. *Harvard Business Review,* **78:6**, 86-96.

Partyka, J.G. and Hall, R.W. (2000). On the road to service. *OR/MS Today,* August, 26-35.

Simchi-Levi, D., P. Kaminsky and E. Simchi-Levi (2000). *Designing and Managing the Supply Chain: Concepts, Strategies and Cases,* Irwin/McGraw-Hill: New York.

Smilowitz, K.R. (2001). *Design and Operation of Multimode, Multiservice Logistics Systems,* Institute of Transportation Studies Ph.D. Dissertation, University of California at Berkeley, UCB-ITS-DS-2001-04.

Ukovich, W., F. Baita and R. Pesenti (1998). Dynamic routing-and-inventory problems: a review. *Transportation Research,* **32A**, 585-596.

Van Hoek, R.I. (2000). The role of third-party logistics providers in mass customization. *International Journal of Logistics Management,* **11**, 37-46.

Vaughan, R. (1984). Approximate formulas for average distances associated with zones. *Transportation Science*, **18**, 231-244.

Zhong, H.S. (2002). Ph.D. Dissertation, University of Southern California.

This page intentionally left blank

16 REVENUE MANAGEMENT

Garrett J. van Ryzin and Kalyan T. Talluri

16.1 Introduction

Revenue management (RM), refers to the collection of strategies and tactics by which airlines (and other transportation providers) manage demand for their services. This chapter surveys the methods used to perform this demand management function. While revenue management today is applied in a wide range of industries, our focus here is on airline and other transportation RM problems. The chapter is based on excerpted material from our forthcoming book, Talluri and van Ryzin, 2002, *The Theory and Practice of Revenue Management.*

Airline History

The starting point for revenue management was the Airline Deregulation Act of 1978. Passage of the act led to rapid change in the airline industry. New low-cost airlines entered the market and tapped into an entirely new – and vast – market for discretionary travel. The potential of this market was embodied in the rapid rise of PeopleExpress. The airline started in 1981 with a cost-efficient operation and fares 50–70% lower than the major carriers. By 1984, its revenues were approaching $1B, and at the close of 1984 PeopleExpress posted a profit of $60M, its highest profit ever (Cross, 1997).

The result was a significant migration of price-sensitive discretionary travelers from major airlines to the new, low-cost competitors. Yet the major airlines had strengths that these new entrants lacked. They offered more frequent schedules, service to more cities, and an established brand-name and reputation. For many business travelers, schedule convenience and service was (and is still) more important than price, so the threat posed by low-cost airlines was less acute in the business-traveler segment of the market.

Nevertheless, the cumulative losses in revenue from this shift in traffic to low-cost airlines was badly damaging the profits of major carriers. Among major airlines, a strategy to combat this trend and recapture the leisure passenger was needed.

Robert Crandall, American Airline's CEO at the time, is widely credited with the breakthrough in solving this problem (the low fares and expanded service created by this new competition were less of a "problem" in the eyes of consumers and regulators). He recognized that the seats that were going unsold on many of American's flights were already being "produced" at a very low cost. This is because the vast majority of the costs of a flight (capital costs, wages, fuel) are fixed, and the marginal cost to carry an additional passenger is almost zero. As a result, American could in fact afford to "compete on cost" with the upstarts using its surplus seats.

However, two problems had to be solved in order to execute this strategy. First, American needed to identifying the "surplus" seats on each flight. The scheme would not be profitable if a sale of a low price seat displaced one of their high-paying business customers.[1] Second, they had to ensure that American's business customers did not switch and buy the new low price products targeted at discretionary, leisure customers.

American solved these two problems using a combination *of purchase restrictions* and *capacity-controlled fares.*[2] Discounts had to be purchased 30 days in advance of departure and required a seven-day minimum stay, restrictions that prevented most business traveler from utilizing the new low fares. At the same time, American limited the number of such seats sold on each flight (it capacity-controlled the fares).

Initially, the capacity controls American used were based on setting aside a fixed portion of seats on each flights for the new low-fare products. However, as American gained experience with its Super-Saver fares, it realized that a more intelligent approach to capacity control was need to realize their full potential. American's operations research staff therefore embarked on the development of what became known as the "DINAMO" system – the *Dynamic Inventory Allocation and Maintenance Optimizer.* Based on massively detailed statistical forecasts and optimization models, DINAMO determined automatically the number of seats to reserve for discounted products on each individual departure. DINAMO was, in essence, the first large-scale RM system.

DINAMO was implemented in full in January of 1985 along with a new fare program – entitled "Ultimate Super Saver Fares". The DINAMO system allowed American to be much more aggressive in its pricing strategy. They now could post very low fares on a large number of individual flights, confident of their capability to accurately control the discounts on every individual departure. Indeed, this feature – of pricing aggressively and competitively at an aggregate, market level, while controlling capacity at a tactical, individual-departure level – still characterizes the practice of RM in the airline industry today.

PeopleExpress was especially hard hit by this move, as American repeatedly matched or beat their prices in every market they served (Cross, 1997). PeopleExpress's annual profit fell from an all time high in 1984 to a loss of over $160M by 1986. In just two short years they went bankrupt, and in September 1986 the company was sold to Continental Airlines. This experience was repeated throughout the industry, and airlines that did not have RM capabilities quickly scrambled to get them.

As a result of this history, the practice of revenue management in the airline industry today is both pervasive and mature. Indeed, it's no exaggeration to say that a large, modern airline today would quite simply not be able to operate profitably *without* revenue management. For example, American Airlines' estimates that its RM practices generated $1.4 billion in additional incremental revenue over a three year period starting around 1988, Smith et al., 1992, a figure comparable to the airline's total profits over this period.

Components of an RM System

Here, we briefly describe the generic components of a RM system at a high level. This serves to introduce the various components involved and gives an overview of the information flows, controls and design of the overall system.

Revenue management processes generally consist of the following four steps:

(1) **Data Collection** Collect and store relevant historical data (prices, demand, causal factors).
(2) **Estimation and Forecasting** Estimate the parameters of the demand model. Forecast demand and other relevant quantities, like no-show and cancellation rates.
(3) **Optimization** Find the optimal set of controls controls for the sale of inventory (allocations, bid-prices, overbooking limits) to be applied until the next re-optimization.
(4) **Control** Control the sale of inventory using the optimized controls. This is done either through the firm's own transaction-processing systems (internet, call centers) or through shared distribution systems (e.g. computer reservations systems).

The revenue management process typically involves cycling through these steps at repeated intervals. The frequency with which each step is performed is a function of many factors that include the volume of data and how fast it changes, the type of forecasting and optimization methods used and the relative importance the resulting decisions.

Overview of Topics

This Chapter mainly surveys the Steps 3 and 4 above, and even this coverage is somewhat abbreviated. Complete details as well as a treatment of forecasting and implementation issues are provided in Talluri and van Ryzin, 2002. McGill and van Ryzin, 1999 also provides an overview and annotated bibliography of the published academic literature in the field through 1998.

We first address the problems of capacity control. Section 16.2 looks at capacity controls for a single resource (seats on a single flight) which is sold to differentiated demand classes – the so-called "single-leg" problem in airline RM. Section 16.3 looks at the same capacity control decisions, but in a setting in which products require multiple resources – called the "network" setting. In the airline industry, the main motivation for network methods are to control availability of discount classes at the origin-destination (O&D) level rather than the flight (leg) level. Finally, Section 16.4 considers overbooking decisions and their relation to capacity controls.

16.2 Single-resource Capacity Control

The single-resource (single-leg) problem addresses optimally allocating capacity of a single resource to different demand segments. We begin by surveying the nontechnical issues surrounding this problem in the airline industry. The remainder of the section focuses on various models and methods for making capacity control decisions.

Fare Classes

As mentioned, the single-resource problem has its roots in the emergence of capacity-controlled discounted airfares shortly after the airline industry was deregulated in the U.S. in the mid 1970's. Today, most airlines offer discounts based on a relatively stable set of restrictions. These include advance purchase restrictions of 7, 14, 21 and 30 days, the requirement to stay a Saturday night and nonrefundability and/or penalties for changes in the itinerary after purchase. Various combinations of these restrictions are used to define different fare products.

In most airline computer reservation systems (CRSs), the allocation of seats is controlled through the use of fare class codes – or simply *fare classes.* Fare classes are not necessarily related to the "class of service" the passenger receives, though in some cases they are (e.g. "F" designate first class while "Y" designates coach/economy class). More importantly, they indicate different discount levels that are available for each cabin of service. However, fare classes are still an aggregation of actual fares, and thus different fares are frequently sold within the same fare class. This means that the revenue generated by a fare class is often quite variable. Every airline is free to chose its own fare class designations, though commonly "Y" designates a full fare, and letters "Q", "B" and "M" designate various discounted classes.

Types of Controls

Reservation systems may provide different mechanisms for controlling fare class availability. These mechanisms are typically deeply imbedded in the software logic of the reservation system itself and can be quite expensive (if not impossible) to change as a result. Therefore, the control mechanism itself is frequently an important practical constraint faced when implementing revenue management. Here we review the most common types of controls.

Booking Limits Booking limits are controls that limit the amount of capacity that can be sold to any particular demand class at a given point in time. For example, a booking limit of 18 on Class 2 indicates that at most 18 units of capacity can be sold to customers in Class 2. Beyond this limit, the class would be "closed" to additional Class 2 customers. This limit of 18 may be less than the physical capacity, if we want to reserve capacity for higher revenue classes.

Booking limits are either *partitioned* or *nested*: A *partitioned booking limit* logically divides the available capacity into separate blocks – one for each demand class – which can be sold only to the designated class. For example, with 30 units to sell, a partitioned booking limit may set a booking limit of 12 units for Class 1, 10 units for Class 2 and 8 units for Class 3. If the 12 units of Class 1 capacity are used up, Class 1 would be closed regardless of how much capacity is available in the remaining buckets. This could obviously be undesirable if Class 1 has higher revenues than Classes 2 and 3.

With a *nested booking limit,* the capacity available to different classes overlaps in a hierarchical manner – with higher revenue classes having availability to all the capacity reserved for lower-revenue classes (and perhaps then some). Let the nested booking limit for Class *j* be denoted b_j. With the same 30 units of capacity, the nested booking limits could be $b_1 = 30$ (all the available capacity), $b_2 = 18$

and $b_3 = 8$. (See Figure 16.1.) We would accept at most 8 Class 3 customers, at most 18 Class 2 customers, and as many Class 1 customers as possible. Nested booking limits avoid the problem of capacity being simultaneously unavailable for a high-revenue class yet available for lower-revenue classes. Most reservations systems that use booking limit controls, quite sensibly, use nested rather than partitioned booking limits for this reason.

Protection Levels Protection levels are in many ways equivalent to booking limits. A protection level specifies an amount of capacity to reserve (protect) for a particular demand class or set of classes. Again, protection levels can be nested or partitioned. In the nested case, protection levels are defined for sets of demand classes – again ordered in a hierarchical manner according to revenue. Suppose Class 1 is the highest revenue class, Class 2 the second highest, etc. Then the protection level j, denoted y_j, is defined as the amount of capacity to save for Classes $j, j-1,\ldots,1$; that is, for Classes j and higher (in terms of revenue order). Figure 16.1 shows the relationship between protection levels and booking limits. The booking limit for Class j, b_j, is simply the capacity minus the protection level

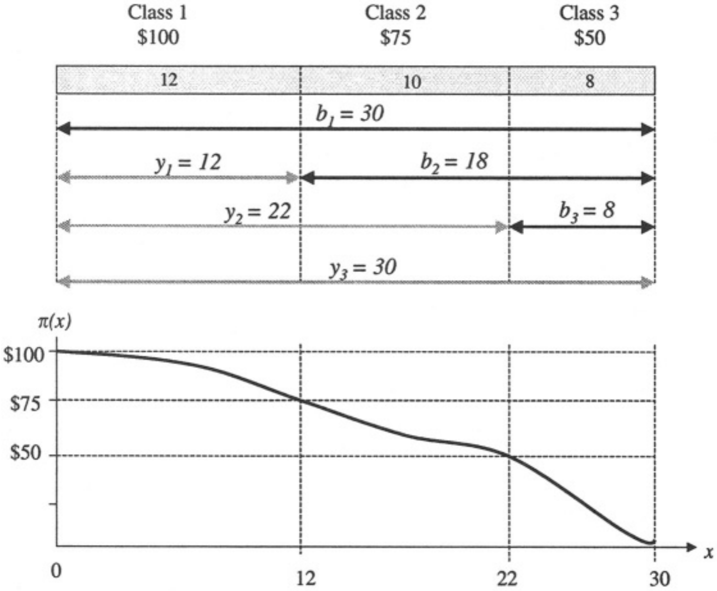


Figure 16.1 The Relationship Between Booking Limits, b_j, Protection Levels, y_j, and Bid Prices $\pi(x)$

for Classes $j - 1$ and higher. That is,

$$b_j = C - y_{j-1}, \quad j = 2, \ldots, n$$

where C is the capacity. For convenience, we define $b_1 = C$ (e.g. the highest revenue class has a booking limit equal to the capacity) and $y_n = C$ (all classes have a protection level equal to capacity).

Bid Prices What distinguishes bid-price controls from both booking limits and protection levels in that they are revenue-based rather than capacity-based controls. Specifically, a bid-price control sets a threshold price (which may depend on variables such as the remaining capacity or time), such that a request is accepted if its revenue exceeds the threshold price and rejected if its revenue is less than the threshold price.

As shown in Figure 16.1, bid prices can usually be used to implement the same nested allocation policy as booking limits and protection levels. In our example in Figure 16.1, the bid price $\pi(x)$ is plotted as a function of the remaining capacity x. When there are 12 or fewer units remaining, the bid price is over $75 but less than $100, so only Class 1 demand is accepted. With 13 to 22 units remaining, the bid price is over $50 but less than $75 so only Classes 1 and 2 are accepted. With more than 22 units of capacity available, the bid price drops below $50 so all three classes are accepted. (Note that the bid price must be adjusted after each sale to reflect the current remaining capacity to achieve this control.)

Static Models

In this section, we examine so-called *static,* single-leg models of the capacity control problem. In these models, demand from each class is assumed to arrive in separate, non-overlapping intervals, with low revenue classes arriving before high fares.[3] The term "static" here is somewhat of a misnomer, because demand does arrive sequentially over time – albeit in stages ordered from low-revenue to high-revenue demand. However, this term is now standard and helps distinguish this class of models from *dynamic* models that allow arbitrary arrival orders.

The assumption that demand for each class arrives in nonoverlapping intervals is an approximation of most real reservation processes; in reality, demand for classes typically overlaps in time. However, this assumption leads to a simple model, is often approximately true, and represents something of a "worst-case" (conservative) assumption. Also, the optimal controls that emerge from the model apply – at least heuristically – in cases where demand comes in arbitrary order. For all these reasons, the static model is quite popular in practice.

The static models also assumes that the demand for different classes is statistically independent. (Brumelle et al., 1990 however do consider a 2-class static model with dependent demand.) Largely, this is done for analytical convenience, however the assumptions is partially justified on practical grounds, since in practice the parameters of the demand distribution are based on detailed forecasts for individual flights. To the extent that there are systematic factors affecting all demand classes (e.g. seasonalities), these are usually picked up in the forecast and become part of the *explained* variation in demand in the model (e.g. the differences in the means and variance on different days). Nevertheless, one has to worry about possible residual dependence in the *unexplained* variation in demand (i.e. the noise terms). Even these correlations, however, are partially captured in practice through frequent cycles of reforecasting and reoptimizing prior to departure mentioned above.

The static models also assumes risk neutrality, which is usually well-justified in practice since a firm implementing RM typically makes many such decisions for a large numbers of flights over time.

Finally, the models here do not consider overbooking. Overbooking is addressed separately in Section 16.4. In practice, the problems are typically decomposed; the overbooking limits are set first and the resulting *virtual capacity* (physical capacity plus the *overbooking pad*) is then used in a capacity control model. This is the perspective taken in this section, though we address combined overbooking and capacity control models in Section 16.4.

The earliest paper on the static models is Littlewood, 1972. Another early applied paper is Bhatia and Parekh, 1973. But there are close connections to earlier work on the stock-rationing problem in the inventory literature by Kaplan, 1969 and Topkis, 1968. Optimal policies for the $n > 2$ case were obtained in close succession (using slightly different methods and assumptions) in by Brumelle and McGill, 1993, Curry, 1990, Robinson, 1995 and Wollmer, 1992.

Littlewood's Two-class Model The simplest single-resource model was introduced by Littlewood, 1972. The model assumes two product classes, with associated revenues $r_1 > r_2$. The capacity (or virtual capacity) is C and it is assumed that there are no cancellations or overbooking. Demand for class k is denoted D_j and its distribution is denoted by $F_j(x)$. Demand for Class 2 arrives first. The central problem is to decide how much of Class 2 demand to accept prior to seeing the realization of Class 1 demand.

The optimal decision can be argued informally using a simple marginal analysis: Suppose we have x units of capacity remaining and receive a request from Class 2. If we accept the request, we collect revenues of r_2. If we do not accept it, we will sell seat x (the marginal seat) at r_1 if and only if demand for Class 1 is x or higher. That is, if $D_1 \geq x$. Thus the expected gain from reserving the *x-th* seat (the *expected marginal value*) is $r_1 P(D_1 \geq x)$. Therefore, it makes sense to accept a Class 2 request as long as the current revenue exceeds this marginal value. Equivalently, if and only if

$$r_2 \geq r_1 P(D_1 \geq x).$$

This decision rule can be implemented using either protection levels, booking limits or bid prices. Note the right hand side above is decreasing in x. Therefore, there will be an optimal protection level, denoted y_1^*, such that we accept Class 2 if the remaining capacity exceeds y_1^* and reject it if the remaining capacity is y_1^* or less. Formally, y_1^* satisfies

$$r_2 \geq r_1 P(D_1 \geq y_1^*) \text{ and } r_2 < r_1 P(D_1 \geq y_1^* + 1).$$

If the distribution of D_1 is continuous, then the optimal protection level y_1^* is given by the simpler expression

$$r_2 = r_1 P(D_1 > y_1^*), \tag{16.1}$$

which is known as *Littlewood's rule.* Setting a protection level of y_1^* for Class 1 according to Littlewood's rule is an optimal policy for the two-class, static model. Equivalently, setting a booking limit of $b_2^* = c - y_1^*$ on Class 2 demand is optimal. Alternatively, we can use a bid-price control with the bid-price set at $\pi(x) = r_1 P(D_1 \geq x)$.

Static, Multiple-class Models We next consider the general case of $n > 2$ product classes. Again, we assume that demand for the n classes arrives in n stages, one for each demand class, with classes arriving in reverse order of their revenue values. Let the classes be indexed so that $r_1 > r_2 > \ldots > r_n$. Hence, Class n (the lowest revenue) demand arrives in the first stage (Stage n), followed by Class $n - 1$ demand in Stage $n - 1$, etc. with the highest price class (Class 1) arriving in the last stage (Stage 1). Since there is a one-to-one correspondence between stages and classes, we will index both stages and classes by j. Demand and capacity are most often assumed to be discrete, but we will also use continuous capacity and demand where it helps simplify the analysis or optimality conditions.

Dynamic Programming Formulation This problem can be formulated as a dynamic program in the Stages j with state variable, x, being the remaining capacity.

At the start of each Stage j, the demand $D_j, D_{j-1}, \ldots, D_1$ has not been realized. Within Stage j, the following sequence of events occurs:

1. The realization of the demand D_j occurs.
2. We observe the realized value D_j and decide how much of this demand, denoted u, to accept – subject to the constraint that we can't accept more demand than has arrived, and we can't accept more than the remaining capacity x. So, $0 \leq u \leq \min\{D_j, x\}$.
3. The revenue $r_j u$ is collected, and we proceed to the start of Period $j - 1$ with a remaining capacity of $x - u$.

Thus, the decision in Stage j is made with perfect information about D_j; however, we know only the distribution of future demand (the "forecast" of demand) for the remaining stages.

Note this sequence of events does not necessarily reflect how demand is realized and decisions are made in most real reservations processes. Indeed, it represents a best-case situation in which we know the demand D_j perfectly when making our Stage-j decision. However, as we show below, the optimal control in fact does not require any information about D_j and can be implemented with simple threshold-type rules.

To proceed, let $V_j(x)$ denote the value function at the start of Stage j. That is, $V_j(x)$ gives the maximum expected revenue that can be obtained entering Stage j with x units of capacity remaining. Once the value D_j is observed, the value of u is chosen to maximize the current Period-j revenue plus the revenue-to-go, or

$$r_j u + V_{j-1}(x - u),$$

subject to the constraint $0 \leq u \leq \min\{D_j, x\}$. The value function entering Stage j, $V_j(x)$, is then the expected value of this optimization with respect to the demand D_j. Hence, the Bellman equation is[4]

$$V_j(x) = E\left[\max_{0 \leq u \leq \min\{D_j, x\}} \{r_j u + V_{j-1}(x - u)\}\right], \tag{16.2}$$

with boundary conditions

$$V_{n+1}(x) = 0 \quad x = 0, 1, \ldots, C.$$

The values u^* that maximize the right-hand-side of (16.2) for each j and x form an optimal control policy for this model.

Optimal Policy: Discrete Demand and Capacity We first consider the case where demand and capacity are discrete. To analyze the form of the optimal control in this case, define

$$\Delta V_j(x) \equiv V_j(x) - V_j(x-1).$$

$\Delta V_j(x)$ is the *expected marginal value of capacity* at Stage j – i.e. the expected incremental value of the x-th unit of capacity. A key result for determining the structure of the optimal control concerns how these marginal values change with capacity x and the Stage j.

Proposition 1 *The marginal values $\Delta V_j(x)$ of the value function $V_j(x)$ defined by (16.2) satisfy:*

(i) $\Delta V_j(x+1) \le \Delta V_j(x) \quad \forall x, j$
(ii) $\Delta V_{j+1}(x) \ge \Delta V_j(x) \quad \forall x, j$

That is, the marginal value is decreasing in the remaining capacity at each Stage j, and at a given capacity level, x, the marginal value increases in the number of stages remaining. These two properties are intuitive and also greatly simplify the control.

To see this, consider the optimization problem at Stage $j+1$. From (16.2) and the definition of $\Delta V_j(x)$, we can write

$$V_{j+1}(x) = V_j(x) + E\left[\max_{0 \le u \le \min\{D_{j+1}, x\}} \left\{\sum_{z=1}^{u} (r_{j+1} - \Delta V_j(x+1-z)\right\}\right],$$

where we take the sum above to be empty if $u = 0$. Since $\Delta V_j(x)$ is decreasing in x by Proposition 1, Part (i), it follows that the terms in the sum, $r_{j+1} - \Delta V_j(x+1-z)$, are decreasing in z. Thus, it is optimal to increase u (to keep adding terms) until the terms become negative or the upper bound D_{j+1} is reached, whichever comes first.

The resulting u^* can be expressed in terms of optimal protection levels y_j^*. Define the optimal protection level for Classes $j, j-1, \ldots, 1$ (Class j and higher in the fare order) by

$$y_j^* \equiv \max\{x : r_{j+1} < \Delta V_j(x)\}, \quad j = 1, \ldots, n-1. \tag{16.3}$$

(The optimal protection level $y_n^* \equiv C$ by convention.) The optimal control is then

$$u^* = \min\{x - y_j^*, D_{j+1}\}. \tag{16.4}$$

The quantity $x - y_j^*$ is the remaining capacity in excess of the protection level, which is the maximum capacity we are willing to sell in Stage j. Thus, the optimal

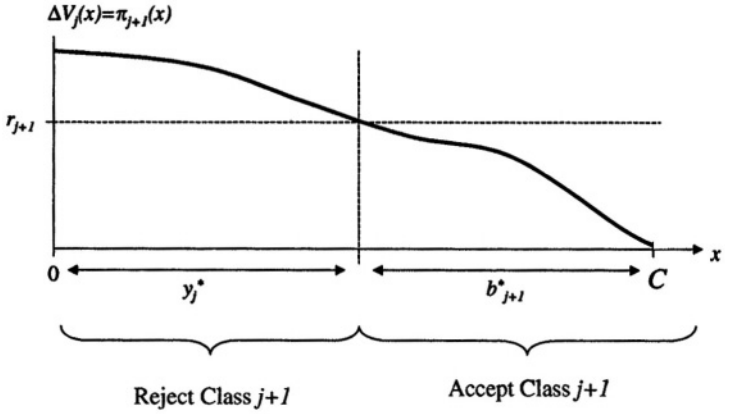


Figure 16.2 Illustration of the Optimal Protection Level y_j^* in the Static Model

policy is to keep selling capacity to Class $j+1$ in Stage $j+1$ until this limit is reached. The situation is shown in Figure 16.2.

In practice, therefore, we can simply post the protection level y_j^* in a reservation system and accept requests first-come, first-serve until the capacity threshold y_j^* is reached or the period ends, whichever comes first. The optimal protection-level control, therefore, involves only simple rules and requires no information about the demand D_{j+1} within Period $j+1$, yet it produces the same optimal decision had we known D_{j+1} exactly at the start of the period.

Part (ii) of Proposition 1 can be used to determine how the protection levels are related. It implies that if $r_1 > r_2 > \cdots > r_n$, the protection levels are ordered,

$$y_1^* \leq y_2^* \leq \ldots \leq y_n^*.$$

This fact is easily seen from Figure 16.2, because if r_{j+1} increases with j and the curve $\Delta V_j(x)$ decreases (shifts down) with j, then the optimal protection level y_j^* will shift to the left (decrease). Together, this ordering produces the nested-protection-level structure.

One can also use booking limits in place of protection levels to achieve the same control. Optimal nested booking limits are defined by

$$b_j^* \equiv C - y_{j-1}^*, \quad j = 2, \ldots, n, \tag{16.5}$$

with $b_1^* \equiv C$.The optimal control in Stage $j+1$ is then to accept

$$u^* = \min\{b_{j+1} - (C-x), D_j\}.$$

Note $C - x$ is the total capacity sold prior to Stage $j + 1$ and b_{j+1} is the booking limit for Class $j + 1$, so $b_{j+1} - (C - x)$ is the remaining capacity available for Class $j + 1$. The optimal booking limit is also shown in Figure 16.2.

Finally, the optimal control can also be implemented through a table of bid prices. Indeed, if we define the Stage-$j + 1$ bid prices by

$$\pi_{j+1}(x) \equiv \Delta V_j(x), \tag{16.6}$$

then the optimal control is

$$u^* = \max\{z : r_{j+1} \geq \pi_{j+1}(x - z)\}.$$

In words, we accept the z-th request in Stage $j + 1$ if the revenue r_{j+1} exceeds the bid price value $\pi_{j+1}(x - z)$ of the z-th unit of capacity that is allocated. In practice, this control can be implemented by storing a table of bid prices and processing requests by sequentially comparing the revenue to the table values corresponding to the remaining capacity.

We summarize these results in the following theorem:

Theorem 1 *For the static model defined by (16.2), the optimal control can be acheived using either: (1) Nested protection levels defined by (16.3), (2) nested booking limits defined by (16.5), or (3) bid price tables defined by (16.6).*

Optimality Conditions for Continuous Demand Next, we consider the case where capacity is continuous and demand at each stage has a continuous distribution. In this case, the dynamic program is still given by (16.2), however D_j, x and u are now continuous quantities. The analysis of the dynamic program is slightly more complex than in the discrete-demand case, but many of the details are quite similar and are omitted. However, one of the chief virtues of the continuous model is that it leads to simplified expressions for the optimal vector of protection levels $y^* = (y_1^*, \ldots, y_n^*)$. We state the basic result without proof. (See Brumelle and McGill, 1993 for a proof.)

First, for an arbitrary vector of protection levels y, define the following $n - 1$ *fill events*

$$\begin{aligned} A_j(y, D) \equiv \{&D_1 > y_1, D_1 + D_2 > y_2, \ldots, \\ &D_1 + \cdots + D_j > y_j\}, \quad j = 1, \ldots, n - 1. \end{aligned} \tag{16.7}$$

Then $A_j(y, D)$ is the event that demand-to-come in Stages $1,2,\ldots,j$ exceeds the corresponding protection levels. A necessary and sufficient condition for y^* to be an optimal vector of protection levels is that it satisfy the $n - 1$ equations

$$P(A_j(y^*, D)) = \frac{r_{j+1}}{r_1}, \quad j = 1, 2, \ldots, n - 1. \tag{16.8}$$

That is, the j-th fill event should occur with probability equal to the ratio of Class-(j +1) revenue to Class 1 revenue. As it should, this reduces to Littlewood's rule in the $n = 2$ case, since $P(A_1(y^*,D)) = P(D_1 > y_1^*) = r_2/r_1$.

The conditions (16.8) can also be used to define an adaptive algorithm for determining optimal protection levels, as shown by van Ryzin and McGill, 2000. The method starts by fixing an arbitrary protection levels y. Then, based on whether the events $A_j(y,D), j = 1,2,\ldots$ occur or not, a simple adjustment to y is made (i.e. if the event $A_j(y,D)$ occurs, the protection level y_j is increased and if $A_j(y,D)$ does not occur, y_j is decreased in a controlled fashion). Repeated application of these simple updates is proven to converge to the optimal protection levels y^* under stationary demand conditions. (See van Ryzin and McGill, 2000.)

Computing Optimal Nested Allocations A common sentiment among some practitioners is that optimal nested allocations are complex to compute. However, computing these values is in fact quite easy and efficient algorithmically. There are two basic approaches: Dynamic programming and Monte Carlo integration.

The first approach is based on using the DP recursion (16.2) directly and requires that demand and capacity are discrete – or in the continuous case that these quantities can be suitably discretized. The inner optimization in (16.2) is simplified by using the optimal protection levels y_{j-1}^* from the previous stage. Thus, substituting (16.4) into (16.2) we obtain

$$V_j(x) = E[r_j \min\{D_j,(x-y_{j-1}^*)^+\} + V_{j-1}(x - \min\{D_j,(x-y_{j-1}^*)^+\})],$$

from which y_j^* is determined using (16.3). This procedure is repeated starting from $j = 1$ and working backwards to $j = n - 1$.

For discrete demand distributions, computing the expectation above for each state x requires evaluating at most $O(C)$ terms since $\min\{D_j,(x-y_{j-1}^*)^+\} \leq C$. Since there are C states (capacity levels), the complexity at each stage is $O(C^2)$. The critical values y_j^* can then be identified from (16.3) in log (C) time by binary search since $\Delta V_j(x)$ is nonincreasing. Indeed, since we know $y_j^* \geq y_{j-1}^*$, the binary search can be further constrained to values in the interval $[y_{j-1}^*, C]$. Therefore computing y_j^* does not add to the complexity at Stage j. Since these steps must be repeated for each of the $n - 1$ stages (stage n need not be computed as mentioned above), the total complexity of the recursion is $O(nC^2)$.

The second approach to computing optimal protection levels is based on using (16.8) together with Monte Carlo integration. It is most naturally suited to the case of continuous demand and capacity, though the discrete case can be computed (at least heuristically) with this method as well. The idea is to simulate a large number M of demand vectors $D = (D_1,\ldots,D_n)$. One then progressively sorts through these values to find thresholds $y_1, y_2, \ldots, y_{n-1}$ that approximately satisfy (16.8). Details can be found in Robinson, 1995 and Talluri and van Ryzin, 2002.

Dynamic Models

Dynamic models relax the assumption that fare classes arrive in a strict order. Instead, they allow for an arbitrary order of arrival, with the possibility of overlapping arrivals of several fare classes. While at first this seems like a strict generalization of the static case, the dynamic models require the assumption of Markovian (e.g. Poisson) arrivals to make them tractable. As a result, they cannot model different levels of variability in demand, which one can do rather easily in the static case. Indeed, this constraint on the distribution of demand is the main drawback of dynamic models in practice. Also, dynamic models require an estimate of the pattern of arrivals over time (called the "booking curve"), which may be difficult to estimate in certain applications. Thus, the choice of dynamic versus static models essentially comes down to a choice of which set of approximations is more acceptable and which data are available in any given application.

Other assumptions of the static model are retained: Demand is assumed to be independent over time and the seller is assumed to be risk neutral. Again, the justification for these assumptions follows the same reasoning as in the static-model case.

The dynamic model was first analyzed by Lee and Hersh, 1993. Lautenbacher and Stidham, 1999 provided a unified analysis of both the static and dynamic single-resource models. See Liang, 1999 for an analysis of a continuous time version of the dynamic model.

Formulation and Structural Properties In the simplest dynamic model, we again have n demand classes with associated revenues $r_1 \geq r_2 \geq \cdots \geq r_n$. Let $r = (r_1, \ldots, r_n)$ denote the vector of revenues. Then there are T total periods and t indexes the periods, with the time index running forward (e.g. $t = 1$ is the first period and $t = T$ is the last period). Since there is no longer a one-to-one correspondence between periods and classes, we use separate indices; t for periods and j for classes.

In contrast to the static model, in each period we assume that at most one arrival occurs.[5] The probability of an arrival of Class j in Period t is denoted λ_t^j, and we let $\lambda(t) = (\lambda_1(t), \ldots, \lambda_n(t))$. The assumption of at most one arrival per period implies that $\lambda(t)$ must satisfy

$$\sum_{j=1}^{n} \lambda_j(t) \leq 1.$$

This condition can be satisfied in practice by discretizing time into sufficiently small intervals so that the arrival probabilities in any period are small. In general, the periods need not be of the same duration. Note also the arrival probabilities may vary with t, so the mix of classes that arrive may vary over time. (In particular, we do not require lower-revenue classes to arrive earlier than higher revenue classes.)

Dynamic Program As before, x denotes the remaining capacity and the value function in Period t is denoted $V_t(x)$. Let $R(t) = r_j$ if a demand for Class j arrives in Period t and $R(t) = 0$ otherwise. Note that $P(R(t) = r_j) = \lambda_t^j$. Let $u = 1$ if we accept the arrival (if any) in and $u = 0$ otherwise. We seek to maximize the sum of current revenue and the revenue-to-go, or

$$R(t)u + V_{t+1}(x - u).$$

The Bellman equation is therefore

$$\begin{aligned} V_t(x) &= E[\max_{u\in\{0,1\}} \{R(t)u + V_{t+1}(x - u)\}] \\ &= V_{t+1}(x) + E[\max_{u\in\{0,1\}} \{(R(t)u - \Delta V_{t+1}(x))u\}]. \end{aligned} \tag{16.9}$$

where $\Delta V_{t+1}(x) = V_{t+1}(x) - V_{t+1}(x-1)$ is the expected marginal value of capacity in Period $t + 1$. The boundary conditions are

$$V_{T+1}(x) = 0, \quad x = 0, 1, \ldots, C.$$

Optimal Policy An immediate consequence of (16.9) is that if a Class j request arrives, so that $R(t) = r_j$, then it is optimal to accept the request if and only if

$$r_j \geq \Delta V_{t+1}(x).$$

Thus, the optimal control can be implemented using a bid-price control where the bid price is equal to the marginal value,

$$\pi_t(x) = \Delta V_t(x). \tag{16.10}$$

Revenues that exceed this threshold are accepted, those that do not are rejected.

Also, as in the static case, an important property of the value function is that it has decreasing marginal value $\Delta V_t(x) = V_t(x) - V_t(x-1)$. Namely,

Proposition 2 *The increments $\Delta V_t(x)$ of the value function $V_t(x)$ defined by (16.9) satisfy:*

(i) $\Delta V_t(x+1) \leq \Delta V_t(x) \quad \forall x, t$
(ii) $\Delta V_{t+1}(x) \leq \Delta V_t(x) \quad \forall x, t$

This proof is omitted, but is a natural and intuitive property since one would expect the value of additional capacity at any point in time to have a decreasing marginal benefit and the marginal value at any given remaining capacity x to decrease with time.

As a consequence, the optimization on the right hand side of (16.9) can also be implemented as a nested allocation policy, albeit one that has time-varying protection levels (or booking limits). Specifically, we can define time-dependent optimal protection levels

$$y_j^*(t) = \max\{x : r_{j+1} < \Delta V_{t+1}(x)\}, \quad j = 1, 2, \ldots, n-1 \qquad (16.11)$$

that have the usual interpretation that $y_j^*(t)$ is the capacity we protect for Classes $j, j-1, \ldots, 1$. Then the protection levels are nested, $y_1^*(t) \leq y_2^*(t) \leq \cdots \leq y_{n-1}^*(t)$, and it is optimal to accept Class j if and only if the capacity remaining exceed $y_{j-1}^*(t)$. The situation is illustrated in Figure 16.3.

Finally, time-dependent nested booking limits can be defined as usual by

$$b_j^*(t) \equiv C - y_{j-1}^*(t), \quad j = 2, \ldots, n, \qquad (16.12)$$

That the booking limits and protection levels depend on time in this case essentially stems from the fact that the demand-to-come varies with time in the dynamic model. The change in demand-to-come as time evolves effects the opportunity cost and therefore the resulting protection levels. As a practical matter, however, the value function is not likely to change much over short periods of time. So fixing the protection levels computed by a dynamic model and then updating them periodically (as is done in most RM systems in practice), is typically near optimal. Still, the time-varying nature of the protection levels is a key difference between static and dynamic models.

