Urban and Regional Transportation Modeling

Essays in Honor of David Boyce

Edited by

Der-Horng Lee

New Dimensions in Networks

Urban and Regional Transportation Modeling

NEW DIMENSIONS IN NETWORKS

Series Editor: Anna Nagurney, *John F. Smith Memorial Professor, Isenberg School of Management, University of Massachusetts at Amherst, USA*

Networks provide a unifying framework for conceptualizing and studying problems and applications. They range from transportation and telecommunication networks and logistic networks to economic, social and financial networks. This series is designed to publish original manuscripts and edited volumes that push the development of theory and applications of networks to new dimensions. It is interdisciplinary and international in its coverage, and aims to connect existing areas, unveil new applications and extend existing conceptual frameworks as well as methodologies. An outstanding editorial advisory board made up of scholars from many fields and many countries assures the high quality and originality of all of the volumes in the series.

Titles in the series include:

Supernetworks
Decision-Making for the Information Age
Anna Nagurney and June Dong

Innovations in Financial and Economic Networks
Anna Nagurney

Urban and Regional Transportation Modeling
Essays in Honor of David Boyce
Edited by Der-Horng Lee

Urban and Regional Transportation Modeling

Essays in Honor of David Boyce

Edited by

Der-Horng Lee
National University of Singapore

NEW DIMENSIONS IN NETWORKS

Edward Elgar
Cheltenham, UK • Northampton, MA, USA

Published by
Edward Elgar Publishing Limited
Glensanda House
Montpellier Parade
Cheltenham
Glos GL50 1UA
UK

Edward Elgar Publishing, Inc.
136 West Street
Suite 202
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Cataloguing in Publication Data
Urban and regional transportation modeling : essays in honor of David Boyce / edited by Der-Horng Lee.
p. cm. – (New dimensions in networks)
1. Urban transportation–Planning. 2. Urban transportation–Mathematical models. 3. Regional planning. I. Lee, Der-Horng, 1967- II. Boyce, David E. III. Series.

HE305.U683 2004
388.4′01′1–dc22 2003064255

ISBN 1 84376 306 0

Printed and bound in Great Britain by MPG Books Ltd, Bodmin, Cornwall

Contents

Contributors

Åke E. Andersson Department of Infrastructure and Planning, Royal Institute of Technology, Stockholm, Sweden.

David Emanuel Andersson Graduate Institute of Regional Development and Management, Leader University, Tainan, Taiwan.

Hillel Bar-Gera Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er Sheva, Israel.

David Boyce Department of Civil and Materials Engineering, University of Illinois at Chicago, Chicago, Illinois, USA.

K.S. Chan Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hong Kong, China.

Huey-Kuo Chen Department of Civil Engineering, National Central University, Chungli, Taiwan.

Cheng-Yi Chou DHL International Inc., Taipei, Taiwan.

June Dong School of Business, State University of New York at Oswego, Oswego, New York, USA.

Sven Erlander Department of Mathematics, Linköping University, Linköping, Sweden.

Joaquim J.M. Guilhoto Regional Economics Applications Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

Timothy D. Hau School of Economics and Finance, University of Hong Kong, Hong Kong, China.

Yang He Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona, USA.

Geoffrey J.D. Hewings Regional Economics Applications Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

Tschangho John Kim Department of Urban and Regional Planning, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA.

William H.K. Lam Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hong Kong, China.

Torbjörn Larsson Department of Mathematics, Linköping University, Linköping, Sweden.

Hong K. Lo Department of Civil Engineering, Hong Kong University of Science and Technology, Hong Kong, China.

David Lucas Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona, USA.

Jan T. Lundgren Department of Science and Technology, Linköping University, Norrköping, Sweden.

Lars-Göran Mattsson Department of Infrastructure, Royal Institute of Technology, Stockholm, Sweden.

Claudio Meneguzzer Dipartimento di Costruzioni e Trasporti, University of Padova, Padova, Italy.

Qiang Meng Department of Civil Engineering, National University of Singapore, Singapore.

Pitu Mirchandani Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona, USA.

Anna Nagurney Eugene M. Isenberg School of Management, University of Massachusetts at Amherst, Amherst, Massachusetts, USA.

Peter C. Nelson Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA.

Michael Patriksson Department of Mathematics, Chalmers University of Technology, Göteberg, Sweden.

Hsiao-Chi Peng DHL International Inc., Taipei, Taiwan.

Joseph Raj Oracle Corporation, Redwood Shores, California, USA.

Gopalakrishnan Rajaraman School of Civil Engineering, West Lafayette, Indiana, Purdue University, USA.

Bin Ran Department of Civil and Environmental Engineering, University of Wisconsin at Madison, Madison, Wisconsin, USA.

Paulo Resende Ibmec Business School, São Paulo, Brazil.

Nagui M. Rouphail Department of Civil Engineering, North Carolina State University, Raleigh, North Carolina, USA.

Clas Rydergren Department of Science and Technology, Linköping University, Linköping, Sweden.

B.F. Si Department of Civil and Structural Engineering, Hong Kong Polytechnic University, Hong Kong, China.

Lina Sjölin Inregia AB, Stockholm, Sweden.

Hongyu Sun Department of Civil and Environmental Engineering, University of Wisconsin at Madison, Madison, Wisconsin, USA.

Rohit Syal Wilbur Smith Associates, New Haven, Connecticut, USA.

W.Y. Szeto Department of Civil Engineering, Hong Kong University of Science and Technology, Hong Kong, China.

Andrew P. Tarko School of Civil Engineering, West Lafayette, Indiana, Purdue University, USA.

Elliott A. Torres Borland Software Corporation, Chocago, Illinois, USA.

Michael Wegener Spiekermann & Wegener, Urban and Regional Research, Dortmund, Germany.

Huw C.W.L. Williams Department of City and Regional Planning, Cardiff University, Wales, UK.

K.I. Wong Department of Civil Engineering, University of Hong Kong, Hong Kong, China.

S.C. Wong Department of Civil Engineering, University of Hong Kong, Hong Kong, China.

J.H. Wu TJKM Transportation Consultants, Pleasonton, California, USA.

Heng Xiao Department of Civil and Environmental Engineering, University of Wisconsin at Madison, Madison, Wisconsin, USA.

Hai Yang Department of Civil Engineering, Hong Kong University of Science and Technology, Hong Kong, China.

Ding Zhang School of Business, State University of New York at Oswego, Oswego, New York, USA.

Preface

It was with great eagerness and joy that colleagues and former students of Professor David Boyce came together to produce this festschrift to celebrate his career. Here, we hope not only to honor his many professional contributions and accomplishments in the fields of regional science and transportation modeling, but to acknowledge the gifts he gave to all of us through the many years he has acted as our teacher, mentor, and colleague. Having had the good fortune to study and earn my PhD under Professor Boyce, I was altogether thrilled when Professor Tschangho John Kim at the University of Illinois at Urbana-Champaign asked me to act as editor for this project.

I first met Professor Boyce in 1993 at the annual meeting of the Transportation Research Board (TRB). I was introduced by my master thesis advisor, Professor Huey-Kuo Chen, himself a former PhD student of Professor Boyce. And it was not long thereafter that I was enrolled in a PhD program with Professor Boyce at the University of Illinois-Chicago. I earned my degree in 1996, but Professor Boyce's influence continues with me to this day.

From the perspective of a student – and I am comfortable in saying that all of his academic children would agree – it was immediately apparent that Professor Boyce has great passion for his work. And that passion is infectious in his students. That love of learning was passed down to each of us; it would have been impossible to resist, as Professor Boyce took an interest in each one of us as individuals, both personally and professionally. Perhaps this is best exemplified through his consistent encouragement and support of his students to pursue our own interests and areas of study, and not to take on projects simply because they held an interest for him. He has constantly encouraged our independence and the ideal of excelling on our own. That he was always ready to support, to listen, to encourage, to guide, and to help us in innumerable ways is, I truly think, what makes Professor Boyce so special. In effect, he teaches each of his students the 'art of learning', which is the greatest thing a teacher can teach a student.

It is our further good fortune that Professor Boyce does not consider his work over once a student receives a diploma. Strong professional and personal ties established in graduate school continue even as his students move on to their own careers. In a sense, every student becomes a part of the big 'Boyce Family'. Even if it has been years since last studying with him, we, his former students, know that we can still call on him for advice, to talk over an idea, or just to catch up.

And his work continues. It was a great privilege to act as Professor Boyce's host at the National University of Singapore (NUS) in January and February 2003. Professor Boyce spent seven weeks with us as visiting professor, during which time he offered a course on travel forecasting. It was a pleasure to witness his NUS postgraduates experiencing the invaluable mentoring so many of us have benefited from firsthand over the years. Unfortunately, as it was only for seven weeks, NUS students got only a glimpse of the love of learning and dedication to students that are so central to Professor Boyce as a teacher.

Professor David Boyce's unlimited patience, continuous inspiration, honorable attitude toward life, and his work have created a shining landmark for all of us.

As editor of this festschrift, I want to express my appreciation to all of the contributing authors scattered around the globe. I also want to thank all the referees for their careful reviews of submitted manuscripts, and for working with me on the extremely tight timeframe – even during the holiday season. I must thank Mr Alan Sturmer at Edward Elgar for his administrative support and assistance in making this project possible. Special thanks go to Professors Tschangho John Kim and Anna Nagurney for their advice, comments, and suggestions during the preparation of this festschrift. Special thanks also go to Dr Maya Tatineni (who was a fellow PhD student with me under Professor Boyce in the early 1990s) for taking time from her busy schedule to proofread manuscripts and for sharing her thoughts and feelings about Professor Boyce. I would also like to thank my students and Professor Boyce's postgraduates at the Intelligent Transportation and Vehicle Systems Laboratory at the National University of Singapore – Mr Weizhong Zheng, Miss Lan Wu, Miss Liying Song, Mr Yueping Sun, and Mr Nan Liu – who were actively involved in the editing process. And finally, but certainly not least, I want to thank Professor David Boyce for all he has brought to the profession, to his colleagues, and to his students.

Der-Horng Lee
National University of Singapore
April 2003

Biography and appreciation of Professor David Boyce

Tschangho John Kim

It gives me great pride and is an honor to summarize Professor David Boyce's scholarly and professional activities and to offer an appreciation for his contributions to the fields of transportation science, regional science and urban planning. He has been a dear friend, teacher, colleague, co-author, co-principal investigator, and roommate in many hotels while participating in international conferences.

During 37 years of research and teaching, Professor Boyce addressed key methodological issues related to metropolitan transportation and land-use planning. His early monograph, *Metropolitan Plan Making*, published in 1970, critically examined the experience with the land-use and travel forecasting models during the 1960s. Recognizing that these methods lacked an adequate scientific basis, he has since devoted himself to the formulation and solution of urban travel and land-use forecasting models as constrained optimization problems and related constructs, which synthesize elements of network analysis and modeling, stochastic discrete choice theory and entropy-based methods.

Through this research he concluded that the conventional travel forecasting paradigm, widely known as the four-step travel forecasting procedure, may be seen to be a counterproductive concept. By focusing research on individual elements of daily travel decisions, mainly represented as having fixed travel times and costs, the conventional point of view obscures the overall equilibria and interdependence of travel choices. To offer an alternative perspective, Professor Boyce rigorously formulated, implemented, estimated and validated large-scale, integrated models of travel behavior. His ongoing research offers an alternative, both to the conventional viewpoint, and to newer initiatives, some of which also lack a rigorous scientific foundation. He also extended this integrated approach to the study of regional economies, interregional commodity flows and freight transportation systems.

In addition to this primary research theme, from 1986 to 1996, Professor Boyce was an early innovator of in-vehicle dynamic route guidance systems, one element of the emerging field of intelligent transportation systems. This research culminated in his leading a multi-university team that performed development and evaluation tasks for the ADVANCE Project, a large-scale field test of a prototype route guidance system, in conjunction with Motorola, Inc., and federal and state transportation departments. In this role he also conducted theoretical and modeling studies of the performance of route guidance systems on urban road networks.

Following the completion of his PhD in the field of regional science in 1965, Professor Boyce has provided institutional support and leadership in various roles to the Regional Science Association International (RSAI) in North America, Europe and Asia. During

1969–89, he organized its North American Meetings; he also served as Secretary, 1969–78, and presently serves as Archivist. He was a co-editor of *Environment and Planning A*, and an associate editor of *Transportation Science*. In addition, he directed a National Science Foundation workshop on transportation research, and has served on many editorial boards in regional science and transportation.

In recognition of his research and service contributions to the field of regional science, in 2000 he was awarded the Founder's Medal of the RSAI; in 2002, he was elected as one of four Inaugural Fellows of the RSAI. In 2000, he received the University of Illinois at Chicago (UIC) Inventor of the Year and the UIC College of Engineering Faculty Research Awards for his contributions to transportation modeling and algorithms. He also received the University of Illinois Alumni Association's UIC Flame Award for Teaching Excellence in 2001. He is a Registered Professional Engineer in the State of Ohio, a Life Member of the American Society of Civil Engineers, and an Emeritus Member of the Transportation Research Board. He has published 165 journal articles, books, book chapters and reports during the past 38 years.

To list all of Professor Boyce's professional activities would be prohibitively voluminous; the following are highlights excerpted from his curriculum vitae.

1961 BS	Civil Engineering	Northwestern University
1963 MCP	City and Regional Planning	University of Pennsylvania
1964 MA	Regional Science	University of Pennsylvania
1965 PhD	Regional Science	University of Pennsylvania
1966–1968	Assistant Professor of City and Regional Planning, University of Pennsylvania, Philadelphia	
1968–1977	Assistant Professor, Associate Professor, Professor of Regional Science and Transportation, University of Pennsylvania, Philadelphia	
1972–1973	Senior Visiting Fellow, University of Leeds, UK	
1977–1988	Professor of Transportation and Regional Science, Department of Civil Engineering, University of Illinois at Urbana-Champaign; secondary appointment in Department of Urban and Regional Planning; Chair, Regional Science Program in the Graduate College, 1980–88	
1978–1994	Associate Editor: *Transportation Science*	
1978, 1985	Guest Editor: *Transportation Research*	
1979–1988	Co-Editor: *Environment and Planning A*	
1988–1996	Director, Urban Transportation Center and Professor of Transportation and Regional Science in the College of Urban Planning and Public Affairs, University of Illinois at Chicago	
1997–2003	Professor of Transportation and Regional Science, Department of Civil and Materials Engineering, University of Illinois at Chicago	

Introduction

This festschrift contains 20 chapters addressing issues from regional science, travel forecasting, transportation planning, transportation network modeling, intelligent transportation systems and so on. This collection reflects the breadth of Professor Boyce's involvement and research interests in these areas.