We summarize these results in the following theorem:

Theorem 2 *For the dynamic model defined by (16.9), the optimal control can be acheived using either: 1) Time-dependent nested protection levels defined by (16.11), 2) time-dependent nested booking limits defined by (16.12), or 3) bid price tables defined by (16.10).*

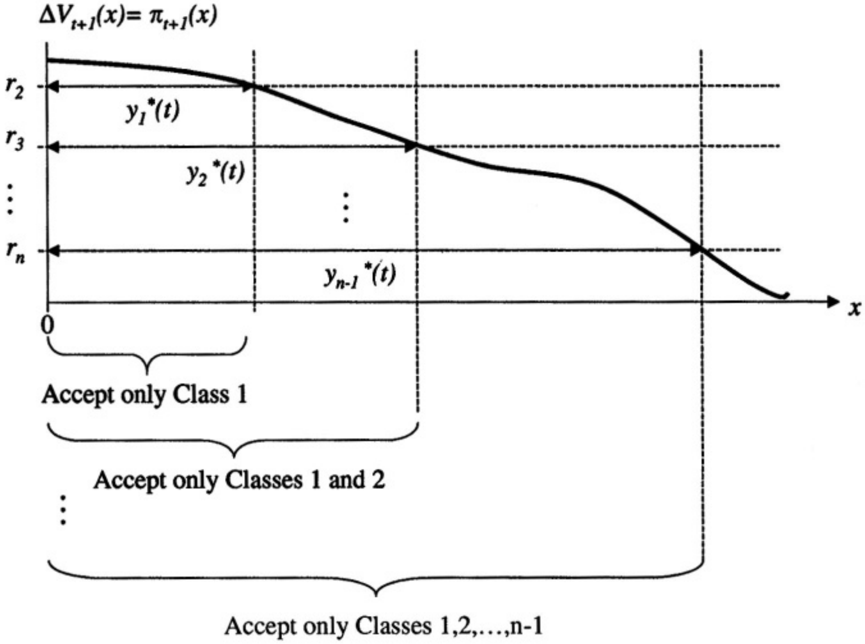

$\Delta V_{t+1}(x) = \pi_{t+1}(x)$
r_2
r_3
r_n
$y_1^*(t)$
$y_2^*(t)$
$y_{n-1}^*(t)$
0
x
Accept only Class 1
Accept only Classes 1 and 2
Accept only Classes 1,2,…,n-1

Figure 16.3 Illustration of the Optimal Protection Level $y_j^*(t)$ in the Dynamic Model

Computation Computationally, the dynamic model is solved by substituting the optimal policy into (16.9). This yields the recursion

$$\begin{aligned} V_t(x) &= V_{t+1}(x) + E\left[(R(t)u - \Delta V_{t+1}(x))^+\right] \\ &= \sum_{j+1}^{n} \lambda_j(t)(r_j u - \Delta V_{t+1}(x))^+, \quad t = 1, \dots, T, \end{aligned}$$

where $z^+ = \max\{0, z\}$ denotes the positive part of z. Starting with the boundary condition $V_{T+1}(x) = 0, \forall x,$ we proceed with the recursion backward in time t. Each stage t requires $O(nC)$ operations, so the overall complexity is $O(nCT)$. Usually, the value of T is $O(C)$ because in most practical problems the total expected demand is the same magnitude as the capacity and the periods are typically chosen so that there are $O(1)$ arrivals per period. So the complexity in terms of n and C is approximately $O(nC^2)$, which is the same as that of the dynamic program for the static model.

Approximations and Heuristics

As we have seen, computing optimal controls for either the static or dynamic single-resource model is not particularly difficult. Despite this fact, exact optimization models are not widely used in practice. Indeed, most single-leg revenue management systems used in the airline industry use one of several approximate methods to compute booking limits and protection levels.

There are two main reasons for this state of affairs. The first is simply a case of practice in the area being one step ahead of the underlying theory. As mentioned, in the airline industry the practice of using capacity controls to manage multiple fare classes quickly gained popularity following deregulation in the mid 1970's. But this predates the theory of optimal controls by more than a decade. The only known optimal controls in the 70's were Littlewood's results for the 2-class problem. As a result, heuristics were developed for the general n-class problem.

A second reason that heuristics remain widely used is that they are typically simpler to code, quicker to run and in many cases produce controls that are near-optimal. Indeed, many practitioners in the airline industry simply believe that even the modest effort of computing optimal controls is not worth the benefit they provide in improved revenue performance.

Regardless of one's view on the use of heuristics, it is important to understand them. They remain widely used in practice, produce good solutions in most cases, and can also help develop some useful intuition.

We next look at the two most popular heuristics: EMSR-a and EMSR-b, both of which are due to Belobaba. (See Belobaba, 1987a, 1987b, 1989, 1992 and Belobaba and Weatherford, 1996.) Both heuristics are based on the n-class, static, single resource model and assumptions outlined above in Section 16.2. They differ only in how they approximate the problem. As before, classes are indexed so that $r_1 > r_2 > \cdots > r_n$, $F_j(x)$ denotes the cdf of Class j demand and low revenue demand arrives before high revenue demand in stages that are indexed by j as well. That is, Class n demand with revenue r_n arrives in the first stage (Stage n), Class $n-1$ with revenue r_{n-1} arrives in Stage n – 2, etc. Moreover, for ease of exposition we assume that capacity and demand are continuous, and the distribution functions $F_j(x), j = 1, \ldots, n$ are continuous as well, though these assumptions are easily relaxed.

EMSR-a EMSR-a (*expected marginal seat revenue – version a*) is the most widely publicized heuristic for single-resource problem. Despite this fact, it is less popular in practice than its close cousin, EMSR-b, which is curiously quite obscure in the published literature. Generally, EMSR-b provides better revenue performance, though EMSR-a is important to know just the same.

EMSR-a is based on the idea of adding the protection levels produced by applying Littlewood's rule to successive pairs of fare classes. Consider Stage $j+1$ in which demand of Class $j+1$ will arrive with revenue r_{j+1}. We are interested in computing how much capacity to reserve for the remaining demand classes, i.e. classes $j, j-1,...,1$; that is, the protection level, y_j, for classes j and higher. Suppose there was only one class remaining, call it Class k. Then we would use Littlewood's rule (for continuous demand) and reserve capacity y_j^k for Class k, where

$$P(D_k > y_j^k) = \frac{r_{j+1}}{r_k}. \tag{16.13}$$

Repeating for each future demand class $k = j, j-1, \ldots, 1$ we could compute how much capacity we would reserve for each such class *in isolation.* The idea of EMSR-a, then, is simply to add up these individual protection levels to approximate the total protection level y_j. That is, set the protection level y_j, using

$$y_j = \sum_{k=1}^{j} y_j^k, \tag{16.14}$$

where y_j^k is given by (16.13). As always, given capacity C one can immediately define equivalent booking limits $b_j = C - y_{j-1}$. One then repeats this same calculation for each Stage j.

EMSR-a is certainly simple and has an intuitive appeal. However, it is not hard to show that it can be excessively conservative and produce protection levels that are larger than optimal in certain cases. This is because adding the individual protection levels y_j^k ignores the statistical averaging effect ("pooling effect") produced by aggregating demand across classes.

EMSR-b EMSR-b (*expected marginal seat revenue – version b*) is an alternative single-resource heuristic that avoids the lack-of-pooling defect in EMSR-a mentioned above. EMSR-b is again based on an approximation that reduces the problem at each stage to two classes, but in contrast to EMSR-a, the approximation is based on aggregating *demand* rather than aggregating *protection levels.* Specifically, the demand from future classes is aggregated and treated as one class with a revenue equal to the weighted average revenue. The aggregation step is simplified if we assume demand is normally distributed, so the distribution of the aggregated demand can be easily determined from the means and variances of demand for each class.

Consider Stage $j+1$ in which we want to determine protection level y_j. Define the aggregated demand to come by

$$S_j = \sum_{k=1}^{j} D_k,$$

and let the weighted average revenue from classes $1,\dots j$, denoted $\bar{R}_j$, be defined by

$$\bar{R}_j = \frac{\sum_{k=1}^{j} r_k E[D_k]}{\sum_{k=1}^{j} E[D_k]}, \tag{16.15}$$

where $E[D_j]$ denotes the mean of Class j demand. Then the EMSR-b protection level for Class j and higher, y_j, is chosen so that

$$P(S_j > y_j) = \frac{r_{j+1}}{\bar{R}_j}. \tag{16.16}$$

For example, it is common when using EMSR-b to assume demand for each Class j is independent and normally distributed with mean μ_j and variance σ_j^2, in which case

$$y_j = \mu + z_\alpha \sigma$$

where $\mu = \sum_{k=1}^{j} \mu_k$ is the mean and $\sigma^2 = \sum_{k=1}^{j} \sigma_k^2$ is the variance of the aggregated demand to come at Stage j+1 and $z_\alpha = \Phi^{-1}(1-r_{j+1}/\bar{R}_j)$ (recall $\Phi^{-1}(x)$ is the inverse of the standard normal cdf). Again, one repeats this calculation for each j.

EMSR-b clearly captures the pooling – or statistical averaging – effect that is lacking in EMSR-a. This is the key advantage of the approximation.

Numerical Example A simple numerical example gives some sense of the protection levels and revenues produced by these two approximations. The example we consider is based on a slightly modified instance of the data reported in Wollmer, 1992. There are four fare classes and demand is assumed to be normally distributed. The data for the example are shown in Table 16.1 along with the protection levels produced by EMSR-a, EMSR-b and the optimal policy. Note that there is a considerable discrepancy between the protection levels computed under the heuristics, both compared to each other and to the optimal protection levels.

The revene performance of the methods from a simulation study are shown below in Table 16.2. Capacity was varied from 80 to 150 to create demand factors (ratio of total mean demand to capacity) in the range 1.7 to 0.9. The percentage

Table 16.1 Static Single Resource Model: EMSR Example

		Demand statistics		*Protection levels (y_j)*		
j	r_j	μ_j	σ_j	*OPT*	*EMSR-a*	*EMSR-b*
1	1050	17.3	5.8	16.7	16.7	16.7
2	567	45.1	15.0	42.5	38.7	50.9
3	534	39.6	13.2	72.3	55.6	83.1
4	520	34.0	11.3			

Table 16.2 Revenue Performance for EMSR Example

		OPT	*EMSR-a*		*EMSR-b*	
C	*DF*	*Rev.*	*Rev.*	*%Sub. Opt.*	*Rev.*	*%Sub. Opt.*
80	1.70	49,666	49,515	0.30%	49,463	0.41%
90	1.51	54,846	54,721	0.23%	54,560	0.52%
100	1.36	60,063	59,985	0.13%	59,786	0.46%
110	1.24	65,112	65,078	0.05%	64,881	0.35%
120	1.13	69,916	69,904	0.02%	69,764	0.22%
130	1.05	73,975	73,973	0.00%	73,898	0.10%
140	0.97	77,177	77,174	0.00%	77,143	0.04%
150	0.91	79,544	79,544	0.00%	79,534	0.01%

suboptimality is also reported (one minus the ratio of policy revenues to optimal revenues). Note for this example that EMSR-a is slightly better than EMSR-b, though both perform quite well; the suboptimality gap of EMSR-b reaches a high of 0.52%, while the maximum suboptimality of EMSR-a is only 0.30%. However, this behavior is somewhat problem specific. For example, using this same example but with a slight change in the revenue values accross the different classes results in EMSR-b doing better than EMSR-a.

Group Arrivals

Group arrivals can pose additional complications. A group request is a single request – but one for multiple units of capacity (e.g. a family of four traveling together). We will briefly describe this case, but omit any detailed formulations because the basic ideas follow readily from what we have seen thus far (and the more complicated ideas are beyond the scope of this text).

If groups can be *partially accepted* – that is, given a request for $m > 1$ units we can sell any quantity u in the range $0 \leq u \leq m$ (and more importantly the customer is willing to buy any amount in this range, which is not unusual among tour operators) – then there is little impact on the single-leg models discussed above. Indeed, the static model (16.2) can be thought of as a group model where in each period one "large group" of size D_j arrives, because we can sell u units, where $0 \leq u \leq \min\{D_j, x\}$ and x is the total available capacity. Thus, with groups that can be partially accepted, we only need to keep track of the aggregate demand for each class and the formulations are essentially the same as in the traditional case.

The real complication in group arrivals occurs when groups must be *completely accepted* – that is, given a request for $m > 1$ units we can sell only all m units or none at all. This seemingly modest change has a profound impact on the structure of optimal allocation policies. First, we must specify the distribution of group sizes and we must model how much demand we have from groups of various sizes. But this in itself does not pose too much of a theoretical difficulty. The bigger problem is that the value function may not be concave (i.e. the marginal value of capacity may in fact increase), so using booking limits, protection levels or bid prices may not be optimal.

Again, we will omit the details here, but essentially, the requirement to completely accept or reject groups creates a combinatorial (bin packing) phenomenon in allocating capacity. The resulting nonmonotonicities in the value function mean that optimal policies are considerably more complex than in the case where groups can be partially accepted. The problem is addressed in Brumelle and Walczak, 1997, Kleywegt and Papastavrou, 1998 and Van Slyke and Young, 2000.

Buy-up, Diversion and Discrete Choice Models

A key assumption in the above single-resource models is that demand for each of the product classes is completely independent of the capacity controls being applied by the seller. That is, it is assumed that the likelihood of receiving a request for any given product does not depend on which other products are available at the time of the request. Needless-to-say, this is a somewhat unrealistic assumption. For example, in the airline case the likelihood of selling a full fare ticket may very well depend on whether a discount fare is available at the same time and the likelihood that a customer buys at all may depend on the lowest available fare. When customers buy a higher fare when a discount is closed it is called *buy-up*; when they chose another flight when a discount is closed it is called *diversion.*

Clearly, such consumer behavior could have important revenue management consequences and should ideally be considered when making control decisions.

We next look at some heuristic and exact methods for incorporating consumer choice behavior in single-resource problems.

Buy-up Factors One approach to modeling consumer choice behavior that works with the two-class model is to include buy-up probabilities – also called "buy-up factors" – in the formulation. (See Belobaba and Weatherford, 1996.)

The approach works as follows: Consider the simple two-class static model and recall that Littlewood's rule (slightly restated) is to accept demand from Class 2 if and only if

$$r_2 \geq r_1 P(D_1 \geq x), \tag{16.17}$$

where x is the remaining capacity. That is, if the revenue from accepting Class 2 exceeds the marginal value of the unit of capacity required to satisfy the request. Now suppose that there is a probability q that a customer for Class 2 will buy a Class 1 fare if Class 2 is closed. The net benefit of accepting the request is still the same, but now rather than losing the request when we reject it, there is some chance the customer will sell-up to Class 1. If so, we earn a net benefit of $r_1 - r_1 P(D_1 \geq x)$ (the Class 1 revenue minus the expected marginal cost). Thus, it is optimal to accept Class 2 now if $r_2 - r_1 P(D_1 \geq x) \geq q r_1 (1 - P(D_1 \geq x))$, or equivalently if

$$r_2 \geq (1 - q) r_1 P(D_1 \geq x) + q r_1.$$

Note that the RHS side above is strictly larger than the RHS of Littlewood's rule (16.17), which means that the above rule is more likely to reject Class 2 demand. This is intuitive, because with the possibility of customers up-grading to Class 1, we should be more eager to close Class 2.

The difficulty with this approach is that it does not extend very neatly to more than two fare classes – at least not in an exact way – because the probability that a customer buys Class i given that Class j is closed depends not only on i and j, but on which other classes are also available. That is, with more than two fare classes the consumer faces a multinomial rather than a binary choice.

However, one can at least heuristically extend the buy-up factor idea to EMSR-a or EMSR-b since these heuristics approximate the multi-class problem using the two-class model. For example, EMSR-b can be extended to allow for a buy-up factor by modifying the equation for determining the protection level y_j, (16.16), as follows

$$r_{j+1} = (1 - q_{j+1}) \bar{R}_j P(S_j > y_j) + q_{j+1} \hat{r}_{j+1}, \tag{16.18}$$

where q_{j+1} is the probability that a customer of Class $j + 1$ buys-up to one of the Classes j, $j - 1,...,1$, $\bar{R}_j$ is the weighted average revenue from these classes as defined by (16.15) and $\hat{r}_{j+1} > r_{j+1}$ is an estimate of the average revenue received given that a Class $j + 1$ customer buys-up to one of the classes j, j – 1...,1 (e.g. $\hat{r}_{j+1} = r_j$ if customers are assumed to buy-up to the next-lowest price class). Again, the net effect of this change it to increase the protection level y_j and thus close down Class $j + 1$ earlier than one would do under the traditional EMSR-b rule.[6]

While this modification to EMSR-b provides a simple heuristic way to incorporate choice behavior, it is nevertheless a somewhat ad hoc adjustment to an already heuristic approach to the problem. However, it has proved useful as a rough-cut approach for incorporating choice behavior in practice.

Discrete Choice Models We next look at a single-resource problem in which consumer choice behavior is modeled exactly using a general discrete choice model. The approach is from Talluri and van Ryzin, 2001. (See also Algers and Besser, 2001 and Andersson, 1989 for an application of similar discrete choice models at SAS.) In contrast to the heuristic approach of buy-up factors, this model provides a more theoretically sound approach to incorporating choice behavior. It also provides insights into how choice behavior affects the optimal availability controls.

Model Definition As in the traditional dynamic model, time is discrete and indexed by t, with the indices running forwards in time. In each period there is at most one arrival. The probability of arrival is denoted by λ, which we assume for ease of exposition is the same for all time periods t. There are n fare classes and we let $N = \{1, \ldots, n\}$ denotes the entire set of fare classes. Each fare class $j \in N$ has an associated revenue r_j, and without loss of generality we index fare classes so that $r_1 \geq r_2 \geq \cdots r_n \geq 0$.

However, in this model the choice of an arrival is not fixed but is an outcome of the fare the customer selects. This is modeled as follows: In each period t, the seller chooses a subset $S_t \subseteq N$ of fare classes to offer. When the fares S_t are offered, the probability that a customer chooses Class $j \in S_t$ is denoted $P_j(S_t)$. We let $j = 0$ denote the no-purchase choice; that is, the event that the customer does not purchase any of the fares offered in S_t. $P_0(S_t)$ denotes the no-purchase probability. The probability that a sale of Class j is made in Period t is therefore $\lambda P_j(S_t)$, and the probability that no sale is made is $\lambda P_0(S_t) + (1 - \lambda)$. Note this last expression reflects the fact that having no sales in a period could be due either to no arrival at all or an arrival that does not purchase.

The only condition imposed on the choice probabilities $P_j(S)$ it that they define a proper probability function. That is, for every set $S \subseteq N$, the probabilities satisfy

$$P_j(S) \geq 0 \quad \forall j \in S$$
$$\sum_{j \in S} P_j(S) + P_0(S) = 1.$$

The following running example will be used to illustrate the model and analysis:

Example 1 *An airline offers three fare products, Y,M and Q with fares of $800, $500 and $450 respectively. The probability of purchasing each fare given the set S of available fares is given by Table 16.3.*

Formulation and Optimal Policy A dynamic program for this model is formulated as follows: Let C denote the total capacity, T denote the number of time periods, t denote the current period and x denotes the number of remaining inventory units. Define the value function $V_t(x)$ as the maximum expected revenue obtainable from periods $t, t+1, \ldots, T$ given that there are x inventory units remaining at time t. Then the Bellman equation for $V_t(x)$ is

$$\begin{aligned} V_t(x) = \max_{S \subseteq N} \Big\{ \sum_{j \in S} \lambda P_j(S)(r_j + V_{t+1}(x-1)) \\ + (\lambda P_0(S) + 1 - \lambda) V_{t+1}(x) \Big\} \\ = \max_{S \subseteq N} \Big\{ \sum_{j \in S} \lambda P_j(S)(r_j - \Delta V_{t+1}(x)) \Big\} + V_{t+1}(x), \qquad (16.19) \end{aligned}$$

Table 16.3 Choice Probabilities $P_j(S)$, Probability of Purchase, $Q(S)$ and Expected Revenue, $R(S)$ for Example 1

S	$P_Y(S)$	$P_M(S)$	$P_Q(S)$	$Q(S)$	$R(S)$	*Dominated?*
$\{Y\}$	0.3	0	0	0.3	240	No
$\{M\}$	0	0.4	0	0.4	200	Yes
$\{Q\}$	0	0	0.5	0.5	225	Yes
$\{Y, M\}$	0.1	0.4	0	0.5	280	Yes
$\{Y, Q\}$	0.3	0	0.5	0.8	465	No
$\{M, Q\}$	0	0.4	0.5	0.9	425	Yes
$\{Y, M, Q\}$	0.1	0.4	0.5	1	505	No

where $\Delta V_{t+1}(x) = V_{t+1}(x) - V_{t+1}(x-1)$ denotes the marginal cost of capacity in the next period, and have used the fact that for all S, $\sum_{j\in S} P_j(S) + P_0(S) = 1$. The boundary conditions are

$$V_{T+1}(x) = 0, \quad x = 0, 1, \ldots, C. \tag{16.20}$$

The problem (16.19) at first seems to have very little structure. However, a sequence of simplifications provides a good characterization of the optimal policy.

The first simplification is to write (16.19) in more compact form as

$$V_t(x) = \max_{S\subseteq N} \{\lambda(R(S) - Q(S)\Delta V_{t+1}(x))\} + V_{t+1}(x), \tag{16.21}$$

where

$$Q(S) = \sum_{j\in S} P_j(S) = 1 - P_0(S)$$

denotes the total probability of purchase and

$$R(S) = \sum_{j\in S} P_j(S) r_j$$

denotes the total expected revenue from offering set S. Table 16.3 gives the values $Q(S)$ and $R(S)$ for our Example 1.

The second simplification is to note that not all $2^n - 1$ subsets need to be considered when maximizing the right hand side of (16.21). Indeed, the search can be reduced to only those sets that are not *dominated.* We omit a formal definition, but roughly a set T is said to be *dominated* if a randomization of other sets S produces an expected revenue that is strictly greater than $R(T)$ with no increase in the probability of purchase $Q(T)$ (or at least the same revenue $R(T)$ with a probability of purchase strictly lower than $Q(T)$). That dominated sets can be eliminated from consideration is quite intuitive from (16.21); a dominated set provides less revenue $R(T)$ than other sets and incurs a higher probability of consuming capacity $Q(T)$ (and hence incurs a higher opportunity cost $Q(S)\Delta V_{t+1}(x)$ in (16.21)).

For Example 1, Table 16.3 shows which sets are nondominated, namely the sets $\{Y\}$, $\{Y,Q\}$ and $\{Y,Q,M\}$. That these sets are nondominated follows from inspection of Figure 16.4, which shows a scatter plot of the value $Q(S)$ and $R(S)$ for all subsets S. Note from this figure a nondominated set is a point that is on the "efficient frontier" of the set of points $\{Q(S), R(S)\}, S \subseteq N$. Here, "efficiency" is with respect to the trade-off between expected revenue, $R(S)$, and probability of sale, $Q(S)$.

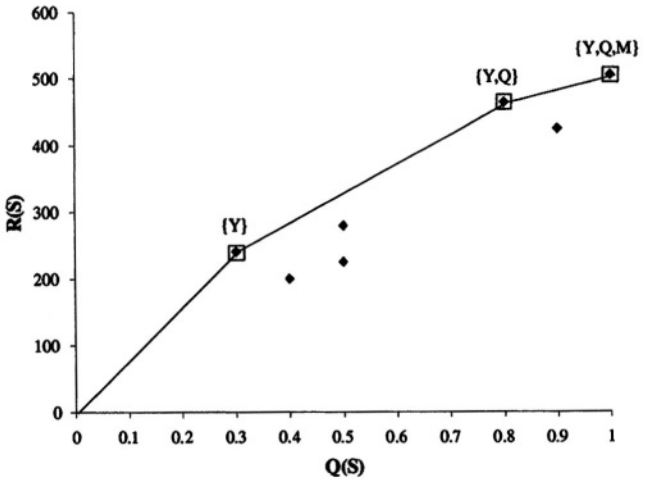


Figure 16.4 Scatter plot of $Q(S)$ and $R(S)$ for Example 1 (nondominated points are enclosed in squares and labeled)

The third simplification is to note that the nondominated sets can be easily ordered. Indeed, let m denote the number of nondominated sets. These sets can be indexed $S_1, \ldots, S_m$ such that both the revenues and probabilities of purchase are monotone increasing in the index: That is, if the collection of m nondominated sets is indexed such that $Q(S_1) \leq Q(S_2) \leq \cdots \leq Q(S_m)$, then $R(S_1) \leq R(S_2) \leq \cdots \leq R(S_m)$ as well. The proof of this fact is again omitted, but it is easy to see intuitively from Figure 16.4. For our Example 1, note from Table 16.3 that the three nondominated sets can be ordered $S_1 = \{Y\}$, $S_2 = \{Y, Q\}$ and $S_3 = \{Y, Q, M\}$ with associated probabilities of purchase $Q_1 = 0.3$, $Q_2 = 0.8$ and $Q_3 = 1$ and revenue $R_1 = \$240$, $R_2 = \$465$ and $R_3 = \$505$.

Henceforth, we assume the nondominated sets are denoted $S_1, \ldots, S_m$ and are indexed in increasing revenue and probability order. Also, to keep notion simpler we let $R_k = R(S_k)$ and $Q_k = Q(S_k)$, and note R_k and Q_k are both increasing in k. So the Bellman equation can be further simplified to

$$V_t(x) = \max_{k=1,\ldots,m} \{\lambda(R_k - Q_k \Delta V_{t+1}(x))\} + V_{t+1}(x), \tag{16.22}$$

When expressed in terms of the sequence $S_1, \ldots, S_m$ of nondominated sets, the optimal policy has a quite simple form as shown by the following theorem (the proof is omitted):

Theorem 3 *An optimal policy for* (*16.19*) *is to select a set* k^**from among the m nondominated, ordered sets* $\{S_k : k = 1, \ldots, m\}$ *that maximizes* (*16.22*). *Moreover, for a fixed t, the largest optimal index*k^* *is increasing in the remaining capacity x, and for any fixed x,* k^* *is increasing in time t.*

For example, applying Theorem 3 to Example 1, we see that the nondominated sets $S_1 = \{Y\}$, $S_2 = \{Y, Q\}$ and $S_3 = \{Y, Q, M\}$ would be used as follows: with very large amounts of capacity remaining, S_3 is optimal – i.e. all three fare classes are opened. As capacity is consumed, at some point we switch to only offering S_2 – i.e. Class *M* is closed and only *Y* and *Q* are offered. As capacity is reduced further, at some point we close Class *Q* and only offer Class *Y* (i.e. set S_1 is used).

Optimality of Nested Allocation Policies The optimization results above have significant implications for the optimality of nested allocation policies. The notion of dominance and Theorem 3 can be used to provide a quite complete characterization of cases in which nested allocation polices are optimal. They also can be used to provide conditions under which the optimal nesting is by fare order.

We begin with a precise definition of a nested allocation policy in this context:

Definition 1 *A control policy is called a nested policy if there is an increasing family of subsets* $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_m$ *and an index,* $k_t(x)$, *that is increasing in x, such that set* $S_{k_t(x)}$ *is chosen at time t when the remaining capacity is x.*

Though this is a somewhat abstract definition of a nested policy, it is in fact the natural generalization of nested allocations from the traditional single-resource models and implies an ordering of the classes based on when they first appear in the increasing sequence of sets S_k. That is, Class *i* is considered “higher” than Class *j* in the nesting order if Class *i* appears earlier in the sequence. Returning to Example 1, we see that the the nondominated sets are indeed nested according to this definition because $S_1 = \{Y\}$, $S_2 = \{Y, Q\}$ and $S_3 = \{Y, Q, M\}$ are increasing. Class *Y* would be considered the highest in the nested order, followed by Class *Q* and then Class *M*.

If the optimal policy is nested in this sense, then we can define optimal protection levels $y_k^*(t), k = 1, \ldots, m$, such that classes lower in the nesting order than those in S_k are closed if the remaining capacity is less than $y_k^*(t)$, just as in the traditional single-resource case. The optimal protection levels are defined by

$$y_k^*(t) = \max\{x : R_k - Q_k \Delta V_{t+1}(x) > R_{k+1} - Q_{k+1} \Delta V_{t+1}(x)\},$$
$$k = 1, 2, \ldots, m - 1.$$

Nested booking limits can also be defined in the usual way based on these protection levels, $b_k(t) = C - y_{k-1}(t)$.

As shown in Talluri and van Ryzin, 2001, the notion of dominated sets can also be used to characterize when nesting by fare order is optimal. However, the details are beyond the scope of this chapter. The conditions confirm, for example, that for the independent demand model of the traditional dynamic single-resource problem, nesting by fare order is optimal. They also show that nesting by fare order is optimal when the choice probabilities are determined by a multinomial logit (MNL) discrete choice model.

16.3 Network Problems

We next examine the problem of capacity control on a network of resources; for example, managing the capacities of a set of flights in a hub-and-spoke airline network with connecting and local traffic. The dependence among the resources in such cases is created by customer demand; customers may require several resources simultaneously (e.g. two connecting flights) to satisfy their needs. Thus, limiting availability of one resource may cause a loss of demand for complementary resources. This in turn creates dependencies among the resources that necessitates making control decisions at the network level. In the airline industry, network revenue management is also called “the passenger mix problem” or “O&D (origin-destination) control”.

Simulation studies of airline hub-and-spoke networks have shown that there can be significant revenue benefits from using network methods over single-resource methods. (See Belobaba and Lee, 2000, Belobaba, 2001 and Williamson, 1992.) In terms of industrial practice, the potential improvements have been sufficient to justify significant investments in network revenue management systems within the airline industry, hotel industry and elsewhere. However, network revenue management poses significant implementation and methodological challenges. On the implementation side, network revenue management vastly increases the complexity and volume of data that one must collect, store and manage. On the forecasting side, it requires a massive increase in the scale of the forecasting system, which now must produce forecasts for each individual itinerary and price-class combination – which we will call a *product* – at each point in the booking process. Optimization is more complex as well. In the case of a single-resource problem there are many exact optimization methods, but in the network case exact optimization is, for all practical purposes, impossible. Therefore optimization methods necessarily require approximations of various types. Achieving a good balance between the quality of the approximation and the efficiency of the resulting algorithms becomes the primary challenge.

Types of Controls

As with single-resource problems, in network allocation problems there are a variety of ways one can control the availability of capacity. We next look at the major categories of network controls. Most are network versions of the controls used for single-resource problems. But others, virtual nesting in particular, are somewhat unique to the network setting.

Partitioned Booking Limits Partitioned booking limits in networks are an extension ofpartitioned booking limits in the single-resource setting. In the network case, partitioned booking limits allocate a fixed amount of capacity on each resource for every product (i.e. itinerary-fare-class in the airline case) that is offered. These allocated amounts ofcapacity are nonoverlapping – or partitioned; demand for one product has access only to its allocated capacity and no other product may use this capacity.

Partitioned booking limits have a number of defects. First, the number of product combinations in even a modest-size network can be very large. Thus, allocating fixed amounts of capacity to each combination results in dividing the capacity of each resource into a very large number of small allocations. But even if they are not practical, partitioned allocation have an important role to play both theoretically and computationally. Theoretically, they can be used to provide bounds that are useful in characterizing optimal network revenues and optimal network controls. Computationally, they are used in many approximate models that provide inputs to other types of controls. An asymptotic analysis ofpartitioned controls is given by Cooper, 2000.

Virtual Nesting Controls Nested booking limits, of the type we saw in Section 16.2 for the single-resource case, are difficult to translate directly into a network setting. However, the ability of nested controls to dynamically share the capacity of a resource – and thereby recover the pooling economies lost in partitioned controls – is an attractive feature. Thus, it is desirable to have a control that combines these features.

Virtual nesting control – a hybrid of network and single-resource controls – provides one solution. This control scheme was developed by American Airlines beginning in 1983 as a strategy for incorporating some degree of network control within the single-leg nested allocation structure of American's (then leg-based) reservation systems. It was first implemented in SABRE[7] in 1986. (See Smith and Penn, 1988.)

Virtual nesting uses single-resource nested allocation controls at each resource in the network. However, the classes used in these nested allocations are not the fare classes themselves. Rather, they are based on a set of *virtual classes* that do not necessarily correspond to the fares of each product. Products are assigned to a virtual class through process know as *indexing*. This indexing is typically updated over time as network demand patterns change, though typically indexing is not a "real time" process. Nested booking limits (or protection levels) for each resource are then computed using these virtual classes and their associated displacement adjusted revenue estimates.

Virtual nesting has proven to be quite effective and popular in practice, especially in the airline industry. It preserves the booking-class control logic of most airline CRSs yet incorporates network displacement cost information. It therefore provides a nice compromise between leg-level and full network O&D control.

Descriptions of virtual nesting can be found in Belobaba, 1987a; Smith and Penn, 1988; Williamson, 1988; Williamson, 1992 and Vinod, 1989; Vinod, 1995.

Bid-price Controls While nested allocations are difficult to extend directly to networks, network bid-price controls are a simple extension of their single-resource versions described in Section 16.2. In a network setting, a bid-price control sets a threshold price – or *bid-price* - for each resource in the network. Roughly, this bid-price is an estimate of the marginal cost to the network of consuming the next incremental unit of the resource's capacity. When a request for a product comes in, the revenue of the request is compared to the *sum* of the bid prices of the resources required by the product. If the revenue exceeds the sum of the bid prices, the request is accepted; if not, it is rejected.

Bid-price controls were first introduced by Smith and Penn, 1988 and Simpson, 1989. Williamson, 1988 and Williamson, 1992 provides variations of bid-price controls and provides detailed simulation comparisons with other control methods. See also the industry publications of Phillips, 1994 and Curry, 1992. Theoretical properties and an asymptotic analysis of bid-price controls is provided by Talluri and van Ryzin, 1999a.

A General Network Model

We begin with a basic model of the network allocation problem. The network has m resources which can be used to provide n products. We let $a_{ij} = 1$ if Resource i is used by Product j and $a_{ij} = 0$ otherwise. Define the *incidence matrix* $A = [a_{ij}]$. Thus, the j-th column of A, denoted A_j, is the *incidence vector* for products j; the i-th row, denoted A^i, has entries of one for any column j corresponding to a Product j that uses Resource i. We also use the notation $i \in A_j$ to indicate that Resource i is used by Product j and $j \in A^i$ to mean that Product j uses Resource i.

The state of the network is described by a vector $x = (x_1, \ldots, x_m)$ of resource capacities. If Product j is sold, the state of the network changes to $x - A_j$. To simplify our analysis at this stage, we will ignore cancellations and no-shows and assume the capacities represent virtual capacity inflated to account for cancellations and no-shows on each resource.

Time is discrete, there are T periods and the index t represents the current time (time indices run forwards). Within each time Period t, we assume that at most one request for a product can arrive; that is, the discretization of time is sufficiently fine so that the probability of more than one request is negligible. This assumption can be generalized in many of the results below, but is the simplest case to present. It is analogous to the network version of the dynamic single-resource model.

To make the notation more compact, demand in Period t is modeled as the realization of a *single* random vector $R(t) = (R_1(t), \ldots, R_n(t))$. If $R_j(t) = r_j > 0$, this indicates a request for Product j occurred and that its associated revenue is $R_j(t)$; if $R_j(t) = 0$, this indicates no request for Product j occurred. A realization $R(t) = 0$ (all components equal to zero) indicates that no request from *any* product occurred at time t. For example, if we have $n = 3$ products, then a value $R(t) = (0,0,0)$ indicates no requests arrived, a value $R(t) = (120,0,0)$ indicates a request for Product 1 with revenue of \$120. Note by our assumption that at most one arrival occurs in each time period, at most one component of $R(t)$ can be positive (as indicated in the example above). More formally, let $E_n = \{e_0, e_1, \ldots, e_n\}$, where e_j is the j-th unit n-vector and e_0 is the zero n-vector, and define the set $\mathcal{S} = \{R : R = \alpha e, e \in E_n, \alpha \geq 0\}$. Then, $R(t) \in \mathcal{S}$. The revenue $R_j(t)$ associated with Product j may be random as well. The sequence $\{R(t); t \geq 1\}$ is assumed to be independent with known joint distributions in each Period t. When revenues associated with Product j are fixed, we will also denote these by r_j.

Given the current time, t, the current remaining capacity x and the current request $R(t)$, we are faced with a decision: Do we or do we not accept the current request?

Let an n-vector $u(t)$ denote this decision, where $u_j(t) = 1$ if we accept a request for Product j in Period t, and $u_j(t) = 0$ otherwise. In general, the decision to accept, $u_j(t)$, is a function of the remaining capacity vector x and the revenue r_j of Product j, i.e. $u_j(t) = u_j(t, x, r_j)$, and hence $u(t) = u(t, x, r)$. Since we can accept at most one request in any period and resources cannot be oversold, if the current seat inventory is x, then $u(t)$ is restricted to the set $\mathcal{U}(x) = \{u \in E_n : Au \leq x\}$.

The Structure of the Optimal Controls

In order to formulate a dynamic program to determine optimal decisions $u^*(t,x,r)$, let $V_t(x)$ denote the maximum expected revenue-to-go given remaining capacity x in Period t. Then $V_t(x)$ must satisfy the Bellman equation

$$V_t(x) = E\left[\max_{u \in \mathcal{U}(x)} \{R^{\mathrm{T}}(t)u(t,x,R(t)) + V_{t+1}(x - Au)\}\right], \tag{16.23}$$

with the boundary condition

$$V_{T+1}(x) = 0, \quad \forall x. \tag{16.24}$$

Therefore, a control u^* is optimal if and only if it satisfies:

$$u_j{}^*(t,x,r_j) = \begin{cases} 1 & r_j \geq V_{t+1}(x) - V_{t+1}(x - A_j) \text{ and } A_j \leq x \\ 0 & \text{otherwise} \end{cases} \tag{16.25}$$

The control (16.25) says that an optimal policy for accepting requests is of the form: accept revenue r_j for Product j if and only if we have sufficient remaining capacity and

$$r_j \geq V_{t+1}(x) - V_{t+1}(x - A_j),$$

where $R_j(t) = r_j$ is the revenue value of the request for Product j. This reflects the rather intuitive notion that we accept a revenue of r for a given product only when it exceeds the *opportunity cost* of the reduction in resource capacities required to satisfy the request.

One can show that bid prices are not able to acheive the optimal control (16.25) in all cases due to the non-additive nature of the value function. This can be shown via some simple counter examples. (See Talluri and van Ryzin, 2002 for details.) At the same time, one can show that bid-price controls have good asymptotic properties and are in fact asymptotically optimal as the number of seats sold and the demand are increased (proportionately), as shown in Talluri and van Ryzin, 1999a. However, in practical terms, these results are somewhat crude. The real test of network control methods and models in practice must be determined through careful simulation testing.

Approximations Based on Network Models

The exact formulation (16.23) can not be solved exactly for most networks of realistic size. Instead, one must rely on approximations of various types. Most approximation methods proposed to date follow one of two basic (not necessarily mutually exclusive) strategies: The first, which we look at in this subsection, is to use a simplified network model. For example, posing the problem as a static math program. The second strategy, which we look at in the next subsection, is to decompose the network problem into a collection of single-resource problems.

Whichever method is used, it is useful to view all such methods as producing different approximations of the optimal value function. The outputs from these approximations can be used to construct controls of various types, either bid-price controls; partitioned or nested allocations; or virtual nesting controls.

Among the most useful information provided by an approximation method are estimates of displacement costs – or bid prices. These are used either directly in bid-price control mechanisms, or indirectly in other mechanisms like virtual nesting. Given an approximation method M that yields an estimate of the value function $V_t^M(x)$, we can approximate the displacement cost of accepting Product j by

$$V_t^M(x) - V_t^M(x - A^j) \approx \nabla_x^{\mathsf{T}} V_t^M(x) A^j,$$

where $\nabla_x V_t^M(x)$ is the gradient of the value function approximation $V_t^M(x)$, assuming the gradient exists. The bid prices are then defined by

$$\pi_i^M(t,x) = \frac{\partial}{\partial x_i} V_t^M(x)$$

If the gradient does not exist, then $\nabla_x V_t^M(x)$ is typically replaced (at least implicitly) by a subgradient of $V_t^M(x)$. If the approximation is discrete, then first differences are used in place of partial derivatives.

Clearly, one objective for an approximation method is to produce a good estimate of the value function – and more importantly, a good estimate of the displacement costs or bid prices. On the other hand, speed of computation matters as well. The approximation $V_t^M(x)$ may be a static approximation that must be resolved quite frequently in practice to account for changes in remaining capacity x and remaining time t. A static method that is accurate but very computationally complex will therefore be of little use in practice. Thus, one should always keep these two criteria – accuracy and speed – in mind when judging network approximation methods.

The Deterministic Linear Programming Model The *deterministic linear programming* (DLP) method uses the approximation

$$\begin{aligned} V_t^{LP}(x) = \max\, & r^{\mathsf{T}} y \\ & Ay \leq x \\ & 0 \leq y \leq E[D] \end{aligned} \tag{16.26}$$

where recall $D = (D_1, \ldots, D_n)$ is the vector of demand-to-come (demand over the periods t, $t + 1, \ldots, T$) for products j and $r = (r_1, \ldots, r_n)$ is the vector of revenues associated with the n products. The decision variables $y = (y_1, \ldots, y_n)$ represent partitioned allocation of capacity for each of the n products. The approximation effectively treats demand as if it were deterministic and equal to its mean $E[D]$ and makes an optimal partitioned allocation accordingly.

The DLP was among the first models analyzed in the early work of D'Sylva, 1982; Glover et al., 1982; Dror et al., 1988; Simpson, 1989; Williamson, 1988; Williamson, 1992, Wollmer, 1986 and Wong, 1990; Wong et al., 1993. Bertsimas and Popescu, 2001 investigate a variation of the DLP method that corrects for integrality and degeneracy of the LP.

Occasionally, the optimal primal solution to (16.26) is used to construct a partitioned control. More often, the primal allocations are discarded and one uses only the optimal dual variables, π^{LP}, associated with the constraints $Ay \leq x$ as bid prices.

The main advantage of the DLP model is that it is computationally very efficient to solve. Due to its simplicity and speed, it is a popular in practice. The weakness of the DLP approximation is that it considers only the mean demand and ignores all other distributional information.

Despite this deficiency, simulation studies have shown that with frequent reoptimization, the performance of DLP bid prices is quite good, producing higher revenue than both the probabilistic nonlinear programming model and a variety of leg-based EMSR heuristics. (See Belobaba and Lee, 2000; Belobaba, 2001; Williamson, 1988; Williamson, 1992.), though other studies have reported more mixed results. (See Belobaba, 1998.) In general, the performance of the DLP method (like many network methods) depends heavily on the type of network, the order in which fare products arrive and the frequency of reoptimization.[8]

The Probabilistic Nonlinear Programming Model The *probabilistic nonlinear programming* (PNLP) method uses the approximation

$$V_t^{PNLP}(x) = \max \sum_{j=1}^{n} r_j E[\min\{D_j, y_j\}] \tag{16.27}$$

$$Ay \leq x$$

$$y \geq 0$$

where again D_j and r_j are defined as in the DLP case. As in the DLP, the decision variables y_j represent a partitioned allocation of capacity to each Product j and the term $E[\min\{D_j, y_j\}]$ is the expected sales of Product j under this partitioned allocation. While this model results in a nonlinear program, it is a relatively easy one; the objective function is concave and separable in the variables y_j, and the constraints are linear. A variety of specialized algorithms can be designed that make solving the PNLP model feasible for large networks.

Again, one can obtain bid-price values from the dual variables of the PNLP. If the active constraints $Ay \leq x$ are linearly independent at the optimal solution, then $\nabla V_t^{LP}(x)$ exists and is given by the unique vector of optimal dual prices associated with these constraints; if the active constraints are dependent, then multiple optimal dual vectors are subgradients of the function $V_t^{LP}(x)$.

Randomized Linear Programming Randomized linear programming (RLP) is another approach for incorporating stochastic information into the DLP method based on replacing the expected value of D in the constraint (16.26) by the random vector D itself. The expected value of the resulting optimal solution then forms an approximation to the value function. That is, define

$$\begin{aligned} H_t(x,D) = \max r^{\mathsf{T}} y& \\ Ay \leq x& \\ 0 \leq y \leq D& \end{aligned} \tag{16.28}$$

The optimal value $H_t(x,D)$ is a random variable. Let $\pi(x,D)$ denote an optimal vector of dual prices for the set of constraints $Ay \leq x$, and note that $\pi(x,D)$ is also a random vector.

Next, consider approximating the value function by the expected value of $H_t(x,D)$,

$$V_t^{RLP}(x) = E[H_t(x,D)]. \tag{16.29}$$

Note the right hand side corresponds to a "perfect information" approximation, because it reflects a case in which *future* allocations (and revenues) are based on perfect knowledge of the realized demand D. We then use $\nabla_x E[H_t(x,D)]$ as a vector of bid prices.

The RLP approximation (16.29) requires a method to efficiently compute $\nabla_x E[H_t(x,D)]$. One simple approach is to simulate k independent samples of the demand vector, $D^{(1)},\ldots,D^{(k)}$, and solve (16.28) for each sample. Then estimate the gradient using

$$\pi^{RLP} = \frac{1}{k}\sum_{i=1}^{k} \pi(x,D^{(i)}). \tag{16.30}$$

That is, simply *average* the dual prices from k perfect-information allocation solutions on randomly generated demands. Hence the name *randomized linear* *programming* (RLP) method. The randomized linear programming method first mentioned in Smith and Penn, 1988, and later investigated by Talluri and van Ryzin, 1999b.

Approximations Based on Decomposition

Another strategy for generating network controls is to decompose the problem (approximately) into m single-resource problems, each of which may incorporate some network information, but which are nevertheless independent. Formally, one can think of such a decomposition method as follows: an approximation method M decomposes the network problem into m single-resource models, denoted Model $i = 1, \ldots, m$, with value functions $V_t^{M_i}(x_i)$, that depends on the time-to-go t and the remaining capacity x_i of Resource i. These may be constructed by incorporating some static, network information into the estimates. Then, the total value function is approximated by

$$V_t^M(x) = \sum_{i=1}^{m} V_t^{M_i}(x_i).$$

Typically, such approximations are discrete and yield bid prices

$$\pi_i(t,x) = \Delta V_t^{M_i}(x_i), \quad i = 1,\ldots,m.$$

where $\Delta V_t^{M_i}(x_i) = V_t^{M_i}(x_i) - V_t^{M_i}(x_i - 1)$ is the usual marginal expected value produced by Model i.

Decomposition approximations have several advantages relative to network approximations. First, because they are based on single-resource problems, the displacement costs and bid prices are typically dynamic and can be represented as a table of outputs (in the case of dynamic programming models) or simple formulas (in the case of EMSR approximations). Thus, it is easy to quickly determine the effect of changes in both the remaining time t and remaining capacity x on the resulting bid prices. This should be contrasted with network models, which must typically be re-solved to determine the effects of such changes. Second, because they are often based on simple, single-resource models, decomposition methods allow for more realistic assumptions, such as discrete demand and capacity, sequential decision making over time and stochastic dynamic demand.

The primary disadvantage of decomposition methods is that in the process of separating the problem by resources, it can be difficult to retain important network effects in the approximations. However, as we show below, hybrids of the two approaches can be used to try to achieve the benefits of both network and decomposition methods.

Prorated EMSR This idea was first investigated by Williamson, 1992 with EMSR and was called the prorated expected marginal seat revenue (PEMSR) scheme.

The PEMSR schemes involve allocating a portion of the revenue of each product to the resources used by the product. One then solves m single-resource-level models using the EMSR heuristic, though other single-resource models can be used as well. The resulting marginal values from each resource are then used as bid prices in a bid-price control scheme or the allocations are used directly in resource-level nested allocation controls.

Specifically, let $\alpha = (\alpha_1, \ldots, \alpha_m)$ be a non-negative real vector. For each Product j, define new revenues, one for each Resource i in the product, by

$$\bar{r}_{ij} = \frac{\alpha_i}{\sum_{i \in A_j} \alpha_i} r_j \quad i \in A_j.$$

Next, treat each resource i independently as if it received demand D_j, but with reduced revenue $\bar{r}_{ij}$ and solve the corresponding EMSR model. The approximation to the value function is then

$$V_t^{PEMSR}(x) = \sum_{i=1}^{m} V_t^{EMSR_i}(x_i, \alpha),$$

where $V_t^{EMSR_i}(x_i, \alpha)$ denotes the expected revenue of Model i under the allocation α.

Williamson, 1992 investigated several methods for determining the allocation weights α in airline problems, including prorating based on mileage, number of resources and the relative revenue value of local demand on each resource. Her conclusion is that none of these fixed allocations is very robust in general. This is one of the main disadvantages of fixed proration, and hence the idea is not used much anymore in practice.

Displacement Adjusted Virtual Nesting (**DAVN**) While virtual nesting is often viewed as a control strategy – and indeed is used as such in most cases in practice – it can also be viewed as a decomposition approximation to the network value function. Indeed, the marginal values produced by the virtual nesting approximation can be used in a bid-price control scheme which avoids the virtual nesting controls entirely.

DAVN starts with a set of static bid prices – or marginal value estimate – which we denote by $\bar{\pi} = (\bar{\pi}_1, \ldots, \bar{\pi}_m)$. These estimates may be obtained, for example, from one of the various network math programming models presented in

Section 16.3. Given the bid prices $\bar{\pi}$, one then solves a leg-level problem at each resource i as follows:

First, for all Products j that use Resource i, a *displacement adjusted revenue* $\bar{r}_{ij}$ is computed using

$$\bar{r}_{ij} = r_j - \sum_{l \in A_j, l \neq i} \bar{\pi}_l. \tag{16.31}$$

That is, the revenue of Product j on Resource i is reduced by the static bid-price values of the other resources used by Product j. This adjustment is intended to approximate the net benefit of accepting Product j on Resource i.[9]

The next step is clustering – or *indexing*. In this step, the displacement adjusted revenue values on each resource are clustered into a specified number $\bar{c}$ of *virtual classes* – or *buckets* – denoted $c = 1, \ldots, \bar{c}$. The number of virtual classes, $\bar{c}$, is a design parameter, but is typically on the order of 10. It may also vary across resources. The indexing from Product j to Virtual Class c on each resource can be performed using a variety of clustering algorithms. The particular indexing method and clustering criteria are also design decisions and vary from implementation to implementation.

Once the virtual classes are formed, we compute a representative revenue value for each class – usually the demand-weighted average revenue. Then, the distribution of total demand in a virtual class is computed – typically by adding the means and variances of demand-to-come. Next, one solves a multi-class, single-resource problem based on these data. The problem could be solved exactly using the static single-leg model or approximately using an EMSR heuristic. We call this Model i. This procedure yields a set of booking limits (or protection levels) for the virtual classes at each resource i and a value function estimate $V_t^{DAVN_i}(x_i)$.

The resulting DAVN approximation can be used in two basic control strategies. Most often, the control is a booking limit control on the virtual classes. That is, a request for Product j is converted into a request for the corresponding virtual class at each Resource i required by Product j. (Note the virtual class on each of these resources need not be the same.) If the virtual class on each resource is available, the request is accepted. If the virtual class on one or more resources is closed, the request is rejected. Thus, once the indexing from products to virtual classes is performed, the control logic is an independent, nested allocation class-level control at each resource in the network. This is the primary appeal of – and motivation for – the method in the airline industry, because it produces the sort of booking-class-level controls that are widely used by CRSs.

However, DAVN can also be used to produce bid-price controls. The bid price for Resource i is simply given by

$$\pi_i^{DAVN}(t,x) = \Delta V_t^{DAVN_i}(x_i)$$

where as usual $\Delta V_t^{DAVN_i}(x_i) = V_t^{DAVN_i}(x_i) - V_t^{DAVN_i}(x_i - 1)$ denotes the marginal value generated by Model *i*.

Regardless of the control method, typically the network model that was used to generate the static bid prices $\bar{\pi}$ is re-solved and the products are re-indexed periodically as demand conditions change. In the airline industry, for example, the indexing process is a fairly major change to the CRS, so often it is only done on a seasonal basis.

Dynamic Programming Decomposition

Dynamic programming decomposition is very similar in spirit to DAVN. Indeed, the only real difference is that while DAVN takes displacement adjusted revenues and aggregates them into a small number of virtual classes, in dynamic programming decomposition, the revenue and demand remains disaggregated. As with other decomposition methods, there are several possible variations to the basic approach. However, for purposes of illustration we focus on one special case here; specifically, the dynamic single-resource model in which demand for Product *j* arrives in Period *t* with probability $\lambda_j(t)$.

We start the decomposition as in DAVN with a static vector of bid price $\bar{\pi}$. Again, this may be computed in a variety of ways, though typically using one of the network math programming models of Section 16.3. Then, for each Resource *i*, we solve a single-resource dynamic program based on displacement adjusted revenues. That is, for each Product *j* that uses Resource *i*, compute the displacement adjusted revenue

$$\bar{r}_{ij} = r_j - \sum_{l \in A_j, l \neq i} \bar{\pi}_l.$$

Then formulate a dynamic single-resource model for Resource *i* (Model *i*) with arrival rates $\lambda_j(t)$ and revenues $\bar{r}_{ij}$. Let the resulting value function be denoted $V_t^{DPD_i}(x_i)$. The total value function approximation is then

$$V_t^{DPD}(x) = \sum_{i=1}^{m} V_t^{DPD_i}(x_i),$$

and the bid prices are given by

$$\pi_i^{DPD}(t,x) = \Delta V_t^{DPD_i}(x_i), \quad i = 1, \dots, m,$$

where $\Delta V_t^{DPD_i}(x_i)$ is the marginal value from Model *i*.

Because dynamic programming decomposition is so similar to DAVN, the choice of which one to use is most often dictated by the control strategy one wants to use in the end. If the objective is to construct virtual nesting controls, then aggregating and indexing as in DAVN will more accurately match the control strategy. If one is using bid-price controls, then the aggregating and indexing of DAVN is not necessary and a dynamic programming decomposition will tend to yield more accurate bid prices approximations.

Iterative Decomposition Methods Iterative decomposition methods are also closely related to DAVN. The methods were originally based on the EMSR heuristics for the single-resource problem and so were called *iterative prorated EMSR* methods, but there are several variations of the general idea and it can be used with any single-resource model. The approach can be found in several sources, including Williamson, 1988; Williamson, 1992; Phillips, 1994. Here we look at only one version to illustrate the idea.

Iterative DAVN is essentially a method for computing the static bid prices, $\bar{\pi}$, used by DAVN. The motivation for the approach is heuristic but intuitively appealing. Namely, if $\bar{\pi}$ indeed represents the marginal displacement cost vector, then once DAVN is solved, for consistency we should have that

$$\bar{\pi}_i = \pi_i^{DAVN}(t,x), \quad i = 1,\ldots,m,$$

where $\pi_i^{DAVN}(t,x) = \Delta V_t^{DAVN_i}(x_i)$, the marginal value generated by Model i at the current time (time t). That is, the marginal costs produced by the DAVN decomposition should match our static estimate $\bar{\pi}$.

A natural question arises then: What happens if $\pi_i^{DAVN}(t,x) \neq \bar{\pi}_i$? The idea of iterative methods is that if these values do not match, we simply feed back the estimates $\pi_i^{DAVN}(t,x)$ as new static bid prices into the procedure and recompute the DAVN models.

Abstractly, this algorithm produces a mapping, Ψ, from the space of bid-price vectors $\bar{\pi}$ onto itself. That is, $\bar{\pi}^{(k+1)} = \Psi(\bar{\pi}^{(k)})$. The algorithm terminates if it finds a fixed point $\bar{\pi}^*$ of this map, $\bar{\pi}^* = \Psi(\bar{\pi}^*)$. However, whether Ψ is a contraction mapping or not has not been investigated to date for DAVN. For a similar scheme based on prorating revenues by the bid prices, Bratu, 1999 proves this mapping is convergent. Howver, one ought not to read too much into this fact (e.g. a convergence of all bid prices to zero would satisfy this claim), and there are some counter examples that show that the resulting convergent bid prices can be quite bad. (See Talluri and van Ryzin, 2002.)

16.4 Overbooking

Overbooking is somewhat distinct from the core pricing and capacity allocation problems of revenue management. While most of revenue management is concerned with how to *best use* or *price* a given amount of capacity, overbooking is concerned with *how much* capacity to provide in the first place. These two problems are, of course, quite related, and generally both are considered part of revenue management.

From a historical standpoint, overbooking is the oldest – and, in financial terms, among the most successful – of revenue management practices. For example, in the airline industry it is estimated that approximately 50% of reservations result in cancellations or no-shows and, as a result, about 15% of all seats would go unsold without some form of overbooking. (See Smith et al., 1992.) This is to be compared to fare class allocation, which by most estimates accounts for on the order of 5% in incremental revenues. Despite this, many researchers consider overbooking a somewhat mature area and it has received less attention in the research literature than fare class allocation and pricing.

We first review the history and practice of overbooking. We then survey the main methods for making overbooking decisions.

History and Practice of Overbooking

Rothstein's series of articles (Rothstein, 1971; Rothstein, 1975; Rothstein, 1985) provide the best source for the history of overbooking in the airline industry. There was also much lively debate surrounding the oversale auction idea, captured in a series of articles (Simon, 1968; Simon, 1972; Simon, 1993; Simon, 1994). Here, we briefly review this history.

Prior to 1961, intentional overbooking was practiced somewhat clandestinely by U.S. airlines and was not acknowledged publicly. Despite this fact, Rothstein, 1985 reports that as director of OR at American Airlines he "… found much publicly available evidence that all the major airlines were deliberately overbooking." In 1961, the Civil Aeronautics Board (CAB) reported a no-show rate of 1 out of every 10 passengers booked among the 12 leading carriers at that time. The CAB acknowledged that this situation created real economic problems for the airlines. The CAB conducted another study of overbooking in 1965–66. They found that the denied boarding rate at that time was only 7.69 per 10,000 passengers boarded (Civil Aeronautics Board, 1967) and concluded that, overall, overbooking practices benefited the traveling public by lowering the cost of air travel. Thus, as of 1965, overbooking was an officially sanctioned practice, provided it was "carefully controlled," a criterion that was never precisely defined by the CAB.

Table 16.4 U.S. Major Airline Denied Boarding Rates: 1993–2000[a,b,c]

	1993	1994	1995	1996	1997	1998	1999	2000
Boarded $(\times 10^6)$	449	457	460	481	503	514	523	540
Total DB $(\times 10^3)$	683	824	843	957	1,072	1,126	1,070	1,113
Voluntary	632	771	794	899	1,018	1,081	1,024	1,057
Involuntary	51	53	49	58	54	45	46	56
VDB Rate (per 10^4*)*	15.21	18.03	18.33	19.90	21.31	21.03	19.58	19.57
IDB Rate (per 10^4*)*	1.14	1.16	1.07	1.21	1.07	0.87	0.88	1.04

[a] Data are for nonstop scheduled service flights between points within the United States (including territories) by the 10 largest U.S. air carriers, i.e., those with at least 1% of total domestic scheduled-service passenger revenues (Alaska, America West, American, Continental, Delta, Northwest, Southwest, TWA, United, and US Airways). Before 1994, carriers included both majors and national airlines, i.e., airlines with over $100 million in revenue.

[b] Statistics are the number of passengers who hold confirmed reservations and are denied boarding ("bumped") from a flight because it is oversold. These figures include only passengers whose oversold flight departs without them; they do not include passengers affected by canceled, delayed, or diverted flights.

[c] *Source:* U.S. Department of Transportation, Office of the Secretary, Air Travel Consumer Report (Washington, DC: Annual March issues).

In parallel, the CAB also increased the denied boarding penalty to 100% of the coupon. Airlines controlled the percentage of denied boardings and the CAB carefully monitored the denied boarding performance of each airline. The involuntary denied boarding rate is still carefully monitored in the U.S. by the Department of Transportation (DOT) and currently hovers around 1–2 involuntary and 15–20 voluntary denied boardings per 10,000 passengers (see Table 16.4).

In 1968, economist Julian Simon proposed what he called "an almost practical solution to airline overbooking," in which airlines would conduct a sealed-bid "reverse auction" to find passengers willing to accept monetary compensation for being bumped. Simon predicted (rightly so, as initial responses to his letters to airline executives later indicated) that the airlines would object to the scheme. Simon wrote many letters to executives, regulators, policy makers and consumer groups arguing for his "oversale auction" idea. Despite these efforts, he failed to get even one airline to experiment with it on one flight. The scheme continued to flounder until 1977 when Alfred Kahn, an economist, was appointed to head the CAB. Simon wrote to Kahn about his proposal and Kahn liked and largely adopted it under the heading of a "volunteer" bumping plan, as mentioned above. It has since proved very successful in the industry, providing a fair and efficient means of implementing voluntary denied boarding in the airline industry, and is now a widespread practice.

Static Models

We next look at the methodology for making overbooking decisions. The simplest and most widely used methodology is based on *static* overbooking models. In static models, the dynamics of customer cancellations and new reservation requests over time are ignored. Rather, the models simply determine the maximum number of reservations to hold at the current time given cancellation probabilities from the current time until the day of service. This maximum number of reservations, or *overbooking limit,* is then recomputed periodically prior to service to reflect changing cancellation probabilities over time. While more sophisticated dynamic overbooking models have been developed, the simplicity, flexibility and robustness of the simpler static models has made them more popular in practice.

Two types of events impact the overbooking decision: *cancellations* and *no-shows,* with the difference simply related to the timing of the events. A cancellation is a reservation that is withdrawn by a customer strictly prior to the time of service. A no-show is a reservation that is in the system until the time of service, but the customer does not show up at the time of service. Under a static model, the distinction between the two is unnecessary, since a static model assumes a static overbooking limit is set without recourse to adjust it. Thus, all that matters is the probability that a reservation survives to the time of service (the *show demand* as it is sometimes called). In dynamic overbooking models, however, the distinction between no-shows and cancellations is quite important.

As mentioned, static models are typically used to compute overbooking limits – also called *virtual capacities* or *authorization levels* in the airline industry – which in turn are inputs to capacity control models. These static overbooking models are typically resolved periodically to account for changes in the cancellation and no-show probabilities over time, resulting in overbooking limits that vary (typically decline) over time. The maximum number of reservations one can accept at any time is given by the current overbooking limit.

The situation is illustrated in Figure 16.5. The top, wide line is the overbooking limit over time. Solving a static model gives one point on this curve. Overbooking limits are initially high because the probability of a reservation currently in the system cancelling or no-showing prior to the time of service is typically higher when the time to service is further away. At the time of service (T) approaches, the overbooking limits fall. At the same time, reservations are being accumulated in the system over time. The dark line in Figure 16.5 shows that with overbooking in place, the reservations in the system can exceed the capacity $C,$ and we don't stop accepting reservations until the overbooking limit is reached. At that point reservations are rejected. The resulting show demand at time T is (ideally) close to the capacity $C.$ The lighter line shows the same trajectory of reservations without

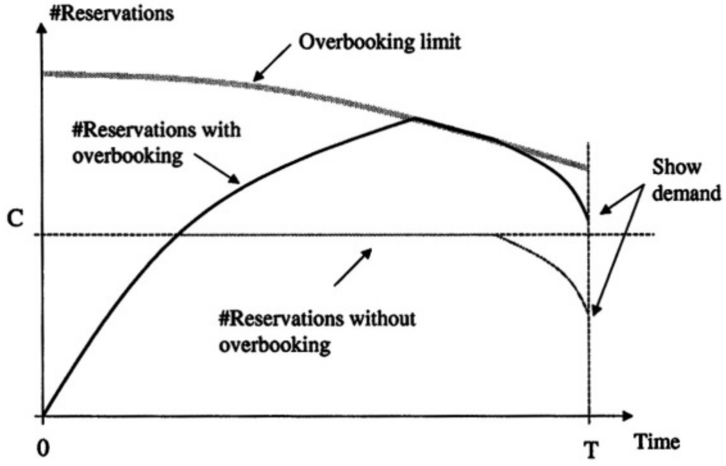

#Reservations
Overbooking limit
#Reservations with overbooking
Show demand
C
#Reservations without overbooking
0
T
Time

Figure 16.5 Illustration of Overbooking Limits and Reservations Over Time

overbooking. In this case, the reservations in the system are truncated at the capacity *C* early on in the booking process. As a result, once reservations start to cancel and no-show, the show demand is significantly less than capacity.

The static overbooking problem first appeared in a pair of papers by Beckmann, 1958; Beckman and Bobkowski, 1958. Other early treatments of the static problem are Taylor, 1962, Thompson, 1961 and Rothstein and Stone, 1967. See also Bierman Jr. and Thomas, 1975 and Shlifer and Vardi, 1975. Dynamic overbooking models (not covered here) are addressed in Chatwin, 1993 and subsequent published articles Chatwin, 1997; Chatwin, 1999.

The Binomial Model

The simplest static model is based on a Binomial model of cancellations in which no-shows are lumped together with cancellations (e.g. a no-show is treated simply as a cancellation that occurs at the day of service).

The following assumptions are made:

- Customers cancel independently of one another.
- Each customer has the same probability of canceling.
- The cancellation probability is Markovian; it depends only on the time remaining to service and is independent of the age of the reservations.

(Martinez and Sanchez, 1970 tested the Markovian property of the binomial model empirically and showed it is a reasonable approximation.)

Let t denote the time remaining until service, C denote the physical capacity, u denote the number of reservations on-hand and q denote the probability that a reservation currently on hand shows up at the time of service ($1 - q$ is the probability they cancel prior to the time of service). Note q is really a function of the time remaining, since in general the more time remaining the more likely it is that customers cancel before the time of service. However, to keep the notation simple we suppress the dependence of q on t.

Under the assumptions stated above, the number of customers who show up at the time of service (the show demand), denoted $Z(u)$, is binomially distributed with p.m.f.

$$\begin{aligned} p_u(z) &= P(Z(u) = z) \\ &= \binom{u}{z} q^z (1-q)^{u-z}, \quad z = 0, 1, \ldots, u, \end{aligned} \tag{16.32}$$

c.d.f.

$$\begin{aligned} F_u(z) &= P(Z(u) \leq z) \\ &= \sum_{k=0}^{z} \binom{u}{k} q^k (1-q)^{u-k}, \end{aligned} \tag{16.33}$$

mean $E[Z(u)] = qu$ and variance $Var(Z(u)) = uq(1 - q)$. It is convenient to work with the compliment of the distribution F_u, denoted by $\bar{F}_u$, which is defined by

$$\bar{F}_u(z) = 1 - F_u(z)$$

Several studies have validated this binomial model of cancellations. For example, in one of the earliest investigations of overbooking, Thompson Thompson, 1961 considers data from 59 flights from Auckland to Sydney operated by Tasman Empire Airways. He eliminated groups of 6 or more since they exhibit much lower cancellation rates and although rare (11 total booking on the 59 flights), can significantly distort the cancellation rate on the flights involved. Parties of 6 or less constituted 99.6% of all bookings; 81% of the remaining were singles; 15% were paired and 4% were parties of 3-6. While the results showed that group cancellation behavior does invalidate the the binomial model for certain cabins at certain time periods, overall he concludes the binomial model adequately fits the data. (Group cancellation effects are discussed further below.)

Overbooking Based on Service Level Criteria One measure of service is the probability of oversale at the time of service, which we call the *Type 1 service* *level.* Given that there are u reservations on hand, this probability is denoted $s_1(u)$ and is given by

$$s_1(u) = \bar{F}_u(C).$$

A more intuitive measure of service is the fraction of customers who are denied service, which we call the *Type 2 service level* and denote by $s_2(u)$. This fraction is given by

$$\begin{aligned} s_2(u) &= \frac{E[(Z(u) - C)^+]}{E[Z(u)]} \\ &= \frac{\sum_{k=c+1}^{u} (k - C)p_u(k)}{u(1-q)}. \end{aligned}$$

Through some algebraic simplification, this simplifies to

$$s_2(u) = \bar{F}_{u-1}(c-1) - \frac{C}{qu}\bar{F}_u(c), \tag{16.34}$$

which is a more convenient formula for computations.[10]

Tabel 16.5 shows the Type 1 and Type 2 service levels for an example with $c = 150$, $q = 0.85$ and varying overbooking limits u^*. Typically, one first specifies a service standard and then numerically searches for the largest booking level u^* satisfying this standard. The resulting u^* is the overbooking limit. The quantity $u^* - C$ (the excess over capacity) is typically referred to as the *overbooking pad.*

Example 2 *Suppose we want an average of no more than* 0.1% *of customers to be denied service* ($s_2(u) \leq 0.001$) *and our capacity is* $C = 150$ *and* $q = 0.85$. *From Table 16.5, we should accept at most* 168 *reservations* ($u^* = 168$). *(Thought* 169 *has a service level only slightly over the standard and might be a candidate as well.) Reservations would then be accepted as long as the number of bookings on hand was less than* u^*. *The overbooking pad would be* $168 - 150 = 18$, *so we would be willing to oversell the capacity by* 18 *units.*

Overbooking Based on Economic Criteria An alternative to setting overbooking limits based on service standards is to use an economic criteria. This approach requires an estimate of the revenue loss from not accepting additional reservations and an estimate of the cost of denied service. We first develop the details of the economic-based model, and then discuss some of the issues involved in estimating the required revenue and cost inputs.

Table 16.5 Binomial and Normal Approximation Overbooking Probabilities and Service Levels: $C = 150$, $q = 0.85$

	Binomial model			*Normal approximation*	
u	$\bar{F}_{u-1}(c-1)$	$s_1(u)$	$s_2(u)$	$s_1(u)$	$s_2(u)$
160	0.00021	0.00019	0.00000	0.00097	0.00001
161	0.00053	0.00048	0.00001	0.00185	0.00002
162	0.00122	0.00111	0.00001	0.00340	0.00003
163	0.00262	0.00239	0.00003	0.00601	0.00006
164	0.00523	0.00480	0.00006	0.01022	0.00011
165	0.00979	0.00904	0.00012	0.01676	0.00020
166	0.01726	0.01603	0.00022	0.02652	0.00033
167	0.02883	0.02691	0.00039	0.04053	0.00053
168	0.04577	0.04294	0.00066	0.05989	0.00084
169	0.06935	0.06539	0.00107	0.08566	0.00127
170	0.10062	0.09533	0.00166	0.11873	0.00188
171	0.14025	0.13351	0.00247	0.15966	0.00270
172	0.18838	0.18015	0.00355	0.20855	0.00377
173	0.24449	0.23484	0.00494	0.26496	0.00514
174	0.30743	0.29654	0.00668	0.32785	0.00685
175	0.37549	0.36365	0.00879	0.39565	0.00891

Let z denote the number of customers who show up on the day of service (the show demand) and let $c(z)$ denote the denied service cost. We shall assume $c(z)$ is an increasing convex function of z. For example, a simple and common assumption in practice is that each denied service costs the provider a constant marginal amount h, in which case

$$c(z) = h(z - C)^{+}. \tag{16.35}$$

An arguably more realistic assumption is to assume strictly increasing marginal costs, reflecting the need to offer higher levels of compensation (or incur higher goodwill costs) as each additional customer is denied service.

Let r denote the marginal revenue of accepting an additional reservation. One could also allow this marginal revenue to vary, but it is a common simplification in practice to consider it fixed, though we discuss this issue further below. Then the total profit from having u reservations on hand is given by

$$\pi(u) = ru - E[c(Z(u))], \tag{16.36}$$

where recall that the random variable $Z(u)$ denotes the number of customers who show up on the day of service out of u reservations. One can show for the binomial model that if $c(z)$ is convex, then $\pi(u)$ is convex in u.[11] Therefore a maximizing u^* is the largest value of u satisfying

$$E[c(Z(u))] - E[c(Z(u) - 1)] \le r.$$

For the constant marginal cost case where $c(z) = h(z - C)^+$, this condition reduces to

$$hqP(Z(u-1) \ge C) \le r. \tag{16.37}$$

This expression can be argued intuitively by noting that when we accept the u-th reservation, we incur a marginal denied boarding penalty of h if and only if: (i) the current reservations on hand consume all the capacity $(Z(u-1) \ge C)$, and (ii) the u-th customer shows up. The left-hand-side above is simply the marginal penalty multiplied by the probability of this event, or equivalently the expected marginal cost. Then u^* is the largest value of u for which the expected marginal cost is less than the marginal revenue.

We can express (16.38) as

$$\bar{F}_{u-1}(C-1) \le \frac{r}{qh}. \tag{16.38}$$

Note this is equivalent to setting a fixed Type 1 service level for a capacity of $C - 1$. For large capacities C, $\bar{F}_u(C-1) \approx \bar{F}_u(C)$, so using economic criteria with constant marginal costs corresponds approximately to specifying a particular Type 1 service level, which provides one justification for using Type 1 service levels.

Table 16.5 displays probabilities for $\bar{F}_{u-1}(C-1)$ for the example $C = 100$ and $q = 0.85$. To illustrate (16.38), suppose the overbooking cost is $h = \$500$ and the marginal revenue is $r = \$100$. Then $r/qh = 0.235$. We see from Table 16.5 that the optimal overbooking limit is then $u^* = 172$.

Static Model Approximations

While the binomial model is quite simple, it is often desirable to have simpler, closed-form expressions for the overbooking limits. We next look briefly as such approximations.

Deterministic Approximation The deterministic approximation simply sets the overbooking limit so that the average show demand is exactly equal to the capacity. That is,

$$u^* = \frac{C}{q}.$$

As simplistic as this is, we have seen implementations where this approximation (or variations of it) are in fact used.

Normal Approximation A popular approximation in practice is the normal approximation, in which $F_u(x)$ is replaced by the normal distribution with mean, μ, and variance, σ^2, chosen to match the binomial, viz

$$\mu_u = qu$$
$$\sigma_u^2 = uq(1-q).$$

The Type 1 service level is then approximated by

$$s_1(u) \approx 1 - \Phi(z_u),$$

where $z_u = \frac{C-\mu_u}{\sigma_u}$ and $\Phi(z)$ is the standard normal distribution $\Phi(z) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z/2} dz$. Let $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}}$ denote the p.d.f. of the standard normal. The Type 2 service level is then approximated by:[12]

$$s_2(u) \approx \frac{\sigma_u}{\mu_u} \left[\phi(z_u) - z_u(1 - \Phi(z_u))\right]. \tag{16.39}$$

The last columns of Table 16.5 show the estimates produced by the normal approximation for our example with $C = 150$ and $q = 0.85$, which are resonably close to the values of the binomial model.

The economic-based overbooking limit (16.38) for the constant-marginal-cost function (16.35) is approximated by chosing u^* to satisfy

$$\Phi_{u^*}(C) = 1 - \frac{r}{qh}. \tag{16.40}$$

Group Cancellations

The presence of groups also has an important effect on cancellation models in practice. If a group decides to cancel, then all reservations are cancelled simultaneously. The resulting positive correlation in cancellations increases the variance of the show demand. When dealing with large numbers of reservations, it is often possible to ignore the effect of groups but with small numbers of reservations, group effects can result in significant deviations from the binomial model.

To gain some sense of the presence of groups in airline reservations, Table 16.6 provides an empirical distribution of group sizes over approximately one half

Table 16.6 Empirical Distribution of Group Sizes[a]

Number in Party	*Count of Passengers*	*Percent*	*Cummulative Percent*
1	198,056	45.1	45.1
2	114,418	26.0	71.1
3	28,641	6.5	77.6
4	25,688	5.8	83.5
5	17,930	4.1	87.6
6	7,134	1.6	89.2
7	2,135	0.5	89.7
8	2,896	0.7	90.3
9	1,125	0.3	90.6
10	4,960	1.1	91.7
>10	36,375	8.3	100.0
Total	439,358	100.0	

[a] Data reported by Rothstein and Stone, 1967.

million airline reservations, showing that about half of reservations are individual reservations while the other half are from groups of two or more.

One simple technique used in practice to adjust for group size is to simply inflate the variance of the show demand by a factor that accounts for group size. For example, if one is using the normal approximation to the binomial model as described above, then the mean estimate, μ_u, is unchanged but the variance estimate, σ_u^2, is modified; that is,

$$\begin{aligned}\mu_u &= qu\\ \sigma_u^2 &= kuq(1-q),\end{aligned}$$

where k is an estimate of the average group size.[13] A more refined technique based on moment generating functions and is discussed in Talluri and van Ryzin, 2002.