Chapter 1, contributed by Huw Williams, examines and illustrates – from a British perspective – four themes in travel behavior and transport modeling: the practical modeling alternatives; the contribution of theory to practical model development; the balance between theoretical sophistication and simplification in practical model design; and issues in the specification and solution of equilibrium models associated with substantive policy issues.

Chapter 2, written by K.I. Wong, S.C. Wong, J.H. Wu, Hai Yang and William Lam, presents a combined distribution, hierarchical mode choice, and assignment network model with multiple user and mode classes. In their chapter, Evans's partial linearization algorithm is proposed to solve the problem. The strategic transportation network in Hong Kong is used as a case study to illustrate the potential applicability of the proposed methodology for solving complex transportation planning problems.

In Chapter 3, Hillel Bar-Gera and David Boyce propose a fixed point formulation to formulate general combined models mathematically, including most models used in practice. The origin-based algorithm is adopted and is proved as a more efficient solution approach than prevailing alternatives in which faster convergence is achieved.

In Chapter 4, Michael Wegener outlines a methodology to model activity patterns, trips and trip chains, destination, mode and route choice of individual travelers in urban regions by time of day, including within-day and period-to-period adjustment of behavior, by microsimulation without iteration, which assigns the individual trips generated in a microscopic activity-based travel forecasting model to a multimodal transport network. It is found that the iteration-free nature of the approach makes it particularly suitable for integrated models of urban land use, transport and environment (LTE).

In Chapter 5, Sven Erlander and Jan T. Lundgren focus their attention on a new notion of cost minimizing behavior by expressing it in terms of the decisions taken by a group of decision makers, and this is introduced in discrete choice modeling. The probability distribution of the log linear form for an individual's choice between the alternatives is derived. It is shown that all standard discrete choice models of log linear type satisfy the conditions for this newly introduced cost minimizing behavior.

Equilibrium traffic signal setting (ETSS) is the problem of determining a joint equilibrium of link flows and signal settings in a road network that operates under traffic-responsive signal control. Claudio Meneguzzer in Chapter 6 proposes a heuristic to imitate the real-world interaction between signal control decisions and route choices. In addition, a new step in Meneguzzer's approach is added to the 'classical' version of the iterative scheme, so as to incorporate the partial driver response feature into the flow-updating

rule. This leads to a more general framework for the analysis of ETSS, in which the 'level of responsiveness' of the network users can be explicitly accounted for.

Lars-Göran Mattsson and Lina Sjölin (Chapter 7) present a stylized model of a generic symmetric city for the simulation policies of road investment and road pricing. The authors consider these two policies as two possible options to relieve congestion problems. It will not only affect the demand for transport in various respects, but may also, in the long run, change the location of activities. This model evaluates transport and land-use effects of congestion pricing and of inner and outer toll rings in the road network.

The study presented in Chapter 8 by Hai Yang, Qiang Meng and Timothy D. Hau investigates the relationships between trans-modal transport pricing and subsidy policy for optimal modal split, and presents optimization models on a bi-modal transportation network. Transport pricing and transit subsidy in the chapter are sought for optimal modal split under the assumption of a transit budget constraint.

In Chapter 9, Hong K. Lo and W.Y. Szeto extend the traditional continuous network design problem by incorporating the time dimension, which enables one to be able to design for the optimal project initiation time, phasing, and financial arrangements over the planning horizon. A single-level optimization program is formulated and two simple numerical examples are set up to compare the performances between the traditional and the proposed approaches. Numerical results show that this extended formulation outperforms greatly the traditional formulation.

There are nine chapters in this festschrift about dynamic network modeling, vehicle routing and navigation, travel time and traffic delay estimation and so on. This reminds us of Professor Boyce's pioneering efforts in developing the vehicle route guidance system (the well-known ADVANCE Project) and dynamic transportation network modeling. The first chapter of this kind is from Professor Boyce's collaborators in the ADVANCE Project, Elliott A. Torres, Peter C. Nelson, Nagui M. Rouphail and Joseph Raj. In Chapter 10, they explore and contrast the use of artificial intelligence techniques against traditional methods in order to improve the efficiency and accuracy of delay estimates for arterial streets. They found that the neural network approach provided additional flexibility that could not be matched by the statistical approach.

Andrew P. Tarko and Gopalakrishnan Rajaraman in Chapter 11 propose a new concept of expected cumulative counts (L curves) that are more suitable for traffic modeling than cumulative counts. The use of L curves is considered for non-FIFO and non-conserved traffic, and the conditions for unbiased estimates of expected travel times of individual vehicles are determined. A method of modeling travel times between two signalized intersections is proposed and tested on three street segments.

In Chapter 12, William H.K. Lam, K.S. Chan and B.F. Si propose a bi-level programming model to investigate under what circumstances the traffic authority should encourage or discourage the implementation of the Advanced Traveler Information Systems (ATIS). A sensitivity-analysis-based solution algorithm is proposed for solving the problem and an example is illustrated.

The contribution from Hongyu Sun, Heng Xiao and Bin Ran (Chapter 13) addresses short-term traffic predictions by using data mining. Traffic data on neighboring links are incorporated into inputs of the local linear model to predict traffic conditions on the individual link of interest.

In Chapter 14, Pitu Mirchandani, Rohit Syal, David Lucas and Yang He focus on on-

line traffic assignment and network loading. In this chapter, an accelerated version of the method of successive averages (MSA) for predicting traffic patterns undergoing network interventions is proposed. Rather than volume-delay functions like the conventional BPR (Bureau of Public Roads) function, route attributes experienced by the travelers are used to track the resulting effect of infrastructure changes. The model is evaluated by a route-based simulation.

Chapter 15 by Tschangho John Kim is about multi-modal routing and navigation cost functions for location-based services (LBS). This chapter focuses on developing functional forms for costs for providing multi-modal routing and navigation services and on searching for feasible directions to solve the functions heuristically, and presents a feasible set of functional forms. The author presents the idea of developing heuristic solution algorithms by developing a node–node adjacency matrix and estimating spatiotemporal link travel time.

June Dong, Ding Zhang and Anna Nagurney (Chapter 16) propose a supply chain network model in the form of a super-network, in which both physical and electronic transactions are allowed and the demands associated with the retail outlets are considered as random. They model the optimizing behavior of the various decision makers, derive the equilibrium conditions, and establish the finite-dimensional variational inequality formulation.

In Chapter 17, Huey-Kuo Chen, Hsiao-Chi Peng and Cheng-Yi Chou, study the dynamics of the joint entropy distribution/assignment (JEDA) problem by developing the dynamic user equilibrium problem with doubly constrained origin–destination/departure time/route choice. A path-based algorithm is proposed and compared with the dynamic version of the modified Evans algorithm in terms of computational efficiency. Their results show that the proposed algorithm is more efficient and thus has great potential for solving large network problems.

In Chapter 18, Torbjörn Larsson, Jan T. Lundgren, Michael Patriksson and Clas Rydergren provide numerical examples of a decision support methodology for strategic traffic management. The decision support methodology is based on an equilibrium model for route choice in a congested urban traffic network. The authors show how the management methodology can be applied to some traffic management scenarios. The traffic networks of Sioux Falls and Linköping are used for computational examples.

In Chapter 19, Paulo Resende, Joaquim J.M. Guilhoto and Geoffrey J.D. Hewings study the free trade and transportation infrastructure in Brazil. Potential limitations imposed by transportation infrastructure are big issues in the development of models analysing the impacts of free trade agreements between countries or regions within countries. In this chapter, an illustration of a potential approach to this problem is illustrated with reference to Brazil in the case of MERCOSUL in South America.

Finally, in Chapter 20, Åke E. Andersson and David E. Andersson study the accessibility and site rents in the C-economy. Note that the C in 'C-economy' stands for several typical features of the post-industrial economy, such as creative, cognitive and computer capacities, culture, and communications. Their empirical analysis by a simplified growth model indicates that regions that are accessible by road transportation and that have an initial advantage in terms of the availability of knowledge capital tend to have both higher income growth and higher expected returns to real estate investments, reinforcing the agglomerative tendencies of the C-economy.

1. Themes in the development and application of transport planning models

Huw C.W.L. Williams*

1. INTRODUCTION

In this chapter I shall examine four themes in travel behaviour and transport modelling. The themes are universal but each tends to emerge in relation to local contexts and states of practice. They will therefore be illustrated from a British perspective. Each is very broad and a comprehensive coverage is not sought. No doubt I shall emphasize those areas I know best but an attempt will be made to link them directly to the central challenges of transportation and regional science with which this festschrift is associated.

The four themes are as follows: the practical modelling alternatives available at any time to address particular questions; the contribution of theory to practical model development; the balance between theoretical sophistication and simplification in practical model design; and issues in the specification and solution of equilibrium models associated with substantive policy issues.

Much of David Boyce's recent work has concerned the theoretical and analytical integrity of models and their solution, on the one hand, and the concern for appropriate practical application in the midst of considerable apparent choice in the United States (Boyce 1998a, 1998b). Boyce has championed the application and rigorous solution of equilibrium models in the long tradition of Beckmann et al. (1956) and has called for greater prescription and guidance by the professional community to bring order to the range of available models *for practice*. The themes examined will relate to these issues and, while broadly in agreement with David Boyce's views, I wish to offer some qualifications based on the practical developments and institutional context in the UK.

To set the scene, in Section 2 I present a broad overview of model developments in British transport planning practice with the view to identifying distinct alternatives, while in Section 3 the contributions of theoretical, rather than specific technical, developments are noted.

Since the seminal works of Wardrop (1952) and Beckmann et al. (1956) there has been a great deal of research into the nature and properties of equilibria in models of congested networks and their basic economics under comparative static changes. It is widely recognized that, for most network investment policies, route switching is the dominant response and the accompanying level-of-service changes contribute by far the largest component of the user benefits – a key model output. Even though travel forecasting is well into middle age, we know considerably less about the wider *responses* of travellers associated with the short- and long-run adjustments to land-use, investment, and demand management policies. Most of the comments in Sections 4 and 5 will relate to

this limited knowledge in relation to the assessment of very traditional policies. In the third theme, I discuss the issues of complexity and simplification in model design with reference to an important and influential UK report (SACTRA 1994) on the nature, measurement and methods of analysis of induced traffic from highway schemes. Here, the assumptions and technical issues surrounding the simplification procedure have had important implications for the development of, and advice on, the use of models in the UK.

The final theme, presented in Section 5, returns to the importance of specification and solution of models, and equilibrium methods in particular, again focusing on induced traffic. I shall take three examples to discuss equilibria and systems effects in multi-modal contexts and identify outstanding research issues in traditional policy areas. A short conclusion (Section 6) draws together the various themes and offers some reflections on David Boyce's suggestions.

2. DEVELOPMENTS IN TRANSPORT PLANNING MODELS IN THE UK

What model alternatives are available at any particular time to address specific questions to the required degree of accuracy? Over the past four decades in the UK, the range of policies addressed, the evaluation frameworks applied, and the precision of information required have, of course, changed very considerably and this expansion in scope continues to pose a significant challenge to the field.[1] In discussing the development of transport models over this period no functional details will be given. For this purpose the reader is referred to the text by Ortuzar and Willumsen (2001) and to IHT (1996). My intention is to identify broad alternatives available *in practice*.

2.1 Early Developments in the UK

Analytical transport planning, embracing a systems approach, was imported into Britain from the United States in the early 1960s. Over the decade or so following the London Traffic Study in 1962 (London County Council, 1964) it was applied in most of the major towns and cities. As many studies were carried out by joint US–UK consultancy teams they drew heavily on the four-stage approach, although local variations were evident depending on the many factors which influence the methods applied in transport planning practice.

By the late 1960s several innovations began to appear in UK studies: household-based category analysis for forecasting trip ends was introduced (Wootton and Pick 1967), and zonal-based regression models for trip generation quickly fell from grace. Generalized costs with value-of-time measures derived from micro-statistical studies of modal choice were embedded into the assignment, modal split and distribution models (Quarmby 1967; Wilson 1969) and became a 'universal mechanism' for the interpretation, analysis and evaluation of policy in terms of changes in price and level-of-service variables. The South East Lancashire North East Cheshire (SELNEC) study was to bring Wilson's work on model theory and structures (Wilson 1967; Wilson et al. 1969) into the mainstream of British transport planning practice. The travel forecasting models adopted were thus

hybrids drawing on disaggregate analyses and market segmentation in a pragmatic way to provide aggregate forecasts.

While the large majority of journey-to-work models were based on the sequence of sub-models G/D/MS/A corresponding to the Generation (G), Distribution (D), Modal Split (MS) and Assignment (A), the other variants G/MS/D/A, G/D–MS/A corresponding to pre-distribution and joint-distribution modal split arrangements were also present. A full discussion of these variants and the means of interfacing the sub-models can be found in Senior and Williams (1977) and Williams and Senior (1977).

The models were invariably successfully calibrated, providing adequate and sometimes impressive statistical fits to base-year trip patterns.[2] In conjunction with assumptions of constancy of trip rates and dispersion parameters, traffic and travel forecasts were established over typically 15 to 25 years for a range of land-use, investment and restraint policies.

However, it was noted that several widely used four-stage model forms with different dispositions of D and MS and means of interfacing these sub-models, could satisfy the calibration criteria and yet produce remarkably different, even pathological, response properties, in some cases implying cross-elasticities of the wrong sign (Senior and Williams 1977; Williams and Senior 1977). This affirmed that a 'good fitting' model, was a necessary but hardly a sufficient condition for sound predictions.

The early emphasis was, as often stated, on highway solutions to peak-hour congestion problems, and the 'predict and provide' philosophy was well in evidence. Increasingly major public transport schemes were considered, for example, in Merseyside, Glasgow, Tyneside and Manchester.

After the first wave of transportation studies came to an end in the mid-1970s the four-stage multi-modal models applied with detailed zoning systems and networks went out of fashion and, at least outside London, relatively little effort was devoted to their support or maintenance.

2.2 Developments in Land-Use–Transport Planning Models

Over the last 25 years or so the development of transport forecasting in the UK has been characterized by sometimes opposing tendencies in response to academic developments and the needs of practice. I would put these as: integration, fragmentation, assimilation of microeconometric models; refinement in network models; simplification; and increased presentational efficiency. These are now briefly considered.

Towards integrated land-use–transport models

The large majority of four-stage transport planning models in the 1960s and 1970s were undertaken in the classical style with exogenous input of land-use variables. The influence of accessibility on land use, if considered at all, was done so informally. The *academic* pursuit of integrated land-use–transport models in the UK dates from the late 1960s and early 1970s drawing together the travel, locational behaviour and stock (al-)location sub-models. These varied from relatively simple five-stage forms in which accessibility measures were 'fed back' directly into the land-use allocation stage to, at their most ambitious, much more extensive and integrated models of cities and regions. Many of these drew on key ideas from Wilson's reformulation of locational models, including that of the Lowry model (Wilson 1970, 1974).