Combined Capacity Control and Overbooking Models

Thus far, we have analyzed the overbooking problem in isolation without considering the interaction of overbooking decisions with capacity controls. We next look at both exact and approximate methods to model cancellations and no-shows together with class allocations. The treatment here is largely based on the work of Subramanian et al., 1999.

Incorporating no-shows or cancellations in either the static or dynamic single-resource model is not too difficult theoretically provide one makes the following set of assumptions:

Assumption 1

(i) *The cancellation and no-show probabilities are the samefor all customers.*
(ii) *Cancellations and no-shows are mutually independent accross customers.*
(iii) *Cancellations and no-shows in any period are independent of the time a reservation was accepted.*
(iv) *The refunds and denied service costs are the samefor all customers.*

As a result of these assumptions, the number of no-shows and the costs incurred are only a function of the total number of reservations on hand. Therefore, we only need to retain a single state variable, and the resulting dynamic programs are only slightly more complex than those presented in Section 16.2.

The most offensive of these assumptions in practice are (i) and (iv). In particular, because cancellation options and penalties are often linked directly to a booking class, cancellation and no-show rates and costs can be vary significantly from one class to the next. Ideally, these differences should be accounted for when making allocation decisions. However, this significantly complicates the problem as discussed below. As already mentioned, Assumption 1-(ii) is often unrealistic, because reservations from people in groups typically cancel at the same time. Assumption 1-(iii) is typically less of a problem in practice and as mentioned above has some empirical support (See Thompson, 1961.).

It is partly for all these reasons that in practice the overbooking problem is separated from the capacity allocation problem. Often, an approximate overbooking model can be solved that is able to relax (at least heuristically) some or all parts of Assumption 1. However, given Assumption 1 the two problems can be combined exactly as shown below.

Exact Methods for No-shows Under Assumption 1 We first consider only no-shows without cancellations. (Recall, a no-show is a customer who does not show up at the time of service, while a cancellation is a reservation that departs the system strictly before the time of service.) Let q_0 denote the probability that a customer with a reservation shows up for service $(1-q_0$ is the no-show probability). By Assumption 1 (i), this probability is assumed to be the same for all customers, and by Assumption 1 (ii) it is assumed to be independent of when the reservation was made.

Let $Z_i = 1$ if customer i shows up for service and $Z_i = 0$ otherwise. Given x reservations on hand at the time of service, the number of customers who show up at time zero (the show demand), denoted $Z(x)$, is then

$$Z(x) = \sum_{i=1}^{x} Z_i.$$

and by Assumption 1-iii) $Z_0(x)$ is a binomial(q_0, x) random variable, with

$$P(Z_0(x) = z) = \binom{x}{z} q_0^z (1 - q_0)^{x-z}, \quad z = 0, 1, \dots, x$$

By Assumption 1 iv), the total cost of denied service is only a function of the show demand z. Let $c(z)$ denote the overbooking cost given z. We will require the $c(z)$ be increasing and convex with $c(0) = 0$. Convexity in cost is quite natural since the marginal cost of denying service to customers tends to increase with the number denied. For example, we could have a simple linear cost h per denied customer in which case $c(z) = h(z - C)^+$ where, as before, C is the capacity. If customers are refunded a class

Given this no-show model, the expected cost of service given that there are x reservations on hand at the time of service, which we denote $V_0(x)$, is given by

$$V_0(x) = E[-c(Z(x)], \quad x \geq 0. \tag{16.41}$$

Stochastic convexity arguments show that $V_0(x)$ is concave in x if $c(\cdot)$ is convex. The above expression then replaces the boundary conditions of the dynamic program for the static and dynamic models. We look at each in turn.

Static Model Consider the static model of Section 16.1, where the classes are ordered $r_1 > r_2 > \cdots > r_n$ and we assume classes arrive in the order of lowest to highest revenue. Classes and stages are indexed by j. The state variable, x, is now defined to be the number of reservations on-hand rather than the remaining capacity.

The Bellman equation (16.2) for the static model is then modified to account for no-shows as follows

$$V_j(x) = E\left[\max_{0 \leq u \leq D_j} \{r_j u + V_{j-1}(x + u)\}\right], \tag{16.42}$$

with boundary conditions (16.41), where here $V_j(x)$ is now interpreted as the expected *net benefit* (expected revenue minus the expected terminal cost) of operating the system from period j onward given that there are x reservations on hand.[14]

Given the concavity of $V_0(x)$, a modification of Proposition 1 shows that the value function $V_j(x)$ in (16.42) is concave in x for all j and x. Since there is no hard capacity constraint in this case, it is more meaningful to express the optimal policy in terms of booking limits. The optimal nested booking limits are given by

$$b_j^* = \min\{x \geq 0 : r_j < \Delta V_{j-1}(x)\}, \quad j = 1, \ldots, n-1,$$

where $\Delta V_{j-1}(x) \doteq V_{j-1}(x) - V_{j-1}(x+1)$ now has the interpretation as the marginal opportunity cost of holding another reservation in Stage $j-1$. It is then optimal to accept Class j if and only if the number of reservations on hand, x, is strictly less than b_j^*.

Dynamic Model Similarly, the optimality equations (16.9) for the dynamic model of Section 16.2 are modified to account for no-shows as follows:

$$V_t(x) = E\Big[\max_{u \in \{0,1\}} \{R_t u + V_{t+1}(x+u)\}\Big], \tag{16.43}$$

where again the state variable x is the number of reservations on hand. The boundary conditions are

$$V_{T+1}(x) = E[-c(Z(x)], \quad x \geq 0. \tag{16.44}$$

Again, it is optimal to accept an arrival of Class j iff

$$r_j \geq \Delta V_{t+1}(x),$$

where again $\Delta V_{t+1}(x) = V_{t+1}(x) - V_{t+1}(x+1)$, has the interpretation as the marginal cost of accepting another reservation.

Note that under this model, one can always justify accepting a sufficiently high revenue r_j, provided the marginal cost $\Delta V_{t+1}(x)$ is finite. This makes perfect economic sense since we should in principle we willing to accept an almost certain denied service cost if some customer is willing to pay enough to compensate us for this cost. For example, if the overbooking cost is linear of the form $c(z) = h(z - C)^+$, then one can show that the marginal cost is never more than h, so any request with revenue greater than h will always be accepted. This property of not having an explicit limit on the number of reservation (only an economically driven limit) has been called "infinite overbooking" by some in the airline industry, since it is in sharp contrast to the usual practice of setting a hard overbooking limit.

Also, it highlights some of the potential suboptimality of using fixed overbooking limits.

Exact Methods for Cancellations under Assumption 1 Cancellations complicate the dynamic program a little more than no-shows, but they are still quite manageable given Assumption 1. Again, we look at the static and dynamic models in turn.

Static Model Let q_j denote the probability that a reservation in the system at the start of period j survives to period $k + 1$. Then $1 - q_j$ is the probability that a reservation cancels in Period j. By Assumption 1 (i), (ii) and (iii) these probabilities are the same and independent for all customers and are independent of the age of the reservation. Let $Z_j(x)$ denote the number of reservations that survive Period j given that there are x reservations on-hand in Period j (so $x - Z_j(x)$ are the number of cancellations in Period j).

The Bellman equation (16.2) for the static model is then modified to account for cancellations as follows

$$V_j(x) = E\left[\max_{0 \leq u \leq D_j} \{r_j y + H_{j-1}(x+u)\}\right], \tag{16.45}$$

with boundary conditions (16.41), where

$$H_{j-1}(x) = E[V_{j-1}(Z_j(x))] = \sum_{z=0}^{x} \binom{x}{z} q_k^z (1 - q_k)^{x-z} V_{j-1}(z)$$

is the expected value function after cancellations in Period j. Again, stochastic convexity arguements show that if $V_{j-1}(z)$ is concave in z, then $H_{j-1}(x)$ is concave in x and hence a modification of the argument in Proposition 1 shows that the value function $V_j(x)$ defined by (16.45) is concave in x.

Again, nested booking limits are optimal with the optimal booking limits given by

$$b_j^* = \min\{x \geq 0 : r_j < H_{j-1}(x) - H_{j-1}(x+1)\}, \quad j = 1, \ldots, n-1,$$

where it is optimal to accept Class j if and only if the number of reservations on hand, x, is strictly less than b_j^*.

Dynamic Model As in the static model, let q_t denote the probability that a reservation in the system at the start of Period t survives to Period $t + 1$, so by Assumption 1 (i), (ii) and (iii) the number of surviving reservations, $Z_t(x)$, is again binomial. Similar to ths static case, the optimality equations for the dynamic model with cancellations become:

$$V_t(x) = E[\max_{u\in\{0,1\}}\{R_t u + H_{t+1}(x+u)\}]. \tag{16.46}$$

where

$$H_{t+1}(x) = E[V_{t+1}(Z_t(x))] = \sum_{z=0}^{x}\binom{x}{z} q_t^z(1-q_t)^{1-z}V_{t+1}(z)$$

is the expected value function after cancellations in Period t. The boundary conditions are given by (16.44).

As a result, it is optimal to accept an arrival of Class j iff

$$r_j \geq H_{t+1}(x) - H_{t+1}(x+1).$$

16.5 Conclusions

Revenue management has come a long was since its birth in the wake of the deregulation of the U.S. airline industry in the 1970's. It is now a highly developed scientific and professional practice in the airline industry. And perhaps more importantly for the future of the field, this airline success has lead to a rapidly growing interest in using revenue management techniques in other industries. Current industry adopters include hotels, car rental companies, shipping companies, television and radio broadcasters, energy transmission companies and apparel retailers. Both industrial practice and research in the field has truly blossomed over the last decade as a result. While the details of RM problems can change significantly from one industry to the next, the focus is always on making better demand decisions – and not manually with guess work and intuition – but rather scientifically with models and technology, all implemented with disciplined processes and systems. This "industrialization" of the entire demand decision-making process is what defines the practice of revenue management today.

We have focused our attention on the core methodology developed for use in the airline industry (and related industries like hotels) over the last 25 years. These and other methods are covered in more depth in Talluri and van Ryzin, 2002, along with a broader range of RM-related problems on dynamic pricing and auctions that are important in other business contexts.

But regardless of the details, the best RM work always strives to balance rigorous science with the complex, real-world concerns of practical implementation. It is this balance of theory and practice that makes revenue management both intellectually challenging and professionally rewarding. It is likely to remain so for many years to come.

Notes

1. As we will show in this chapter, the notation of this sort of *displacement cost* is central to the theory and practice of RM.

2. Though in fact, the use of capacity-restricted discount fares with restrictions predates this American Airlines story. APEX (advanced purchase excursion) fares orginally used in many international markets as early as the 1960's.

3. Though the low-before-high fare assumption is not difficult to relax, as shown by Robinson, 1995.

4. Readers familiar with dynamic programming may notice that this Bellman equation is of the form $E[\max\{\cdot\}]$ and not max $E[\cdot]$ as in many standard texts. The max $E[\cdot]$ form can be recovered by considering the demand D_j to be a state variable along with x. While the two forms can be shown to be equivalent, the $E[\max\{\cdot\}]$ is simpler to work with in many revenue management problems.

5. The assumption of one arrival per period can in fact be relaxed as shown by Lautenbacher and Stidham, 1999.

6. That it increases the protection level about the usual EMSR-b value can be seen by noting that $r_{j+1} = \bar{R}_j P(S_j > y_j)$ in the usual EMSR-b case and $\hat{r}_{j+1} > r_{j+1}$; thus, y_j as to increase to satisfy the equality (16.18).

7. SABRE was American Airlines central reservation system (CRS), subsequently spun off from American's parent, AMR Corp., to become a separate corporation in 1996.

8. Interestingly Cooper, 2000 shows that more frequent reoptimization is not always better; there are cases where reoptimizing the DLP more frequently can actually result in strictly worse revenue performance,

9. Observe that the displacement adjusted revenue could be negative. In this case, Produce j is never accepted on Resource i, and typically we either eliminates Product j from the problem on Resource i or (equivalently) set the displacement adjusted revenue value to zero.

10. As a technical aside, note that one may be tempted to define the Type 2 service level as

$$E\left[\frac{(Z(u)-C)^{+}}{Z(u)}\right],$$

the average fraction denied service, rather than by (16.34). This is wrong, however, because it does not account for the varying number of customers served. For example, if $C = 100$ then it would count a day in which $Z(u) = 1$ and a day in which $Z(u) = 100$ equally as two days with denied service fractions of zero, when in reality the second day represents 100 times as many customers. The renewal-reward theorem leading to (16.34) provides the correct measure of the long-run fraction of customers who are denied service.

11. This follows from stochastic covexity arguments; see Talluri and van Ryzin, 2002 for details.

12. This follows from the fact that if Z is a normal random variable with mean μ and variance σ^2, then

$$E(Z-C)^{+} = \sigma(\phi(z) - z(1-\Phi(z))),$$

where $z = (C-\mu)/\sigma$.

13. Setting k equal to the average group size can be obtained by assuming that all reservations are in groups of exactly size k, in which case with u reservations on hand, there are u/k groups of size k, so the variance of the show demand is $k^2(u/k)q(1-q)$.

14. Note in this case, $V_j(x)$ is a decreasing function of x, since the more reservations we have on hand now, the fewer future opportunities to collect revenue and/or the higher the expected future terminal costs.

References

Algers, S. and Besser, M. (2001). Modelling choice of flight and booking class – a study using stated preference and revealed preference data. *Intl. J. of Services Technology and Management,* 2:28–45.

Andersson, S. E. (1989). Operational planning in airline business – Can science improve efficiency? Experiences from SAS. *European Journal of Operations Research,* 43:3–12.

Beckman, M. J. and Bobkowski, F. (1958). Airline demand: An analysis of some frequency distributions. *Naval Research Quarterly,* 43(5):43–51.

Beckmann, M. J. (1958). Decision and team problems in airline reservations. *Econometrica,* 26:134–145.

Belobaba, P. P. (1987a). *Air Travel Demand and Airline Seat Inventory Management.* PhD thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusettes.

Belobaba, P. P. (1987b). Airline yield management: An overview of seat inventory control. *Transportation Science,* 21:63–73.

Belobaba, P. P. (1989). Application of a probabilistic decision model to airline seat inventory control. *Operations Research,* 37:183–197.

Belobaba, P. P. (1992). Optimal vs. heuristic methods for nested seat allocation. Presentation at ORSA/TIMS Joint National Meeting.

Belobaba, P. P. (1998). PODS results update: Impacts of forecasting on O-D control methods. In *1998 AGIFORS Reservations and Yield Management Study Group Symposium,* Melborne.

Belobaba, P. P. (2001). Revenue and competitive impacts of O-D control: Summary of PODS results. In *First Annual INFORMS Revenue Management Section Meeting,* New York.

Belobaba, P. P. and Lee, S. (2000). PODS update: Large network O-D control results. In *2000 AGIFORS Reservations and Yield Management Study Group Symposium,* New York.

Belobaba, P. P. and Weatherford, L. R. (1996). Comparing decision rules that incorporate customer diversion in perishable asset revenue management situations. *Decision Sciences,* 27:343–363.

Bertsimas, D. J. and Popescu, I. (2001). Revenue management in a dynamic network environment. Working Paper, Massachusettes Institute of Technology.

Bhatia, A. V. and Parekh, S. C. (1973). Optimal allocation of seats by fare. Presentation to AGIFORS Reservations Study Group, Trans World Airlines.

Bierman Jr., H. and Thomas, J. (1975). Airline overbooking strategies and bumping procedures. *Public Policy,* 21:601–606.

Bratu, S. (1999). Network value concept in airline revenue management. Master's thesis, Massachusettes Institute of Technology, Cambridge, Massachusettes. Department of Aeronautics and Astronatics.

Brumelle, S. and Walczak, D. (1997). Dynamic allocation of airline seat inventory with batch arrivals. In *1997 Air Transport Research Group of the WCTR Society Proceedings 3,* Vancouver, Faculty of Commerce and Business Administration, University of British Columbia.

Brumelle, S. L. and McGill, J. I. (1993). Airline seat allocation with multiple nested fare classes. *Operations Research,* 41(1):127–137.

Brumelle, S. L., McGill, J. I., Oum, T. H., Sawaki, K., and Tretheway, M. W. (1990). Allocation of airline seat between stochastically dependent demands. *Transportation Science,* 24:183–192.

Chatwin, R. E. (1993). *Optimal Airline Overbooking.* PhD thesis, Stanford University, Palo Alto.

Chatwin, R. E. (1997). Continuous-time airline overbooking with time dependent fares and refunds. Working paper, Applied Decision Analysis, Inc., Menlo Park, California.

Chatwin, R. E. (1999). Mutiperiod airline overbooking with a single fare class. *Operations Research,* 46:805–819.

Civil Aeronautics Board (1967). *Civil Aeronautics Board Economic Regulations Docket 16563.* Washington, D.C.

Cooper, W. L. (2000). Asymptotic behavior of some revenue management policies. Working Paper, Univ. of Minnesota.

Cross, R. G. (1997). *Revenue Management: Hardcore Tactics for Market Domination.* Broadway Books (Bantam, Doubleday, Dell Publishing Group), New York.

Curry, R. E. (1990). Optimal airline seat allocation with fare classes nested by origins and destinations. *Transportation Science,* 24:193–204.

Curry, R. E. (1992). Real-time revenue management: Bid price strategies for origins/destinations and legs. *Scorecard, Aeronomics Inc., Atlanta,* 2Q:n.a.

Dror, M., Trudeau, P., and Ladany, S. P. (1988). Network models for seat allocation on flights. *Transportation Research,* 22B:239–250.

D'Sylva, E. (1982). O-and-D seat assignment to maximize expected revenue. Technical report, Boeing Commercial Airplane Company, Seattle, Washington. Unpublished internal technical report.

Glover, F., Glover, R., Lorenzo, J., and McMillan, C. (1982). The passenger mix problem in the scheduled airlines. *Interfaces,* 12:73–79.

Kaplan, A. (1969). Stock rationing. *Management Science,* 15:260–267.

Kleywegt, A. J. and Papastavrou, J. D. (1998). The dynamic and stochastic knapsack problem. *Operations Research,* 46:17–35.

Lautenbacher, C. J. and Stidham, S. J. (1999). The underlying markov decision process in the single-leg airline yield management problem. *Transportation Science,* 34:136–146.

Lee, T. C. and Hersh, M. (1993). A model for dynamic airline seat inventory control with multiple seat bookings. *Transportation Science,* 27:252–265.

Liang, Y. (1999). Solution to the continuous time dynamic yield management model. *Transportation Science,* 33:117–123.

Littlewood, K. (1972). Forecasting and control of passenger bookings. In *Proceedings of the Twelfth Annual AGIFORS Symposium,* Nathanya, Israel.

Martinez, R. and Sanchez, M. (1970). Automatic booking level control. In *Proceedings of the Tenth AGIFORS Symposium.*

McGill, J. I. and van Ryzin, G. J. (1999). Revenue management: Research overview and prospects. *Transportation Science,* 33(2):233–256.

Phillips, R. L. (1994). A marginal value approach to airline origin and destination revenue management. In Henry, J. and Yvon, P., editors, *Proceedings of the 16th Conference on System Modeling and Optimization,* New York. Springer-Verlag.

Robinson, L. W. (1995). Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Operations Research,* 43:252–263.

Rothstein, M. (1971). Airline overbooking: The state of the art. *J. Trans. Econ. & Policy,* 5:96–99.

Rothstein, M. (1975). Airline overbooking: Fresh approaches are needed. *Transportation Science,* 2:169–173.

Rothstein, M. (1985). O.R. and the airline overbooking problem. *Operations Research,* 33:237–248.

Rothstein, M. and Stone, A. W. (1967). Passenger booking levels. In *Proceedings of the Seventh AGIFORS Symposium.*

Shlifer, E. and Vardi, Y. (1975). An airline overbooking policy. *Transportation Science,* 9:101–114.

Simon, J. L. (1968). An almost practical solution to airline overbooking. *J. Trans. Econ. & Policy,* 2:201–202.

Simon, J. L. (1972). Airline overbooking: The state of the art – a reply. J. *Trans. Econ. & Policy,* 6:255–256.

Simon, J. L. (1993). The airline oversales auction plan: How it was adopted and how it has fared. In *Fifth IATA Revenue Management Conference,* Montreal.

Simon, J. L. (1994). The airline oversales auction plan: The results. *J. Trans. Econ. & Policy,* 28:319–323.

Simpson, R. W. (1989). Using network flow techniques to find shadow prices for market and seat inventory control. Technical Report Memorandum, M89-1, MIT Flight Transportation Laboratory, Cambridge, Massachusettes.

Smith, B. C. and Penn, C. W. (1988). Analysis of alternative origin-destination control strategies. In *Proceedings of the Twenty Eighth Annual AGIFORS Symposium,* New Seabury, Massachusettes.

Smith, B. C., Leimkuhler, J. F., and Darrow, R. M. (1992). Yield management at american airlines. *Interfaces,* 22:8–31.

Subramanian, J., Stidham Jr., S., and Lautenbacher, C. (1999). Airline yield management with overbooking, cancellations and no-shows. *Transportation Science,* 33:147–167.

Talluri, K. T. and van Ryzin, G. J. (1999a). An analysis of bid-price controls for network revenue management. *Management Science,* 44:1577–1593.

Talluri, K. T. and van Ryzin, G. J. (1999b). A randomized linear programming method for computing network bid prices. *Transportation Science,* 33:207–216.

Talluri, K. T. and van Ryzin, G. J. (2001). Revenue management under a general discrete choice model of consumer behavior. Working paper, Graduate School of Business, Columbia University.

Talluri, K. T. and van Ryzin, G. J. (2002). *The Theory and Practice of Revenue Management.* Kluwer Academic Publishers, Dordrecht, The Netherlands. To be published.

Taylor, C. J. (1962). The determination of passenger booking levels. In *Proceedings of the Second AGIFORS Symposium.*

Thompson, H.R. (1961). Statistical problems in airline reservation control. *Oper. Res. Quart.,* 12:167–185.

Topkis, D. M. (1968). Optimal ordering and rationing policies in a nonstationary dynamic inventory model with *n* demand classes. *Management Science,* 15:160–176.

van Ryzin, G. J. and McGill, J. I. (2000). Revenue management without forecasting or optimization: An adaptive algorithm for determining seat protection levels. *Management Science,* 46:760–775.

Van Slyke, R. and Young, Y. (2000). Finite horizon stochastic knapsacks with applications to yield management. *Operations Research,* 48:155–172.

Vinod, B. (1989). A set partitioning algorithm for virtual nesting indexing using dynamic programming. Technical report, Internal Technical Report, SABRE Decision Technologies.

Vinod, B. (1995). Origin-and-destination yield management. In Jenkins, D., editor, *The Handbook of Airline Economics,* pages 459–468. The Aviation Weekly Group of the McGraw-Hill Companies, New York.

Williamson, E. L. (1988). Comparison of optimization techniques for origin-destination seat inventory control. Master's thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusettes.

Williamson, E. L. (1992). *Airline Network Seat Inventory Control: Methodologies and Revenue Impacts*. PhD thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusettes.

Wollmer, R. D. (1986). A hub-spoke seat management model. Unpublished company report, Douglas Aircraft Company, McDonnell Douglas Corporation.

Wollmer, R. D. (1992). An airline seat management model for a single leg route when lower fare classes book first. *Operations Research,* 40:26–37.

Wong, J. T. (1990). *AirlineNetworkSeatAllocation.* PhD thesis, Northwestern University, Evanston, Ill.

Wong, J. T., Koppelman, F. S., and Daskin, M. S. (1993). Flexible assignment approach to itinerary seat allocation. *Transportation Research,* 27B:33–48.

This page intentionally left blank

17 SPATIAL INTERACTION MODELING

Piet Rietveld
Peter Nijkamp

17.1 Transport Systems Analysis: A Portrait

The ancient Greek philosopher Heraclitos is often quoted for his historical remark that “everything is in motion, except the motion process itself”. Most likely he did not refer to our modern, highly mobile society, but in any case he recognized the phenomenon of perpetual motion. Our age is the age of mobility, in the sense of intensive geographical movement of people, goods, and also information. The past centuries have witnessed an uninterrupted growth in distance traveled. Economic historian Charles van Doren (1992) has made some simple, but intriguing observations on the past and future mobility pattern of modern man. According to his estimates, the average distance that could conveniently be traveled in the year 1800 was 12 miles a day. This figure had already risen to 60 miles a day in 1900, and 300 miles in 2000. Extrapolation of these figures would lead to the expectation that by the year 2100 the average daily distance that could conveniently be bridged would be 1500 miles and by the year 2200 even 7500 miles. Despite some speculation, it ought to be recognized that the growth rate of daily mobility is a surprising phenomenon. The ‘homo mobilis’ has become a dominant species on earth.

It is evident that the growth in mobility is not an autonomous factor, but is dependent on several background factors, such as the rise in income and welfare, the rise in leisure time, demographic developments etc. It is indeed surprising that many countries exhibit more or less the same mobility pattern (see for a presentation of many statistical details Salomon et al. 1993).

It should be added that in most cases transport behavior is not an independent phenomenon with its own intrinsic value, but is dependent on other motives (e.g., work, shopping, recreation etc.). Transport is often seen as a derived demand; the underlying motives originate form other societal or individual objectives. Consequently, the intricate pattern of spatial transport flows finds its origins in human choices and decisions of various kind outside the physical transport system. This observation has also laid the foundation for the so-called activity-based modeling in transportation research.

A major determinant of the growth in transport flows stems from changes in economic structures or in relocation of people or firms. A substantial change in accessibility of cities or regions will have a decisive influence on the size and distribution of transport flows. On the other hand, transport is an input factor in the economic process and, therefore, any change in the costs or productive contribution of transport has a clear impact on the economy, not only national, but also regional or local. Thus, changes in the efficiency of transport systems do not only affect welfare positions of countries or regions, but influence also the relative competitive position of these countries or regions. National-economic and regional-economic development on the one hand, and efficiency changes in the transport system and resulting distributional flow changes on the other hand, are two sides of the same coin. This is also witnessed in the income elasticity of transport (both passenger traffic measured as person-km and freight traffic measured as tonne-km), which in many countries is rather stable and fluctuates around 1. In a broader context, the impact of transport investments on economic growth has been discussed by Aschauer (1989).

At the same time, transportation has become a source of much concern because of its great many externalities, such as decay in quality of life, climate change, congestion, landscape segmentation, and decline in safety. The concept of sustainable transport has gained much popularity in recent years (see Nijkamp et al. 1999) and it has penetrated increasingly policy analysis of transport initiatives, but it has not yet sufficiently been incorporated in transportation modeling in relation to land use (for a review see Hayashi and Roy, 1996).

It is no surprise that in a world dominated by mobility a vast array of transport systems modeling has been developed. These models range from simple single-equation models (e.g., price or income elasticities of transport) to large-scale spatial equilibrium models (see e.g. Van den Bergh et al. 1996). We have also observed the emergence of an array of network models (see Nagurney and Siokos 1997). With the emergence of network models we have also witnessed an increasing use of communication and information models (see e.g., Batten et al. 1995), as well as of models addressing the impact of telematics or intelligent transport systems (ITS) on transportation behavior (see Nijkamp et al. 1996). Network models do not only address the user side, but also the supply impacts of infrastructure (including quality-upgrading devices such as information and communication technology).

An important but often neglected element in network modeling is the presence of externalities. Such externalities may refer to non-market negative externalities such as unpriced environmental effects (see for a review Verhoef 1996), but also to broader positive effects such as club externalities (see e.g. Capello 1994).

Another underrepresented element in network modeling is the dynamics of a spatial and transportation system (see e.g. Nijkamp and Reggiani 1992, 1998). This issue has prompted inter alia research in synergetic behavior, chaos phenomena, catastrophy theory and self-organization. The elegance of such approaches is without any doubt, but the empirical verification and application is still the Achilles heel of such dynamic spatial and transportation models. More recently, this has also led to new forms of network modeling, in particular the use of neural network modeling and genetic algorithms (see e.g. Himanen et al. 1998).

A particular class of transport systems models that has received a prominent place in literature is the family of spatial interaction models. This family comprises a vast array of members. We will present here the most common types of spatial interaction models, followed by a exploration of complementary approaches which might make the use of such models more oriented towards real-world transport issues.

17.2 The Four Stage Model: A Comprehensive Approach.

In this review we will pay special attention to the four-stage transportation model. The reason is not that this model is the ideal standard, but because it has been applied frequently for various policy purposes and because there is not yet an operational alternative for it. The basic components of the four-stage model are:

- Trip generation and attraction (determining the total number of trips attracted by zones of origin and destination)
- Trip distribution (determining the travel flows between zones of origin and destination)
- Modal choice (analysis of the choice of transport mode for combinations of origin and destination
- Assignment (choice of route).

The four-stage model is usually applied at the level of metropolitan areas where some 100-1000 or more zones are distinguished. The primary aim of the model at the time that it was developed was to predict the effects of changes in transport networks (for example the construction of a new express way) on transport flows, and this is still a main aim for its use. We note in passing that in addition to models developed for metropolitan areas also a series of national models have been developed for a sample of European countries with model features that are close to those of the metropolitan areas (see Lundqvist and Mattsson, 2001, for a review).

A simple version of the model is illustrated in Figure 17.1. According to this figure the model has a recursive structure starting from generation and attraction to assignment. In terms of demand and supply factors the first three stages relate to *demand.* The demand for trips is essentially analysed by means of a *conditional* model: the number of trips per origin-destination pair is conditional on the total number of trips per origin generated and a similar structure holds for the number of trips per origin-destination pair that use a certain transport mode. *Supply* factors are included in the model via the various costs (in terms of out-of-pocket costs and time) imposed on users of the network. *Confrontation of supply and demand* takes place in the last phase (the assignment model). When capacity on road links is not sufficient to achieve free flow speeds the time costs of the use of the links will be higher than the free flow travel times and this will induce road users to look for other routes. This is represented in the model by means of the feedback from assignment to earlier phases such as the choice of destination of trips.

The basic model obviously focuses on the demand side of transport markets. The supply side in terms of the provision of transport infrastructure and transport services (public transport) is assumed to be given. This makes the model suitable for policy analyses in terms of 'what if' questions like: how does transport demand respond to measures such as the upgrading of a road, or the increase of frequencies of services. In case the private sector would have a strong influence on these variables, the model would need to be extended in such a way that responses from the supply side would be incorporated. The fact that such additions are not usually implemented reveals that the model has a mainly short-term nature.

An extended version of the model is presented in the other part of Figure 17.1. This model version is more advanced in two respects: it includes a time-of day element and in addition it gives a more refined treatment of choice of trip destination and mode choice. The time of day element is an important addition when transport networks are rather congested and part of the users are flexible in their choice of time to travel. The integrated analysis of the choice of destination and transport mode implies that rigid assumptions on the sequence in which these choices are made are avoided. These two versions of the model indicate that a considerable variety of applications exist. Details of these variations will be discussed in the next sections.

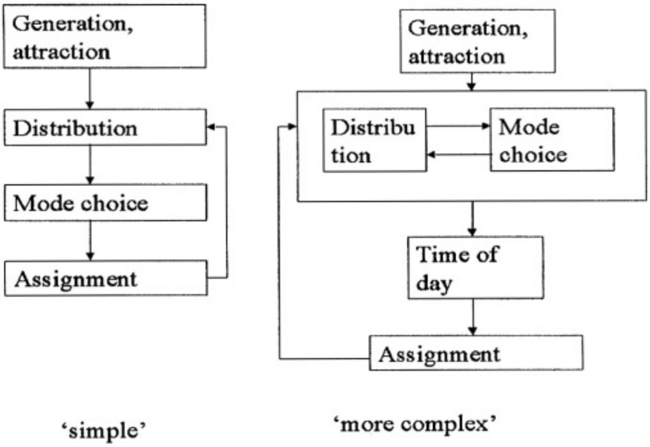


Figure 17.1 Components and Links of Four-Stage Transport Model

The history of the four-stage model has recently been described by Bates (2000) and McNally (2000). Early versions of the model were developed in the USA in the 1950's. Further developments took place in the decades thereafter. The model has been gradually improved and it has become popular all over the world.

At the same time the model has been broadly criticised for its limitations (see for example McNally, 2000). A major point of criticism concerns the assumption that *total travel demand is inelastic*; this is the consequence of using the total number of trips per zone as the starting point of the analysis. Changes in transport costs are assumed not to affect the total number of trips; they only affect choices of destination, mode and route. A related point is that the model is *trip-based* rather than activity based. Thus, the model ignores the spatial temporal choices of people who have to organise their various activities at home and elsewhere. As a consequence complex patterns of timing of home and non-home based trips jointly with the activities themselves emerge that are not dealt with in trip based models in an adequate way (see for example Bhat and Koppelman, Chapter 3). A more specific consequence of this limitation is that the analysis of the *timing of trips* is problematic in trip based models. It is exactly this timing issue that is essential in so many policy applications where congestion plays a role. The introduction of the choice of time of day (peak versus off-peak) is a first step to come to meet this criticism, but it does not go far enough to do justice to the complexities of behavioural choices involved and the functioning of transport networks in congested periods. Another limitation of the four-stage model is the lack of feed back from the transport system on the spatial location of activities. Especially the choice of location of residence by households and firms is assumed to be given in the standard model so that it can be interpreted as a short-term model. Extensions of the basic four-stage model towards integrated transport-land use models are discussed in Section 17.7. Other limitations of early versions of the four stage model concern the lack of feed-back between the various stages and the lack of individual choice elements in it. However, these limitations have in the course of time been overcome by introducing feed backs between the stages and by the introduction of results of discrete choice theory in these models (Chapter 2).

In the next sections we will discuss the components of the four-stage model with the exception of the fourth stage, the assignment model, which is already covered by Chapter 11 in this handbook.

17.3 Trip Generation.

Trip production and attraction are the corner stones of transport modelling. Trips are usually distinguished into home based and non-home based trips. In the case of *home based* trips, the home of the trip maker is either the origin or the destination of the trip. The remaining trips (usually some 25% of the total number of trips) are called *non-home based.* Based on this distinction, the concept of *trip production* can be outlined; with home based trips, trip production relates to the home end of home based trips; in the case of non-home based trips it relates to the origin of the trips. Trip attraction can be defined in a similar way: with home based trips, trip attraction relates to the non-home end of the trips; for non-home based trips it relates to the destination of the trips. It is important to note that data on home based trips are usually better than on non-home based trips. Also, data on the home end of the trip are usually better than data on the activity end of a trip. Hence, research on trip production tends to generate more reliable results than research on trip attraction. Since the total numbers of produced and attracted trips are equal, the usual approach is to use the total number of produced trips as a control total for the total number of attracted trips (Ortuzar and Willumsen, 2001).

In principle, transport models deal with all trips going through part of a certain metropolitan area. This means that not only the trips made by all residents have to be considered, but also visits from residents living outside the metropolitan area. In addition, through traffic has to be considered. In practice the two latter categories tend to be treated in a less refined way than the trips made by residents of the area. As with the trips made by the residents of the area, some underdeveloped or ignored segments can be discerned. First, there is often a lower age limit implying that trips made by children below a certain age are not covered. In addition, it often appears that short trips are ignored or incompletely covered. The background of the latter problem is that most of these short trips will be realised within one zone so that they do not play a large role in the multi-zonal transport models that focus on interzonal trips (Rietveld, 2000). Comparison of trip data in various countries reveals that there are substantial differences between countries in trip production rates (see for example Salomon et al., 1993). A substantial part of these differences can be explained by differences in efforts to include short distance trips. The importance of including short distance trips obviously depends on the aim of the analysis. Especially when one is interested in the role of non-motorised transport in transport systems, or in issues of traffic safety (pedestrians being an important group of victims) short distance trips should not be ignored. However, even when the focus would be on much more general themes where non-motorised transport does not seem to play an important role, the issue of underreporting of short distance trips deserves attention. The reason is that one may expect that there will be a certain degree of substitution between short distance trips and long distance trips. Therefore, under-representation of short distance trips makes the total number of trips more elastic for changes in the transport system. Thus, the usual assumption that trip production is inelastic in a zone is more difficult to defend when short trips are not well covered.

A final consideration is that in addition to passenger transport also freight transport should be included, at least when one wants to take into account road congestion since trucks and private cars make use of the same infrastructure. Given the longer trip distances of lorries compared with private cars, the share of lorries tends to be higher on express ways than on other road types. With shares of up to 30% of all vehicles on certain express ways, the role of lorries cannot be ignored, especially since per vehicle they use more road capacity so that their contribution to congestion is above average. There is an extensive literature on freight transport modelling (see for example Chapters 12 and 13, d' Este, 2000) but at the level of multizonal metropolitan modelling freight models have not reached the same level of spread and sophistication compared with personal transport models. One of the reasons is that freight transport is part of complex logistical chains where production, warehousing and distribution activities take place in an integrated manner. Therefore, there is not just origin-destination traffic, which is relatively easy to model, but also multi-stop trips of delivery vans have to be considered that are involved in pick-up and delivery operations.

A standard element in all transport models is the distinction between trip purposes. Purposes usually covered in transport models are work, social & recreation, shopping, education, business and others (see also Salomon, Bovy and Orfeuil, 1993). Table 17.1 gives a view on the trip frequencies of these purposes for The Netherlands. The distinction according to purposes is important because for some of them behavioral choices such as changes of destination or changes in terms of departure time are much more flexible than for others. Compare for example work at a fixed location and given working times with shopping at a shopping centre of choice at some time of the day).

Table 17.1 gives an overview of trip rates for several purposes for a selection of time slots. It demonstrates that the morning peak is less peaked than the afternoon peak. And it also appears that trips rates between 10.00 and 12.00 are higher than between 7.00 and 9.00. The morning peak stands out because of its high share of work and education. Both activities usually imply high scheduling costs. Shopping is the most frequently mentioned purpose during the afternoon peak. Social visits and recreation dominate the evening travel pattern. Thus from the trip frequencies themselves one cannot infer that express way congestion is high between 7.00 and 9.00. The reason is that congestion only becomes visible when also the elements of trip distribution, modal choice and route choice are taken into account.

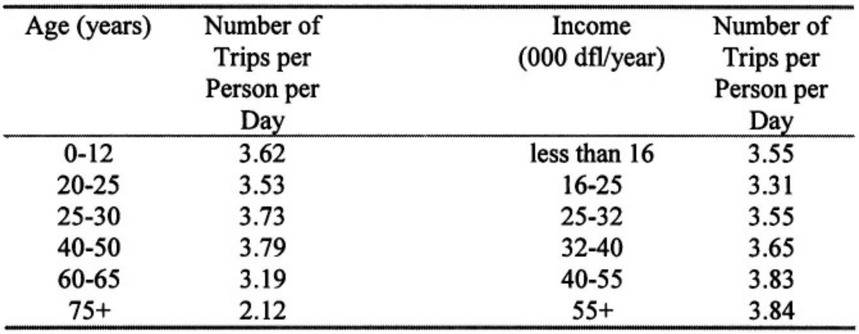
	Time Period			
	7.00-9.00	10.00-12.00	16.00-18.00	18.00-20.00
Work	16	3	11	3
Business	1	2	2	0
Social Visit	1	6	10	8
Shopping	2	16	14	5
Education	10	2	1	0
Recreation	1	2	8	9
Others	9	12	12	8
Total	41	44	57	33

Source: CBS

Table 17.1 Trip Frequencies (Number of Trips per 100 Persons per Day) for Some Time Slots According to Purpose (The Netherlands, 1999)

Trip frequencies depend on a good number of individual features such as age, income, educational level and car ownership. We list some results in Table 17.2.

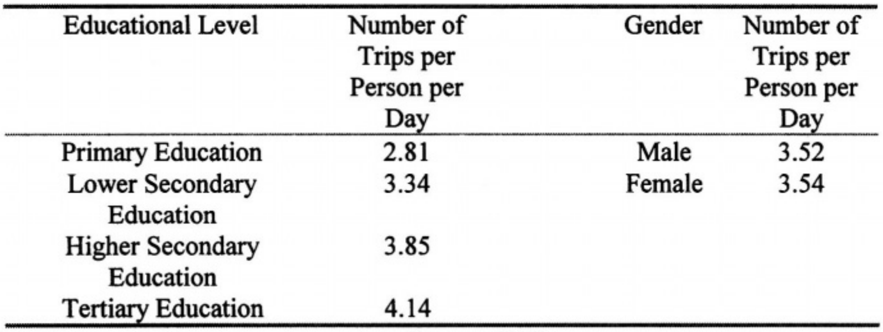
The table shows that trip frequencies are rather insensitive to age for most age groups. Only after the age of 50 years a strong decline can be noted. There appears to be positive relation between income and trip frequency. The only exception is that the lowest income group has high frequencies. Table 17.3 shows that educational level appears to be an important determinant of trip frequencies: people with high educational attainment travel much more than people with little education. The gender difference appears to have remarkably little impact on the average trip frequencies.



Age (years)	Number of Trips per Person per Day	Income (000 dfl/year)	Number of Trips per Person per Day
0-12	3.62	less than 16	3.55
20-25	3.53	16-25	3.31
25-30	3.73	25-32	3.55
40-50	3.79	32-40	3.65
60-65	3.19	40-55	3.83
75+	2.12	55+	3.84

Source: CBS

Table 17.2 Number of Trips Per Person Per Day for Income Groups and a Selection of Age Groups (The Netherlands, 1999).



Educational Level	Number of Trips per Person per Day	Gender	Number of Trips per Person per Day
Primary Education	2.81	Male	3.52
Lower Secondary Education	3.34	Female	3.54
Higher Secondary Education	3.85		
Tertiary Education	4.14		

Source: CBS

Table 17.3 Number of Trips per Person per Day for Educational Levels and Gender Groups (The Netherlands, 1999).

It should be noted that the above figures only relate to the total number of trips so that trip purposes are ignored. For specific trip purposes of course much larger differences among the various groups can be observed. A second observation is that the above results ignore interrelationships between features of people. Since education and income are correlated, part of the positive correlation between income and trip frequency may stem from the education effect. This calls for the estimation of propensities to make trips based on micro-data. Combination of such propensities with detailed data on the composition of zones would allow one to predict trip production per zone. This means essentially that rather detailed data are needed on the composition of population per zone according to various classes. Ortuzar and Willumsen (2001) discuss a number of issues in this respect, such as the problem that one may easily end up with untractable numbers of classes. Another point of interest is that for practical applications one would need updating procedures when detailed information on trip frequencies in the past has to be combined with more recent, but incomplete information. Another issue in trip production models is whether classifications should be based on individual features (as presented above) or on household features.

An important point of discussion is whether the number of trips made depends on the location of the zone. As Table 17.4 shows there are only small differences between zones in terms of their degree of urbanisation. One might expect that since in highly urbanised areas accessibility of various services is high so that the number of trips made is also higher. This indeed appears to be the case with shopping, sport and recreation. The pattern for shopping suggests that in non urban areas people plan their shopping activities in a more careful way so that they go to shops less frequently. A similar pattern seems to exist for sport and recreation. However, the total number of trips does not reveal a consistent pattern. The background is of course that the populations in these types of areas will be rather different. For example the figures on work related trips indicate that participation on the labour market is smaller in highly urbanised areas (probably due to demographic factors) so that the number of work trips is smaller there. As Ortuzar and Willumsen (2001) indicate the evidence of an impact of accessibility on trip rates is rather weak. The figures in the table suggest that for some trip purposes such an impact is indeed relevant, but that it is difficult to trace for the aggregate of all trips.

The above discussion was focussed on trip production. The literature on trip attraction is much smaller since trip attraction data are less accurate. Nevertheless the issue of trip attraction is important, not only for the purpose of modelling per se, but also for many practical and policy purposes. For example, for the development of shopping areas or the planning of schools, the issue of trip attraction is of utmost importance. It is plausible that the attraction of zones for education or shopping trips depends to a much stronger extent on accessibility than the production of zones for these trips. This is an issue that will be further discussed in the context of the next section, which is on trip distribution.

Degree of Urbanisation	Number of Trips per Day: Work	Number of Trips per Day: Shopping	Number of Trips per Day: Sport and Recreation	Number of Trips per Day: Total
very strongly urban	0.51	0.89	0.47	3.44
strongly urban	0.53	0.88	0.44	3.57
moderately urban	0.53	0.83	0.44	3.61
slightly urban	0.54	0.77	0.44	3.61
non urban	0.54	0.71	0.42	3.47

Source: CBS (2000)

Table 17.4 Trip Frequencies per Person per Day for a Selection of Trip Purposes, Distinguished According to Degrees of Urbanisation (The Netherlands, 1999).

17.4 Trip Distribution.

Trip distribution concerns the volumes of flows between zones of origins and of destination. In trip distribution models the flow between two zones depends on:

- Properties of the zone of origin i (X_i, for example, population size)
- Properties of the zone of destination j (Z_j, for example employment)
- Transport cost indicator between zones i and j (c_{ij}, for example distance, travel time, monetary costs or generalised costs).

A possible formula for the flow T_{ij} between zones i and j would be:

$$T_{ij} = f(X_i)\, g(Z_j)\, h(c_{ij}) \tag{17.1}$$

Note that this formulation does not include constraints on the total number of trips per zone of origin or destination. Therefore it is called the *unconstrained model.* For $f(X_i)$ and $g(Z_j)$ the usual formulations are $X_i^{\ a}$ and $Z_j^{\ b}$. Thus, a and b can be interpreted as elasticities: for example, when X_i increases with 1%, the flow between i and j increases with a %. The function $h(c_{ij})$ represents the impact of transport costs on spatial interaction. It is known as the deterrence function and indicates the sensitivity of spatial interaction to transport costs.

Form of the Deterrence Function.

For the deterrence function common forms are the exponential form and the power form:

$$h(c_{ij}) = \exp(-\beta c_{ij}) \tag{17.2}$$

or

$$h(c_{ij}) = c_{ij}^{-\gamma} \tag{17.3}$$

The parameters β and γ have different interpretations (see for example Fotheringham and O'Kelly, 1989). In the case of the power function, γ can be interpreted as an elasticity: when the costs of interaction between i and j increase with 1 %, the flow between the two zones will decrease with γ%. Thus, the power function implies a constant relationship between relative changes in costs and relative changes in flows. With the exponential form, the elasticity is not constant because the relative change in flows is related to the absolute change in costs, not to the relative change. Thus the relative effect of a cost increase of say $ 1 of a trip does not depend on the current level of the trip in this case. The two functions have been illustrated in Figure 17.2.

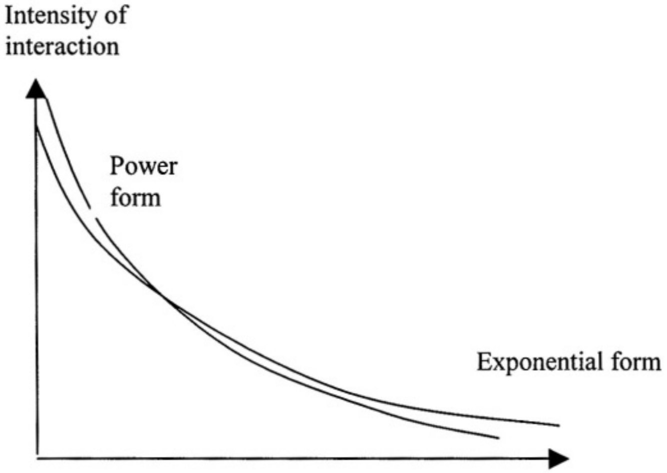


Figure 17.2 Illustration of Power Form and Exponential Form of Deterrence Function.

A notable difference between the two functional forms is that the power function goes to infinity for very low levels of costs, whereas the exponential form remains finite. The consequence is that intra-zonal transport may become difficult to deal with the power function because outcomes become very sensitive to small errors in transport costs within a zone (usually little is known about intra-zonal transport costs). Similarly, also flows between nearby zones may become sensitive for errors in cost estimates. The exponential form does not lead to these problems. For very high costs a similar but reversed problem occurs. Both formulas imply that flows tend to zero when costs become very high, but in this cost range the exponential formula is much more sensitive to changes in costs than the power formula. Because of these reasons Fortheringham and O'Kelly (1989) indicate that as a rule of thumb the power formula is most suitable for spatial interaction where long distances dominate (for example aviation or migration), whereas the exponential formula is more suitable for short distance spatial interactions (for example shopping behaviour). This means also that in the context of modeling metropolitan transport the exponential form will most probably be the more attractive candidate.

The power function specification has the feature that γ is independent from the unit ofmeasurement of c_{ij}. In the exponential form this is no longer true: for example, when costs would be measured in terms of dollars in stead of in cents, this would entail that the β parameter has to be multiplied with 100. This means that one cannot automatically transfer the values of β from one context to the other.

Despite these differences between the two formulas it should be noted, however, that they are more closely linked than is often thought. The point is that in the case of heterogeneity when individuals vary according to the β parameter of the exponential function, if this parameter has a Gamma-distribution, the resulting aggregate deterrence function can be approximated with a power function (See Choukroun, 1975 and Fotheringham and O'Kelly, 1989).

In the literature several ways can be detected to use more general forms for the deterrence function. A first possibility would be to formulate flexible forms such as the following step function with a constant e_k for each step k:

$$h(c_{ij}) = e_k \text{ for all } c_{ij} \text{ in a range k defined by } l_k < c_{ij} < l_{k+1} \qquad (17.4)$$

This obviously yields a very flexible form since the number of parameters equals the number of travel cost classes distinguished. On the other hand when these classes are very narrow empirical applications may lead to improbable (for example non-monotone decreasing) patterns of the estimated parameters e_k.

Another solution sometimes proposed is to use a combination of the power form and the exponential form:

$$h(c_{ij}) = \exp(-\beta c_{ij})\ c_{ij}^{-\gamma} \qquad (17.5)$$

This is of course a more flexible form than the two standard forms, but it does not yield a solution for the tendency that the power function grows to infinity for low transport costs and that the exponential function tends to be sensitive for small cost changes with high transport costs since the product of the two functions will take on board both problems. A more promising form is the use of the Box-Cox transformation (see Box and Cox, 1964):

$$\begin{aligned} c_{ij}^{(b)} &= (c_{ij}^{b}-1)/b && \text{when } b \neq 0 \\ c_{ij}^{(0)} &= \ln c_{ij} && \text{when } b=0 \end{aligned} \qquad (17.6)$$

Thus, by using this special form a variable can be shown to vary between a linear (b=l) and a logarithmic (b=0) specification. Note that with the exponential form of the deterrence function as used above in $h(c_{ij}) = \exp(-\beta c_{ij}^{(b)})$ the value b=0 implies the power form, whereas b=l leads to the exponential form. Thus, by introducing the parameter b the exponential and the power forms become special cases of a more general set of functional forms (see Table 17.5).

Box-Cox Parameter b in $h(c_{ii}) = \exp(-\beta c_{ii}^{(b)})$	**Form of Deterrence Function:**
b=0	**Exponential Form**
0<b<1	**Intermediate Case**
b=1	**Power form**

Table 17.5 Form of Deterrence Function Depending on Box-Cox Parameter.

As demonstrated by Fik and Mulligan (1998) and Ortuzar and Willumsen (2001) the use of the Box-Cox transformation often shows that the parameters a and b are significantly different from 0 and 1. As a result the values of the β parameters may be strongly affected. The obvious reason why the Box-Cox specification often remains unused is that it leads to intrinsically non-linear forms so that as a result OLS approaches cannot be used. Maximum likelihood estimation is the approach to arrive at estimations of the model parameters in this case.

Double, Single and Unconstrained Models

After this discussion of the functional form of the deterrence function, we address another issue in the distribution model, i.e. the link between the flow volumes T_{ij} and the stock volumes X_i and Z_j. Consider the form

$$T_{ij} = X_i^{a} Z_j^{b} h(c_{ij}). \tag{17.7}$$

In these forms one often observes values for a and b that are close to 1. Then, a uniform increase of 1% in the X and Z variables (for example population, income, number of workplaces, etc) will lead to the conclusion that the flows will increase by about 2%. This is an implausible conclusion since it would imply that as metropolitan areas grow, traffic would grow twice as fast, which should be considered as extreme. Therefore there is a need for alternative specifications of the distribution function that do not give rise to such implausible results. The standard way to do so is to provide a link between the trip distribution and trip generation models in the following way. Let O_i and D_j denote the total number of trips leaving and entering zones i and j[1]. Then the distribution function can be formulated as:

$$T_{ij} = O_i A_i D_j B_j h(c_{ij}) \tag{17.8}$$

Where

$$\begin{aligned} \Sigma_i T_{ij} &= D_j \\ \Sigma_j T_{ij} &= O_i \end{aligned} \tag{17.9}$$

The factors A_i and B_j have been added to ensure that the definitional requirements are met that all trips leaving and entering a certain a zone add up to their totals determined in the trip generation model. They are usually interpreted as balancing factors. After rewriting the following expressions can be obtained for the balancing factors:

$$\begin{aligned} A_i^{-1} &= \Sigma_j B_j D_j h(c_{ij}) \\ B_j^{-1} &= \Sigma_i A_i O_i h(c_{ij}) \end{aligned} \tag{17.10}$$

Thus it appears to be impossible to find explicit solutions for the balancing factors. Iterative procedures have to be used to arrive at values for A_i and B_j. An interpretation that is often used for these factors is that A_i^{-1} and B_j^{-1} are interpreted as accessibility indicators. See for reviews Rietveld and Bruinsma (1998) and Reggiani (1998). Note that in the present model an increase in O_i and D_j with 1% leads to an increase in the flows of 1%. Thus by the introduction of the balancing factors the problem that flows grow systematically faster than stocks has been removed. The above formulation is known as the *doubly constrained model.* An obvious underlying assumption of this model is that the trip generation model can be used to produce reliable estimates of the total number of trips entering and leaving the respective zones. Typical examples of trip purposes for which the doubly constrained model may apply are social visits and work trips (when the number of jobs in zones is given). We note in passing that whereas in most applications the origin zones i and destination zones j will be of the same type, this is not necessarily the case. For example it may well be that residential zones are different from employment zones; a similar case holds for hospital visits.

There may be cases where it is not so evident that for both origins and destinations the total number of trips can be imposed beforehand. Consider for example shopping trips. Whether or not a zone will attract many trips of this type will depend on its accessibility. Therefore for shopping one may expect that the *production constrained model* makes sense. This means that

$$T_{ij} = A_i \, O_i \, h(c_{ij}) \tag{17.11}$$

where

$$A_i^{-1} = \Sigma_j \; h(c_{ij}) \tag{17.12}$$

This formulation implies that the total number of visits to shops in zone j just follows as Σ_i T_{ij} after the model has been solved to arrive at the flows. The model predicts that in a zone j with a large number of nearby residences (high O_i and low transport costs c_{ij}) there will be many shopping trips[2]. The impact of accessibility A_i^{-1} of the zones of origin means that nearby residential zones with a high level of accessibility will have a lower impact on shopping visits in j. The reason is that residents in these highly accessible zones apparently have ample opportunity to visit other shopping zones at short distances. We note in passing that Table 17.4 in the preceding section suggests that also the total number of trips made by the residents in a zone is not exogenous but depends on accessibility. In Section 17.6 we will discuss other formulations of the spatial interaction models that have this property.

Similar to production constrained models, there may also be *attraction constrained models.* In this case the model reads:

$$T_{ij} = B_j \, D_j \, h(c_{ij}) \tag{17.13}$$

Where

$$B_j^{-1} = \Sigma_i \, h(c_{ij}) \tag{17.14}$$

A possible example of an attraction constrained model is the case of modeling the location of residence of students who are studying at particular educational institutions with limited capacity, or of the choice of hospital by patients[3].

Estimation of the unconstrained model by means of ordinary least squares (OLS) after logarithmic transformation of the variables is straightforward. However, in the case of the constrained models a more complex estimation results. As shown by Cesario (1975) OLS can still be used in the production or attraction constrained model, but there is a need to add origin (or destination) specific constant terms. Concerning the doubly constrained model Sen and Soot (1981) have solved the problem of separating the estimation of the parameter of the distance deterrence function and the computation of the balancing factors by using $(T_{ij}/T_{ii}).(T_{ji}/T_{jj})$ as the starting point of the estimation procedure. An alternative approach is that the interaction models are estimated by means of Maximum Likelihood methods. This has as an additional advantage that the problem of zero flows disappears. The OLS regressions rely on logarithmic transformations and therefore zero flows cause difficulties. Especially when there are many zones and many groups of travellers, the probability of zero flows will be considerable. When maximum likelihood methods are used, the zero flow observations can be accommodated by using the Poisson model.

We end this section by noting that the various combinations of constraints considered have strong implications for the elasticity of total travel demand. As already hinted above, in the case of doubly constrained models total travel demand is given when it is measured terms of the total number of trips entering or leaving each zone. In the unconstrained model the total number of trips is entirely flexible. Thus, the chosen constraints have considerable impact on the elasticity of travel demand in terms of total number of trips. In Section 17.7 a class of more general models will be discussed which has the large advantage that the outcomes do not depend on the a priori chosen constraint regime. The difference is probably smaller when one considers the total *distance travelled* instead of the total *number of trips* because in that case even the doubly constrained model would allow substantial changes in total number of kilometres travelled when the generalised transport costs would change. The issue of elasticities will be further discussed in the context of modal choice (Section 17.5).

17.5 Modal Choice

It is customary to distinguish three clusters of variables that affect modal choice:

1. Features of the individual.
2. Features of the modal alternatives.
3. Features of the choice context.

Among the *individual features* to be considered are factors such as the availability of a car, possession of driver's license, income and household type.

The *modal features* relate to factors such as price or costs (possibly distinguished between fuel costs, toll costs, parking costs, fares) travel time (in the case of public transport possibly distinguished between in vehicle time, waiting time, access time and egress time), comfort, reliability, availability and other qualitative indicators.

Finally the *choice context* is relevant. For example, the time of the day may have an impact on the attractiveness of transport modes (for example, during the evening non-motorised traffic may be considered to give rise to extra social risks). Also the trip purpose may play a role because activities such as working are confronted with stricter scheduling compared with social visits. Weather conditions have an impact on the attractiveness of certain transport modes. During winter-time the use of bicycles for commuting is less than during the summer.

The choice of a transport mode starts with the formation of the choice set (see for example Ortuzar and Willumsen, 2001, Punj and Brookes, 2001 and Exel and Rietveld, 2002). The choice set depends partly on objective factors such as the existence of a railway station in a city or ownership of a car by a household. In the long run these factors may change, for example when households buy an additional car. Also car pooling as an alternative may deserve longer-term preparation since partners have to be found.

Another part of choice set formation relates to information and perception issues. People may be uninformed or mis-informed on travel alternatives. In addition travellers may impose threshold values on certain attributes so that for a certain origin-destination connection they would, for example, never consider an alternative with a travel time longer than 1 hour. Fotheringham and O'Kelly (1989) propose to include the dimension of choice set formation by a stochastic process implying that alternatives with a low attractiveness are not included in the choice set. Then observed choice probabilities p_j that alternative j is selected are the result of two choice processes: q_{1j}, the probability that alternative j is included in the choice set, and the probability q_{2j} that alternative j is selected from the choice set, given that it is in the choice set (see also Lerman, 1984, and Morikawa, 1996).

The problem is that in many cases we have little information on the two sub-processes so that when both are governed by the same criteria and strategies of weighing these (compensatory versus non-compensatory) it will be difficult to distinguish them. Exel and Rietveld (2002) show that when there is explicit information on the availability of the modal alternative ('could you have made this trip by transit') there is no problem in differentiating the two. They find that the two steps are not entirely governed by the same considerations.

Consider the choice between two modes: private transport (car) and public transport (bus). The choice between the two is governed by the factors mentioned above such as travel time, comfort and price. Note that for those who do not have a car the choice probability is 1 for public transport: these travellers can be considered as captives. For the other travellers the choice is governed by the utility function:

$$U_j = a_j + bt_j \qquad \text{for } j=1,2 \qquad (17.15)$$

where all factors except travel time are incorporated into a_j; note that b is negative.

Then, using the simple logit model (see Chapter 2) the probability that the public transport alternative (j=2) is chosen equals:

$$p_2 = 1/[1+\exp\{a_1-a_2 + b(t_1-t_2)\}]. \tag{17.16}$$

An illustration is given in Figure 17.3. It shows that when travel time increases, the probability of a choice for public transport will tend to zero when the difference in travel times between the two modes will get very large.

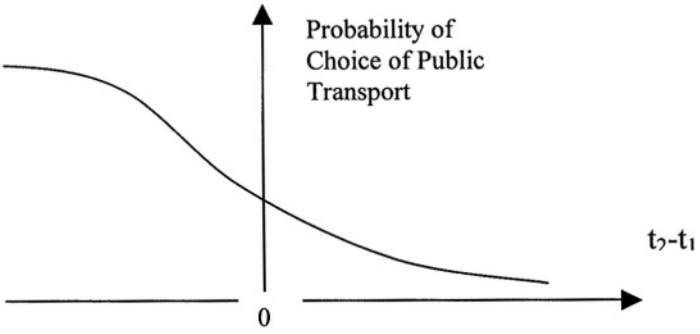


Figure 17.3 Probability of Choice of Public Transport as a Function of the Difference Between Travel Time in Public Transport and Car (t_2-t_1)

Note, however, that the observed choice probabilities for various combinations do not only depend on the trade-offs considered in (16), but also on the share of the population that has access to alternative 1. Therefore, it will not be possible to estimate directly the parameters of the utility function above on the basis of observed modal shares without information on the availability of modes. This obviously holds true for aggregate data, but also in the case of individual data this has to be taken into account. For example in research on the choice of carrier in aviation research it is usually taken for granted that the passenger has a choice between the available flights. However, when some planes are fully-booked this is not a realistic assumption. Along similar lines the observed choice for a transport mode may be governed by incidental factors having an impact on availability of modes. For example the car is not available for use (it happens to be used by another household member, or needs reparation), strikes occur from time to time in public transport, and adverse weather conditions (inland navigation may be impossible when rivers are frozen).

We conclude that actual behaviour does not always provide a reliable indication of the preferences governing modal choice. Stated choice is therefore an attractive alternative. Note that stated choice experiments can also help to avoid the perception problems mentioned above. Another attractive property of stated choice approaches is that it may overcome problems of multicollinearity. It often happens that the features of alternatives are strongly correlated (for example travel time and price). In that case estimates of the relative importance of price versus travel time in modal choice would lead to unreliable results. It is no surprise therefore that stated choice has become a popular way to study modal choice (cf. Henscher, 1994).

The standard way to analyse stated preference data on modal choice is the use of limited dependent variable models such as the logit model, the probit model and extensions. These models entail the estimation of utility functions governing individual choice behaviour. They are surveyed by Ben Akiva and Bierlaire in Chapter 2. Once these models have been estimated they can be used to derive some parameters that have high policy relevance: the value of travel time and choice elasticities.

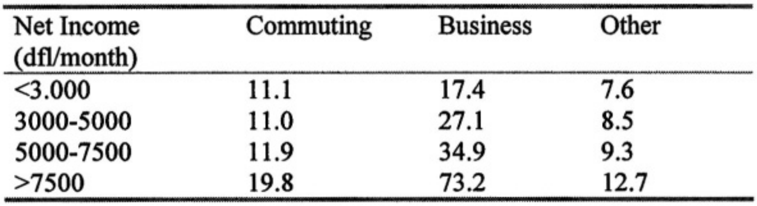
Value of Travel Time.

For the measurement of the *value of travel time,* consider the utility function

$$V_j = a_j + b.p_j + d.t_j \tag{17.17}$$

Then the measurement dimensions of the parameters b and d are [utility per money] and [utility per time]. The absolute meaning of the parameters has little relevance, but the ratio d/b has the dimension [money per time] and can be interpreted as a value of travel time, accordingly. This is a very relevant input for policy studies because it provides a key for the monetarisation of time gains and time losses in cost benefit analysis (see also Small, 1992). It appears that in many infrastructure projects time gains of passengers play a decisive role in the balance between benefits and costs. In a review of value of time studies from the UK, Wardman (1998) finds that the value of time depends on several factors: individual characteristics, modal features and the choice context. Personal factors play a role: people with higher incomes have higher opportunity costs and may be expected to have higher values of time. This implies that in a context of increasing incomes, values of time may be expected to increase. It also means that as people get richer they pay more attention to the quality aspects of trips rather then to the direct monetary expenses.

Also the features of the modal alternatives are important: values of time depend on travel modes. This is no surprise since the levels of comfort are different. An obvious example is the comparison between a railway versus an airline trip between two destinations, *both* taking four hours. The difference in comfort will lead to a higher valuation of the time costs for the flight compared with the railway trip. In addition, the value of time will depend on trip purpose. Trips with a business purpose tend to have a higher value than commuting trips which in their turn have a higher value of time than trips made for social visits. An example for the Netherlands is given in Table 17.6.[4]



Net Income (dfl/month)	Commuting	Business	Other
<3.000	11.1	17.4	7.6
3000-5000	11.0	27.1	8.5
5000-7500	11.9	34.9	9.3
>7500	19.8	73.2	12.7

Bron: HCG (1990), update .

Table 17.6 Values of Travel Time Savings (dfl/hour) for Various Income Groups and Trip Purposes (1999).

Another finding from comparative analysis is that the values of time appear to depend on the specific choice context. For example, when the alternatives used in the Stated Preference research are phrased in terms of tolls to be paid the resulting trade-offs appear to be higher than when the costs relate to fuel (Wardman, 1998). The probable reason is that tolls are paid immediately, whereas fuel is bought from time to time.

Another aspect of value of time studies is that various components of travel time deserve attention. In addition to in vehicle time, total travel time relates to walking time, waiting time, search time, etc. For example, McCarthy (2001) finds that the value of walking time is about four times as high as the value of in vehicle time. Also travel time variability (uncertainty) is important. These findings have important implications for scheduling policies of public transport operators. For example, policies aiming solely at improving the speed of the services by reducing the number of bus stops may be counterproductive when these would imply higher access times. A broader review of these issues is given by Wardman (2001)

Price Elasticity of Travel Demand.

Elasticities are of equal relevance for policy purposes because they reveal how changes in features of modal alternatives lead to changes in choice probabilities. An elasticity is defined as the relative change in the use of a mode divided by the relative change in a feature of that mode (for example the price). The choice probability p_j of an alternative j directly depends on its price j, for example in the ordinary logit model:

$$p_j = \exp V_j / [\text{sum}_{j'} \exp V_{j'}]$$

Thus, when the parameters of $V_{j'}$ have been estimated it is possible to compute the elasticity of the choice probability with respect to price or any other feature. The literature has yielded a large number of estimates of price elasticities of travel demand in the mean time. Comparative analyses of these can be found among others in Oum et al. (1992), Goodwin (1992), Kremers et al. (2002) and Brons et al. (2002). It is important to pay attention to the exact definition of the dependent variable when comparing price elaticities of travel demand. In the context of modal choice the dependent variable is actually a choice probability, which is closely related to the number of trips made. However, in other models the elasticity relates to the number of kilometres travelled, so that also the other elements of travel decisions play a role, in particular trip generation and trip distribution. Therefore, one may expect that elasticities based on discrete choice analysis of modal choice are closer to zero than elasticities based on total number of kilometres travelled. This is indeed confirmed in the literature (Oum et al., 1992, Kremers et al., 2002).

Another important reason for differences between price elasticities is the difference between short and long-run estimates. Brons et al. (2002) find that long-run estimates are more elastic than short-run estimates. The background is that in the short-run there are fewer possibilities to adjust to price changes (for example, jobs and dwellings are fixed in the short run). Trip purposes are also important. A general finding is that business traffic is least elastic, whereas social visits are most price elastic. For commuting intermediate values are found. One of the reasons for this pattern is that the costs of business and commuting travel are not entirely borne by the traveller himself. In many countries institutional arrangements exist such that such that (part of) the travel costs are paid by the employer. Many employers face the problem that the transaction costs of dealing with these transport expenses are high so that simple arrangements are made implying that there are little incentives for employees to reduce travel.

Other factors that have an impact on price elasticities concern the number of alternatives and the specific way in which payment takes place. When the number of alternatives is small, one may expect travel demand to be relatively inelastic compared with the situation where the number of alternatives is large. Thus, demand for transport in an urban context (where often many alternatives exist) tends to be more elastic than in contexts where the choice set is small.

Another point that deserves attention is that price elasticity estimates can be based on both revealed and stated preference data. Kremers et al. (2002) find a tendency that according to stated preference estimates demand is more elastic compared with revealed preference estimates. This may be an indication that, in stated preference studies, people exaggerate their responsiveness to price changes. On the other hand, as has been indicated already above, the difference may also relate to estimation problems with revealed choice data.

The way of payment is also important. For example, public transport demand is often rather elastic in terms of ticket prices compared with demand for travel by car. The reason is that with public transport the ticket price is 100% of the total price, whereas with car use there are many other price components (car ownership, fuel costs, maintenance, parking, etc.). Consider the case that total monetary costs of a trip are c_{tot} and c_l is the contribution of a certain cost component. Let E_{tot} be the elasticity of travel demand with respect to total costs. Then it is not difficult to demonstrate that the elasticity E_l with respect to this cost component follows as:

$$E_l = (c_l/c_{tot}) \,.\, E_{tot} \tag{17.18}$$

Thus, cost shares play an essential role in the link between total price elasticities and specific price elasticities.

The above results on values of time and prices elasticities indicate that there is a large variation in outcomes between the various studies. This raises the issue of transferability of estimation results. Meta analyses as reported above are useful tools for this purpose.

Transferability does not only relate to elasticities and values of travel time, but also to parameters of the utility functions underlying these. An important finding in transport modeling is that transferability of the model parameters is difficult. In particular it appears that mode-specific constants such as a_j in equation (15) are needed to arrive at plausible modelling outcomes for modal choice. The implication is that although one may have the ambition to make all quality aspects of transport modes explicit in utility functions, in practice one will still need mode specific constants. These constants appear to absorb various situation specific circumstances that cannot be transferred from one situation to the other. Note that when these constants would be close to zero, the utility model can easily be used to predict the market potential for a new transport mode (for example the introduction of a metro system in a metropolitan area where only buses and cars are employed). This was the main idea behind the classic contribution of Quant and Baumol (1966) who introduced a model for travel demand in terms of abstract modes where mode specific coefficients did not play a role. It is a pity that this vision has not been realised thus far, because it limits the usefulness of the existing transport models for policy purposes. It implies that existing models cannot directly be used for the analysis of the introduction of a new transport modes, but that one should first carry out stated preference type of interviews in order to get meaningful estimates of the market potential for the new mode.

17.6 Transport – Land Use Linkages

There has been a long debate in the transportation and land use literature on the relationship between transportation behavior and the geographical location of economic activities. Do land use and location determine transport flows or do transport networks determine the location of economic activity? Much of the debate has taken the form of a chicken and egg dilemma and has neglected the difference between slow and fast motion. A well-specified dynamic land-use transportation model would be able to encapsulate both phenomena simultaneously.

Nevertheless, in many urban and infrastructure planning debates we often observe a one-sided perspective. For example, in the new urbanism we witness much emphasis on the causality from land use to travel behavior, with the consequence that the policy handles are sought in land use planning. The problem here is that human behavior does not obediently follow the logic imposed by the land-use transportation causality, because of inertia in human behavior and limits to flexibility regarding spatial relocation. It ought to be recognized that – even though transportation is regarded by individuals and society at large as a problem in terms of congestion, air quality, quality of life or safety – the ultimate outcomes of individual and collective choices (in terms of trip generation, trip length, modal split or route choice) do not necessarily reflect an optimal result of the transport system in terms of long-term environmental sustainability.

There is a main point of difference between environmental aspects and land use aspects in transport systems. The environmental aspects can mainly be considered as an outcome of transport systems in combination with travel demand. On the other hand, land use aspects are a determinant of travel demand: as explained in the section on trip generation, the number of trips produced and attracted by zones depends strongly on land use in the zones. However, although this land use is assumed to be exogenous in many urban transport models, it is not necessarily so because changes in transport systems may lead to changes in land use. This calls for an integrated analysis of land-use and transport. Reviews of integrated transport – land use models can be found in Webster et al. (1988), Webster and Paulley (1990), Wegener (1991), Hayashi and Roy (1996) and Martinez (2000)

In the context of the present paper we will discuss a specific example of a transport-land use model that is closely linked to the transport models discussed in the section on distribution models. It can actually be formulated as a generalisation of this model (see Alonso, 1978 and De Vries et al., 2002). In this section we discussed the various cases of the double, single and non-constrained model. A general formulation that would cover all these models would be:

$$
\begin{gathered}
T_{ij} = O_i A_i D_j B_j h(c_{ij}) \\
O_i = A_i^{-\alpha} X_i \\
D_j = B_j^{-\beta} Z_j \\
\Sigma_i T_{ij} = D_j \\
\Sigma_j T_{ij} = O_i
\end{gathered}
\tag{17.19}
$$

In this information X_i and Z_j denote exogenous features of origin and destination zones. It can easily be verified that the following special cases hold (see Table 17.7).

	$\alpha = 0$	$\alpha = 1$
$\beta = 0$	Double Constrained model	Attraction Constrained Model
$\beta = 1$	Production Constrained Model	Unconstrained Model

Table 17.7 Links Between Integrated Transport Land Use Model and Spatial Interaction Models.

The new element of the above model is obviously that the total number of trips produced and attracted by a zone is dependent on accessibility in a flexible way. This means that changes in transportation systems, leading to changes in accessibility indicators A_i and B_j will in their turn lead to changes in activities such as shopping, housing and hence to changes in the use of land in the pertaining zones. Note that the double constrained formulation is a special case where such feedbacks of transport costs on zonal activities are rules out. However, an implication of this formulation is that also single constrained models imply a transport land use interaction. An obvious example is shopping. The model shows that as shopping areas get better accessibility they will also attract more customers so that one may expect expansion of shops. The large advantage of the above model compared with the standard single and double constrained models is that there is no need to impose a priori the degree of the constraint. Between the extremes of a proportional response and a non-response with respect to accessibility this formulation also allows intermediate elasticities, implying values for α and β between 0 and 1.

17.7 New Perspectives

The above described classes of transportation models are essentially members of the family of spatial interaction models. Such models map out the geography of movement and are a prominent approach to the analysis of spatial flows (see for an overview Nijkamp and Reggiani 1992). A popular set of spatial interaction models is based on gravity theory. The use of such physics-based models can be justified due to its link (or formal analogy) with utility theory, either at an aggregate (systemic) level or on individual (choice-maker's) level. Parallel approaches can be found in entropy theory and in Alonso's theory of movement as outlined in Section 17.6. It can be demonstrated that also these models have their roots in standard utility and cost principles, while also a parallel can be drawn to linear programming models. The family of spatial interaction models comprises at present indeed a wide variety of different but largely complementary models, such as micro-simulation models, random utility models, activity-based choice models and spatial search models. The close orientation of spatial interaction models to micro-utility models has several major advantages, such as a firm behavioural underpinning, the possibility of a multi-level aggregation, the opportunity to organize panel studies in a a longitudinal context, the possibility to include categorical measurement of response variables and the possible inclusion of multi-actor effects (e.g., congestion). For the study of dynamic spatial interaction models, in particular in the context of catastrophy and chaos models, the reader in referred to Nijkamp and Reggiani (1992).

A promising direction of research in the field of spatial interaction modeling is the integration of trip modeling and activity modeling. This will lead to a more explicit treatment of the duration element of activities in transport models. It will also provide a good basis for a further introduction of scheduling issues of trips and activities in transport models. Scheduling of trips and activities is an important way to avoid bottleneck congestion in metropolitan areas. To make the present spatial interaction models more relevant for the analysis of congestion problems the themes of scheduling and activity based modeling deserve more attention than they usually receive.

References

Aschauer, D.A. (1989) Is public expenditure productive?, *Journal of Monetary Economics,* **33**, 177-200

Bates, J (2000) The history of demand modelling, in: D.A. Henscher and K.J. Button (eds.), *Handbook of Transport Modelling,* Pergamon, Amsterdam, pp. 11-33.