Several integrated land-use–transport models, of comparative static and incremental dynamic form, have emerged out of research programmes conducted at a number of British centres, with LILT, MEPLAN, TRANUS and DELTA(/START) representing prominent examples (DSP&MEP 1999). A class of optimization models with an embedded Lowry-type mechanism is also represented (Coelho and Williams 1978; Wilson et al. 1981). Over time, these embraced a much greater economic content both in the representation of individual behaviour, drawing on the discrete choice framework, and in interfacing demand and supply in modelling housing and land markets.

The structure of several of these models has been widely reviewed and compared as part of the ISGLUTI study (Webster et al.,1988; Paulley and Webster 1991); by Wilson (1998), and as part of the deliberations of the Standing Advisory Committee on Trunk Road Assessment (SACTRA 1999) into the links between transport infrastructure and the space economy (DSP&MEP 1999). Some models remained academic prototypes, some were discontinued, while others have been widely applied both in this country and abroad in sub-regional development studies and in assessing the impact of major infrastructure projects. Several models, which are either stand-alone integrated forms, or combinations of existing land-use and transport models brought together for specific purposes, are available in the UK to address policies in strategic and tactical contexts. Their recent use in the current generation of multi-modal studies is noted below.

Fragmentation

While the search for suitable integration inevitably gave rise to greater model complexity, in the changed world of the late 1970s and 1980s the contexts of application broadened considerably to that of forecasting transport impacts at sub-regional scale down to local level in land-use, investment, demand management and traffic management model applications. Again, the particular information required for broad strategic and tactical purposes dictated the available model options. To address these, transport consultants developed a 'tool-kit' approach to provide a flexible arrangement of sub-models which could be selected and combined on a 'pick-and-mix' basis according to the problem at hand. Individual sub-models were embellished as required and specialist consultancies emerged to provide expertise in particular areas.

Assimilation of microeconometric models and stated preference methods

As noted above, British practice had by the late 1960s assimilated the generalized cost models and their parameters derived from microstatistical studies. Multi-modal applications proved to be fertile ground for the application of the newly emerging discrete choice approach. In this regard, Andrew Daly's work, first at LGORU (Local Government Operational Research Unit) and subsequently at Hague Consulting Group (now RAND Europe) in model development, parameter estimation and software design (Daly and Zachary 1978; Daly 1987, 1992) is particularly noteworthy.

Although well established in principle, relatively few consultants (world-wide) have brought to fruition the initial aspirations of the 'disaggregate model' programme in which a complete microeconometric model is subject to specification, estimation, aggregation and equilibration to examine a full range of urban transport policies.[3] More typically, multinomial and nested logit models have been widely applied in econometric studies of modal choice and slowly absorbed *alongside* other components in the software tool-kit.[4]

Undoubtedly, one of the most significant developments in travel forecasting over the last 20 years has been the greater use of stated preference methods for estimating the parameters of discrete choice utility models. British academics and consultants have been particularly active in this field. The potential of the stated preference (SP) approach was noted in the early 1980s (Sheldon and Steer 1982) and led to rapid development (Bates 1988; Fowkes 1991). Sometimes complementary application with revealed preference (RP) enabled analysts to exploit the strengths and avoid the weaknesses of each approach (Wardman 1988). Both RP and SP approaches were widely applied in estimating utility models for light rail in the 1980s and 1990s, and SP in estimating the demand for new rail stations (Fowkes and Preston 1991).

Value-of-time (VoT) research did much to establish the credentials of discrete choice models in the UK. The parameters which emerged from early RP work in Leeds in the late 1960s, provided the foundations for transport system appraisal until the Department of Transport[5] commissioned a major VoT study in the mid-1980s (MVA/ITS/TSU 1987). This was an amalgam of several investigations of behaviour in route and modal choice contexts in different parts of the country to examine variation in the VoT, and test the compatibility of revealed and stated preference methods for parameter estimation. The 1980s work set the scene for a further set of studies in the mid-1990s (HCG/Accent 1996; Gunn et al. 1996), which drew on earlier Dutch work, and included qualitative/quantitative research (see Wardman 1998, for a review of UK VoT work).

Refinement of network models

The refinement of link-based network models to treat a greater complexity of the supply side and more realistic movement through the network quickly developed to address traffic management applications (for example, SATURN and CONTRAM). At its outset the former was viewed as a 'modern assignment program' (Van Vliet 1982), embodying representations of network processes at different levels of resolution to address the relevant policy context. After 20 years of development the suite has much extended capabilities including equilibrium assignment with elastic demand. It is the main context in which rigorous algorithmic approaches to network equilibrium with a variety of demand functions are currently being pursued in the UK (Willumsen et al. 1993; Van Vliet and Hall 1998; Bates et al. 1999).

While microsimulation models of junction control have existed for many years, their application to wider urban networks is a relatively recent development. In the UK the PARAMICS model developed by SIAS has been applied to towns and cities as diverse as Lanark, London and Los Angeles (Druit 2000). Druit asks the pertinent question 'If we never had had a system of traffic modelling, and were to start inventing something today, what would we come up with?'. His response is, unsurprisingly, 'It is difficult to imagine that this would not be centred on the modelling of *individual* vehicle movements as in contemporary microsimulation'. The TRANSIMS system is, of course, an even more ambitious extension to address both the intricacies of the supply and demand sides. Even among professionals, who are *au fait* with the underlying mathematics, attitudes vary from deep suspicion, through to declared enthusiasm.

Public transport networks too have been greatly refined both on the demand side in accommodating nested sub-modal choice and on the supply side with the representation of crowding in public transport assignments.

Simplification

While the five-stage models represented a move to ever greater conceptual and operational complexity, the need for simplification particularly in strategic studies was also evident and it was achieved in various ways: establishing hierarchical systems with resolution varying with level; limiting the number of zones; simplifying the supply characteristics; omitting one or more response mechanisms; or reducing the calibration burden by expressing the model in pivot-point form and importing elasticities. An example of the latter is the incremental nested logit model (Bates et al. 1987). Another key feature in reducing the cost of transport and traffic studies was the rejuvenation of ageing trip matrices through traffic counts (Willumsen 1982).

Presentational effectiveness

A further most significant development over the last two decades has been the vast increase in computing power and the associated improved means of interrogating the outputs of models using GIS (geographic information systems) platforms and, in microsimulation contexts, the use of dynamic displays. These are widely encountered in diverse applications from accessibility studies to junction designs and are, of course, dangerously beguiling!

2.3 (Back) into an Era of Multi-modal Studies

Over the last decade many UK cities have actively considered demand management (through a variety of regulatory and fiscal instruments) and in the unhelpful and harsh commercial climate of deregulated bus transport, some form of light rail transit (LRT) for congestion relief or regeneration objectives. Although LRT systems are present only in Newcastle (Tyneside), London Docklands, Manchester, Sheffield, West Midlands and Croydon, consultants' reports have been prepared to explore their prospects in many other large UK towns and cities. As noted above, microeconometric models in conjunction with stated and revealed preference methods have been widely applied in this context.

Several cities, including Edinburgh, Bristol and Glasgow, are actively considering some form of area traffic restraint and all have been justified by extensive modelling exercises. The first urban congestion charging scheme in the UK was introduced in Durham in 2002 to cover its historic centre, but by far the most significant is that of London, which introduced cordon pricing in February 2003. A summary of the models supporting the London policy can be found in Bates et al. (1996).

The response to congestion problems on the inter-urban network and those trunk roads for which the Department of Transport has responsibility has traditionally been to seek solutions within a single mode – highway – framework. In 1998 the government announced a major initiative which some have heralded the final demise of the 'predict and provide' approach. Twenty-two studies of major corridors and parts of conurbations are to be applied in three tranches. 'New thinking' is to be applied requiring a whole range of policy instruments, land use, investment, pricing and a range of 'soft measures' – workplace travel plans, school travel plans, teleworking, cycling and walking strategies, for which case studies, the transfer of experience from other areas and expert judgement will prove important means of estimating impact. So far, only three multi-modal studies have been completed but we can anticipate that the nature and economics of final recommendations will dependent crucially on the acceptability of some form of congestion charging.

At the present time, the context within which local authorities are conducting their local transport plans, and the Department of Transport addressing inter-urban congestion, is the government's Ten Year Plan (until 2010). Within both urban and inter-urban contexts the Department for Transport offers substantial guidance on the development of appropriate models (DETR 2000; DfT 2002). This is not, however, narrowly prescriptive and has allowed considerable discretion in the local selection of methodology according to local contexts. In the multi-modal studies, in particular, this has resulted in a range of land-use and transport models in applications and in proposals.

2.4 What Options Are Available?

The history of transport planning models in the UK is one of evolution and fragmentation and the key response to the diversity of current practice is a 'tool-kit' approach. Many innovations, some much more significant than others, have been assimilated into the tool-kit. Not surprisingly innovations have been particularly welcomed when changes are small and they have least disturbed existing model software.

In a British context I would put the broad composition of the tool-kit as follows:

- integrated land-use–transport models;
- five-/four-stage approach with simplifications (one or more responses being suppressed);
- incremental (pivot point) models with imported elasticities (including direct elasticity, logit and nested logit forms);
- microeconometric studies involving multinomial and nested logit models;
- alternative approaches to analysing preferences (revealed and stated preference methods); and
- link-junction-based network models based on equilibrium algorithms and/or (micro-)simulation.

The range of choices is perhaps more apparent than real, often amounting to minor variations and simplifications of the more complex four- or five-stage model, or separate sub-models 'bolted together'. Very few consultancies will have the full range as part of their model software suites. Given the increasingly specialized requirements of different approaches and study designs it is not unusual for different consultancies to pool expertise. The Department for Transport, through its publication guidance and the award of contracts, plays a role in endorsing what are acceptable components of the toolkit.

3. THE CONTRIBUTION OF THEORY TO PRACTICAL MODEL DEVELOPMENT

3.1 The Role of Theory in Practice

The use of the above cross-sectional models within the transport planning process was, of course, informed by theoretical developments, often by default and sometimes more

explicitly than others. In principle, the theory of models should assert itself in four contexts: by providing a *description and explanation of current travel patterns* (base context); by providing guidance on the *development of behaviour over time* (forecasting context); by providing guidance on the appropriate form of model to be used as a *stimulus–response relationship* (policy testing); and by providing a link between the model as a predictor of response and evaluation measures to *appraise* alternatives (evaluation context).

I shall discuss the contribution of three phases of theoretical development in the UK: the entropy maximization/information theory approach; and behavioural explanations associated with the random utility discrete choice and activity-travel approaches. Much has been written on these different paradigms and I shall limit myself to summarizing their contribution to UK practice.

3.2 The Entropy/Information Theoretic Approach

The entropy/information theoretic approach brought some order to the wide range of spatial interaction models through classification according to the availability of independent estimates of trip ends. It also allowed, for the first time, the interpretation of dispersion in location (distribution), modal split and route split models to be unified (Wilson 1969, 1970). Here, share models could be viewed in terms of the maximum statistical likelihood of the probability distributions being observed, consistent with observable properties of those distributions and subject to any other imposed external constraints.

More generally, the approach brought analytical rigour and consistency to the derivation of location models, in individual spatial markets and as part of wider urban models (Wilson 1974). As the model formulation was expressed as an extremal problem it also provided the basis for rigorous solution and calibration of trip-based models.

Although the work provided the basis for interfacing cross-sectional or longitudinal models of travel budgets with estimates of cross-sectional trip patterns (which would allow the trip length parameters to be independently estimated for forecasting) this was restricted to academic studies (Wilson 1974).

While the traditional four-stage approach employed empirically derived trip length distributions for spatial interaction models and diversion curves were applied for modal split and (occasionally) route split, an emerging British style, through Wilson's formulation, was based on analytical models of the linked logit form. These were similar in form to the random utility discrete choice models being developed in the United States, which provided the clue to the establishment of a complete theory of the four-stage model structure based on the nested logit function (Williams 1977).

3.3 Emergence of Discrete Choice Theory

From the start there was always a latent behavioural basis for each of the four-stage models although this tended to be informal. Under the micro-approach the generation, distribution, modal split and assignment models came to be directly associated with frequency (f), location (l), modal choice (ms), and route choices (r), respectively, and with the decomposition of joint probabilities into marginal and conditional forms on the basis of 'sequential' or 'simultaneous' decision making.

The microeconometric approach offered a 'first principles' attack on aggregate forecast-

ing based on methodological individualism in which revealed behaviour was the outcome of a rational choice process.[6] The source of variability in individual trip making was attributed to observable and non-observable attributes, their distribution over the population determining the analytic form of models. The economic basis of the approach allowed a natural integration of choice models with consumer surplus welfare measures.

At any particular time, the structure and detailed form of existing models provides a touchstone for an emerging theory. Indeed, as Ortuzar (2001) has observed, both the multinomial logit model and the nested logit functional forms had been applied in analytic specifications prior to their derivation within the random utility discrete choice framework. Just as the multinomial logit model had been widely used (for example by Wilson in the UK) prior to its derivation within utility theory by McFadden (1973), so too had the nested form, the first application of which is attributed to the work of Ben-Akiva (1973).

The derivation and application of the multinomial logit model within the framework of random utility theory (McFadden 1973; Domencich and McFadden 1975) provided an attractive marriage between microeconomics and the statistics of individual travel choice data. The role of theory was found in the derivation of forms for the representative utility functions which allowed interpretation in terms of individual consumption behaviour and imposed restrictions on acceptable parameter signs.

In turn, the derivation of the nested logit form (Williams 1977; Daly and Zachary 1978) and its generalization to the general extreme value (GEV) class of models by McFadden (1978) addressed the problems of consistency and structural ambiguity in the overall demand function. As Ortuzar (2001) notes, the theory provided consistency with utility maximization in which the structure of a model was determined by the variance–covariance matrix of the utility distributions. It also resulted in inequalities between the parameter values in the various nests which allowed a model to be screened for behavioural consistency (essentially models with inappropriate cross-elasticities could be eliminated). Finally, it allowed the derivation of exact economic benefit measures consistent with the random utility basis of the model. Such measures included 'log sum' variables for logit-type models (Williams 1976, 1977).

The derivation of the nested logit model[7] thus provided a *post hoc* rationalization for, and suitable amendments to, the conventional structures G/D/MS/A, G/MS/D/A and G/D–MS/A which now could be discussed in terms of particular structures of the utility functions over multiple dimensions of choice. The parameter inequalities allowed the ordering of nested sub-models to be subject to empirical test. The theory thus provided necessary restrictions on the use of the model in policy testing and a successful integration with the evaluation context.