Batten, D.F., J.Casti and R. Thord (eds.) (1995) *Networks in Action,* Springer-Verlag, Berlin.

Bergh, J.C.J.M. van den, P. Nijkamp and P. Rietveld (eds.) (1996) *Recent Advances in Spatial Equilibrium Modelling* Springer-Verlag, Berlin.

Brons, M., E. Pels, P. Nijkamp, P. Rietveld (2002) Price elasticities of demand for passenger air travel, *Journal of Air Transport Management,* forthcoming.

Capello, R. (1994) *Spatial Economic Analysis of Telecommunications Network Externalities,* Averbury, Aldershot, UK,

CBS (Central Bureau of Statistics) (2000) De mobiliteit van de Nederlandse bevolking, The Hague.

Cesario, F (1975) Linear and non-linear regression models of spatial interaction, *Economic Geography,* **51**, 69-77.

Choukroun, J-M (1975) A general framework for the development of gravity-type trip distribution models, *Regional Science and Urban Economics,* **5**, 177-202.

Doren, Ch. van (1992) *A History of Knowledge,* Ballantine Press, New York.

Este, G. d' (2000) Urban freight movement modeling, in: D.A. Henscher and K.J. Button (eds.), *Handbook of Transport Modelling,* Pergamon, Amsterdam, pp. 539-552.

Exel, J. van, and P. Rietveld (2002) Could you also have made this trip by public transport? Determinants of choice set formation, Vrije Universiteit, Amsterdam.

Fik, T.J., and G.F.Mulligan (1998) Functional form and spatial interaction models, *Environment and Planning, A,* **30**, 1497-1507.

Fotheringham, A.S. and M.E. O'Kelly (1989) *Spatial Interaction models: formulations and applications,* Kluwer, Dordrecht.

Goodwin, Ph. (1992), A review of new demand elasticities with special reference to short and long run effect of price changes, *Journal of Transport Economics and Policy,* **36**, 155-169.

Hayashi, Y., Roy, J. (eds.) (1996) *Transport, Land Use and the Environment.* Dordrecht: Kluwer, pp. 103-124.

HCG, (Hague Consulting Group), (1990) The Netherlands value of time study: final report, Hague Consulting Group, Den Haag.

Henscher, D.A. (1994) Stated preference analysis of travel choices: the state of practice, *Transportation,* **21** (2), 106-134.

Himanen, V., P. Nijkamp, and A. Reggiani (eds.) (1998) *Neural Networks in Transport Applications,* Ashgate, Aldershot, UK.

Kim, T.J. (1979) Alternative transportation modes in a land-use model: a general equilibrium approach, *Journal of Urban Economics,* **6**(2), 197-215

Kim, T.J., (1997) A combined land-use transportation model when zonal travel demand is endogenously determined, *Transportation Research B,* 1983, **17B**(6), 449-462

Kremers, H., P. Nijkamp and P. Rietveld (2002) A meta-analysis of price elasticities of transport demand in a general equilibrium framework, *Economic Modeling,* forthcoming.

Lerman, S.R, Recent advances in disaggregate demand modelling, in: M. Florian (ed.), *Transportation Planning Models,* North Holland, Amsterdam.

Lundqvist, L. and L.-G. Mattsson (eds.) (2001) *National Transport Models,* Springer, Berlin.

McCarthy, P.S. (2001) *Transportation Economics,* Blackwell Publishers, Oxford.

McNally, M.G. (2000) The four step model, in: D.A. Henscher and K.J. Button (eds.), *Handbook of Transport Modelling,* Pergamon, Amsterdam, pp. 35-52.

Morikawa, T. (1996) A hybrid probabilistic choice model with compensatory and non-compensatory choice rules, in: D. Henscher, J. King and T. Oum (eds.) World Transport Research, World Conference for Transportation Research, Sydney.

Nagurney, A., and S. Siokos (1997) *Financial Networks, Statics and Dynamics,* Springer, Berlin.

Nijkamp, P., S. Rienstra and J. Vleugel (1999) *Transportation Planning and the Future,* John Wiley, New York.

Nijkamp, P., and A. Reggiani (1998) *The Economics of Complex Spatial Systems,* Elsevier, Amsterdam.

Nijkamp, P., G. Pepping and D. Banister (1996) *Telematics and Transport Behaviour,* Springer-Verlag, Berlin

Nijkamp, P., and A. Reggiani (1992) *Interaction, Evolution and Chaos in Space,* Springer-Verlag, Berlin.

Ortuzar, J. de D., and L.G. Willumsen, (2001) *Modelling Transport,* Wiley, New York.

Oum, T.H., W.G, Waters and J.S. Jong (1992) Concepts of price elasticities of transport demand and recent empirical estimates, *Journal of Transport Economics and Policy,* **26**, 139-154.

Punj, G. and R. Brookes (2001) Decision constraints and consideration-set formation in consumer durables, *Psychology and Marketing,* **18**, 843-863.

Quant, R. and W.J. Baumol (1966) The demand for abstract transport modes: theory and measurement, *Journal of Regional Science,* **6**, 13-26.

Reggiani, A. (1998) Accessibility, trade and locational behaviour, Ashgate, Aldershot.

Rietveld, P. and F. Bruinsma (1998) *Is Transport Infrastructure Effective?,* Springer, Berlin.

Rietveld, P. (2000) Non-motorised modes in transport systems: a multimodel chain perspective for The Netherlands, *Transportation Research D: Transport and Environment,* **5**, 31-36.

Rho, J.H., T.J. Kim, and L. Lundqvist, (1993) Integrated land-use transportation model: application to Chicago and outline for Stockholm, in T.R. Lakshmanan and P. Nijkamp (eds.), *Structure and Change in the Space Economy,* Springer-Verlag, Berlin-Heidelberg.

Salomon, I., P. Bovy and J.P. Orfeuil (1993) *A Billion Trips a Day,* Kluwer, Dordrecht.

Sen, A., and S. Soot (1981) Selected procedures for calibrating the generalized gravity model, *Papers of the Regional Science Association,* **48**, 165-176.

Verhoef, E., (1996) *The Economics of Regulating Road Transport,* Edward Elgar, Cheltenham.

Vries, J.J., P. Nijkamp, and P. Rietveld, (2001) Alonso's theory of movement, *Journal of Geographical Systems,* **3** (3), 233-256

Wardman, M. (1998) The value of travel time, a review of British evidence, *Journal of Transport Economics and Policy,* **32**, 285-316.

Wardman, M. (2001) A review of British evidence on time and service quality valuations, Transportation Research part E, **37**, 107-128

Webster, F.V., Bly, P.H., Paulley, NJ. (eds.) (1988): *Urban Land-Use and Transport Interaction: Policies and Models,* Aldershot: Avebury.

Webster, F.V., Paulley, N.J. (1990): An international study on land-use and transport interaction. *Transport Reviews* **10**, 287-322.

Wegener, M., R.L. Mackett, and D.C. Simmonds, (1991): One city, three models: comparison of land-use/transport policy simulation models for Dortmund. *Transport Reviews* **11,** 107-29.

Wegener, M. (1996): Reduction of CO2 emissions of transport by reorganisation of urban activities. In: Hayashi, Y., Roy, J. (eds.): *Transport, Land Use and the Environment.* Dordrecht: Kluwer, 103-124.

Endnotes

[1] The values of O_i and D_j may be assumed to be given here: they result from the trip generation model.

[2] A more general formulation of the model would allow for the inclusion of a factor Z_j in the model that reflects the attractiveness of zone j for shopping activities.

[3] Along similar lines as in the production constrained model a factor X_i may be added to take into account origin specific factors spatial interaction patterns.

[4] Note that the differentiations mentioned here with respect to variations in value of time correspond closely with the dimensions mentioned at the beginning of this section on modal choice.

18 PRINCIPLES OF TRANSPORT ECONOMICS

Richard Arnott and Marvin Kraus

18.1 Introduction

Traditionally, courses in transport economics were divided into three sections: demand, supply, and regulation. The section on demand focused on empirical work estimating mode-specific demand elasticities and on short-term demand forecasting using discrete choice models. The section on supply concentrated on applied work estimating mode-specific cost functions, but also contained some discussion of the technology of congestion. And the section on regulation considered both the positive and normative aspects of regulation in different transport industries. The normative part applied textbook microeconomic theory to identify market failures and to derive optimal corrective policy, and the positive part applied the same body of theory to examine the effects of alternative regulatory policies.

This chapter will have a considerably narrower scope, focusing on the application of microeconomic theory to resource allocation in the transportation sector. There are two basic issues. How should transportation be priced? And how should capacity be determined? There are correspondingly two basic principles. First, economic agents make socially efficient decisions when they face the full social costs of their decisions and derive the full social benefit from them; typically, this entails pricing at short-run marginal social cost or, where this is not practicable, pricing as close to short-run marginal social cost as is practicable. Second, capacity should be expanded to the point where the social benefit from additional capacity equals the social cost. Of course, the application of these principles in different contexts entails a host of subtleties; otherwise, transport economics would be a short subject.

Happily there is now more exchange of ideas between transport economists and transportation scientists than there was a generation ago. Transportation scientists are coming to appreciate the power of prices, and transport economists are beginning to realize that there is more to their job than deriving first-order optimality conditions in unrealistically simple models. While it is narrowing, there remains a gulf between the two fields. To some extent, this derives from the different roles transportation scientists and transport economists play in the policy/project evaluation process. But more important are methodological differences in training. Engineering training stresses the practical, producing something that does the job well, which requires paying attention to detail. Economics training stresses the theoretical — how to conceptualize problems — in the belief that practical knowledge will come through experience. Since different conceptualizations entail different models, economic reasoning tends to be model-based. Furthermore, economists tend to reason from the general to the specific, and from the simple to the complex. They start with simple models that illuminate general principles and then add complications that do not undermine the basic principles but affect how they are applied.

The central issue in most of economics is the appropriate role of government. Most mainstream economists believe that most economic activity should be organized through markets, with the government intervening only when the market, for clearly-identified reasons, fails. Microeconomists use as points of reference models in which markets perform perfectly so that no government intervention is required, then introduce realistic complications and constraints which induce market failures, and then derive the minimal government intervention needed to correct or at least alleviate those market failures, so that the market assisted by government policy achieves constrained efficiency in resource allocation.

This is the approach we shall follow in this chapter. Section 18.2 will present the traditional highway pricing and investment model. It will start by developing quite thoroughly the simplest variant of the model, and will then add realistic complications one by one. Section 18.3 will examine the highway bottleneck model, whose focus is on trip timing. Section 18.4 will investigate how the traditional model is adapted to deal with mass transit. Given the narrow scope of the essay, many other important topics in transportation on which the application of economics can throw light — such as regulation, transportation and land use, scrappage and maintenance, privatization, and logistics — shall be ignored.

18.2 The Traditional Highway Pricing and Investment Model

There are two broad classes of models in microeconomics — general equilibrium and partial equilibrium. A general equilibrium model provides an exhaustive description of a self-contained abstract economy, while a partial equilibrium model describes only a part of an economy — often a market — treating as implicit what is happening in the rest of the economy. General equilibrium models are conceptually preferable since they account explicitly for everything, but provide a distracting amount of detail, such as the effect of widening a road on the cost of producing a jar of peanut butter. Fortunately, economists have a good understanding of the circumstances under which partial equilibrium analysis is "valid" — gives the same results as would be obtained from a corresponding general equilibrium analysis — and so frequently adopt the short cut of employing a partial equilibrium model. The model to be presented in this section — of a single road link — is partial equilibrium. The main requirement for the conclusions derived from analyzing the model to be valid is that the rest of the economy be efficient — that throughout the rest of the economy, the marginal social benefit of every economic action equal its marginal social cost. Suppose, for example, that a driver who is deterred from taking a trip by a slight increase in congestion uses the time and money saved to eat a peanut butter sandwich. The fact that he consumes peanut butter sandwiches up to the point where the marginal social benefit equals the marginal social cost means that the social benefit from his eating that extra sandwich equals the social cost. The net social benefit therefore equals zero, and can be ignored in welfare analyses. By the same token, all of the millions of other incremental adjustments that economic agents make to that slight increase in congestion can be ignored.

The Model

Consider a one-way highway link from point A to point B. Individuals are identical, and conditions on the road are uniform over time (travel on the road is steady state) and space.

At any point in time, there will be a certain number of vehicles on any one-mile stretch of road — the traffic *density.* The hourly rate at which vehicles pass by any point on the road — the traffic *flow* — is denoted by N. Flow is the road's output variable, since it is the rate at which trips from A to B are produced. Travel time from A to B is T minutes per vehicle. As density increases, so does travel time. The relationship between the two is assumed to be technological and such that, for the relatively low densities to which the analysis is applied, flow increases with density.[1,2] The maximum flow on the road is termed its capacity, and denoted by s.

Figure 18.1 displays the relationship between travel time and flow over the range of densities to which the analysis is applicable. Density is zero at point D and increases steadily for left-to-right movements along the curve. Most empirical analyses have found travel time to be a convex function of flow, and that is assumed in the analysis. An improvement of the road increases capacity, and lowers travel

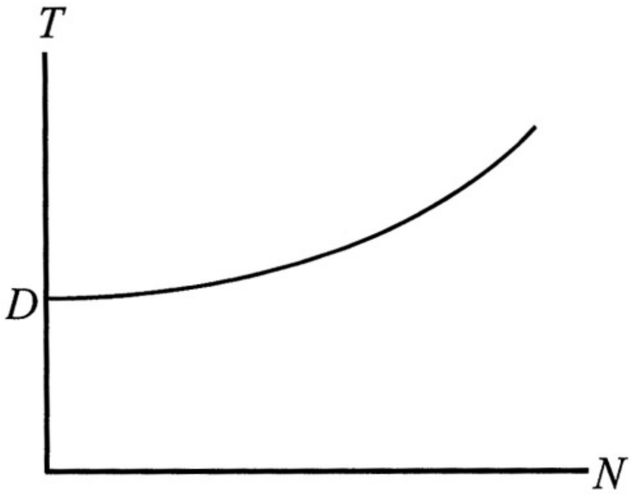


Figure 18.1 Travel Time Versus Flow

time for each level of flow. The traditional analysis assumes a technological relationship between travel time, flow, and capacity in which travel time depends on flow and capacity only through their ratio, *N/s,* which is termed the *volume-capacity ratio.* Accordingly,

$$T = T(N/s),\quad T'(\cdot) > 0,\quad T''(\cdot) > 0\,. \tag{18.1}$$

We assume a constant vehicle operating cost of *c* and a given value of travel time, *v*. The time plus operating cost of a trip will be referred to as the *user cost,* and written as

$$f = f(N/s) \equiv c + vT(N/s)\,, \tag{18.2}$$

where from (18.1), $f'(\cdot) > 0$ and $f''(\cdot) > 0$. Thus, the total user costs per hour corresponding to a flow *N* per hour are

$$C(N,s) \equiv Nf = Nf(N/s)\,. \tag{18.3}$$

Note that $f(\cdot)$ is homogenous of degree zero in N and s, and that $C(\cdot)$ is homogenous of degree one.

The marginal social cost of N is the increase in total hourly user costs from increasing the flow rate by one unit:

$$\partial C/\partial N = f + N\partial f/\partial N. \tag{18.4}$$

This has two components. The first, f, is the user cost of the extra vehicle per hour; the second, $N\partial f/\partial N$, is the *congestion externality* — the increase in the total hourly user costs of the existing or *inframarginal* trip takers from being slowed by the increase in traffic. Economists use the term *externality* to refer to an *external cost,* a cost that an economic agent's actions engender but which is external to that agent — that agent does not pay for.

The relationship between user cost and marginal social cost (msc) is displayed in Figure 18.2. The congestion externality at a given flow rate equals the vertical distance between marginal social cost and user cost at that flow rate. To illustrate traffic congestion, a user cost function of the form

$$f(N/s) = a + b(N/s)^{\gamma} \tag{18.5}$$

is frequently posited. Other functions fit the data better, but this function is used because γ has a nice interpretation as the *congestibility* of the road and because the ratio of the congestion externality to the congestion cost incurred by a traveller, $f(N/s) - f(0)$, equals γ. Thus, if $\gamma = 5$, and if a traveller's trip cost increases from $1 to $2 as a result of congestion, that traveller's presence on the road imposes a cost of $5 on other travellers.

Sadly, the microeconomist's assumption that economic agents are completely selfish performs remarkably well. In the traffic congestion context, this corresponds to the assumption that, in making her travel decisions, an individual considers only her user cost — the cost she herself bears directly — and not the cost she imposes on others, the congestion externality. In the absence of a toll, the individual will travel to the point where the marginal private benefit of the last trip equals the user cost. Since the marginal social cost exceeds the user cost (and under the additional assumptions that marginal social benefit equal marginal private benefit, and that marginal benefit is decreasing in the amount of travel), she will travel too much — beyond the socially optimal level of travel for her, for which marginal social benefit equals marginal social cost.

Economists have a number of standard policy remedies for dealing with externalities. The preferred remedy for traffic congestion is to charge the individual

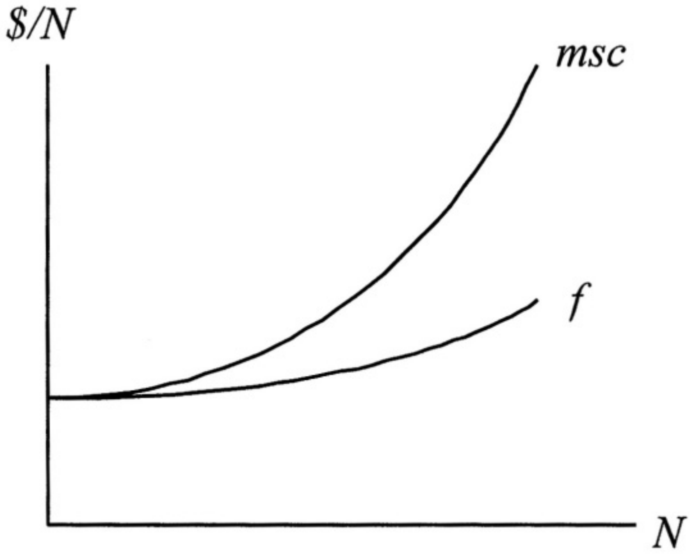


Figure 18.2 User Cost (f) and Marginal Social Cost (msc) Versus Flow

for the congestion externality she imposes on others. This will result in her facing the social cost of a trip, and hence choosing the socially optimal amount of travel for her. The externality is thereby internalized. We shall examine this remedy in more detail shortly.

Above we have described the congestion technology. There are two other elements in the analysis, capacity costs and demand. The capacity cost function, $K(s)$, indicates the hourly cost of building the road so that it has capacity s, employing the least-cost construction method and holding factor prices fixed. For the moment, we make only the natural assumption that $K'(\cdot) > 0$. The capacity cost function and the user cost function together characterize the supply side of the model. Total hourly costs are given by $C(N,s) + K(s)$.

The demand side of the model is characterized simply by a *demand function,* which relates the flow demand for trips to trip price, P:[3]

$$N = N(P); \tag{18.6}$$

it is natural to assume that $N'(\cdot) < 0$. We shall also work with the *inverse demand function P* = *P*(*N*). An individual's trip price equals the user cost plus any fee she is charged for taking the trip, which we term the toll and denote by τ. Thus,

$$P = f(N/s) + \tau. \tag{18.7}$$

The next step in the analysis is to bring together the supply and demand sides of the market to determine market equilibrium. Substituting (18.7) into (18.6),

$$N = N(f(N/s) + \tau). \tag{18.8}$$

Equation (18.8) implicitly defines the equilibrium flow rate as a function of τ and s:

$$N = \hat{N}(\tau, s). \tag{18.9}$$

Under our assumptions that $N'(\cdot) < 0$ and $f'(\cdot) > 0$ it is straightforward to show that $\hat{N}_\tau < 0$ and $\hat{N}_s > 0$, which accords with intuition. Substituting (18.9) back into (18.7) gives a corresponding expression for the equilibrium trip price:

$$P = f(\hat{N}(\tau,s)/s) + \tau \equiv \hat{P}(\tau, s),$$

where $\hat{P}_\tau > 0$ and $\hat{P}_s < 0$.

We now ask the questions: What is the socially optimal level of the toll? And what is the socially optimal level of capacity?

First-best Analysis

Recall our earlier discussion of the conditions under which partial equilibrium welfare analysis is valid — basically that the welfare effects from the spillovers into other markets due to changes in the "market for travel from A to B" net out, which occurs when the rest of the economy is efficient. We assume these conditions to hold. We also assume that there are no constraints on the operation of the market for travel from A to B other than those captured in the supply and demand functions. Under these circumstances, we conduct what is termed first-best partial equilibrium analysis.

We wish to choose road capacity and the level of the toll so as to maximize social welfare. For the conditions under which partial equilibrium is valid, we may ignore the repercussions of changes in s and τ on the rest of the economy. The maximization of social welfare then reduces to the maximization of hourly *social surplus* from the road, which equals the direct social benefits from the road minus the direct social costs. The direct social costs are simply the total costs identified earlier, the sum of user costs and capacity costs. The direct social benefits are given by the area under the demand curve up to the equilibrium trip flow. To see this, note that a point on the demand curve indicates the marginal private benefit of a trip, since the individual who takes that trip is willing to pay the corresponding price but no more. The demand curve is therefore the marginal *private* benefit curve, and the area under the demand curve up to a specified level of flow the total private benefit associated with that level of flow (this is strictly correct only in the absence of income effects, which we have assumed). Two additional assumptions are required to interpret the area under the demand curve as the total *social* benefit. The first is that there be no consumption externalities — that only the person taking a trip derive utility or disutility from his doing so (smoking is an example of a consumption externality). The second, which is considerably more contentious, is that the social planner values a dollar's worth of benefits equally whether it goes to a rich person or a poor person. One rationale for this is that interpersonal utility comparisons are not scientific, so that the only defensible assumption is that different individuals' benefits be weighted equally; another is that the social planner has exercised her authority to redistribute income in lump-sum fashion to the point where everyone's marginal social benefit from income is equalized. We employ the assumption to simplify the exposition. The analysis can be extended straightforwardly to allow for the planner to attach different social valuation to a dollar going to different people, but the results include additional terms reflecting equity considerations.

The objective function is therefore

$$\max_{\tau,s} \int_0^{\hat{N}(\tau,s)} P(n)dn - C(\hat{N}(\tau,s),s) - K(s). \qquad (18.10)$$

The first-order condition with respect to τ is

$$(P - \partial C/\partial N)\hat{N}_\tau = 0.$$

Since $\hat{N}_\tau < 0$, this implies that

$$P = \partial C/\partial N; \qquad (18.11)$$

the toll should be set so that each driver pay a trip price equal to its marginal social cost. Since this result holds regardless of whether capacity is optimized, it demonstrates the optimality of *short-run* marginal cost pricing. Substituting (18.4) and (18.7) into (18.11) yields

$$\tau = N\partial f/\partial N. \qquad (18.12)$$

Price equals user cost plus the toll, while marginal social cost equals user cost plus the congestion externality. Thus, price equals marginal social cost implies that the toll equal the congestion externality.

This result is displayed in Figure 18.3. The socially optimal number of trips occurs where the marginal social benefit of a trip equals the marginal social cost, and is therefore characterized by the point of intersection of $P(N)$ and $\partial C/\partial N$, point E. In the absence of a toll, individuals take trips up to the point where marginal private benefit equals marginal private cost, which in the absence of a toll equals the user cost. Thus, the no-toll equilibrium is characterized by the point G.[4] Because individuals do not pay for the congestion externality they impose on others, there is excessive travel. Now suppose that a toll is imposed equal to the congestion externality, evaluated at the socially optimal number of trips — the height EF in the figure. This causes the marginal private cost to shift up by EF, and hence to intersect the demand curve at E. Thus, imposition of a toll equal to the congestion externality evaluated at the socially optimal number of trips induces the socially optimal amount of travel.

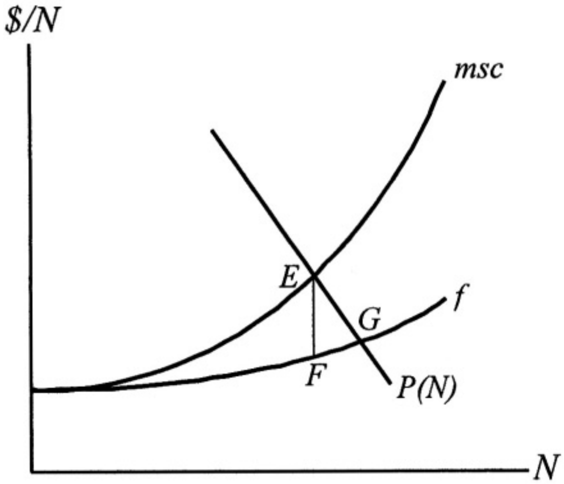


Figure 18.3 Socially Optimal Number of Trips (E) and User Optimal Number of Trips (G)

This result is robust. The model can be generalized to incorporate heterogeneous individuals, multiple modes, networks, etc., If tolls are applied to all links (and nodes, where there is nodal congestion) of a network equal to the corresponding congestion externalities, the socially optimal amount of travel will occur throughout the network. Individuals therefore make socially optimal choices with respect to not only the amount of travel but also route and mode. This accords with the general principle that when rational, self-interested individuals face the full social costs of their actions (no production externalities) and derive the full social benefit from them, they will make efficient decisions. The result also extends to the case where demand varies over time, so that congestion and hence the congestion externality and the optimal toll are higher in peak than in off-peak periods. This result is known as *peak-loadpricing.*

One qualification is in order. Our analysis assumed that congestion is anonymous in the sense that all individuals impose the same congestion externality. But in fact the magnitude of the congestion externality a driver imposes depends on how he drives and the type of vehicle he drives. If the toll can be personalized so that each driver pays for the congestion externality he imposes, then all drivers continue to make socially efficient decisions.[5] But if tolls cannot be fully personalized so that, for instance, bad drivers pay no more than good drivers, full efficiency is no longer achieved. In this case, we say that there is a constraint on the extent to which tolls can be differentiated, which precludes attainment of the full or first-best optimum. The planner will then choose the level of the toll to maximize social welfare subject to the constraint on the extent of toll differentiation, which is an exercise in the theory of the second best. Under the constrained efficient toll, bad drivers will drive too much and good drivers too little. We shall examine second-best optimality in some detail in the next subsection.

Return to the first-best problem, (18.10). The first-order condition with respect to capacity is

$$(P - \partial C/\partial N)\hat{N}_s - \partial C/\partial s - K'(s) = 0, \qquad (18.13)$$

where $\partial C/\partial s$ denotes the partial derivative of $C(\hat{N}(\tau, s), s)$ with respect to its second argument. If the toll is set so that (18.11) holds, (18.13) reduces to

$$\partial C/\partial s + K'(s) = 0, \qquad (18.14)$$

which is the first-order condition for producing the equilibrium number of trips at minimum cost, and is therefore a production efficiency condition. Equation (18.14) states that capacity should be expanded to the point where the reduction in total user costs, the marginal social benefit of capacity, equals the increase in capacity costs, the marginal social cost of capacity.

The highway authority incurs capacity costs of $K(s)$ and receives toll revenues of τN. Under (first-best) optimal pricing and investment, what proportion of capacity costs is financed out of toll revenues? This is known as the *self-financing* question and owes its basic results to Mohring and Harwitz (1962) and Strotz (1965).

Initially, we do not assume the user cost function to be homogenous of degree zero in N and s. Define $TC(N,s) = C(N,s)+ K(s)$ and $AC(N,s) = TC(N,s)/N$. $TC(\cdot)$ and $AC(\cdot)$ are the *short-run* total and average cost functions. The *long-run* total and average cost functions are

$$LRTC(N) = \min_{s} \; TC(N,s) \tag{18.15}$$

and $LRAC(N) = LRTC(N)/N$, and are the lower envelopes of the short-run total and average cost functions, respectively. Long-run marginal cost is given by $LRMC(N) = \partial LRTC(N)/\partial N$. With N and s as the choice variables, the planning problem is

$$\max_{N,s} \int_0^N P(n)dn - TC(N,s). \tag{18.16}$$

Maximizing (18.16) with respect to N gives

$$P(N) - TC_N(N,s) = 0 \tag{18.17}$$

— users should face a trip price equal to short-run marginal cost, whether or not capacity is optimal. Maximizing (18.16) with respect to s is equivalent to minimizing $TC(N,s)$, which from (18.15) results in costs of $LRTC(N)$. Following (18.17) then results in *long-run* marginal cost pricing, because N must now solve

$$\max_{N} \int_0^N P(n)dn - LRTC(N) \tag{18.18}$$

which yields the first-order condition

$$P(N) - LRMC(N) = 0. \tag{18.19}$$

Thus,

$$\begin{aligned} N^*\tau^* - K(s^*) &= N^*(\tau^* + f(N^*,s^*)) - (K(s^*) + N^*f(N^*,s^*)) \\ &= N^*P^* - N^*LRAC(N^*) \\ &= N^*(LRMC(N^*) - LRAC(N^*)) \quad \text{(using (18.19))} \end{aligned} \tag{18.20}$$

where *'s indicate values at the first-best optimum. Equation (18.20) indicates that, at the first-best optimum, toll revenues exceed/equal/fall short of capacity costs according to whether long-run marginal cost exceeds/equals/falls short of long-run average cost, or equivalently whether long-run average cost is (locally) rising, flat, or falling. We state this as:

Proposition 1. At the first-best optimum, toll revenues exceed/equal/fall short of capacity costs according to whether long-run average cost is (locally) rising, flat, or falling.

Proposition 1 is illustrated for the case of locally decreasing long-run average cost in Figure 18.4. From (18.19) the optimal number of trips N^* occurs where the demand curve intersects the long-run marginal cost curve. The optimal level of capacity, s^*, is such that average cost is minimized for that number of trips, and the corresponding short-run average cost curve, drawn as $SRAC(N,s^*)$, is tangent to the long-run average cost curve at N^*. The corresponding short-run marginal cost curve intersects *LRMC* at N^* and also goes through the minimum point of the short-run average cost curve. The user cost curve corresponding to optimal capacity is drawn as $f(N,s^*)$, and the vertical distance between $SRMC(N,s^*)$ and $f(N,s^*)$ at N^*, *YX*, gives the congestion externality at the optimum.

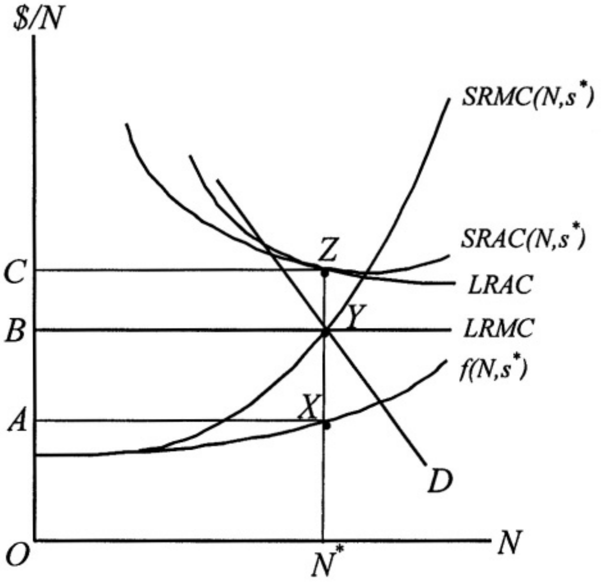


Figure 18.4 At the First-Best Optimum, Toll Revenues Fall Short of Capacity Costs When Long-Run Average Cost is (Locally) Falling.

The optimal toll equals the congestion externality at the optimum, and so toll revenue equals the area *ABYX*. Total cost, meanwhile, can be calculated as average cost times the number of trips, area $OCZN^*$, and total user costs as individual user cost times the number of trips, $OAXN^*$. Total capacity cost equals the difference between these two areas, area $ACZX$. Thus, toll revenue is insufficient to finance optimal capacity, which accords with the stated result since at N^* long-run average cost is locally decreasing.

Consider now the special case of constant long-run average cost. Since $LRMC = LRAC$ in this case, with marginal cost pricing the value of output PN equals $LRTC$, implying that profits, defined as the value of output minus total costs, are zero. Thus,

$$PN = C + K \Rightarrow PN - C = N\tau = K.$$

This line of argument indicates that the self-financing result stated in Proposition 1 is essentially the same as the well-known result in competitive equilibrium theory that with constant long-run average cost and competitive pricing (price equals marginal cost), a competitive firm's long-run profits equal zero.

When the user cost function is assumed to be homogenous of degree zero in N and s, one can say even more. Euler's Theorem states that for a function $g(x_1,...,x_n)$ which is homogenous of degree λ:

$$\sum_{i=1}^{n} x_i \frac{\partial g}{\partial x_i} = \lambda g.$$

Applying this theorem gives

$$N\partial f/\partial N + s\partial f/\partial s = 0. \tag{18.21}$$

From (18.3), $\partial C/\partial s = N\partial f/\partial s$; combining this with (18.14) and (18.21) yields

$$N^2\partial f/\partial N - sK'(s) = 0.$$

Finally, substituting (18.12) into the above equation gives

$$N\tau = sK'(s) = \varepsilon(s)K(s), \quad \text{where } \varepsilon(s) = \frac{K'(s)s}{K(s)}. \tag{18.22}$$

The expression on the left-hand side is toll revenue and the expression on the right-hand side capacity costs times the elasticity of capacity costs with respect to capacity. Thus, we have

Proposition 2. When the user cost function is homogenous of degree zero in N and s, the ratio of the revenue from the optimal toll to the cost of providing optimal capacity equals the (local) elasticity of capacity costs with respect to capacity.

This result indicates that with homogeneous user costs, the degree to which the road is self financing is determined by returns to scale in capacity provision. With decreasing (increasing) costs in capacity provision, the revenue raised from the optimal toll falls short of (exceeds) the amount needed to pay for the optimal capacity. Also,

Corollary 1. When the user cost function is homogenous of degree zero in N and s and there are constant costs of capacity, with optimal tolling and capacity the road is self-financing.

These are central results in transportation economics, and merit scrutiny. An obvious implication is that, in the first best, the magnitude of the government subsidy a transport mode merits depends on the degree of returns to scale in capacity provision, which in turn points to the importance for policy purposes of obtaining accurate empirical estimates of the degree of returns to scale in capacity provision. The way in which our model is specified masks an important subtlety. Since we worked with cost functions defined with factor prices fixed, the elasticity of capacity costs with respect to capacity, a property of a cost function, is simply the reciprocal of the degree of returns to scale in capacity provision, a property of a production function. But when factor prices vary with the scale of capacity provision, this simple relationship no longer holds. Which is crucial to the degree of self financing in the first-best optimum, the scale properties of the cost function or of the production function? This is the subject of a recent paper by Berechman and Pines (1991), who demonstrated that the answer is the production function. In other words, the correct way to determine the degree of self financing in the first-best optimum is to *hold factor prices fixed.*

In our model, capacity is a one-dimensional variable. But in actual policy situations, the degree of scale economies will depend on how scale is expanded — increasing the capacities of existing links versus adding new links, or in the context of mass transit, providing more frequent versus denser service. It can be shown (Kraus (1981)) that production efficiency entails equalizing the degree of scale economies on all active margins on which scale can be expanded. Thus, when production efficiency is realized, which it is in the first best, there is no ambiguity in the measurement of scale economies. In real-world situations, production efficiency cannot be expected, so the relevant measure of scale economies depends on how scale is to be expanded.

The self-financing results are robust, just as the optimal tolling results were earlier. The reason we have treated the first best at such length is that it provides a very valuable point of reference in deriving second-best results, which are the essence of optimal policy in realistic situations.

Second-best Analysis

First-best policy analysis examines the policy choices of a benevolent planner who is constrained only by resources and technology; in the previous section, the planner took as given the congestion technology, the technology of capacity provision, and resource availability. Since the analysis was partial equilibrium, resource availability was captured by prices, which in a first-best analysis are assumed to accurately capture the corresponding social opportunity cost.

In second-best policy analysis, be it partial equilibrium or general equilibrium, the planner faces other constraints as well, and these affect policy choices. These additional constraints can take many forms. Transactions costs, broadly interpreted, may rule out certain policy options; for example, it is often argued that congestion tolling on city streets is infeasible because implementation costs are excessive. Since transactions costs are difficult to quantify, they are normally treated implicitly rather than explicitly by ruling out certain policy options. There may be political constraints on the planner's actions; for instance, if he is an elected official, he may consider only those policies consistent with his reelection. The planner may face informational constraints. The best-known example of this (which was highlighted in James Mirrlees' and William Vickrey's Nobel Prize citations) is the limits to redistribution. Recall that one defense of the assumption that a dollar's worth of benefits is valued the same by the planner to whomever it goes was that the planner is at liberty to redistribute in lump-sum fashion, taking from the rich and giving to the poor until the social value of a dollar given to each is equalized. Such lump-sum redistribution presupposes that the planner is able to identify who is needy. Typically the planner cannot do this directly but must infer an individual's need from some signal over which the individual has some control. The most frequently identified signal is income. But if the government redistributes via income taxation, it will distort individuals' labor-leisure choices, which should be accounted for in the policy analysis. The list could be extended virtually *ad infinitum.*

In partial equilibrium analysis, these additional constraints may impinge directly either on the market under analysis or in some other market. The former case is easier to deal with since the general equilibrium repercussions of policy intervention in the market under examination can be ignored, on the assumption that rest of the economy is undistorted. The latter case is trickier. Suppose that the policy intervention is in market A while the distortion occurs in market B, with the rest of the economy undistorted. Then in policy analysis it is necessary to examine markets A and B simultaneously; repercussions in the rest of the economy can be ignored.

The theory of the second best was originally developed in the context of optimal taxation, but was soon applied to transport policy and remains central to its study.

Before proceeding to the theory of the second best in the context of transportation, we introduce the concept of deadweight loss. Deadweight loss, which is also known as efficiency cost or loss (and excess burden in the context of taxation), is the dollar cost of a distortion, or equivalently the loss in social surplus. Consider a single market in an undistorted economy. The optimal level of output occurs where marginal social cost equals marginal social benefit — Q^* in Figure 18.5. The social benefit provided by Q^* units of output equals the area under the marginal social benefit curve (which coincides with the demand curve under the assumptions for which partial equilibrium analysis is valid) up to Q^*, area $OBCQ^*$ in the diagram. Ignoring fixed costs, the social cost of Q^* units of output is the area under the marginal social cost curve up to Q^*, area $OACQ^*$ in the diagram. Thus, social surplus at the undistorted optimum is ABC, which has the interpretation as the net benefit to society from Q^* units of the good being provided rather than none. Suppose now that the market is supplied by a profit-maximizing monopolist who is constrained to charging all consumers the same price. The profit-maximizing output rule is then to choose output so that marginal revenue (with a linear demand curve, the marginal revenue curve is also linear and has the same price intercept and twice the slope of the demand curve) equals marginal cost, Q^m in the diagram. The monopolist's profit equals OP^mXQ^m (revenue) minus $OAYQ^m$ (cost). Social surplus equals social benefit $(OBXQ^m)$ minus social cost $(OAYQ^m)$: $ABXY$. The deadweight loss due to monopoly is the loss in social surplus under monopoly compared to the social optimum: $DWL = SS^{opt} - SS^m = ABC - ABXY = YXC$, the familiar monopoly deadweight loss triangle. The deadweight loss may be more simply and intuitively derived as follows: The distortion due to monopoly derives from a less than socially optimal level of output. Starting from the monopoly level of output, increase output one unit at a time from Q^m to Q^* and calculate the gain in social surplus. Increasing output by one unit from Q^m to $Q^m + 1$ increases social benefit by the height X', social cost by the height Y', and social surplus by the area $YXX'Y'$. Increasing output from Q^m to Q^* therefore increases social surplus by the area YXC.

Second-best policy minimizes the loss in social surplus due to distortions. The first application of second-best analysis to transportation was to the optimal pricing of mass transit when congestion tolling cannot be applied to auto travel (Lévy-Lambert (1968), Marchand (1968)). To begin, we make the assumptions that travel from A to B by auto (superscript a) and by mass transit (superscript m) are perfect substitutes (which implies that an individual chooses whichever mode is cheaper),

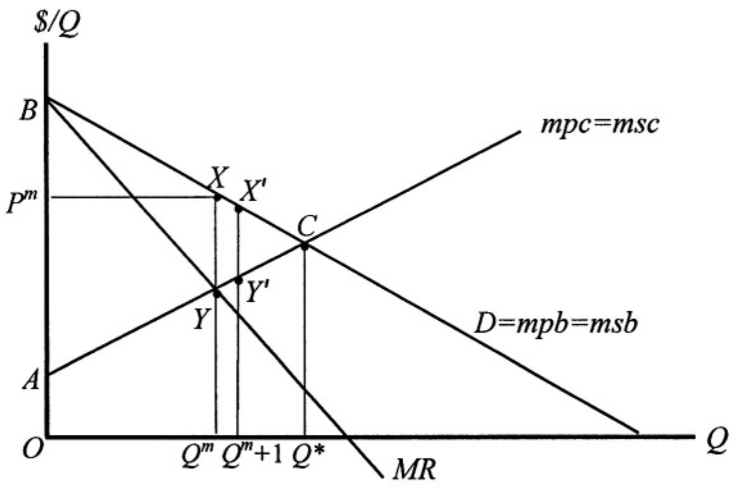


Figure 18.5 The Deadweight Loss from Monopoly

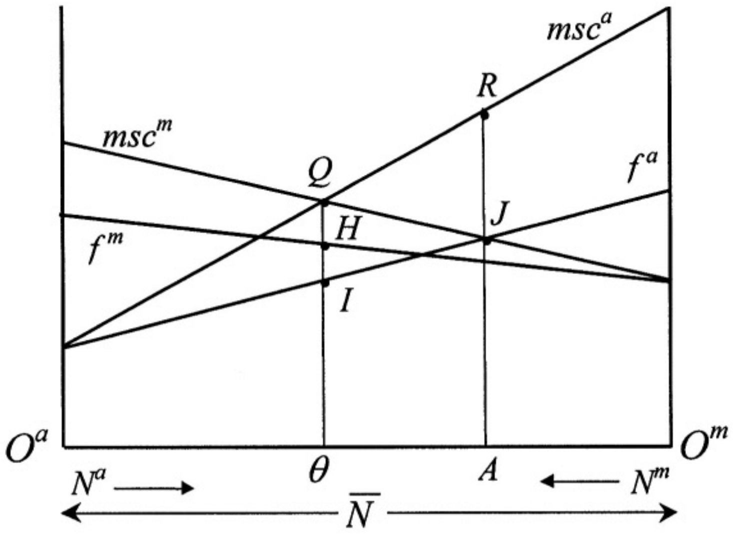


Figure 18.6 The Second-Best Transit Fare When Total Travel Demand is Completely Inelastic

that there is no congestion interaction between the two modes, and that the total demand for travel is completely inelastic — independent of price — and fixed at $\overline{N}$. Figure 18.6 displays the user cost and marginal social cost curves for each mode, with the mass transit cost curves drawn backwards and the origin for the mass transit mode placed such that $N^a + N^m = \overline{N}$. The total social cost of travel from A to B is minimized by allocating travellers over the two modes such that the marginal social cost of travel on each is equalized — the allocation θ. For any other allocation, the total social cost of travel can be reduced by transferring a traveller from the mode with the higher marginal social cost to the mode with the lower marginal social cost. Now suppose that auto congestion tolls cannot be imposed so that the price of an auto trip is the user cost, but that first-best pricing is employed in mass transit so that the mass transit price is given by the mass transit marginal social cost curve. Individuals will distribute themselves across the two modes such that trip prices are equalized, resulting in the allocation *A*. Because auto travel is underpriced $(P^a < MSC^a)$, too many individuals choose to travel by auto; the corresponding deadweight loss is *QRJ*. To correct this inefficient allocation of travel over the two modes, the mass transit fare should be reduced until the optimum allocation of travel between the two modes is achieved, which requires a negative transit fare of *IH*.[6] In this simplified situation, the first-best optimum can be achieved by this second-best policy.

Now consider the same model but with total travel demand sensitive to price. Reducing the mass transit fare below the mass transit congestion externality will then have two effects. First, as before, doing so will improve the efficiency of the modal split. But second, since both modes will then be underpriced, there will be excessive travel. One may envision one deadweight loss triangle associated with an inefficient modal split and another associated with excessive overall travel. The mass transit fare should be chosen to minimize the sum of the areas of the two deadweight loss triangles. This illustrates the principle that may be stated as "Two small deadweight loss triangles are better than one large one" or "The best way to deal with a distortion is typically to introduce an offsetting distortion."

Here, we consider only the simple case in which the two modes are perfect substitutes in demand and mass transit is not subject to congestion.[7] At the first-best optimum, the number of trips is such that the marginal social benefit of a trip equals the marginal social cost, and the modal split is such that the marginal social cost of a mass transit trip equals the marginal social cost of an auto trip. In Figure 18.7, these conditions are fulfilled with *OA* auto trips and *AB* mass transit trips, and would be achieved by setting a toll *XY* for auto and a zero mass transit fare; the corresponding social surplus would be *ZXWV*. Now suppose auto travel is untolled and that the mass transit fare is held at zero. All trips are made by car, the equilibrium number of trips is *OU*, and the social surplus is $ZQV - QRT$. The deadweight loss is *XWTR*, which can be decomposed into *XWF*, the deadweight loss due to an inefficient modal split, and *WFRT*, the deadweight loss from excessive travel. Now lower the mass

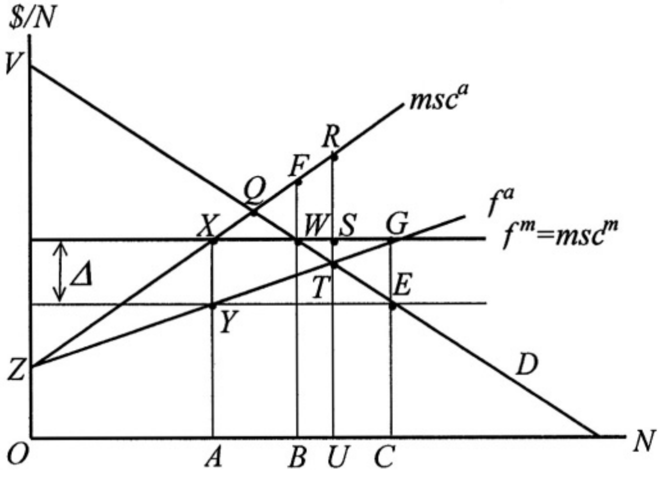


Figure 18.7 The Second-Best Transit Fare When Total Travel Demand Depends on Price

transit fare by *Δ*, which results in an efficient modal split but even more excessive travel. The deadweight loss is *WGE,* all of which is due to excessive travel. Now raise the mass transit fare a small amount. Doing so reduces the deadweight loss from excessive travel but has only a second-order effect on the deadweight loss from an inefficient modal split. Continue raising the mass transit fare until the marginal decrease in deadweight loss from excessive travel equals the marginal increase in deadweight loss from an inefficient modal split. This is the *second-best* mass transit fare.

A similar analysis can be applied to another problem in which there is a single mode, say auto, but a peak and an off-peak period. In the first best, the toll in each period is set equal to that period's congestion externality. Suppose, however, that the toll cannot be varied between peak and off-peak periods. What is the second-best toll with this *uniform toll constraint*? It should be raised to the point where the marginal decrease in deadweight loss from underpricing peak travel equals the marginal increase in deadweight loss from overpricing off-peak travel.

Another important second-best problem arises when there is a *deficit constraint.*[8] Consider, for example, a bridge whose construction is characterized by decreasing costs. With first-best optimal capacity and first-best tolling, the bridge would operate at a loss. Suppose, however, that by statute the bridge is required to break even. What is the second-best policy with this constraint? Shortly we shall consider how optimal capacity should be modified when pricing is distorted. But for the moment, we assume that the bridge has been built at first-best optimal capacity, and examine how the peak and off-peak tolls should be set to maximize social surplus subject to the break-even constraint. To simplify, we assume that peak travel occurs one-half the time, and off-peak travel the other half, and that peak and off-peak demands are independent. Denoting peak-period variables with a 1 subscript, and off-peak variables with a 2, the planner's problem is:

$$\max_{N_1,N_2} \int_0^{N_1} P_1(n_1)dn_1 + \int_0^{N_2} P_2(n_2)dn_2 - C(N_1,N_2,s^*) - K(s^*)$$
$$+ \lambda(P_1(N_1)N_1 + P_2(N_2)N_2 - C(N_1,N_2,s^*) - K(s^*)), \quad (18.23)$$

where s^* is the capacity that is common to the peak and off-peak periods, $C(N_1, N_2, s^*) \equiv N_1 f(N_1/s^*) + N_2 f(N_2/s^*)$ is total user cost over the demand cycle, and $K(s^*)$ is capacity cost over the cycle. The first-order conditions are

$$P_i - \partial C/\partial N_i + \lambda(P_i + N_i P_i'(N_i) - \partial C/\partial N_i) = 0; \quad i = 1, 2. \quad (18.24)$$

Rewriting (18.24):

$$(1+\lambda)(P_i - \partial C/\partial N_i) = -\lambda N_i P_i'(N_i) = -\lambda P_i / \varepsilon_i; \quad i = 1, 2, \quad (18.25)$$

where ε_i is the (absolute value of) elasticity of demand. Taking the ratio of (18.25) for $i = 1$ to $i = 2$ gives

$$\frac{(P_1 - \partial C/\partial N_1)/P_1}{(P_2 - \partial C/\partial N_2)/P_2} = \frac{1/\varepsilon_1}{1/\varepsilon_2}. \quad (18.26)$$

Equation (18.26) is the famous *Ramsey pricing formula,* which states that the proportional (to price) markup of price over marginal cost for the two time periods should be in inverse proportion to their demand elasticities. The intuition is as follows. The deadweight loss from a distortion derives from the change in behavior it induces. If demand is completely inelastic in say the peak period, raising the peak-period price above marginal social cost generates no distortion, and so the extra revenue needed to finance the bridge deficit should be raised from peak-period travellers. More generally, the efficient way to finance the bridge deficit is to set the prices such that the marginal deadweight losses from raising an extra dollar of revenue in the peak and off-peak periods are equal, which entails Ramsey pricing.

A couple of remarks are in order before moving on to discuss second-best optimal capacity. First, we remind the reader that the preceding analysis of second-best pricing has ignored distributional considerations. This is a defensible but not an innocuous assumption. If transit passengers are disproportionately poorer than auto users and if lump-sum redistribution is infeasible, the planner may want to take equity considerations into account in pricing. This can be done at varying levels of sophistication. The simplest is to apply distributional weights, not only to different groups in the population but also to the revenue raised by the government, with the distributional weight accorded government revenue depending on the distributional effects of government expenditure. Second, by making a number of simplifying assumptions, we have made second-best pricing appear straightforward and intuitive. But when the simplifying assumptions are relaxed, second-best pricing analysis is not only complex but also full of subtleties. For example, the analysis needs to take into account what information the government has and what policy instruments it has at its disposal. A good introduction to the theory is provided in Atkinson and Stiglitz (1980).

Second-best optimal capacity. To start, let us consider the classic second-best optimal capacity problem in the context of urban transportation, which was stated and partly solved by Wheaton (1978) and then more fully solved by Wilson (1983). Consider a single road in isolation. Solve for optimal capacity under the assumption that efficient pricing is employed and term this first-best optimal capacity. Now suppose that the road cannot be tolled, so that trip price equals user cost. What is optimal capacity taking this constraint into account — what is second-best optimal capacity, and how is it related to first-best optimal capacity?

To conceptualize this problem, imagine starting at the first best, with first-best pricing and first-best optimal capacity, and then removing the toll. Should the road be widened or narrowed? There are two effects. First, removing the toll will result in more cars and more congestion; *holding fixed the number of cars at this level*, the marginal benefit from increasing capacity is higher than in the first best. However, widening the road will lower trip price even further, which will cause the already excessive traffic to be even more excessive. The increase in traffic that occurs with a road expansion is termed the *latent demand* it induces.

This tradeoff is shown diagramatically in Figure 18.8. Expanding the road causes the user cost function to shift down from uc_0 to uc_1 and the marginal social cost function to shift down from msc_0 to msc_1. Consider first the benefit from the road expansion with optimal tolling. The equilibrium shifts from *B* to *B',* which results in an increase in social surplus (ignoring the additional capacity costs) — which is the benefit from expanding the road — of *ABB'*. With no toll, the equilibrium shifts from *C* to *C'*. Initially, the social surplus is *AHB– BGC*; after the road expansion, it is *AHB' – B'G'C'*. Thus, the gain in social surplus is *ABB' +*(*BGC– B'G'C'*)*= ABB'* + (*BGIB' –CIG'C'*). *BGIB'* is the amount by which the benefit from expanding capacity is higher in the no–toll equilibrium than

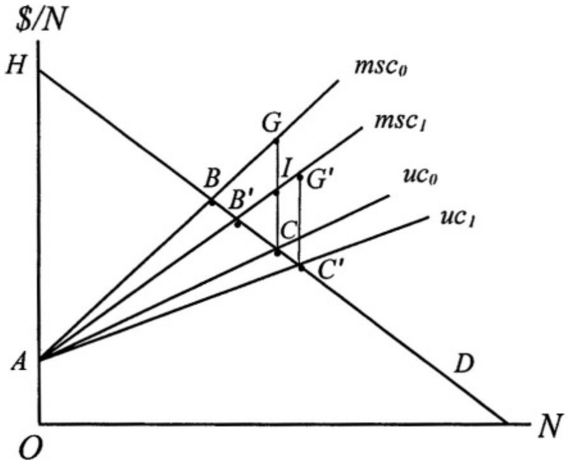


Figure 18.8 Changes in Equilibrium Resulting From Roadway Expansion

with the optimal toll, holding fixed the number of cars at the no-toll equilibrium. *CIG'C'* is the increase in deadweight loss due to the underpricing of car travel in the no-toll equilibrium ascribable to the latent demand induced by the road expansion. While the figure portrays linear demand and cost curves, the reader can visualize that with nonlinear curves, whether the analog to *BGIB'* is larger or smaller than the analog to *CIG'C'* depends in a complicated way on the curvature properties of the demand and cost curves.

We may obtain precise results with local, algebraic analysis. Specifically, we ask: Is the marginal social benefit (*msb*) from an infinitesimal road expansion larger or smaller when the toll is raised an infinitesimal amount? The answer is complicated, depending on the sign of

$$\frac{dmsb}{d\tau} = -\frac{Ne\varepsilon P(1+\gamma)}{s(P+e\varepsilon)^2} - (e-\tau)\frac{d}{d\tau}\left(\frac{dN}{ds}\right), \tag{18.27}$$

where e is the congestion externality, $\gamma \equiv (N/s)f''/f'$ the elasticity of marginal user cost, and other variables as previously defined. One important insight is that in the neighborhood of the first-best optimum (where $e = \tau$), the road should be widened when the toll is reduced; this is because the deadweight loss due to the latent demand generated by the toll decrease is of second order. Another is that since the expression depends on, among other things, how the demand elasticity varies with trip price, on which the empirical literature provides scant information, *local* analysis provides little practical guidance concerning the relative magnitudes of first- and second-best optimal capacity. Wilson however derives a *global* condition under which optimal capacity with underpricing exceeds the first-best level, that he argues can be expected to hold in practice.

The above analysis was complex enough. Calculation of second-best capacity in practical situations is even more difficult because of *network effects.* A good place to start in considering network effects is a road network *with optimal tolling everywhere on the network.* In this special case, *cost-benefit analysis of a link expansion need consider only that link.* The link expansion will change flows throughout the network, but since the marginal social benefit of a trip equals the marginal social cost everywhere on the network, the changes in flows induced by the link expansion have no effect on social surplus and can hence be ignored (This is an example of the Envelope Theorem). To determine whether a small link expansion is desirable, all that is required is to compare the direct benefit from the link expansion, ignoring the induced change in traffic on that and all other links, with the cost of the expansion.

Now consider the other extreme where no links of the network are tolled, in which case the marginal social benefit of a trip on a link does not generally equal the marginal social cost. Then the changes in flow throughout the network induced by a link expansion will affect social surplus and hence the benefit from the link expansion. There is no simple way around the problem. One must calculate social surplus in the pre-expansion network equilibrium, then estimate the post-expansion network equilibrium and calculate the corresponding social surplus, and then compute the benefit from the link expansion as the corresponding gain in social surplus.

The literature has identified numerous traffic paradoxes (see Arnott and Small (1994) for a simple exposition), where the addition of a link to a network or the expansion of an existing link would lower social surplus even if the improvement were costless, A useful way to think about these paradoxes is to decompose the change is social surplus generated by an improvement into a *direct benefit* and an *indirect cost.* The direct benefit is the gain in social surplus that would be achieved if all link flows were held fixed. The indirect cost is the loss in social surplus deriving from the changes in link flows induced by the improvement. The paradoxes illustrate that, with inefficient pricing, it is quite possible for the indirect cost to exceed the direct benefit. With efficient pricing, however, indirect cost is zero — for the reasons given earlier — so that traffic paradoxes never occur.

This raises an important point. If the prices are right, to determine the optimal capacity of a link the planner needs information pertaining only to that link; in contrast, if they are distorted, to determine simply whether an incremental expansion of a link is desirable, the planner needs to forecast how the link expansion alters flows through the entire network. Thus, if efficient congestion pricing is employed, substantially less information is needed to correctly evaluate transport investment projects.

Other second-best policy issues. The recent literature has focused on a number of other second-best policy issues. One is second-best optimal pricing given that only a subset of roads — typically urban freeways — can be tolled. Until recently, work on this problem assumed identical individuals. Based on numerical simulation with realistic parameter values, this work came to the pessimistic conclusion that the lion's share of the efficiency gains from tolling are lost when only partial coverage tolling is applied. The more recent literature (Verhoef and Small (2000)) treats heterogeneity in individuals' value of time and finds, again in carefully-calibrated simulations, that partial coverage tolling achieves a significantly larger proportion of the potential gains from tolling. The reason is that travellers with a high value of time benefit considerably from the tolling of urban freeways.

Another form of partial coverage congestion tolling is *cordon pricing,* under which a toll is charged to enter the central city. Singapore and the major Norwegian cities have cordon pricing, and it is being implemented in other cities. A major problem with cordon pricing is that many drivers choose to circumvent the toll, resulting in a substantial increase in congestion just outside the cordon. One study of the Singapore cordon found that the increase in travel costs outside the cordon actually exceeded the reduction in travel costs inside it. Recently, a Cambridge University team headed by David Newbery has been attempting to ascertain the determinants of optimal cordon placement for a selection of UK cities, using microsimulation with actual street networks (Santos, Newbery and Rojey (2001)).

One of the arguments against congestion tolling is that it is regressive since it essentially causes travellers to pay with money rather than with time. Urban transport economists have investigated two classes of schemes to make congestion tolling more progressive. The first (Small (1992a, 1993)) is to earmark the toll revenue and spend it in a way that disproportionately benefits the poor, typically by improving mass transit. The second (Verhoef and Small (2000)) is to deliberately congestion toll only a subset of roads or subset of lanes of a freeway. Those with high values of time choose the tolled alternative, those with low values of time the untolled alternative.

Thus far in the analysis, we have assumed that outside the transport sector the economy is free of distortions. But there are two distortions in the rest of the economy that are so obvious and so important that they should not be ignored. The first is that the revenue to finance government expenditure is raised in a distortionary manner. This is sometimes treated by assuming that the social cost of raising a dollar of revenue, which is termed the *marginal cost of funds,* exceeds $1.00 (a typical figure assumed is \$1.30). If mass transit is run at a deficit of *S*, the social cost of the revenue raised to finance the deficit is *mS*, where *m* is the marginal cost of funds. The second is related — the distortion in the labor/leisure choice caused by the income tax. Treating this explicitly essentially endogenizes the marginal cost of funds. By increasing the cost of working, congestion tolling exacerbates the labor-leisure distortion induced by the income tax. Calibrated numerical examples (Parry and Bento (1999), Calthrop et al. (2000)) suggest that this effect may be so quantitatively important that second-best congestion tolls — taking the labor/leisure distortion into account — may be negative.[9]

The asymmetric information revolution has not yet had a major impact in transportation economics, except through consideration of the interaction between income taxation and congestion tolling. There are however other policy contexts in which asymmetric information is clearly of central importance and to which it should therefore be applied. One is the design of optimal tendering mechanisms for the awarding of infrastructure construction and maintenance contracts and of optimal incentive mechanisms for performance under these contracts. Another is mass transit regulation. We have already considered a number of reasons why transit authorities do not simply first-best/marginal-cost price: equity, underpriced auto travel, and distortionary taxation. Another is that if the government were to commit to covering a transit authority's deficits, the transit authority would have no incentive to lower costs by introducing transport innovations, resisting union demands for higher wages, etc., or to improve service quality. In other areas, *incentive regulation* (Laffont and Tirole (1993)), which designs regulatory policy taking into account the incentives created for managers, has had a major impact; the same can be expected for urban mass transit.

No discussion of the application of second-best theory in the context of urban transportation would be complete without a discussion of *cost-benefit theory* and practice. Drèze and Stern (1985) provides an elegant treatment of cost-benefit theory. The basic idea is that with distortions, the social opportunity cost of commodities (including factors), which are termed *shadow prices,* deviate from market prices. Based on a general equilibrium model of the economy, which includes all distortions and a well-specified objective function, the central planner calculates the shadow prices of all commodities as the Lagrange multipliers on the corresponding market-clearing constraints in the optimization problem. He transmits these prices to project assessors, such as transit authorities, and instructs them to accept all projects that generate a shadow profit — a profit where shadow rather than market prices are employed. Most cost-benefit practice in transportation falls far short of this ideal, but does often replace two particularly important market prices with their shadow prices — shadow wage rates replacing market wage rates, and the social rate of discount replacing the interest rate. The network effects identified earlier are typically taken into account in the evaluation of large transportation projects, as should be done, but usually not in the evaluation of small projects, which is a serious shortcoming of current cost-benefit practice.

18.3 The Bottleneck Model

With the exception of the peak-load problem, all the analysis we have presented thus far is *atemporal.* The traditional model assumes that traffic is in a steady state, while the peak-load problem assumes that traffic is in different steady states during the peak and off-peak periods, with the two periods being related if at all only through cross-price effects in demand. This is unsatisfactory. Capacity is built for the peak of the rush hour, not for the steady state; furthermore, traffic evolves in a particular way over the rush hour, with the rush hour lengthening as the overall volume/capacity ratio increases. Thus, it is obviously desirable to develop an economic model of non-steady-state, rush-hour traffic. There are two difficulties in developing such a model. The first is to obtain an analytically tractable model of non-stationary-state traffic flow — not an easy task; the second, to incorporate individuals' decisions concerning when to travel.

Such a model was originally presented by William Vickrey, the dominant figure in the history of transport economics and a Nobel Prize winner, in a 1969 paper. The model, now known as the *bottleneck model,* was further developed by Arnott et al. (1993) and Braid (1989) and has subsequently been extended in numerous ways and has generated a host of insights. The first of the difficulties mentioned above was dealt with by modeling non-steady-state traffic flow as a queue behind a traffic bottleneck of fixed flow capacity. This representation may or may not be empirically accurate, depending on context, but does succeed in capturing the essentials of non-steady-state traffic flow at a conceptual level. The second difficulty was dealt with by making an assumption analogous to the Wardrop (1952) condition for route choice. The Wardrop condition is that individuals will choose what route to take so as to minimize the generalized trip price (which includes the value of time and tolls payable); the analogous Vickrey condition is that individuals will choose *when* to travel so as to minimize the generalized trip price defined to include as well the costs of travelling at inconvenient times, which have come to be termed *schedule delay costs.*

Let us now turn to the basic model. A fixed number, N, of identical commuters must travel from point A (home in the suburbs) to point B (work downtown) in the morning rush hour. All commuters have the same work start time, t^*. A and B are connected by a single road on which the only congestion that occurs is at a single bottleneck of fixed flow capacity s. If the arrival rate at the bottleneck exceeds s, a queue develops. The fact that the bottleneck's capacity is of the flow type means that not all commuters can arrive at work at t^*; practically all commuters have to incur schedule delay costs. To simplify the analysis somewhat, we assume that late arrival is prohibitively costly. We also ignore vehicle operating costs. As a result, when deciding when to travel, a commuter will trade off travel time costs, time early costs, and where applicable toll costs.

The natural case to start with is the no-toll equilibrium. Each commuter will choose to depart so as to minimize the generalized trip price, which now consists of travel time and time early costs. An *equilibrium departure pattern* is defined to be one in which no commuter can reduce his trip price by changing his departure time. Since commuters are identical, this implies that the trip price is uniform over the *departure interval.* To simplify the algebra, we assume that trip price, which equals user cost since there is no toll, is linear in travel time and time early:

$$P(t)=C(t)=\alpha T(t)+\beta(t^{*}-(t+T(t))), \tag{18.28}$$

where t is the departure time, $P(t)$ and $C(t)$ trip price and user cost as a function of departure time, $T(t)$ travel time, α the value of travel time, and β the value of time early.[10] Thus, $\alpha T(t)$ is travel time cost, $t + T(t)$ arrival time, $t^{*}-(t+T(t))$ time early and $\beta(t^{*}-(t+T(t)))$ time early cost. To further simplify the algebra, we ignore free-flow travel time. Accordingly, a commuter arrives at the bottleneck as soon as he leaves home, and arrives at work as soon as he gets. through the bottleneck. His only travel time is therefore time spent in the queue behind the bottleneck. Let $Q(t)$ be the number of cars in the queue (the queue length). Since the flow capacity of the bottleneck is s,

$$T(t)=\frac{Q(t)}{s}. \tag{18.29}$$

Substituting (18.29) into (18.28) yields

$$P(t)=C(t)=\frac{(\alpha-\beta)Q(t)}{s}+\beta(t^{*}-t). \tag{18.30}$$

The equilibrium trip price condition is $P(t)=\overline{P}$ over the departure interval. Substituting this condition into (18.30) gives an equation which indicates how queue length must evolve over the departure interval in order to satisfy the equal trip price condition. All that remains is to solve for the equilibrium departure interval and the equilibrium trip price. First, the bottleneck must be used to capacity over the departure interval; otherwise, a person could depart after the first person to depart, experiencing no travel time cost and lower time early cost, which would violate the Vickrey condition. Second, the last person to arrive must arrive exactly on time; if he were to arrive earlier, a commuter departing after him but still arriving early would experience a lower time early cost and no larger a travel time cost, which would violate the Vickrey condition; and late arival is not permitted. Let $\underline{t}$ be the time of the first departure, and $\bar{t}$ the time of the last departure. Since the first person to depart faces no queue and hence no travel time,

$$\bar{P} = \beta(t^* - \underline{t}); \tag{18.31a}$$

and since the queue must start at $\underline{t}$ and end at t^*,

$$t^* - \underline{t} = N/s; \tag{18.31b}$$

also, since the last person to depart arrives exactly on time,

$$\bar{t} + \frac{Q(\bar{t})}{s} = t^* . \tag{18.31c}$$

Combining these equations with (18.30) implies

$$Q(t) = \frac{\beta s}{\alpha - \beta}(t - \underline{t}) \quad \text{for } t \in [\underline{t}, \bar{t}]. \tag{18.32}$$

Over the departure interval, since queue length evolves according to

$$\dot{Q}(t) = r(t) - s \tag{18.33a}$$

where $r(t)$ is the departure rate,

$$r(t) = \frac{\alpha s}{\alpha - \beta} . \tag{18.33b}$$

The no-toll equilibrium is portrayed diagrammatically in Figure 18.9. The vertical distance between the cumulative departures and cumulative arrivals schedules at time t is the queue length at that time, and the corresponding horizontal distance is the queuing time for a commuter departing at t. The queue length increases linearly over the departure interval. The key insight is that the queue length evolves to satisfy the Vickrey condition, which in turn determines the equilibrium pattern of departure times.

Total user costs are $N\bar{P} = \beta N^2/s$, with all individuals having the same user cost $\bar{P} = \beta N/s$. Because of the linearity of the cost function in (18.28), total travel time costs equal total time early costs. Marginal social cost is $\frac{d(N\bar{P})}{dN} = \frac{2\beta N}{s}$. Several features of the equilibrium are noteworthy. First, in equilibrium user cost is a

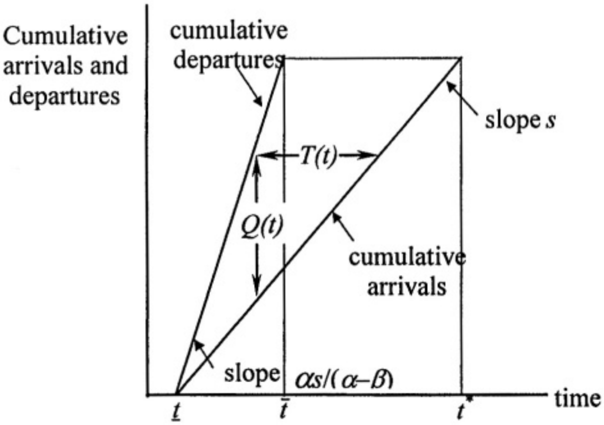


Figure 18.9 The No-Toll Equilibrium in the Bottleneck Model

function of *N/s,* just as it is in the basic static model. Thus, the basic static model may be interpreted as the reduced form of a more complex model which treats endogenously the equilibrium time pattern of congestion over the rush hour. Second, marginal social cost and the congestion externality are independent of departure time. This is remarkable since it might seem that the first person to depart generates the largest congestion externality by causing all drivers to face a queue length which is one car longer, with the last person by the same reasoning generating no congestion externality. The fallacy in this reasoning is that the addition of a commuter at any departure time causes the equilibrium departure time distribution to adjust such that the increase in social cost caused by the additional commuter is independent of when he departs; for example, an extra driver departing just before $\underline{t}$ causes all other commuters to face a queue that is $\beta/(\alpha-\beta)$ cars longer but no change in arrival time, and hence generates a congestion externality of $\beta N/s$, while an extra driver departing just before $\tilde{t}$ causes everyone to depart $1/s$ earlier, generating the same congestion externality. Third, the model endogenizes the length of the rush hour, thereby accounting for peak spreading.

Consider now the social optimum, again with a fixed number N of identical commuters. The planner chooses the departure pattern so as to minimize total user costs, which equal total travel time costs plus total time early costs. Total user costs are certainly minimized if each of these two components is minimized. A departure rate of s from $\underline{t}$ to t^* minimizes both. No queue forms so that total travel time costs equal zero. And since the length of the arrival interval is minimized, the bottleneck being utilized to capacity over the rush hour, and since the arrival interval is as close to t^* as possible subject to the constraint of no late arrivals, total time early costs are minimized, equaling $\beta N^2/2s$.

Thus, in the social optimum, total time early costs are the same as in the no-toll equilibrium, and total user costs are one-half those in the no-toll equilibrium. The deadweight loss in the no-toll equilibrium therefore equals the total travel time costs, which equal one-half of total user costs. The model is very simple but indicates starkly the potentially very sizeable efficiency gains that could be achieved in switching from the no-toll equilibrium departure pattern to the socially optimal one.

The socially optimal departure pattern can be supported as an equilibrium by imposing a time-varying toll of $\tau(t) = \beta(t - \underline{t})$ over the departure interval. Essentially, travel time costs are replaced by toll costs, so that trip price is the same in the social optimum as in the no-toll equilibrium. In contrast to the no-toll equilibrium, however, in the social optimum on average one-half of the trip price is collected as toll revenue, which can be redistributed or spent on expanding capacity so as to make everyone better off.

Optimal first- and second-best capacity can be analyzed straightforwardly. Suppose that there is a constant marginal cost of providing capacity, k. Then first-best capacity minimizes $\beta N^2/2s + ks$, while second-best capacity minimizes $\beta N^2/s + ks$. Thus, second-best capacity is larger by a factor of $\sqrt{2}$. This is consistent with the results obtained from (what we have argued may be interpreted as) the reduced-form model treated in the previous section, since our assumption of a fixed number of commuters precludes latent demand.

Because of its simplicity, the bottleneck model has been enriched analytically in numerous ways to treat elastic demand, simple networks, heterogeneous commuters (differing in α, β *and* t^*), uncertainty in demand and realized capacity, and driver information systems, and has been applied to examine the full range of second-best problems, including situations where there are constraints on the time variaiton of the toll (the optimal uniform and one-step toll, etc.) and on the set of roads that can be tolled (see Arnott et al. (1998) for a review).

Analyzing the class of bottleneck models has contributed considerably to the understanding of the *economics* of non-stationary-state traffic flow. But is the congestion technology it assumes sufficiently realistic to make it useful as a traffic planning tool? The evidence is not yet in. Small (1992b) provides evidence that average travel time on a section of highway is approximately a linear function of flow, and Daganzo et al. (1999) in a recent paper provides strong evidence of queues developing behind freeway bottlenecks that have stable locations and stable discharge rates, both of which are consistent with the bottleneck model. However, there are many observed traffic flow phenomena that cannot be explained by the bottleneck technology — gridlock, which derives from the physical length of queues that are ignored in the bottleneck model; the deceleration of cars on approaching a bottleneck and their acceleration on leaving it; turbulence just beyond entry ramps; the nodal congestion at intersections; the congestion on city streets caused by cruising for parking; etc. Nevertheless, it seems likely that dynamic network simulation models of non-stationary-state traffic flow with treatments of congestion that enrich the basic bottleneck technology will be heavily used by the next generation of traffic planners. An example is QUATUOR, a simulation model of traffic congestion in the Paris metropolitan area that is being developed by a team being headed by André de Palma at the University of Cergy-Pontoise.

Thus far in the chapter, we have presented the central, canonical model of transport economics, emphasizing economic principles and intuition. This model has proved to be very valuable, providing a unified conceptualization that accommodates most aspects of transport economics, and because of its analytical simplicity, permitting a rich and varied set of extensions and applications. At the same time, the power of the model has caused transport economic theorists to slight aspects of transportation that do not fit neatly into its conceptualization, and to hark on principles at the expense of the practical detail that is so necessary in policy application. One example is the treatment of congestion. By treating congestion as technological, they have ignored behavioral aspects of congestion, which their training makes them well-suited to analyze; for instance, how traffic flow depends on individual driving decisions which are based in part at least on the tradeoff between travel speed (relative to the mean speed of traffic) and the probability of accident. By focusing on simple specifications of link flow congestion, they have tended to overlook stock (e.g., parking) and nodal (e.g., intersection) congestion and to ignore the complexity and variety of congestion phenomena, and therefore to considerably underestimate the practical difficulties of applying their panacea — congestion pricing. And by concentrating on individual travel, especially in the urban context, they have neglected freight transportation. Another example is the treatment of traffic-related pollution and traffic accidents. Transport economists have tended to subsume both under the rubric of congestion and to apply the canonical model in analyzing them, thereby ignoring their particularities.

18.4 Mass Transit

Most of urban transport economic theory has been developed in the context of urban auto travel. This section investigates models of mass transit, with the aim of demonstrating that the general principles developed in the previous sections in the context of auto travel carry over to mass transit.

There are three obvious differences between mass transit and auto travel. First, in mass transit, unless all passengers are perfectly informed of timetables and trains run exactly on time, waiting time at transit stops needs to be considered, for which there is no analog in auto travel. Second, while in auto travel there were only two policy variables, the toll and capacity, in mass transit there are the fare, number of trains, train capacity, scheduling, and perhaps others. Third, in auto travel, the short run and long run have natural interpretations, with capacity being fixed in the former and variable in the latter, but in mass transit, with so many margins of policy choice, the distinction is not as clear cut.