For those with a behavioural predisposition, the derivation of logit and nested logit models allowed a discrete choice reformulation of residential, employment and service location models to be established (Lerman 1975; McFadden 1978; Williams and Senior 1978; Anas 1982; and DSC&MEP 1999). It also provided a behavioural basis for the commuting and service trip making which integrated the basic employment, housing and service sectors in the Lowry model (Coelho and Williams 1978; Wilson et al. 1981).

The probabilistic discrete choice framework based on random utility theory was thus to provide the basis for reinterpreting and refinement of existing models and a range of new developments, some well beyond the transport sector. Ironically, in the late 1970s, by

making existing model forms behaviourally consistent, it laid bare the necessary simple microbehavioural assumptions required and thereby made it vulnerable to attack from much more precise descriptions of the decision-making context and process.

3.4 The Activity–Travel Framework

The development of the activity–travel framework (hereafter ATF) by Jones, Heggie and their colleagues at the Oxford University Transport Studies Unit, brought to fruition the work of earlier activity theorists, most notably Torsten Hagerstrand and F. Stuart Chapin Jr. By viewing travel as the outcome of activities in an institutional (household) perspective in which constraints, interdependencies, activity scheduling and coordination in space and time were considered essential to understand decisions and the basic motivation for travel, it provided a much richer description and explanation of travel behaviour (Jones 1979; Jones et al. 1983). The conceptual framework presented was one of considerable sophistication and, in principle, subsumed all travel behaviour models that went before it. Indeed, Jones and Heggie saw a typology of models determined by the strength of constraints and household interdependencies. The existing trip models of the random utility kind occupied a rather small portion of that framework. The limitations of trip/generalized cost-based forecasting models and the challenges posed by current societal changes and demand management policies have recently been reviewed by Jones (2002).

After 25 years of research into the microbehavioural aspects of individual and household travel, the ATF approach has failed to make a major impact on the mainstream of transport planning models in the UK. The reasons are not profound. The trip to tour transition was not straightforward and it was (and is) theoretically more challenging. Indeed, for many years it was not obvious how complex activity-time substitution within a tour-based framework could be effected within a practical framework. Isolated applications of tour-based modelling can be found in the UK, such as those by Polak et al. (1993) and Polak and Jones (see Jones 2002), to examine the implications of time shifting in response to road pricing in London. The RAND model also uses a tour-based approach (see note 3).

Despite the slow take-up of the approach, I would suggest that the basic tenets of the framework are now widely accepted. From a conceptual viewpoint, it has certainly provided an improved description of travel and use of time, their variation within the population, and wider range of potential adaptation to policy. Because it provides a richer description of travel behaviour and range of response mechanisms, this suggests to many that the impact of policies may be more subtle and substantial than hitherto considered (see, for example, Jones 2002).

In the early 1980s there was little common ground in either the ends or means of travel behaviour analysis. Exponents of the microeconometric approach largely saw their task in terms of *prediction* based on relatively simple behavioural hypotheses. Exponents of the ATF were more focused on providing a richer *explanatory* basis and improving current forecasting methods was not a priority.[8] Over the past 15 years the field has progressed towards a synthesis, with econometricians increasingly trying to incorporate constraints and key explanatory variables within tour-based approaches, while ATF researchers are increasingly embracing preferences through stated preference methods. While there have been many advances over the last decade which draw on the different behavioural tradi-

tions in modelling decision making within different time frames (see, for example, Ben-Akiva et al. 1996; Bowman and Ben-Akiva 2001; Bhat 2002; Miller and Salvini 2002) practical implementation remains slow. Certainly, UK researchers have yet to engage fully in the development of fifth generation behavioural forecasting models.

4. COMPLEXITY AND SIMPLIFICATION: A CAUTIONARY TALE

4.1 Introduction

All models are simplifications, by definition, and the balance between representational complexity and practical requirements is often a fine one, governed by the information sought, the study budget, and the state of the art, however contentious that might be. Government agencies and private consultancies seek practical modelling tools which are believed to be robust for the purpose, avoiding unjustified elaboration and mis-specification from overly simplistic forms.

As noted in Section 2, not infrequently, the basis for simplification is the suppression of one or more behavioural mechanisms or travel responses (setting the corresponding partial elasticities to zero). In the classical four-stage approach, for example, these were the assumptions that the frequency elasticity of demand and the cross-elasticity of demand between time periods were zero, leaving only the possibility of substitution between routes, modes and locations.

My third theme concerns the search for appropriate simplification in cases where more sophisticated demand specifications might have resulted in different decisions being made over the choice or viability of a scheme and even might have had different implications for policy development. I have taken a traditional policy of highway investment to provide the link with the past. If we consider the understanding of route choice and assignment models to be relatively secure, then the successful modelling of investment policies rests on the understanding and representation of induced traffic.[9]

4.2 Application of Fixed Matrix Approaches: A Simplified Approach

It is not uncommon in highway policy appraisals throughout the world to apply simplified approaches to the analysis of relatively small road schemes unless there are compelling reasons to adopt more sophisticated approaches. Certainly, in the UK, four- or five-stage models were regarded as overelaborate for the appraisal of small trunk road schemes outside urban areas. For this reason the UK Department of Transport has long applied simple methods based on a 'fixed matrix' (FM) assumption to determine the impact and economic implications (user benefits) from most trunk road schemes.

In the FM approach, the only response available to travellers is to switch route, all other responses being assumed negligible. Here, the trip matrix evolves over time in response to exogenous factors, including changes in land use, car ownership and the real cost of travel, but is insensitive to any travel time changes resulting from the policy itself. Increasingly, such zero-elasticity methods were applied in the appraisal of quite large schemes, some in peri-urban contexts.

On the face of it, application of FM methods might be regarded as foolhardy given the possibility of induced traffic and the wide availability of much more sophisticated models capable of greater precision. To condemn the practice on these grounds alone, however, would be harsh. It has always been appreciated by analysts that expansion of infrastructure *would* induce additional traffic (or vehicle miles) under a fall in equilibrium travel times – on the basis of simple economic arguments. However, it was thought that such additional traffic or vehicle miles generated by responses other than route switching would, first, be very small and, second, contribute *positively* to scheme benefits. FM estimates of user benefits were therefore seen as robust and inherently conservative. If a scheme could be justified by cost benefit analysis on the basis of the FM assumptions it would have an even stronger justification under variable trip matrix (VM) methods which accommodated other demand responses.[10]

4.3 SACTRA's Landmark Report: Roads and the Generation of Traffic

Following alarming national traffic forecasts for 2025 based on 1989, the apparent underprediction of traffic which accompanied some larger road schemes, and the prospect of many schemes operating in congested conditions for much of their lifetime, the government sought advice from SACTRA on the limitations of the FM approach. Could increasing congestion lead to suppression[11] of demand and subsequent induced traffic accompanying capacity expansion? Could these mechanisms significantly undermine the benefits to existing road users and result in an *upward* bias of FM estimates?

In their report addressing the existence and implications of induced traffic, SACTRA brought together a number of important issues on the design of models and their simplification. What the report did was to draw together considerable empirical evidence of highway scheme impacts, and the results of model applications, to examine the likely incidence and effects of induced traffic (see SACTRA 1994; Coombe 1996). Not surprisingly, the extent of additional traffic and/or vehicle miles and the degree to which user benefits are undermined by suppression and/or induction of demand is strongly dependent on the specification of the response mechanisms in the demand function. Those which give rise to *additional* trips will result in greater external effects (on existing trips) than those which involve substitution of *existing* journeys through retiming, relocation, modal switching and so on (see, for example, Williams and Lai 1991; Williams and Lam, 1991; Williams et al. 1991).[12]

While SACTRA did not identify the likely sizes of each potential behavioural response, which are difficult to measure, the committee concluded that significant quantities of induced traffic might occur in the following circumstances:

- where the elasticity of demand with respect to generalized cost is high;
- where the scheme results in a relatively large change in the travel cost;
- where the network is operating in a congested state for much of its lifetime.

Normally these would relate to urban and peri-urban situations as well as estuary crossings; strategic interurban schemes and motorway widening.[13] In such circumstances FM specifications might lead to significant *underestimation* of traffic on the network and corresponding *overestimation* of user benefits.

The committee report makes many recommendations for the development of improved

variable matrix procedures for scheme and strategic level assessment and evaluation. Where possible and justified by the scale of the problem it advocated that appraisal should be undertaken with a full range of responses.[14] As an interim measure the committee advocated the use of existing traffic assignment models in conjunction with simple elasticity models in which the range of elasticities used in sensitivity analysis reflected available knowledge.

It further proposed that extant four-stage models should be scrutinized by the Department of Transport and advice given on good practice; that strategic models should be audited to establish the credibility of their response properties; and that a wide range of research should be established on appropriate values of elasticities. It also placed admirable emphasis on the enhancement of the department's before-and-after monitoring of highway schemes to provide more information on induced traffic.

4.4 The Advice of the Department of Transport

The Department of Transport[15] acted swiftly on the SACTRA recommendations issuing two versions of a Guidance note. The second version (February 1997) is part of the *Design Manual for Roads and Bridges* (12.2.2) and contains advice on the form and application of road scheme appraisal methods. A guiding principle of the advice is the appropriate design of models to avoid overelaboration but be sufficiently rich in detail for the problem at hand.

The advice (Highway Agency 1997), supported by additional research on improved elasticities, spells out the order of the many individual response mechanisms in some detail and the circumstances in which they might be modelled either collectively or individually. The department reaffirmed that changing route is considered the dominant response in almost all cases, and contended that retiming of journeys is common and may be important in many cases. The likely impacts of the several responses other than route selection – retiming, modal switching, redistribution, frequency and land-use changes – are then used to classify contexts which are amenable to particular forms of model simplification. Specifically, circumstances are established in which it is believed to be appropriate to apply fixed matrix methods; simplified elasticity methods (in which all responses other than route selection are subsumed into a single parameter model); and more complex model forms requiring a fuller range of responses.

The department, along with several academic critics, was mindful of the potential *downward* bias of user benefit estimates from the application of simple generalized cost elasticity methods (which contain the FM approach as a limiting case) which do not distinguish between new journeys and changes in vehicle miles travelled from substitution associated with existing journeys.[16]

The department also provided advice on short- and long-term generalized cost elasticities which might be applied and used in sensitivity analyses. This advice is subject to ongoing research and, given its importance, inevitably attracts critical debate. Goodwin (1992, 1995, 1997) supports a higher range with a rule of thumb that long-run elasticities are double short-term values. He cites research on the effects of measures such as road space allocation, public transport improvements, travel plans, personalized marketing, and congestion charging to suggest that 'individual behaviour is much more variable and responsive than assumed'.

4.5 The Legacy of SACTRA (1994)

The SACTRA report was one of a number of influential studies in the mid-1990s which questioned the methods of appraising road schemes, particularly those in or near urban areas, and the potential for unanticipated induced traffic. Another influential report with an increased emphasis on energy consumption and environment impacts was that produced by the US Transport Research Board (TRB 1995). The report also inspired wider European work (Van Vuren and Daly 1996).

The SACTRA report and the following work undertaken by and on behalf of the UK Department of Transport, thus drew attention to potential mis-specification of widely applied methods; it clarified a range of response mechanisms and sought to put values on the corresponding elasticities; and it prescribed circumstances in which simplifications might be justified. It emphasized the importance of travel time switching as an important response and this has encouraged much research on this component of choice. Importantly, in the present discussion, it emphasized a process of model audit to identify good practice, and scheme monitoring for improving knowledge of response. Above all, the work provided valuable advice on the design of models and drew attention to contentious issues in the available knowledge of *response* to that most common of policies, the highway investment – knowledge which was once taken for granted.

5. THE SPECIFICATION AND EQUILIBRATION OF TRANSPORT PLANNING MODELS: SELECTED APPLICATION CONTEXTS

5.1 Introduction

For most of its history the four-stage model of the type discussed in Section 2 has been applied with relatively little concern for the convergence properties and accuracy of the algorithms designed for achieving overall equilibria between the demand for, and supply of, transport services. While external 'feedback' of costs – a term which Boyce views with some dismay – from the assignment stage to remaining response mechanisms has been discussed and undertaken for 30 years, the importance of appropriate algorithms has emerged only in the last decade or so (Boyce et al. 1994; Boyce 1998b). This is of some importance given that local and central governments are willing to allocate very large capital sums on the basis of a cost–benefit calculus, the assumption of constant unit values of time, and the ability to calculate and evaluate what are often *very* modest changes in travel time.[17] The UK Department of Transport has therefore, as a continuation of an interest and concern for induced traffic, invested in research into improving the specification and solution of equilibrium models[18] (see Bates et al. 1999 for a discussion of the broad research strategy).

All the topics considered in my final theme were discussed in, inspired by, or emerged as a reaction to the SACTRA report (1994). I consider three application contexts in which I believe that more research into the specification and solution of equilibrium models is justified. At their heart, all concern the prediction and implications of induced traffic from its many sources. The following contexts are considered:

- equilibrium states in multi-modal systems under investment policies;
- the value of highway investments under different pricing regimes; and
- the effect of capacity reallocation in the context of elastic demand.

5.2 Investment in Multi-modal Systems

The use of multi-modal equilibrium models is advocated for highway and public transport investment schemes in which significant congestion relief is likely (DETR 2000; DfT 2002). In the case of LRT schemes, for example, the prediction of de-congestion benefits arising from modal substitution is prominent in the case for government funding. The systematic treatment of induced traffic in this context appears to be a relatively recent concern. What can be affirmed is that the models developed to date for appraising LRT schemes have not generally dealt with model convergence. In other words, such models do not incorporate 'feedback by which new costs were extracted with LRT included and fed back into the modal split model in an iterative fashion until a pre-defined level of convergence was achieved' (DfT 2002; Annex D of Part 1). It would be of interest to explore the sensitivity of such benefits to model specification and solution algorithms to examine what simplifications are justified. If only modal switching as a source of induced demand is considered, accurate solution for this specification is well within the current state of the art.

Of more theoretical interest is the nature and calculation of short- and long-term equilibria in multi-modal systems in which supply side responses of public transport operators to highway investment are directly considered. This relates directly to the unresolved status of the Downs–Thomson paradox (Downs 1962, 1992; Thomson 1977; Mogridge 1990, 1997) which rests on the existence of latent demand and subsequent release of induced traffic. Can highway investments, in the presence of competing public transport services subject to external economies of scale, be *absolutely* counterproductive and lead to an increase in equilibrium travel times on both modes? Although this does not constitute conventional wisdom, the SACTRA report recognized the *potential* significance of the effect and called for more research.

While it is straightforward to demonstrate how the paradox arises if individuals take the minimum cost path in multi-modal hyper-networks (see, for example, Mackie and Bonsall 1989), will this still arise if mode-switching behaviour is governed by logit-type share models? The limiting case of zero dispersion suggests that paradoxical behaviour must be part of the *general* solution. The nature of equilibrium states in multi-modal systems has been subject to a range of simulation and analytical studies under particular simplifications (see, for example, Bly et al. 1987; Williams et al. 1991; Williams 1998). Using a simplified model of modal competition under decreasing public transport user costs, Williams (1998) has demonstrated analytically that equilibrium states form distinct regimes according to the values of the direct and cross-elasticity of modal demand, and the extent of scale economies on the public transport mode. The paradox will arise when these parameters satisfy certain inequality conditions and will disappear otherwise. The existence and stability of such equilibrium states in networks, and their response to policies under models which embody a *full range* of induced traffic effects is, in the author's view an unresolved research question, as is the existence of such regimes in real cities.

5.3 Value of Investments under Different Pricing Regimes

The SACTRA report itself did not go unscathed. Foster (1995) brushed it aside as effectively addressing the wrong question. For given capacity expansion in congested conditions the benefit, and therefore the case for support of a scheme, will typically be *less* under variable demand estimates compared with fixed (zero elasticity) demand. This, he stressed, tells us nothing about the optimal investment in roads. Instead, he saw the existence of substantial amounts of induced traffic as evidence of *under*investment in capacity (subject to the full accounting for environmental effects). He urged for the problem to be taken back into the traditional heart of transport economics, and re-expressed as the determination of optimal capacity expansion under different road pricing regimes. While this topic has been studied for 40 years with single link models (see, for example, Thomson 1970; Small et al. 1989), wider network applications for elastic demand models are much more recent.

The system of prices needed to convert a user optimum equilibrium to a social welfare optimal, corresponding to the two principles of Wardrop (1952), has received much theoretical and increasing practical interest in the design of practical road pricing schemes. For a recent investigation, see Boyce et al. (2001). Using the SATURN model with elastic demand, Williams et al. (2001a,b) have examined the extent to which the value of an investment is upwardly biased because of its evaluation with respect to highly inefficient (unpriced) reference states. They examined the implications of charging policies for congestion, traffic growth and the value of investments (expressed in terms of user and emission benefits) over a 20-year period for various highway schemes within a Cardiff network. With the selected elastic demand model the results were sensitive to the base level of congestion, the elasticity of demand and the nature of the policy. For the selected policies and demand and cost functions, it was found that the benefit of an investment for a network under free use could exceed that under optimal pricing by up to 25 per cent, although it would typically be considerably less than this. It is not known how such results would differ with models which distinguished the different sources of induced traffic, particularly travel time switching, modal and locational substitution.

5.4 Can Capacity Reduction 'Degenerate' Traffic?

Just as the SACTRA report (1994) led to concern in the UK for the extent to which road capacity expansion could generate additional traffic and vehicle miles, it also inspired a study of the converse effect. To what extent could road capacity reallocation (in favour of high capacity vehicles or cycle lanes) lead to a 'degeneration' of traffic from the wide range of possible responses? On standard micro-assignment calculations under zero demand elasticity, reduced capacity would imply heavy congestion and lengthening queues, to the extent that such policies would be very unlikely to be considered by traffic planners. Empirical observation, however, suggests that cities seldom experience prolonged gridlock and that, even in the short term, travellers adapt to significant changes in capacity.

Cairns et al. (1998) undertook wide-ranging empirical studies of capacity reduction and reallocation in urban contexts. They concluded, on the basis of before-and-after studies of a wide range of international schemes and enforced capacity changes that on average, about 15 per cent of traffic could be said to have 'evaporated' through a variety

of response mechanisms other than route switching. They concluded that the effects of capacity reallocation would be less disruptive and have fewer disbenefits than conventional wisdom (based on zero demand elasticity) would suggest.

As part of the study, MVA (1998) make a valiant attempt to give guidance in a field where relatively little of relevance to UK cities is known. They show how FM assignments can be used to give information about the need to apply more sophisticated models incorporating time shifting and other responses.

6. CONCLUSION

I draw the following conclusions from the above discussion.

First, the development of applied travel forecasting models in the UK over the past 40 years shows many points of progress in the specification, estimation and the presentation of the information they generate. The degree of choice for practitioners at any particular time, including today, is however sometimes more apparent than real. The differences between consultants' proprietary model software, where it has been designed to serve the same purpose, are often relatively superficial. In view of the wide range of contexts, applied models are often minor variants (perhaps simplifications) of more complex forms. Nevertheless there are alternatives available both in the specification of models and in the use of data, and these have been noted.

In the UK and elsewhere, consultants have adopted a tool-kit approach and the Department of Transport is increasingly providing significant advice on the acceptability of analysis methods, particularly where the government is involved in resource allocation.

Second, significant conceptual progress has also been made and we now have a sophisticated framework within which travel behaviour and response can be described and explained. Within specific theoretical paradigms we understand a little better the gaps in knowledge and some of the challenges, and that represents considerable progress. However, in my view, the theoretical base of the *practical* forecasting models for both traditional and novel policy interventions still awaits appropriate behavioural refinement. I attribute the relatively slow progress, in part, to the persistence of a very restricted view of the validity of models.

It is now recognized that models, such as the logit and nested logit forms, can be generated from a range of theoretical assumptions (see, for example, Boyce 1998a). This should not necessarily be interpreted as a strength because these are essentially *post hoc* rationalizations of traditional forms which have developed a life of their own. Just as theory has had implications for practice, in turn established models have clearly influenced subsequent theoretical developments. Ideally, alternative theories would offer distinct predictions which could be subject to empirical test. Before-and-after studies, always valuable, are the key to testing alternative propositions and will be indispensable as we will need to assess different models.

The third theme concerned the importance of models as response functions and considered the variation of their predictions to alternative specifications in circumstances where the scale of an investment or available study resources might not justify complex model forms. I believe that the SACTRA report and subsequent work have much wider implications for the assessment of travel forecasting methods and available alternative

approaches. The emphasis on model auditing, identification of good practice, establishing appropriate simplifications, and accumulation of knowledge from before-and-after studies to assess model validity, tied directly to model use, are crucial messages.

Finally, I have used the equilibrium multi-modal context to draw attention to what I believe are important research issues in the specification and application of equilibrium models. Even if models have deficiencies in their specification we are usually committed to their accurate solution because the outputs of the process – often the *changes* in volumes and travel times accompanying policy intervention – are typically *very* modest. I share David Boyce's concern that insufficient concern is given to the consistent and accurate solution of equilibrium models.

It is seldom the case that transport professionals disagree on the direction of change accompanying a policy. The Downs–Thomson effect is a rare case in our field where this is so and is particularly appealing to those who see road development having no significant part to play in sustainable transport systems. Inexplicably, the longer-term responses of public transport operators, in both regulated and deregulated environments, have not been widely studied in the context of *network* investments. I believe that this and the wider issues of capacity changes under different pricing regimes, merit further research, again drawing on as full a range of behavioural mechanisms as possible.

To conclude, transport planners will always operate within an environment of changing policy and evaluation contexts, imperfect and incomplete information, imprecise theories and challenging study budgets, and we can expect many tensions to persist. The scale of public money allocated to address technical issues of forecasting in the UK is not great and pales by comparison with that devoted to the development of TRANSIMS. The UK is in a good position to learn from the TRANSIMS experience. The above comments suggest that a 'one size fits all' approach might not be particularly useful and that much thought should be given to distinct application contexts where different model designs are appropriate.

David Boyce's call for greater standardization of models and advice on best practice has much to recommend it and I believe that many of the above comments support that view. I can see three things which might militate against the attempt:

- By its nature, the state of the art is evolving and would inevitably be contested by both academics and consultancies[19] (many of which are highly innovative). Reputations and commercial interests are at stake!
- Diversity of context itself will require specification of a potentially large range of models to address the large and growing range of policies. International standardization might prove very problematic in these circumstances.
- In the light of this range of application contexts and study budgets, consultants might wish to retain flexibility to apply models of differing degrees of complexity.

In any attempt at standardization, the potential danger of inhibiting innovation and the possibility of stagnation should be acknowledged.

I believe that the logical consequences of David Boyce's comments, supported by the above comments, are the following:

- there should be a clarification of the criteria by which different *operational* models should be compared;

- an attempt should be made to audit models with the intention of identifying good practice;
- a programme should be established which involves the comparative assessment of different models and their predictions with respect to a range of policies. Common databases should be selected on the availability of before-and-after information accompanying specific policies to allow a close scrutiny of model validity.

I believe that strengthening the empirical basis for models as stimulus–response functions and broadening the notion of model validity to embrace forecasting and response contexts, will be essential if we are to discriminate between competing modelling paradigms and assess the contribution of innovations. In this regard, the transfer of experience and the accumulation of before-and-after case study material will be as important as ever.

NOTES

* I have benefited from conversations or correspondence with John Bates, Peter Bonsall, Andrew Daly, Hugh Gunn, Brian Martin, Juan de Dios Ortuzar, John Polak, Stuart Porter, David Simmons and Luis Willumsen. I alone am responsible for any errors of fact, interpretation or omission.

1. In this chapter I shall not consider the changing nature of the land-use and transport planning processes in the UK, an excellent account of which is given by Banister (2002). Others have addressed the evolving role of models within the British transport planning system (Bonsall 1998; Mackett 1998).
2. Although it sometimes proved necessary for ad hoc insertion of often large parameters (K-factors) into spatial interaction models in order to achieve this.
3. For a discussion of the practical distinction between the four-stage approach and the disaggregate model approach, see IHT (1996). To the author's knowledge, some of the most ambitious applications at national, regional and urban level have been by RAND Europe (formally Hague Consulting Group). The RAND approach applies nested micro-models subjected to equilibration. The approach models primary destination tours, detours are then modelled separately. Here, acknowledgement is also due to US researchers, particularly Ben-Akiva and McFadden, whose work in the Bay Area pointed the way for the series of models developed in Holland from the late 1970s onwards.
4. An example in point is the TRIPS suite of MVA (1998) which provides a flexible set of models in logit, nested logit and incremental logit form to address individual or grouped data.
5. As part of the APRIL suite, the incremental nested logit demand function was used to assess the changes in time of day, mode, destination and frequency which accompanied congestion charging in London (see Williams and Bates 1993).
6. Although decision making is usually considered to be one of utility maximization over relatively simple utility functions to attain analytical tractability, this may readily be relaxed to embrace a wide range of decision rules such as satisficing, at the expense of requiring simulation as a solution method (Williams and Ortuzar 1982; McFadden 2001).
7. Its properties are still discussed and debated in the literature (see Carrasco and Ortuzar 2002; McFadden 2001; and Bhat 2002).
8. The behavioural study of travel behaviour reveals the same tensions and complementarity between different approaches to the study of general household consumption. A good discussion of the contribution and limitations of neoclassical and institutional approaches to consumer theory can be found in Himmelweit et al. (1998).
9. Essentially, the retiming of journeys, travel to new destinations for the same purpose, modal switching, changes in car occupancy, changes in activity allocation, changes in the linkage of trips, increasing the frequency of some journeys, changes in car ownership, changes in residential and/or employment location, changes in the pattern of land use.
10. Two main classes of error are involved in the derivation of user benefits from FM methods. Type I errors arise because the travel cost in the reference state will be higher than those under a more responsive demand function; while Type II errors arise because, under the investment, the equilibrium generalized cost in congested conditions will normally be higher than those under zero elasticity (see Williams and Yamashita 1992a).

11. I use this term to refer to *all* those behavioural mechanisms included in the study of traffic induction which might deter a person from travelling at a particular part of the highway system at a particular time.
12. The net effects of induced traffic on *user* benefits are determined by the relative sizes of a direct positive contribution of any new traffic which receive a benefit from an investment, on the one hand, and the negative external (to themselves) effect which is inflicted on existing traffic, on the other. However, in the case of *environmental* effects, such as emission (dis)benefits, any additional traffic or vehicle miles will contribute to disbenefit due to *both* a direct contribution from the additional vehicle miles and the indirect external contribution of any changes in speed on existing traffic (Williams and Moore 1990; Williams et al. 2001a).
13. The *theoretical* basis for SACTRA's conclusions is unremarkable. It relates directly to the concerns of transport economists in the 1960s and 1970s for latent demand and elastic generation of trips which might accompany investment in highways (see, for example, Thomson 1970). Indeed, simplified equilibrium models may be used to generate analytical results for the interrelationship between induced traffic, traffic growth, congestion, and user benefits from highway investment (see, for example, Williams and Moore 1990; Williams and Yamashita 1992a). The economic arguments arising from SACRA are summarized by Mackie (1996).
14. The specific recommendation (SACTRA 1994, para. 13.23) was 'where these [four-stage transportation] models exist, in areas where trunk road schemes are planned, the calibration and validation . . . is scrutinised by the Department and, if proved satisfactory, they are used in the appraisal of those schemes. Where necessary, existing models should be enhanced, so that they are able to estimate all the important demand responses to road provision, including trip frequency and choice of time of travel'.
15. The government department responsible for transport has changed its name several times over this period. To avoid confusion I shall refer simply to the Department of Transport.
16. Not withstanding this deficiency, the author has supported the view that, *in the absence of more complete models*, a single generalized cost elasticity model should be used and applied in a sensitivity mode with elasticity values ranging from 0 (the FM limit) to −0.5. He would also support the assertion that response mechanisms, other than route switching, will *in congested conditions* over the lifetime of a scheme depress the present value of benefits by between 10 and 30 per cent (Williams et al. 1991; Williams and Yamashita 1992b). The application of a nested incremental model with partial elasticities would, as a generalization of the simple elasticity form, be a more satisfactory approach.
17. While some schemes of national and regional importance can give rise to substantial time savings, many road schemes in mature networks, result in an average time saved *to those receiving a benefit* of the order of fractions of a minute to a few minutes (Welch and Williams 1997). The aggregate user benefit is the sum of a large number of small time savings all of which are assumed to have the same *unit* value.
18. The Department of Transport (DfT 2002) is currently undertaking research into a model system DIADEM (Development of Integrated Assignment and Demand Modelling) to develop improved equilibrium models and algorithms.
19. Here I make the distinction between consultants and clients for their services, often local authority transport planners. In the UK, modelling by local authorities is increasingly contracted out to national and international consultants, and a particular problem is the retention of appropriate skills at the local level.