We start with a stripped-down version of Mohring's (1972) classic model of steady-state urban bus travel. Buses travel round a circle of arbitrary radius. *N* passengers arrive at the circle per mile-hour, each of whom travels *m* miles. It is assumed that they do not know the bus timetable, and so arrive at a uniform rate. Buses themselves are uncongestible. The congestion that occurs derives from the time it takes a passenger to get on and off the bus, δ. Each passenger therefore boards the first bus that passes by after she arrives at the circle. A passenger's value of waiting time is v_w and value of transit time v_t.[11]

The transit authority operates *b* buses per mile at a cost per bus-hour (which includes operating and capital costs) of *c*. A bus' speed when in motion is *v*. The service frequency *f* is endogenous.

The transit operator's short-run problem is to choose the efficient bus fare, and its long-run problem is to choose the number of buses as well. Consider first the short-run problem. With some elasticity of demand, short-run efficiency entails each passenger facing a trip price equal to short-run marginal (social) cost. Since trip price equals the user cost plus the fare while short-run marginal cost equals the user cost plus the congestion externality, this condition is equivalent to the fare being set to equal the congestion externality. To obtain the congestion externality, we first derive short-run total variable costs (*TVC*) per mile-hour, which comprise total waiting time costs (*TWC*) and total transit time costs (*TTC*) per mile-hour. Now

$$TWC = \frac{Nv_w}{2f}, \tag{18.34}$$

since the average wait for a bus is 1/2*f*. Also,

$$TTC = Nv_t m(N\delta/f + 1/v); \tag{18.35}$$

since each bus picks up and drops off *N*/*f* passengers per mile, its travel time per mile is $N\delta f + 1/v$, resulting in a transit time cost of $v_t m(N\delta f + 1/v)$ per passenger.

Finally, since there are b buses per mile and bus travel time per mile is $N\delta f + 1/v$, the time headway between buses is $(N\delta/f + 1/v)/b$, so that service frequency solves

$$(N\delta/f + 1/v)/b = 1/f.$$

This yields

$$f = (b - N\delta)v \equiv f(N,b), \tag{18.36}$$

where $f_N < 0, f_b > 0$. Using (18.36) in (18.34) and (18.35) gives

$$TVC = \frac{Nv_w}{2f(N,b)} + Nv_t m(N\delta/f(N,b) + 1/v), \tag{18.37}$$

which yields

$$AVC = UC = \frac{v_w}{2f(N,b)} + v_t m(N\delta/f(N,b) + 1/v) \tag{18.38}$$

and

$$SRMC = UC - \frac{Nv_w f_N}{2f^2} + \frac{v_t mN\delta}{f} - \frac{N^2 v_t m\delta f_N}{f^2}. \tag{18.39}$$

The last three terms in (18.39) constitute the congestion externality. The first is the waiting time externality, which operates via service frequency; the second is the direct transit time externality that derives from a passenger delaying the other Nm/f passengers on a bus by δ when getting on and off the bus; and the third is another component of the transit time externality, which operates via service frequency. Appending a demand side $N = N(P)$, where P denotes the trip price, yields the short-run equilibrium.

In order to solve the long-run problem, b must be set to minimize total costs. Writing total costs as

$$\begin{aligned} TC(N,b) &= TVC + cb \\ &= \frac{Nv_w}{2f(N,b)} + Nv_t m(N\delta/f(N,b) + 1/v) + cb \end{aligned} \tag{18.40}$$

highlights the tradeoff in the determination of the efficient number of buses. Buses should be added up to the point where the marginal benefit of a bus, deriving from the increase in frequency of service it induces, just covers the marginal cost c.

Minimizing (18.40), after substitution of (18.36), with respect to b yields[12]

$$b = N\delta + \sqrt{(Nv_w + 2N^2 v_t m\delta)/2cv}. \tag{18.41}$$

Substituting this into (18.40) yields the long-run total cost relationship:

$$LRTC(N) = N(\delta c + v_t m/v) + \sqrt{2c(Nv_w + 2N^2 v_t m\delta)/v}, \tag{18.42}$$

from which it can be seen that there are decreasing long-run average costs which operate through waiting time. There are two important insights from the preceding analysis. The first is that the mass transit problem has essentially the same reduced form as the canonical auto model; the short-run cost function can be written as a function of the number of travellers and capacity variables, which in the preceding analysis was simply the number of buses; the long-run cost function can be written as a function of only the number of travellers; and the same diagrammatic analysis applies. The second important insight is that mass transit travel is characterized by *economies of density.* In the model, an increase in passenger density led to increased service frequency and correspondingly reduced waiting time. But there are other sources of economies of density in mass transit systems. Increasing density makes it optimal to decrease the spacing between stops and to increase the number of routes, both of which lead to reductions in walking time, and to operate longer trains and larger buses, which take advantage of technological returns to scale in vehicle construction and operation.

We demonstrated earlier that first-best efficient pricing and capacity provision with economies of scale leads to operation at a financial loss. This provides one rationale for subsidizing mass transit. Other rationales include equity (since mass transit travellers are on average poorer than auto travellers) and second-best considerations, which were discussed earlier. Thus, there are economically sound justifications for mass transit systems operating at a loss. This does not mean, of course, that all losses are defensible on economic grounds. Indeed, a considerable proportion of the losses of actual transit systems are waste, deriving from political patronage, union labor, and inefficiency deriving from poor incentives.

The model presented in this section was pedagogical in nature. Actual mass transit systems are evidently considerably more complicated. Running a bus system, for example, entails choices with respect to not only the number of buses but also their capacity, comfort, acceleration, servicing and cleaning, as well as routing, scheduling, and scrappage. The basic principles still apply, however.

In the previous section, in our examination of highway bottlenecks, we showed that the steady-state highway model can be interpreted as a reduced form representation of a dynamic model which accounts for the pattern of traffic over the rush hour. Kraus and Yoshida (2002) develops a dynamic model of an urban rail system and show that an analogous result holds.

18.5 Conclusion

Economists are experts in the efficient allocation of scarce resources. Crudely put, they are trained how to maximize the size of the pie whose division is decided by a combination of special interest groups operating through government, firms with market power, and anonymous market forces. Policy decision making in transportation – from traffic signing to freeway pavement thickness to parking policy – has tended to be dominated by engineers who until recently at least have not typically been well trained in economics. The result has been grossly inefficient transportation policy; examples include irrational pricing (such as removing tolls from a bridge when it has paid for itself), the excessive building of highways due to using faulty cost-benefit analysis which ignores that a considerable portion of the potential benefits are dissipated through the combination of the underpricing of auto travel and latent demand, the application of uniform engineering standards to widely different roads, and a failure to incorporate flexibility into transport system design due to improperly accounting for uncertainty.

Economists, in turn, have themselves to blame for their lack of influence in transport policy circles. Through preoccupation with simple models which were designed to elucidate basic principles and not to confront nuts-and-bolts policy issues, their advice is often too abstract and divorced from practical detail to be helpful in specific policy applications. Their mantra of "congestion pricing," without thought being given to the political, engineering, and practical problems associated with its implementation, is a prime example. Their advice would be more useful if they were to become better acquainted with the transportation science and traffic engineering literatures and to devote more of their attention to policy at the level of detail at which most transport policy decisions are made.

This chapter has presented the central principles of transport economics through a series of very simple models which illustrate the principles starkly. The world is far more complex than the models, and most policy is made at a level of detail that the models ignore. Knowledge of these general economic principles is essential for efficient and enlightened policy decisions. But so too is the knowledge contained in the traffic engineering and transportation science literatures and the practical experience of traffic engineers and transportation scientists. Hopefully in the years ahead, improved communication and cooperation between economists on the one hand and traffic engineers and transportation scientists on the other will lead to better transport policy.

18.6 Notes

1. Economists term travel under these conditions *congested* travel, and travel under conditions with higher densities for which flow decreases with density *hypercongested* travel. This terminology differs from that employed by traffic engineers, who term the former uncongested travel and the latter congested travel.

   Much ink has been spent by economists in attempting to extend the steady-state analysis to hypercongested travel. The current view is that hypercongestion is an intrinsically non-steady-state phenomenon to which steady-state analysis cannot fruitfully be applied.
2. These relationships were first suggested by Walters (1961). Johnson (1964) expands on his treatment.
3. Economists distinguish between *compensated* and *uncompensated* demand functions. A rise in the price of a good makes an individual worse off. Her uncompensated demand function describes how her quantity demanded changes with price, with no compensation for being made worse off from the price rise. Her compensated demand function describes how her quantity demanded changes with price, when she is compensated by an amount such that the price rise with the compensation makes her neither better nor worse off. We shall ignore the distinction between the two types of demand functions, which is strictly valid only when the demand for trips is independent of income. In microeconomic theory, demand functions are typically derived from utility functions. Here we treat the demand function as a primitive.
4. What we have termed "the optimum," transportation scientists refer to as the "system optimum." And what we have termed the "no-toll equilibrium," transportation scientists refer to as the "user optimum."
5. Thus, for example, our analysis extends to trucks transporting freight. Transport economists tend to be rather cavalier in their treatment of motorized vehicles other than autos, treating them as so many auto-equivalents. The form of the congestion interaction between autos and other motorized vehicles is actually considerably more complex.
6. Treating the impracticability of employing a negative fare would introduce another constraint.
7. Sherman (1971) relaxes both of these assumptions and allows the modes to have interdependent cost functions.
8. Both this and the uniform toll constraint are analyzed in Mohring (1970).
9. These examples assume that travel is more complimentary to work than to leisure, which is by no means obvious.
10. We assume that $\alpha > \beta$, which accords with empirical evidence in Small (1982).
11. There is a tendency to treat the value of time as a parameter which is exogenous to the transport policy maker. It should be kept in mind, however, that since an individual's travel time cost on a trip is the product of the value of

time and travel time, travel time cost can be halved not only by halving travel time but also by halving the value of time, which can in principle be achieved by making the trip more enjoyable.

12. Equation (41) implies that as N increases, b should increase somewhat more than proportionately than $\sqrt{N}$. When the effect of an increase in N on bus travel time is ignored, (41) simplifies to a pure square-root relationship, which is known as the *square-rootprinciple* (Mohring (1972)).

18.7 References

Arnott, R., de Palma, A. and Lindsey, R. (1993). A structural model of peak period congestion: A traffic bottleneck with elastic demand. *American Economic Review* **83**, 161-179.

Arnott, R., de Palma, A. and Lindsey, R. (1998). Recent developments in the bottleneck model. *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* (K. Button and E. Verhoef, eds.), Edward Elgar, Aldershot, England.

Arnott, R. and Small, K. (1994). The economics of traffic congestion. *American Scientist* **82**, 446-455.

Atkinson, A.B. and Stiglitz, J.E. (1980). *Lectures on Public Economics.* McGraw-Hill, New York.

Berechman, J. and Pines, D. (1991). Financing road capacity and returns to scale under marginal cost pricing. *Journal of Transport Economics and Policy* **25**, 177-181.

Braid, R.M. (1989). Uniform versus peak-load pricing of a bottleneck with elastic demand. *Journal of Urban Economics* **26**, 320-327.

Calthrop, E., Proost, S. and Van Dender, K. (2000). Optimal road tolls in the presence of a labor tax. Working paper.

Daganzo, C.F., Cassidy, M.J. and Bertini, R.L. (1999). Possible explanations of phase transitions in highway traffic. *Transportation Research* **33A**, 365-379.

Drèze, J. and Stern, N. (1985). The theory of cost-benefit analysis. *Handbook of Public Economics,* Vol. **2** (A.J. Auerbach and M. Feldstein, eds.), North-Holland, Amsterdam.

Johnson, M.B. (1964). On the economics of road congestion. *Econometrica* **32**, 137-150.

Laffont, J-J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation.* MIT Press, Cambridge, MA.

Lévy-Lambert, H. (1968). Tarification des services à qualité variable: Application aux péages de circulation. *Econometrica* **36**, 564-574.

Kraus, M. (1981). Scale economies analysis for urban highway networks. *Journal of Urban Economics* **9**, 1-22.

Kraus, M. and Yoshida, Y. (2002). The commuter's time-of-use decision and optimal pricing and service in urban mass transit. *Journal of Urban Economics* **51**, 170-195.

Marchand, M. (1968). A note on optimal tolls in an imperfect environment. *Econometrica* **36**, 575-581.

Mohring, H. (1970). The peak load problem with increasing returns and pricing constraints. *American Economic Review* **60**, 693-705.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *American Economic Review* **62**, 591-604.

Mohring, H. and Harwitz, M. (1962). *Highway Benefits: An Analytical Framework.* Northwestern University Press, Evanston, IL

Parry, I. and Bento, A.M. (1999). Revenue recycling and the welfare effects of road pricing. *Policy Research Working Paper* 2253. World Bank, Washington, DC.

Santos, G., Newbery, D.M. and Rojey, L. (2001). Static vs. demand sensitive models and the estimation of efficient congestion tolls: An exercise for eight English towns. *Transportation Research Record* **1747**, 44-50.

Sherman, R. (1971). Congestion interdependence and urban transit fares. *Econometrica* **39**, 565-576.

Small, K.A. (1982). The scheduling of consumer activities: Work trips. *American Economic Review* **72**, 467-479.

Small, K.A. (1992a). Using the revenues from congestion pricing. *Transportation* **19**, 359-381.

Small, K.A. (1992b). *Urban Transportation Economics.* Harwood Academic, Chur, Switzerland.

Small, K.A. (1993). Urban traffic congestion: A new approach to the Gordian knot. *Brookings Review* **11**, 6-11.

Strotz, R.H. (1965). Urban transportation parables. *The Public Economy of Urban Communities* (J. Margolis, ed.), Resources for the Future, Washington, DC.

Verhoef, E.T. and Small, K.A. (2000). Product differentiation on roads: Constrained congestion pricing with heterogeneous users. Working paper.

Vickrey, W.S. (1969). Congestion theory and transport investment. *American Economic Review Proceedings* **59**, 251-260.

Walters, A.A. (1961). The theory of measurement of private and social cost of highway congestion. *Econometrica* **29**, 676-699.

Wardrop, J.G. (1952). Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers,* Part II, Vol. **1**, 325-378.

Wheaton, W.C. (1978). Price-induced distortions in urban highway investment. *Bell Journal of Economics* **9**, 622-632.

Wilson, J.D. (1983). Optimal road capacity in the presence of unpriced congestion. *Journal of Urban Economics* **13**, 337-357.

BIOGRAPHIES

Moshe E. Ben-Akiva (Chapter 2) is the Edmund K. Turner Professor of Civil and Environmental Engineering at the Massachusetts Institute of Technology (MIT) and Director of the MIT Intelligent Transportation Systems Program. He has developed discrete choice methods and behavioral demand model systems and has supervised the development of two traffic simulators: MITSIMLab and DynaMIT. Dr. Ben-Akiva has co-authored over one hundred and fifty published research papers and two books, including the textbook *Discrete Choice Analysis,* published by MIT Press in 1985. He is the recipient of honorary doctor degrees from the Université Lumière Lyon, France and the University of the Aegean, Greece. Dr. Ben-Akiva serves as the Editor-in-Chief of *Transport Policy: The Journal of the World Conference on Transport Research Society,* and as Associate Editor for *Transportation Science.* Dr. Ben-Akiva has worked as a consultant in industries such as transportation, telecommunications, financial services and energy. He is a Principal and member of the Board of Directors of Cambridge Systematics and a Senior Advisor to RAND Europe.

Richard Arnott (Chapter 18) received his S.B. in civil engineering from M.I.T. in 1969 and his Ph.D. in economics from Yale in 1975, and is currently Professor of Economics at Boston College. His research has focused on urban economics, and in recent years on the economics of urban traffic congestion. Having completed a long-term research project with Andre dePalma and Robin Lindsey elaborating the bottleneck model of rush-hour traffic dynamics, he is currently working on parking policy and the interaction between urban land use and traffic congestion. His interest in urban transportation was kindled by Marvin Mannheim and later rekindled by William Vickrey.

Cynthia Barnhart (Chapter 14) is a Professor in the Civil and Environmental Engineering Department and serves as Co-Director of the Center for Transportation and Logistics at the Massachusetts Institute of Technology. She has developed and teaches courses including Carrier Systems, Optimization of Large-Scale Transportation Systems, Airline Schedule Planning and the Airline Industry. Her research activities have focused on the development of planning models and algorithms to improve carrier operations, particularly airlines. Her work has been published in several books and scholarly journals. She has served as an Associate Editor for *Operations Research* and *Transportation Science,* as a Board member for INFORMS, and as a liaison between the INORMS Transportation Science Section and the INFORMS Aviation Applications Special Interest Group. She has been awarded the Mitsui Faculty Development Chair, the Junior Faculty Career Award from the General Electric Foundation and the Presidential Young Investigator Award from the National Science Foundation.

Martin Beckmann (Chapter 9) is Professor Emeritus of Economics at Brown University and Professor Emeritus of Applied Mathematics at the Technical University of Munich. He has authored 14 books, his latest *Lectures on Location Theory,* to be published by Springer-Verlag in 1999. As a research associate of the Cowles Commission, the University of Chicago, he worked under T.C. Koopmans on *Studies in the Economics of Transportation* (with C.B. McGuire and C. Winsten) and continued his interest in Operations Research and Transportation during his active years at Yale University (1956-59), Brown University (1959-89), the University of Bonn (1962-69) and the Technical University of Munich (1969-89). From 1969-89 he was a consultant to the Transportation Science Department at the General Motors Research Laboratories. He makes his home in Providence, Rhode Island, and Munich, Germany.

Dr. Chandra R. Bhat (Chapter 3) is an Associate Professor of Civil Engineering at The University of Texas at Austin, where he teaches courses in transportation systems analysis and transportation planning. Dr. Bhat has contributed toward the development of advanced econometric techniques for travel behavior analysis, including discrete choice models, discrete-continuous econometric systems, and duration models. His research interests include land-use and travel demand modeling, policy evaluation of the effect of transportation control measures on mobility and mobile-source emissions, marketing research of competitive positioning strategies for transportation services, air travel behavior modeling, individual activity pattern analysis, and use of non-motorized modes of travel. Dr. Bhat serves on the editorial boards of *Transportation Research B* and *Transportation.* He is also on the editorial review board of the *International Journal of Operations and Quantitative Management.* He is the Chairman of the TRB Committee on Passenger Travel Demand Forecasting (A1C02), and serves on several other Transportation Research Board Committees. He is the secretary and treasurer of the International Association of Travel Behavior Research (IATBR), and is a member of the Board of Directors of this association. Dr. Bhat received his PhD in Civil Engineering from Northwestern University in 1991.

Michel Bierlaire (Chapter 2) is Maître d'Enseignement et de Recherche at the Ecole Polytechnique Fédérale in Lausanne (EPFL), and Research Affiliate at the Massachusetts Institute of Technology (MIT). He teaches operations research, simulation and optimization to undergraduate, graduate and post-graduate students.

Dr. Bierlaire has been active in transportation research for the last 10 years, focussing on mathematical models for transportation demand modeling, and on real-time systems for Intelligent Transportation Systems. He is the author of HieLoW and Biogeme, specialized packages for the estimation of discrete choice models. Dr. Bierlaire is also a member of the Editorial Advisory Board of *Transportation Research B,* and a reviewer for *Transportation Science* and the *Journal of Mathematical Programming..*

Lawrence Bodin (Chapter 12) is a Professor Emeritus of Decision and Information Technologies in the Robert H. Smith School of Business at the University of Maryland in College Park. His main research interests are in the areas of network optimization, large scale optimization models, transportation planning and logistics, vehicle routing and scheduling and the use of multi-criteria decision analysis in sports applications. He has consulted for many organizations including the United States Postal Service, Federal Express and United Parcel Service. He has served on the editorial boards of several professional journals and has presented tutorials and workshops on vehicle routing and geographic information systems at numerous meetings. Profesor Bodin received his PhD in Industrial Engineering and Operations Research from the University of California at Berkeley in 1967.

Arnab Bose (Chapter 7) is a Senior Research Engineer at Real-Time Innovations, Sunnyvale, California. He received his B.Tech. in EE from Indian Institute of Technology, Kharagpur in 1996. He received his M.S. and Ph.D., both in EE from University of Southern California, in 1998 and 2000, respectively. His research interests include object-oriented modeling, design and development of control systems for embedded and real-time systems, distributed systems, intelligent vehicle and highway systems, hybrid and discrete event systems. Dr. Bose has authored several technical papers and reports and has published in the *Society of Automotive Engineers Journal* and the *Transportation Research Record.* Dr. Bose was awarded the Best Presentation Award at the 1999 American Control Conference and was a recipient of the Dean's Doctoral Merit Fellowship from the School of Engineering at University of Southern California. He is a member of the IEEE Computer Society and the IEEE Control Systems Society.

Michael Cassidy (Chapter 6) is Assoicate Professor in the Department of Civil and Environmental Engineering at University of California at Berkeley. He received a doctorate in Civil Engineering (majoring in Transportation Engineering) from the University of California, Berkeley in 1990. He served for 3.5 years as an Assistant Professor of Civil Engineering at Purdue University in West Lafayette, Indiana before joining the Berkeley faculty in 1994. His research interests are in transportation operations, particularly the empirical study of highway traffic.

Amy Cohn (Chapter 14) received her doctorate in Operations Research from the Massachusetts Institute of Technology in 2002. She subsequently joined the faculty at the University of Michigan, where she is an assistant professor in the department of Industrial and Operations Engineering. Her current research focus is on modeling and solution techniques for large-scale problems in transportation and logistics.

Teodor Gabriel Crainic (Chapter 13) is Professor of Operations Research in the Dept. of Management and Technology of the Université du Québec à Montréal, and adjunct Professor at the Dept, of Computer Science and Operations Research of the Université de Montréal and the Dept. of Quantitative Logistics of Molde College, Norway. His research interests are in operations research models, exact and metaheuristic methods, and planning tools applied to transportation, logistics, e-business, and telecommunications, as well as the study of parallel computing and its impact on the design of models and algorithms. He has authored or coauthored over eighty scientific articles and coauthored STAN, a method and interactive-graphic software for strategic planning of multimodal multicommodity transportation systems used in over 30 organizations in 16 countries. Dr. Crainic co-founded the TRISTAN (TRienial Symposium on Transportation Analysis) series of international meetings and served as Director of the Centre for Research on Transportation (Montréal), president of the Transportation Science Section of INFORMS, and Associate Editor for *Operations Research.* He is North American Editor of the *International Journal of Mathematical Algorithms,* Area Editor for the *Journal of Heuristics,* and serves on the editorial boards of several other operations research and transportation journals.

Mark S. Daskin (Chapter 10) is a Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is the immediate past chair of the department. Dr. Daskin's research interests include facility location models and algorithms, transportation planning, supply chain management, and production planning. He is the author of approximately fifty refereed papers in these fields as well as the text *Network and Discrete Location: Models, Algorithms andApplications* (John Wiley, 1995). He is the immediate past editor-in-chief of *Transportation Science.* In 1989-90 he was a visiting professor in the Department of Statistics and Operations Research at Tel Aviv University funded on a Fulbright Research Grant. He is the editor-in-chief of *IIE* Transactions, serves on the editorial board of advisors of *Transportation Science* and is on the editorial board of *The International Journal ofLogistics Management.* He is the immediate past Vice President of publications of INFORMS, the Institute for Operations Research and the Management Sciences.

Leonard Evans (Chapter 4) is President of *Science Serving Society* (http://www.scienceservingsociety.com). an organization he formed to continue research and other professional activities after completing a 33-year research career with General Motors Corporation. He has a bachelors degree in physics from the Queen's University of Belfast, Northern Ireland, and a doctorate in physics from Oxford University, England. Dr. Evans' 143 publications appear in 40 different technical journals. In 1991 his widely acclaimed influential book *Traffic Safety And The Driver* was published. His contributions to highway safety have received many honors, including major awards from the National Highway Traffic Safety Administration, the International Association for Accident and Traffic Medicine, the Association for the Advancement of Automotive Medicine, the Human Factors and Ergonomics Society, and General Motors. He is president of the International Traffic Medicine Association, former president of the Association for the Advancement of Automotive Medicine, a fellow of the Human Factors and Ergonomics Society, a fellow of the Society of Automotive Engineers, a fellow of the Association for the Advancement of Automotive Medicine, and a member of the National Academy of Engineering.

Michael A. Florian (Chapter 11) is Professor of Computer Science and Operations Research at the University of Montreal. He served as the first Director of the Centre for Research on Transportation and continues to collaborate in its research activities. He has worked in various areas of network analysis and optimization methods. His major contributions are in the area of network equilibrium methods and their applications. He supervised the development of EMME/2 and STAN, which are interactive-graphic packages for transportation planning which are used intensively in practice. He has authored more than 130 published research papers and edited three books. He has been elected to the Royal Society of Canada in 1990 and was awarded the Robert D. Herman Lifetime Achievement Award by the Transportation Science section of INFORMS in 1998. He was named Honorary Professor by the Shanghai University of Science and Technology in 1999 and was awarded an Honorary Doctorate by the University of Linköping in 2000. He is currently a member of the Editorial Advisory board of the journals *Transportation Science, Network and Spatial Theory* and Associate Editor of *International Transactions* in *Operations Research* and *Transport Policy.* His professional activities include consulting assignments in the field of transportation planning to organizations on five continents.

Randolph W. Hall (Chapters 1,5,15) is Chairman of the Daniel J. Epstein Department of Industrial and Systems Engineering at University of Southern California. He also serves as Associate Dean for Research in the School of Engineering at University of Southern California. Dr. Hall was previously the founding director for the METRANS University Transportation Center. He has also held research and faculty positions at PATH, University of California at Berkeley and

General Motors. He holds a Ph.D. in Transportation Engineering and a B.S. in Industrial Engineering and Operations Research, both from University of California at Berkeley. He has published extensively on logistics and transportation operations, and is the author of *Queueing Methods for Services and Manufacturing.* He also has extensive consulting experience in architectural design of computing and communication systems for transportation. Dr. Hall has served as the chair of the Transportation Research Board's Transportation Network Modeling committee. He is editor of the *Intelligent Transportation Systems Journal,* and is on the editorial boards for *Computers and Industrial Engineering* and *Institute of Industrial Engineers Transactions.*

Donald Hearn (Chapter 11) is Professor and Chair of Industrial and Systems Engineering and Co-Director of the Center for Applied Optimization at the University of Florida. He received an undergraduate degree in physics at the University of North Carolina as a Morehead Scholar and received Masters and Ph.D. degrees from Johns Hopkins University in management science and operations research. His teaching includes decision modeling and methods, nonlinear optimization and large-scale optimization. In addition to the University of Florida, he has taught at M.I.T. and has given short courses at the University of Rome and the Royal Institute of Technology in Stockholm. His research interests include applied optimization and transportation science. Recent work has concerned the development of efficient algorithms for models that arise in production planning, urban traffic assignment and water management. He is founding editor of *OPTIMA,* the newsletter of the Mathematical Programming Society, associate editor of *Computational Optimization and Applications* and a past associate editor of *Operations Research.* He is author/co-author of over 60 refereed articles, co-editor of the recent books *Large-Scale Optimization: State of the Art* and *Network Optimization,* and co-editor of the Kluwer book series *Applied Optimization.*

Petros A. Ioannou (Chapter 7) is a Professor of Electrical Engineering-Systems and Director of the Center of Advanced Transportation Technologies at University of Southern California. He received the B.Sc. degree with First Class Honors from University College, London, England and the M.S. and Ph.D. degrees from the University of Illinois, Urbana, Illinois. He is the author/co-author of 5 books and over 300 research papers in dynamics and control, neural networks, and intelligent transportation systems. In 1984 he was a recipient of the Outstanding Transactions Paper Award for "An Asymptotic Error Analysis of Identifiers and Adaptive Observers in the Presence of Parasitics," which appeared in the *IEEE Transactions on Automatic Control.* Dr. Ioannou is the recipient of a Presidential Young Investigator Award. He has been an Associate Editor for the *IEEE Transactions on Automatic Control,* the *International Journal of Control* and *Automatica* and he is a fellow of IEEE. He has served as a technical consultant with Lockheed, Ford Motor Company, Rockwell International, General Motors.

Ellis L. Johnson (Chapter 14) is the Coca-Cola Chaired Professor in the School of Industrial and Systems Engineering. He received a B.A. in mathematics at Georgia Tech and a Ph.D. in operations research at the University of California. Before joining Georgia Tech in 1995, he was at IBM's T.J. Watson Research Center for 26 years. There, he founded and managed the Optimization Center from 1982 until 1990, when he was named IBM Corporate Fellow. In 1980-1981, he was at the University of Bonn, Germany, as recipient of the Alexander Von Humboldt Senior Scientist Award. In 1984, he received the George Dantzig Award for his research in mathematical programming. In 1986, he was awarded the Lanchester Prize for his paper with Crowder and Padberg. In 1988, he was elected to the National Academy of Engineering. In 2000, Dr. Johnson won the INFORMS John Von Neumann Theory Prize. From 1990 to 1995, he began teaching and conducting research at Georgia Tech, where he co-founded and co-directed the Logistics Engineering Center with Professor George Nemhauser. His research interests in logistics include crew scheduling and real-time repair, fleet assignment and routing, distribution planning, network problems, and combinatorial optimization.

Diego Klabjan (Chapter 14) is an assistant professor at the University of Illinois, Urbana-Champaign. After obtaining his doctorate from the School of Industrial and Systems Engineering of the Georgia Institute of Technology in 1999, in the same year he joined the Department of Mechanical and Industrial Engineering at the University of Illinois. He is the recipient of the first prize of the 2000 Transportation Science Dissertation Award. He is serving as the vice-president of the INFORMS Aviation Applications Section. His research is focused on airline operations research, integer programming and parallel computing.

Frank S. Koppelman (Chapter 3) is Professor of Civil Engineering and Transportation at Northwestern University, where he has taught since 1975. He has worked on the development of activity based demand models and the integration of econometric and market research methods to enhance understanding of travel behavior and models for both urban and intercity travel. Dr. Koppelman is principal investigator of Northwestern University's participation in a multi-university research program to develop advanced models of traveler behavior as a component of enhanced transportation planning models. Professor Koppelman holds a Ph.D. and B.S. in Civil Engineering (Transportation) from the Massachusetts Institute of Technology (MIT) and an MBA from the Harvard University Graduate School of Business Administration (HBS). He is active in the Transportation Research Board, where he is past-Chairman of the Committee on Travel Demand Analysis and Forecasting and he was Associate Editor of *Transportation Research-B.*

Marvin Kraus (Chapter 17) is a Professor of Economics at Boston College, where he has been on the faculty since 1972. He received a B.S. in Mathematics from Purdue University in 1967 and a Ph.D. in Economics from the University of Minnesota in 1973. He has authored or coauthored numerous articles on various aspects of transportation economics, with a particular focus on optimal pricing and investment in urban transportation.

Vittorio Maniezzo (Chapter 12) is a researcher at the Department of Computer Science at the University of Bologna, Italy. His main research interests include exact and heuristic algorithms for combinatorial problems, such as the quadratic assignment problem, the vehicle routing problem, project scheduling problems and frequency assignment problems. His scientific papers have appeared in various international journals in operations research and computer science. Dr. Maniezzo has consulted for several companies in Italy. Dr. Maniezzo.received his Ph.D. in Computer Science Engineering from the Politecnico of Milan in 1994.

Aristide Mingozzi (Chapter 12) is an Associate Professor of Operations Research at the Department of Mathematics of the University of Bologna, Italy. His major fields of research include exact combinatorial optimization methods for variants of the vehicle routing problem, crew scheduling problems, project scheduling problems and two-dimensional cutting problems. Professor Mingozzi specializes in solving these NP-hard problems using mathematical programming techniques based on innovative formulations of these problems. Professor Mingozzi is the author of many scientific papers that have been published in international journals and presentations at many professional society meetings. Professor Mingozzi has also consulted for many companies in Europe. Professor Mingozzi received his Ph.D. in Operations Research from the University of London in 1984.

George L. Nemhauser (Chapter 14) is the A. Russell Chandler Professor in the Schoolof Industrial and Systems Engineering and an Institute Professor at the Georgia Institute of Technology, where he has been since 1985. He has served ORSA as Council Member, President, and Editor of *Operations Research,* and is the Past Chairman of the Mathematical Programming Society. He is the founding Editor of *Operations Research Letters,* and co-editor of *Handbooks of Operations Research and Management Science.* Dr. Nemhauser received his Ph.D. in Operations Research from Northwestern University in 1961, and joined the faculty of the Johns Hopkins University as Assistant Professor of Operations Research and Industrial Engineering. In 1970, he was appointed Professor of Operations Research and Industrial Engineering at Cornell University and Leon Welch Professor in 1984. He served as School Director from 1977 to 1983. Dr. Nemhauser's honors and awards include membership in the

National Academy of Engineering, Kimball medal and Lanchester prize (twice) and Morse lecturer of ORSA.

Peter Nijkamp (Chapter 17) is professor in regional and urban economics and in economic geography at the Free University, Amsterdam. His main research interests cover plan evaluation, multicriteria analysis, regional and urban planning, transport systems analysis, mathematical modelling, technological innovation, and resource management. In the past years he has focused his research in particular on quantitative methods for policy analysis, as well as on behavioural analysis of economic agents. He has a broad expertise in the area of public policy, services planning, infrastructure management and environmental protection. In all of these fields he has published many books and numerous articles. He has been visiting professor in many universities all over the world. He is past president of the European Regional Science Association and of the Regional Science Association International. At present, he is vice-president of the Royal Netherlands Academy of Sciences.

Susan H. Owen (Chapter 10) is Engineering Group Manager for the Operations Research Department of General Motors' North American Engineering division. She received her Ph.D. in Industrial Engineering and Management Sciences at Northwestern University. Her research interests and publications are focused in the areas of facility location, mathematical modeling, modern heuristic methods, and scenario planning. Her current work with GM is concentrated on decision support applications for solving resource management and scheduling problems. She is a member of INFORMS.

Markos Papageorgiou (Chapter 8) is Professor and Director of the Dynamic Systems and Simulation Laboratory at the Technical University of Crete. He received the Diplom-Ingenieur and Doktor-Ingenieur (honors) degrees in Electrical Engineering from the Technical University of Munich, Germany, in 1976 and 1981, respectively. In 1988-1994 he was a Professor of Automation at the Technical University of Munich. He is the author of the books *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control* (Springer, 1983) and *Optimierung* (R. Oldenbourg, 1991; 1996), and the editor of the *Concise Encyclopedia of Traffic and Transportation Systems* (Pergamon Press, 1991). His research interests include automatic control, optimization, and their application to traffic and transportation systems and water networks. He is an Associate Editor of *Transportation Research-Part C* and Chairman of the IFAC Technical Committee on Transportation Systems. Dr. Papageorgiou was awarded the 1983 Eugen-Hartmann prize from the Union of German Engineers and received a Fulbright Research and Lecturing award (1997). Dr. Papageorgiou is a Fellow of the IEEE.

Tönu Puu (Chapter 9) was Associate Professor of Economics at Uppsala University from 1964-1971, and is currently Full Professor of Economics at Umeå University, Sweden. He holds a PhD from Uppsala University, Sweden. Dr. Puu has published 100 papers and 10 books on the subjects of: investment, portfolio selection, production, natural resources, spatial economics, nonlinear dynamics, economics of the arts, and the philosophy of science. His latest works are *Mathematical Location and Land Use Theory* (Springer, Heidelberg 1997) and *Nonlinear Economic Dynamics* (4th ed. Springer, Heidelberg 1997). Dr. Puu is also initiator and director of the Nordic Baroque Music Festival.

Piet Rietveld (Chapter 17) is professor in Transport Economics at the Free University, Amsterdam. He has been working on various topics in the field of transport economics and regional economics. This research has been extensively reported in authored and edited books, and in about 250 papers published in scientific journals or as contributions to books. Internationally Piet Rietveld is active as the chairman of NECTAR, a European association of transport experts. He is on the editorial board of several scientific journals in the field of transport and regional development and a member of various advisory committees to the government. He has both chaired the cluster for spatial research of the Netherlands Organisation for Scientific Research (NWO), and the Dutch speaking division of the Regional Science Association (RSA). In 1999 he was awarded the Dr Hendrik Muller prize by the Royal Dutch Academy of Sciences (KNAW) for his scientific work in the field of the spatial sciences.

Garrett J. van Ryzin (Chapter 16) is Professor of Decision, Risk and Operations at the Columbia University Graduate School of Business, where he has been on the faculty since 1991. He received his B.S.E.E. degree from Columbia University, and his S.M. in Electrical Engineering and Computer Science and Ph.D. in Operations Research from the Massachusetts Institute of Technology. His research interests include stochastic optimization, pricing and revenue management and supply chain management. Professor van Ryzin's research has been supported by grants from the National Science Foundation and major corporations, and he has served as a consultant to several leading companies in the area of pricing and revenue management. He is Area Editor for *Operations Research* and is an associate editor for *Management Science* and *Transportation Science.*

Kalyan T. Talluri (Chapter 16) is Associate Professor in the Department of Economics and Business at the Universitat Pompeu Fabra (UPF) in Barcelona, Spain. He did his Ph.D in Operations Research at M.I.T in the area of network design. Subsequently, he worked at USAir in the areas of airline scheduling and fleet planning and revenue management. His recent research has been in the area of pricing and revenue management. He has been teaching at UPF since 1995. He has consulted for many industrial groups in Europe, USA and Asia in the areas of dynamic pricing and revenue management.

Pamela H. Vance (Chapter 14) is an Assistant Professor in the Goizueta Business School at Emory University. Dr. Vance holds Ph.D. and M.S. degrees in operations research, and a bachelors degree in chemical engineering, all from the Georgia Institute of Technology. Her research interests include applying integer programming techniques to large-scale problems arising from applications in transportation. Some specific applications are airline crew scheduling problems, cutting stock problems, integer multicommodity flow problems and multilevel distribution problems. She is a recipent of the National Science Foundation's Early Career Development (CAREER) Award for her work on network design problems arising in transportation. She also serves as an Associate Editor *for Transportation Science* and *Operations Research Letters.*

This page intentionally left blank

INDEX