REFERENCES

Anas, A. (1982), *Residential Location Markets and Urban Transportation: Economic Theory, Econometrics and Policy Analysis with Discrete Choice Models*, New York: Academic Press.

Banister, D. (2002), *Transport Planning*, London: Spon Press.

Bates, J.J. (ed.) (1988), 'Stated preference methods in transport research', *Journal of Transport Economics and Policy*, **22**, 1–137.

Bates, J.J., D.J. Ashley and G. Hyman (1987), 'The nested incremental logit model: theory and application to modal choice', PTRC Summer Annual Meeting, Seminar C, University of Bath, 7–11.

Bates, J.J., D. Coombe, S. Porter and D. Van Vliet (1999), 'Allowing for variable demand in highway scheme assessment', PTRC Summer Annual Meeting.

Bates, J.J., I. Williams and J. Leather (1996), 'The London congestion charging research programme: 4. The transport models', *Traffic Engineering and Control*, **37** (5), 334–9.

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

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

Ben-Akiva, M.E., J.L. Bowman and D. Gopinath (1996), 'Travel demand model system for the information era', *Transportation*, **23**, 241–66.
Bhat, C.R. (2002), 'Recent methodological advances relevant to activity and travel behaviour analysis', Chapter 19 in H.S. Mahmassani (ed.), *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*, Oxford: Pergamon, pp. 381–414.
Bly, P.H., R.H. Johnston and F.V. Webster (1987), 'A panacea for road congestion?', *Traffic Engineering and Control*, **28** (1), 8–12.
Bonsall, P.W. (1998), 'Evolving role of models in transport planning', *Impact Assessment and Project Appraisal*, **16** (2), 100–104.
Bowman, J.L. and M.E. Ben-Akiva (2001), 'Activity-based disaggregate travel demand model system with activity schedules', *Transportation Research*, **35A**, 1–28.
Boyce, D.E. (1998a), 'Long-term advances in the state-of-the-art of travel forecasting methods', in P. Marcotte and S. Nguyen (eds), *Equilibrium and Advanced Transportation Modelling*, Boston/Dordrecht/London: Kluwer Academic, pp. 73–86.
Boyce, D.E. (1998b), *A Practitioner's Guide to Urban Travel Forecasting Models*, Department of Civil and Materials Engineering, University of Illinois at Chicago.
Boyce, D.E., K. Balasubramaniam and X. Tian (2001), 'Implications of marginal cost road pricing for urban travel choices and user benefits', in P. Marcotte and S. Nguyen (eds), *Current Trends in Transportation and Network Analysis*, Dordrecht: Kluwer Academic, pp. 37–48.
Boyce, D.E., Y. Zhang and M.R. Lupa (1994), 'Introducing "feedback" into four-step travel forecasting procedure versus equilibrium solution of combined model', *Transportation Research Record*, **1443**, 65–74.
Cairns, S., C. Hass-Klau and P. Goodwin (1998), *Traffic Impact of Highway Capacity Reductions: Assessment of the Evidence*, London: Landor.
Carrasco, J.A. and J. de D. Ortuzar (2002), 'A review and assessent of the nested logit model', *Transport Reviews*, **22** (2), 197–218.
Coelho, J.D. and H.C.W.L. Williams (1978), 'On the design of land use models through locational surplus maximisation', *Papers of the Regional Science Association*, **40**, 71–85.
Coombe, D. (1996), 'Induced traffic: what do transportation models tell us?', *Transportation*, **23**, 83–101.
Daly, A.J. (1987), 'Estimating "tree" logit models', *Transportation Research*, **21B**, 251–68.
Daly, A.J. (1992), *ALOGIT3.2 User's Guide*, Hague Consulting Group.
Daly, A.J. and S. Zachary (1978), 'Improved multiple choice models', in D.A. Hensher and M.Q. Dalvi (eds), *Determinants of Travel Choice*, Westmead: Saxon House.
David Simmonds Consultancy with Marcial Echenique and Partners Ltd (DSC&MEP) (1999), 'Review of land-use/transport interaction models', DETR, London, www.roads.dft.gov.uk/road-network/sactra/support99/index.htm
Department of the Environment, Transport and the Regions (DETR) (2000), 'Guidance on the Methodology for Multi-Modal Studies (GOMMMS)', www.dft.gov.uk/itwp/mms/index.htm
Department for Transport (DfT) (2002), *Major Scheme Appraisal in Local Transport Plans*, Parts 1, 2, 3, Detailed Guidance on Public Transport and Highway Schemes, Department for Transport, www.local-transport.dft.gov.uk/msapart1/01.htm
Domencich, T. and D. McFadden (1975), *Urban Travel Demand: A Behavioral Analysis*, Amsterdam: North-Holland.
Downs, A. (1962), 'The law of peak-hour expressway congestion', *Traffic Quarterly*, **16**, 393–409.
Downs, A. (1992), *Stuck in Traffic: Coping with Peak-hour Traffic Congestion*, Washington, DC, Brookings Institute.
Druit, S. (2000), 'An introduction to PARAMICS', www.sias.co.uk/sias/paramics/articles/article1.html
Foster, C.D. (1995), 'The dangers of nihilism in roads policy', *Proceedings of the Chartered Institute of Transport*, **4** (2) 22–45.
Fowkes, A.S. (1991), 'Recent developments in stated preference techniques in transport research', Proceedings of the 19th PTRC Summer Annual Meeting, University of Sussex, Brighton, UK.
Fowkes, A.S. and J. Preston (1991), 'Novel approaches to forecasting the demand for new local rail services', *Transportation Research*, **25A** (4), 209–18.

Goodwin, P.H. (1992), 'A review of demand elasticities with special reference to the short and long run effects of price changes', *Journal of Transport Economics and Policy*, **26** (2), 155–70.
Goodwin, P.H. (1995), 'Empirical evidence on induced traffic', *Transportation*, **23**, 35–54.
Goodwin, P.H. (1997), 'Extra traffic induced by road construction: empirical evidence, economic effects and policy implications', in *Infrastructure-induced Mobility*, European Conference of Ministers of Transport, Paris: OECD, pp. 147–200.
Gunn, H.F., M. Bradley and C. Rohr (1996), 'The 1994 National Value of Time Study of Road Traffic in England', PTRC Seminar on the Value of Time, Berkshire, UK. 29–30 October.
Hague Consulting Group and Accent Marketing and Research (HCG/Accent) (1996), 'The value of travel time on UK roads – 1994', prepared for the UK Department of Transport.
Highways Agency (1997), *Design Manual for Roads and Bridges, Vol. 12, Traffic Appraisal of Road Schemes, Section 2, Part 2: Induced Traffic Appraisal*, London: The Stationery Office.
Himmelweit, S., A. Trigg, N. Costello, G. Dawson, M. Mackintosh, R. Simonetti and J. Wells (1998), *Understanding Economic Behaviour: Households*, Milton Keynes: Open University.
Institution of Highways and Transportation (IHT) (1996), *Guidelines for Developing Urban Transport Strategies*, London: IHT, Chapter 6, pp. 128–55.
Jones, P.M. (1979), 'New approaches to understanding travel behaviour: the human activity approach', in D.A. Hensher and P.R. Stopher (eds), *Behavioral Travel Modelling*, London: Croom Helm, pp. 50–80.
Jones, P.M. (2002), 'Setting the research agenda: responses to new transport alternatives and policies', Chapter 1 in H.S. Mahmassani (ed.), *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*, Oxford: Pergamon, pp. 3–22.
Jones, P.M., M.C. Dix, M.I. Clarke and I.G. Heggie (1983), *Understanding Travel Behaviour*, Aldershot: Gower.
Lerman, S.R. (1975), 'A disaggregate behavioral model of urban mobility decisions', PhD thesis, Department of Civil Engineering, MIT, Cambridge, MA.
London County Council (1964), *London Traffic Survey*; Volume 1, Chapters 1–9, LCC, London.
Mackett, R.L. (1998), 'Role of travel demand models in appraisal and policy-making', *Impact Assessment and Project Appraisal*, **16** (2), 91–9.
Mackie, P.J. (1996), 'Induced traffic and economic appraisal', *Transportation*, **23**, 103–19.
Mackie, P.J. and P.W. Bonsall (1989), 'Traveller response to road improvements: implications for user benefits', *Traffic Engineering and Control*, **30**, 411–16.
McFadden, D. (1973), 'Conditional logit analysis of qualitative choice behaviour', in P. Zarembka (ed.), *Frontiers in Econometrics*, New York: Academic Press, pp. 105–42.
McFadden, D. (1978), 'Modelling the choice of residential location', in A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull (eds), *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland, pp. 75–96.
McFadden, D. (2001), 'Disaggregate behavioral travel demand's RUM side: A 30-year retrospective', in D. Hensher (ed.), *Travel Behaviour Research: The Leading Edge*, Amsterdam: Pergamon, pp. 17–64.
Miller, E.J. and P. Salvini (2002), 'Activity-based travel behaviour modelling in a microsimulation framework', Chapter 26 in H.S. Mahmassani (ed.), *Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*, Oxford: Pergamon, pp. 533–58.
Mogridge, M.J.H. (1990), *Travel in Towns: Jam Yesterday, Jam Today and Jam Tomorrow?*, London: Macmillan.
Mogridge, M.J.H. (1997), 'The self-defeating nature of urban road capacity policy', *Transport Policy*, **4** (1), 5–23.
MVA Consultancy (1998), *TRIPS*, Woking, Surrey: MVA House.
MVA Consultancy, ITS University of Leeds, TSU University of Oxford (1987), 'Value of Travel Time Savings', Policy Journals, Newbury, Berks.
MVA (1998), *Traffic Impact of Highway Capacity Reductions: Report on Modelling*, Landor Publishing, London.
Ortuzar, J. de D. (2001), 'On the development of the nested logit model', *Transportation Research*, **35B**, 213–16.
Ortuzar, J. de D. and L. Willumsen (2001), *Modelling Transport*, 3rd edn, London and Chichester: John Wiley & Sons.

Paulley, N. and F.V. Webster (1991), 'Overview of the international study to compare models and evaluate land use and transport policies', *Transport Reviews*, **11**, 197–222.
Polak, J., P. Jones, P. Vythoulkas, R. Sheldon and D. Wofinden (1993), 'Travellers Choice of Time of Travel Under Road Pricing', Report to the UK Department of Transport.
Quarmby, D. (1967), 'Choice of travel mode for the journey to work: some findings', *Journal of Transport Economics and Policy*, **1**, 273–314.
Standing Advisory Committee for Trunk Road Assessment (SACTRA) (1994), *Trunk Roads and the Generation of Traffic*, London: The Stationery Office.
Standing Advisory Committee for Trunk Road Assessment (SACTRA) (1999), *Transport and the Economy*, London: The Stationery Office.
Senior, M.L. and H.C.W.L. Williams (1977), 'Model-based transport policy assessment. Part I: The use of alternative forecasting models', *Traffic Engineering and Control*, **18** (9), 402–6.
Sheldon, R. and J. Steer (1982), 'The use of conjoint analysis in transport research', Proceedings of the 10th PTRC Summer Annual Meeting, Seminar Q, University of Warwick.
Small, K.A., C. Winston and C.A. Evans (1989), *Road Work: A New Highway Pricing and Investment Policy*, Washington, DC: Brookings Institution.
Thomson, J.M. (1970), 'Some aspects of evaluating road improvements in congested areas', *Econometrica*, **38** (2), 298–310.
Thomson, J.M. (1977), *Great Cities and their Traffic*, London: Gollancz (Peregrine).
Transport Research Board (TRB) (1995), 'Expanding metropolitan highways: implications for air quality and energy use', Special Report 245, Transport Research Board, National Research Council, Washington, DC.
Van Vliet, D. (1982), 'SATURN: a modern assignment model', *Traffic Engineering and Control*, **23**, 575–81.
Van Vliet, D. and M. Hall (1998), *SATURN 9.4 User Manual*, Institute for Transport Studies, University of Leeds and W.S. Atkins Planning Consultants, Epsom, Surrey.
Van Vuren, T. and A. Daly (1996), 'Forecasting induced traffic on large-scale transport infrastructure in Europe', Proceedings of the 24th PTRC Summer Annual Meeting, Seminar A,
Wardman, M. (1988), 'Comparison of revealed preference and stated preference models of travel behaviour', *Journal of Transport Economics and Policy*, **22**, 71–91.
Wardman, M. (1998), 'The value of travel time: a review of British evidence', *Journal of Transport Economics and Policy*, **32** (3), 285–316.
Wardrop, J.G. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institute of Civil Engineers*, Part II, **1**, pp. 325–78.
Webster, F.V., P.H. Bly and N.J. Paulley (eds) (1988), *Urban Land-use and Transport Interaction: Policies and Models*, Aldershot: Gower.
Welch, M. and H.C.W.L. Williams (1997), 'The sensitivity of transport investment benefits to the evaluation of small travel-time savings', *Journal of Transport Economics and Policy*, **31** (3), 231–54.
Williams, H.C.W.L. (1976), 'Travel demand models, duality relations and user benefit analysis', *Journal of Regional Science* **16** (2), 147–66.
Williams, H.C.W.L. (1977), 'On the formation of travel demand models and economic evaluation measures of user benefit', *Environment and Planning*, **9A**, 285–344.
Williams, H.C.W.L. (1998), 'Congestion, traffic growth and transport investment: the influence of interactions and multiplier effects in related travel markets', *Journal of Transport Economics and Policy*, **32**, 141–63.
Williams, H.C.W.L. and H.S. Lai (1991), 'Transport policy appraisal with equilibrium models II: Model dependence of highway investment benefits', *Transportation Research*, **25B** (5), 281–92.
Williams, H.C.W.L. and W.M. Lam (1991), 'Transport policy appraisal with equilibrium models I: Generated traffic and highway investment benefits', *Transportation Research*, **25B** (5), 253–79.
Williams, H.C.W.L., W.M. Lam, J. Austin and K.S. Kim (1991), 'Transport policy appraisal with equilibrium models III: Investment benefits in multi-modal systems', *Transportation Research*, **25B** (5), 293–316.
Williams, H.C.W.L. and L.A. Moore (1990), 'The appraisal of highway investments under fixed and variable demand', *Journal of Transport Economics and Policy*, **24**, 61–81.

Williams, H.C.W.L. and J.D. Ortuzar (1982), 'Behavioural theories of dispersion and the mis-specification of travel demand models, *Transportation Research*, **16B** (3), 167–219.
Williams, H.C.W.L. and M.L. Senior (1977), 'Model-based transport policy assessment. Part II: Removing fundamental inconsistencies from the models', *Traffic Engineering and Control*, **10**, 464–9.
Williams, H.C.W.L. and M.L. Senior (1978), 'Accessibility, spatial interaction and the spatial benefit analysis of land-use transportation plans', in A. Karlqvist, L. Lundqvist, F. Snickars and J.W. Weibull (eds), *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland, pp. 253–87.
Williams, H.C.W.L., Vliet D. Van and K.S. Kim (2001a), 'The contribution of suppressed and induced traffic in highway appraisal. Part I: Reference states; Part II: Policy tests', *Environment and Planning*, **33A**, 1057–81 and (II) 1243–64.
Williams, H.C.W.L., Vliet D. Van, C. Parathira and K.S. Kim (2001b), 'Highway investment benefits under alternative pricing regimes', *Journal of Transport Economics and Policy*, **35** (2), 257–84.
Williams, H.C.W.L. and Y. Yamashita (1992a), 'Travel demand forecasts and the evaluation of highway schemes under congested conditions', *Journal of Transport Economics and Policy*, **26** (3), 261–82.
Williams, H.C.W.L. and Y. Yamashita. (1992b), 'Equilibrium forecasts of travel demand and investment benefit measures for congested transport networks', *Proceedings of the PTRC Summer Annual Meeting*, University of Manchester, Vol. P357, 53–76.
Williams, I.N. and J.J. Bates (1993), 'APRIL – a strategic model for road pricing', Proceedings of the 21st PTRC Summer Annual Meeting, Seminar D, London.
Willumsen, L.G. (1982), 'Estimation of trip matrices from traffic counts: validation of a model under congested conditions', Proceedings of the 10th PTRC Summer Annual Meeting, University of Warwick.
Willumsen, L.G., J. Bolland, Y. Arezki and M. Hall (1993), 'Multi-modal modelling in congested networks: SATURN and SATCHMO', *Traffic Engineering and Control*, **34**, 294–301.
Wilson, A.G. (1967), 'A statistical theory of spatial distribution models', *Transportation Research*, **1**, 253–69.
Wilson, A.G. (1969), 'The use of entropy maximising models in the theory of trip distribution, mode split and route split', *Journal of Transport Economics and Policy*, **3**, 108–26.
Wilson, A.G. (1970), *Entropy in Urban and Regional Modelling*, London: Pion.
Wilson, A.G. (1974), *Urban and Regional Models in Geography and Planning*, Chichester and New York: John Wiley.
Wilson, A.G. (1998), 'Land-use/transport interaction models: past and future', *Journal of Transport Economics and Policy*, **32** (1), 3–26.
Wilson, A.G., J.D. Coelho, S.M. Macgill and H.C.W.L. Williams (1981), *Optimization in Locational and Transport Analysis*, Chichester and New York: John Wiley.
Wilson, A.G., A.F. Hawkins, G.J. Hill and D.J. Wagon (1969), 'Calibrating and testing the SELNEC transport model', *Regional Studies*, **3**, 337–50.
Wootton, H.J. and G.W. Pick. (1967), 'A model for trips generated by households', *Journal of Transport Economics and Policy*, **1** (2), 137–53.

2. A combined distribution, hierarchical mode choice, and assignment network model with multiple user and mode classes

K.I. Wong, S.C. Wong, J.H. Wu, Hai Yang and William H.K. Lam*

1. INTRODUCTION

The traditional four-step transportation planning model, consisting of trip generation, trip distribution, modal split, and assignment stages has been used for more than four decades, and is still popular around the world. The four-step model is a top-down sequential process, in which the output from the upper stages forms the input to the lower stages. As the modeling structure is not recursive, the results obtained from the different stages might not be entirely consistent. Moreover, repeated updating and analysis of this top-down sequential process might not guarantee convergence, and hence the problem of inconsistency remains. In the last two decades, this problem has been addressed by many researchers who have sought to obtain consistent results from different stages of transportation models. Formulations have been proposed at different stages or at combinations of stages to integrate and enhance the modeling structure. This is known in the literature as the combined models.

One of the earliest attempts at this approach was to combine the distribution stage with the assignment stage (Florian et al. 1975; Evans 1976). Such efforts were extended to two transportation modes (auto and transit) by Florian and Nguyen (1978), where the travel times of the two modes were not related; and by Florian (1977), where the travel times were interrelated. Alternative formulations of the combined urban location and travel choices using the entropy-constrained methods have been examined extensively by Boyce et al. (1983, 1988). Safwat and Magnanti (1988) made a further extension of the combined model by integrating all stages in sequential demand forecasting, based on the random utility theory of user behavior. Oppenheim (1993, 1995) made interesting extensions of the equilibrium trip distribution/assignment models by considering variable destination costs. On the other hand, Anderstig and Mattsson (1991, 1994) developed an integrated model of residential and employment location based on random utility theory with fixed travel costs and the formulation was solved using a heuristic approach. The model is applied as a policy analysis tool in the Stockholm region to investigate the welfare-maximizing location of housing supply. In contrast with most of the existing combined models that are concerned either with travel and origin (home) or travel and destination (workplace) choices, Yang and Meng (1998) examined a mixed and combined location and travel choice problem in which different types of travelers are modeled using a stochastic user equilibrium (SUE) behavior

model. For instance, some workers may have decided on their residence, but are still seeking a job, while others may have chosen their job, but are seeking a residence; still others may have fixed both their residence and their job locations. Workers of all these types may interact with one another by participating in the same activities and sharing the same transportation network. Meng et al. (2000) also studied a similar problem where destination choices were made with due regard to the travel times in both morning and evening peak periods. Excellent reviews on earlier work in this area were published by Fernandez and Friesz (1983), Boyce (1984) and Boyce et al. (1988), and for more recent works by Boyce and Daskin (1997).

For multi-class problems, Lam and Huang (1992) and Boyce and Bar-Gera (2001a) proposed a combined trip distribution and assignment model for multiple user classes, which was formulated as a convex optimization problem and solved by means of Evans's partial linearization or the Frank–Wolfe algorithm. A comprehensive validation exercise of applying this combined model for the city of Chicago was reported in Boyce and Bar-Gera (2001b). One of the first mathematical formulations of a network equilibrium model with sequential (hierarchical) rather than simultaneous destination and mode choices was proposed by Fernandez et al. (1994), who discussed several approaches to formulating combined modes. Combined trip distribution, mode choice, and assignment models with hierarchical choices were also presented by Abrahamsson and Lundqvist (1999), who determined that mode choice could be conditioned by distribution or vice versa.

More recently, Florian et al. (2002) developed a multi-class, multi-mode variable demand network equilibrium model with hierarchical logit structures. This was applied in the city of Santiago, Chile, to consider a simultaneous equilibrium formulation for distribution, modal split, and assignment, where there is an interaction between the road congestion and transit travel time. The mode choice was modeled by aggregate hierarchical logit structures (see Ortuzar and Willumsen 1994, p. 219) and the destination choice was specified as a multi-proportional entropy-type trip distribution model (or doubly-constrained aggregate gravity model). The problem was formulated using a variational inequality approach, with non-separable cost and demand functions, and was solved by an algorithm based on a Block Gauss–Seidel decomposition approach coupled with the method of successive averages. However, for a special case with separable cost functions, the problem can be formulated as an optimization problem.

In this chapter, we propose an optimization model for combined distribution, hierarchical mode choice, and assignment network model with multiple user and mode classes, in which the problem is formulated as an equivalent mathematical program. Evans's partial linearization algorithm is used to solve the problem. A case study of Hong Kong is conducted to demonstrate the efficiency of this modeling approach, which ensures the consistency of results for transportation planning.

2. PROBLEM DEFINITION AND NOTATION

We consider a congested road network with multiple user and mode classes. Different types of users perceive different values of time and money (for example, high-income work trips or low-income work trips and so on) in the network. The mode classes are categorized into groups so that a mode class is chosen by a user in a hierarchical manner.

The user first selects a group of mode classes as the primary (upper-level) mode choice, and then selects a particular mode class out of that group as the secondary (lower-level) mode choice. Different mode classes are subject to distinct combinations of service area restrictions and cost levels.

Consider a network with a set of nodes, V, and a set of links, A, defined by a graph (V, A). Let $I \subset V$ and $J \subset V$ be the sets of origin and destination zones, respectively. Furthermore, let N and M be the sets of user classes and mode classes in the network, respectively. The set of mode classes M is divided into a set of mutually exclusive groups, $G = (g_1, g_2, \ldots, g_{N_G})$, such that $\bigcup_{\forall i \in N_G} g_i = M$, and for any two groups $g_i, g_j \in G, g_i \cap g_j = \phi$, an empty set, where N_G is the number of groups. These groups can be further classified into road-based mode groups, $G_1 \subset G$, and transit-based mode groups, $G_2 \subset G$, so that $G_1 \cup G_2 = G$ and $G_1 \cap G_2 = \phi$.

We further assume that the zonal trip generations and attractions of each user class are known (they are obtained in the trip generation stage). The origin–destination (O–D) demand matrix for each user class is determined by a doubly constrained gravity-type distribution model. The users of the particular class n who are traveling from zone to zone can choose their mode of transport according to utility values such as the duration of the trip, the mileage cost of the trip, and their preference of transport mode, in different classes of mode alternatives. The following notation will be used throughout the chapter.

2.1 List of Variables

T_{ij}^{nm} = travel demand of user class n and mode class m from zone i to zone j;

$T_{ij}^{ng} = \sum_{m \in g} T_{ij}^{nm}$ = travel demand of user class n and mode group g from zone i to zone j;

$T_{ij}^{n} = \sum_{g \in G} T_{ij}^{ng}$ = total travel demand of user class n from zone i to zone j;

$O_i^n = \sum_{j \in J} T_{ij}^{n}$ = total travel demand (generation) of user class n from origin zone i;

$D_j^n = \sum_{i \in I} T_{ij}^{n}$ = total travel demand (attraction) of user class n to destination zone j;

$f_{ij,k}^{nm}$ = traffic flow of user class n and mode class m from zone i to zone j on path k;

v_a^{nm} = traffic flow of user class n and mode class m on link a;

$v_a = \sum_m \sum_n v_a^{nm}$ = total traffic flows on link a;

t_a = travel time on link a in the road network, which is a function of the total traffic flow on the link, and is assumed to be separable and an increasing function of total traffic flow v_a;

$\delta_{ij,ak}^{nm}$ = 1, if path $k \in R_{ij}^{nm}$ traverses link a, where $k \in R_{ij}^{nm}$ is the set of paths between the O–D pair (i, j) by user class n and mode class m; and
= 0 otherwise;

$c_a^{nm}, m \in g \in G_1$ = generalized cost of travel on link a as perceived by users of class n who choose mode class $m \in g \in G_1$ in the road network, which is expressed as a linear function of travel time and distance on the link,

$$c_a^{nm} = b_0^n t_a(v_a) + \sum_{l \in L} b_l^m d_a^l.$$

This special class of asymmetric cost function has been used in Van Vliet et al. (1986), which showed that an equivalent optimization model can be formulated for the network equilibrium problem;

b_0^n = value of time perceived by the users of class n;

b_l^m = value of one unit of fixed cost component l by mode class m;

d_a^l = fixed cost component l on link a;

$c_{ij,k}^{nm}, m \in g \in G_1$ = generalized cost of travel from zone i to zone j on path k as perceived by users of class n who choose mode class $m \in g \in G_1$,

$$c_{ij,k}^{nm} = \sum_{a \in A} \delta_{ij,ak}^{nm} c_a^{nm}.$$

$\hat{c}_{ij}^{nm}, m \in g \in G_2$ = generalized cost of travel from zone i to zone j as perceived by users of class n who choose mode class $m \in g \in G_2$ in the transit-based network (this generalized cost is assumed to be known, and is independent of the traffic conditions in the road-based network);

$\tilde{u}_{ij}^{nm}$ = minimum generalized cost of travel for the corresponding O–D pair (i, j) by user class n and mode class m (for transit-based modes, the generalized cost is a constant and can be evaluated explicitly),

$$\tilde{u}_{ij}^{nm} = \min\left(c_{ij,k}^{nm}, k \in R_{ij}^{nm}\right), \; m \in g \in G_1$$

$$\tilde{u}_{ij}^{nm} = \hat{c}_{ij}^{nm}, \; m \in g \in G_2;$$

C_{ij}^{nm} = total disutility of travel from zone i to zone j as perceived by users of class n who choose mode class m, $C_{ij}^{nm} = \tilde{u}_{ij}^{nm} + w_i^{nm}$, where w_i^{nm} is a user class, mode class and origin dependent parameter derived from a model calibration process (that is, bias coefficients);

β_1^n = dispersion coefficient (calibration coefficient) for the upper-level logit model as perceived by the users of class n;

β_2^{ng} = dispersion coefficient (calibration coefficient) for the lower-level logit model as perceived by the users of class n who choose mode group g;

θ^n = dispersion coefficient (calibration coefficient) for the gravity-type distribution model as perceived by the users of class n;

L_{ij}^{ng} = log sum of the disutility of travel as perceived by the users of class n who choose mode group g traveling from zone i to zone j,

$$L_{ij}^{ng} = -\frac{1}{\beta_2^{ng}} \ln \sum_{m \in g} \exp\left(-\beta_2^{ng} C_{ij}^{nm}\right);$$

L_{ij}^{n} = log sum of the disutility of travel as perceived by users of class n traveling from zone i to zone j,

$$L_{ij}^{n} = -\frac{1}{\beta_1^{n}} \ln \sum_{g \in G} \exp\left(-\beta_1^{n} L_{ij}^{ng}\right).$$

2.2 Gravity-type Distribution Model

For a particular user class n, given the total travel demands at the origin and destination zones, we have the following trip end constraints,

$$O_i^n = \sum_{j \in J} T_{ij}^n,\ i \in I \subset V,\ n \in N, \tag{2.1}$$

$$D_j^n = \sum_{i \in I} T_{ij}^n,\ j \in J \subset V,\ n \in N. \tag{2.2}$$

The destination choice of the users depends on the perceived minimum cost of the travel, which can be described by the following gravity equation,

$$T_{ij}^n = A_i^n B_j^n \exp\left(-\theta^n L_{ij}^n\right),\ i \in I,\ j \in J,\ n \in N, \tag{2.3}$$

where A_i^n and B_j^n are the balancing factors in the gravity model, and θ^n is a non-negative parameter that can be calibrated from observational data that describes the perception of the corresponding value of cost for user class n.

2.3 Hierarchical Modal Split Model

The travel choice of a user class from an origin zone to a destination zone for a particular mode class is made in a hierarchical manner as shown in Figure 2.1. The mode choice is conditioned by a prior choice of mode group at a higher level, based on the cost of travel and the preference for competing alternatives. We assume a hierarchical logit mode choice structure for the modal split, whereby the selection of one of the groups is made at the upper choice level, whereas the choice of a mode class within the selected group is made at the lower choice level. The logit-type mode choice function specifies the proportion of trips of various mode classes. The proportion of users of class n traveling from zone i to zone j who choose mode group g is defined as:

$$P_{ij}^{ng} = \frac{T_{ij}^{ng}}{T_{ij}^n} = \frac{\exp\left(-\beta_1^n L_{ij}^{ng}\right)}{\sum_{g' \in G} \exp\left(-\beta_1^n L_{ij}^{ng'}\right)} = \frac{\exp\left(-\beta_1^n L_{ij}^{ng}\right)}{\exp\left(-\beta_1^n L_{ij}^n\right)},\ i \in I,\ j \in J,\ n \in N,\ g \in G, \tag{2.4}$$

where

$$L_{ij}^{ng} = -\frac{1}{\beta_2^{ng}} \ln \sum_{m' \in g} \exp\left(-\beta_2^{ng} C_{ij}^{nm'}\right)$$

is the log sum of the travel cost for user class n and mode group g.

Once each user has selected a mode group g, the proportion of users choosing a particular mode class depends on the generalized cost associated with all of the mode classes in the group, which is determined by:

$$P_{ij}^{nm} = \frac{T_{ij}^{nm}}{T_{ij}^{ng}} = \frac{\exp\left(-\beta_2^{ng} C_{ij}^{nm}\right)}{\sum_{m' \in g} \exp\left(-\beta_2^{ng} C_{ij}^{nm'}\right)} = \frac{\exp\left(-\beta_2^{ng} C_{ij}^{nm}\right)}{\exp\left(-\beta_2^{ng} L_{ij}^{ng}\right)},\ i \in I,\ j \in J,\ n \in N,\ m \in g, \tag{2.5}$$

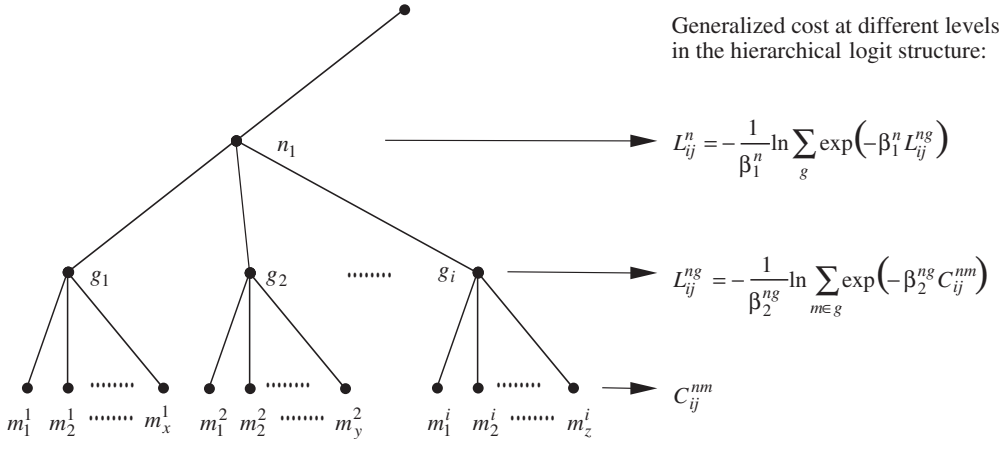


Figure 2.1 Hierarchical logit mode choice structure

where P_{ij}^{nm} is the proportion of users in class n with mode class m within group g traveling from zone i to zone j. Therefore, the total number of users of class n who choose mode class m can be obtained as:

$$T_{ij}^{nm} = T_{ij}^{n} \cdot P_{ij}^{ng} \cdot P_{ij}^{nm},\ i\in I, j\in J, n\in N, m\in g\in G. \tag{2.6}$$

3. MATHEMATICAL PROGRAM

The network equilibrium problem can be formulated as the following mathematical program:

$$\begin{aligned}\text{Minimize } F = &\sum_{a\in A}\int_0^{v_a} t_a(\omega)\mathrm{d}\omega + \sum_{a\in A}\sum_{n\in N}\sum_{m\in g\in G_1}\sum_{l\in L}\frac{b_l^m}{b_0^n}d_a^l v_a^{nm} + \sum_{n\in N}\frac{1}{b_0^n}\left(\sum_{m\in g\in G_2}\sum_{j\in J}\sum_{i\in I}T_{ij}^{nm}\hat{c}_{ij}^{nm}\right)\\
&+\sum_{n\in N}\frac{1}{b_0^n\theta^n}\left[\sum_{j\in J}\sum_{i\in I}T_{ij}^{n}(\ln T_{ij}^{n}-1)\right] - \sum_{n\in N}\frac{1}{b_0^n\beta_1^n}\left[\sum_{j\in J}\sum_{i\in I}T_{ij}^{n}(\ln T_{ij}^{n}-1)\right]\\
&+\sum_{n\in N}\frac{1}{b_0^n\beta_1^n}\left[\sum_{g\in G}\sum_{j\in J}\sum_{i\in I}T_{ij}^{ng}(\ln T_{ij}^{ng}-1)\right] - \sum_{n\in N}\frac{1}{b_0^n\beta_2^{ng}}\left[\sum_{g\in G}\sum_{j\in J}\sum_{i\in I}T_{ij}^{ng}(\ln T_{ij}^{ng}-1)\right]\\
&+\sum_{n\in N}\frac{1}{b_0^n\beta_2^{ng}}\left[\sum_{m\in M}\sum_{j\in J}\sum_{i\in I}T_{ij}^{nm}(\ln T_{ij}^{nm}-1)\right] + \sum_{n\in N}\frac{1}{b_0^n}\sum_{m\in M}\sum_{j\in J}\sum_{i\in I}T_{ij}^{nm}w_i^{nm} \qquad (2.7a)\end{aligned}$$

subject to:

$$\sum_{i\in I}T_{ij}^{n} = D_j^n,\ j\in J,\ n\in N, \tag{2.7b}$$

$$\sum_{j\in J} T_{ij}^{n} = O_i^n,\ i\in I,\ n\in N, \tag{2.7c}$$

$$\sum_{m\in g} T_{ij}^{nm} = T_{ij}^{ng},\ i\in I,\ j\in J,\ n\in N,\ g\in G, \tag{2.7d}$$

$$\sum_{g\in G} T_{ij}^{ng} = T_{ij}^{n},\ i\in I,\ j\in J,\ n\in N, \tag{2.7e}$$

$$\sum_{k\in R_{ij}^{nm}} f_{ij,k}^{nm} = T_{ij}^{nm},\ i\in I,\ j\in J,\ n\in N,\ m\in g\in G_1, \tag{2.7f}$$

$$v_a^{nm} = \sum_{i\in I}\sum_{j\in J}\sum_{k\in R_{ij}^{nm}} \delta_{ij,ak}^{nm} f_{ij,k}^{nm},\ a\in A,\ n\in N,\ m\in g\in G_1, \tag{2.7g}$$

$$v_a = \sum_{m\in g\in G_1}\sum_{n} v_a^{nm},\ a\in A, \tag{2.7h}$$

$$f_{ij,k}^{nm}\geq 0,\ i\in I,\ j\in J,\ k\in R_{ij}^{nm},\ n\in N,\ m\in g\in G_1, \tag{2.7i}$$

$$T_{ij}^{n}>0,\ T_{ij}^{ng}>0,\ T_{ij}^{nm}>0,\ i\in I,\ j\in J,\ n\in N,\ g\in G,\ m\in M, \tag{2.7j}$$

where $b_0^n \neq 0$ for all user classes, which is usually satisfied because the link cost that is perceived by the users is otherwise independent of travel time.

3.1 A Lagrangian Function for the Mathematical Program

We first form a Lagrangian function for the mathematical program as follows:

$$\begin{aligned}
\Pi = &\sum_{a\in A}\int_0^{v_a} t_a(\omega)\mathrm{d}\omega + \sum_{a\in A}\sum_{n\in N}\sum_{m\in g\in G_1}\sum_{l\in L}\frac{b_l^m}{b_0^n} d_a^l v_a^{nm} + \sum_{n\in N}\frac{1}{b_0^n}\left(\sum_{m\in g\in G_2}\sum_{j\in J}\sum_{i\in I} T_{ij}^{nm}\hat{c}_{ij}^{nm}\right)\\
&+\sum_{n\in N}\frac{1}{b_0^n\theta^n}\left[\sum_{j\in J}\sum_{i\in I} T_{ij}^{n}(\ln T_{ij}^{n}-1)\right] - \sum_{n\in N}\frac{1}{b_0^n\beta_1^n}\left[\sum_{j\in J}\sum_{i\in I} T_{ij}^{n}(\ln T_{ij}^{n}-1)\right]\\
&+\sum_{n\in N}\frac{1}{b_0^n\beta_1^n}\left[\sum_{g\in G}\sum_{j\in J}\sum_{i\in I} T_{ij}^{ng}(\ln T_{ij}^{ng}-1)\right] - \sum_{n\in N}\frac{1}{b_0^n\beta_2^{ng}}\left[\sum_{g\in G}\sum_{j\in J}\sum_{i\in I} T_{ij}^{ng}(\ln T_{ij}^{ng}-1)\right]\\
&+\sum_{n\in N}\frac{1}{b_0^n\beta_2^{ng}}\left[\sum_{m\in M}\sum_{j\in J}\sum_{i\in I} T_{ij}^{nm}(\ln T_{ij}^{nm}-1)\right] + \sum_{n\in N}\frac{1}{b_0^n}\sum_{m\in M}\sum_{j\in J}\sum_{i\in I} T_{ij}^{nm} w_i^{nm}\\
&+\sum_{n\in N}\sum_{i\in I}\alpha_i^n\left(\sum_{j\in J} T_{ij}^{n} - O_i^n\right) + \sum_{n\in N}\sum_{j\in J}\beta_j^n\left(\sum_{i\in I} T_{ij}^{n} - D_j^n\right)\\
&+\sum_{g\in G}\sum_{n\in N}\sum_{i\in I}\sum_{j\in J}\lambda_{ij}^{ng}\left(T_{ij}^{ng} - \sum_{m\in g} T_{ij}^{nm}\right) + \sum_{n\in N}\sum_{i\in I}\sum_{j\in J}\lambda_{ij}^{n}\left(T_{ij}^{n} - \sum_{g\in G} T_{ij}^{ng}\right)
\end{aligned}$$

$$+\sum_{n\in N}\sum_{m\in g\in G_1}\sum_{i\in I}\sum_{j\in J}u_{ij}^{nm}\left(T_{ij}^{nm}-\sum_{k\in R_{ij}^{nm}}f_{ij,k}^{nm}\right). \tag{2.8}$$

After differentiating, we have:

$$\frac{\partial\Pi}{\partial f_{ij,k}^{nm}}=\sum_{a\in A}\left[t_a(v_a)+\sum_{l\in L}\frac{b_l^m}{b_0^n}d_a^l\right]\delta_{ij,ak}^{nm}-u_{ij}^{nm},\ i\in I, j\in J, k\in R_{ij}^{nm}, n\in N, m\in g\in G_1, \tag{2.9}$$

$$\frac{\partial\Pi}{\partial T_{ij}^{n}}=\frac{1}{b_0^n\theta^n}\ln T_{ij}^{n}-\frac{1}{b_0^n\beta_1^n}\ln T_{ij}^{n}+\lambda_{ij}^{n}+\alpha_i^n+\beta_j^n,\ i\in I, j\in J, n\in N, \tag{2.10}$$

$$\frac{\partial\Pi}{\partial T_{ij}^{ng}}=\frac{1}{b_0^n\beta_1^n}\ln T_{ij}^{ng}-\frac{1}{b_0^n\beta_2^{ng}}\ln T_{ij}^{ng}+\lambda_{ij}^{ng}-\lambda_{ij}^{n},\ i\in I, j\in J, n\in N, g\in G, \tag{2.11}$$

$$\frac{\partial\Pi}{\partial T_{ij}^{nm}}=\frac{1}{b_0^n\beta_2^{ng}}\ln T_{ij}^{nm}+\frac{1}{b_0^n}w_i^{nm}+u_{ij}^{nm}-\lambda_{ij}^{ng},\ i\in I, j\in J, n\in N, m\in g\in G_1, \tag{2.12}$$

$$\frac{\partial\Pi}{\partial T_{ij}^{nm}}=\frac{1}{b_0^n\beta_2^{ng}}\ln T_{ij}^{nm}+\frac{1}{b_0^n}w_i^{nm}+\frac{1}{b_0^n}\hat{c}_{ij}^{nm}-\lambda_{ij}^{ng},\ i\in I, j\in J, n\in N, m\in g\in G_2. \tag{2.13}$$

3.2 User Equilibrium for Route Choice

For the road-based mode, applying the Kuhn–Tucker conditions to equation (2.9), we have

$$f_{ij,k}^{nm}\left(c_{ij,k}^{nm}-\tilde{u}_{ij}^{nm}\right)=0,\ i\in I, j\in J, k\in R_{ij}^{nm}, n\in N, m\in g\in G_1, \tag{2.14}$$

$$c_{ij,k}^{nm}-\tilde{u}_{ij}^{nm}\geq 0,\ i\in I, j\in J, k\in R_{ij}^{nm}, n\in N, m\in g\in G_1, \tag{2.15}$$

$$f_{ij,k}^{nm}\geq 0,\ i\in I, j\in J, k\in R_{ij}^{nm}, n\in N, m\in g\in G_1. \tag{2.16}$$

These conditions together define a user equilibrium flow pattern for the problem, where $\tilde{u}_{ij}^{nm}=b_0^n u_{ij}^{nm}$ can be interpreted as the minimum travel cost between zone i and zone j for user class n and mode class m.

For the transit-based mode, $m\in g\in G_2$, we assume that the route choice problem of transit modes is analysed by a separate generic transit model, from which the resulting generalized travel cost between each O–D pair for each user class and transit mode class, $\hat{c}_{ij}^{nm}$, is obtained. In this chapter, we focus on the road-based modes and assume that the transit vehicle speeds in the transit-based network are not affected by road-based traffic congestion. The interaction between road- and transit-based modes, can be dealt with by an iterative procedure in which the travel times that are obtained from the road-based sub-network are passed to the transit-based sub-network (if they overlap), and vice versa. However, this aspect of the situation is not studied here.

3.3 Hierarchical Mode Choice Structure

For road-based modes, from the interpretation of minimum travel cost $\tilde{u}_{ij}^{nm} = b_0^n u_{ij}^{nm}$, we can show that $w_i^{nm}/b_0^n + u_{ij}^{nm} = (w_i^{nm} + \tilde{u}_{ij}^{nm})/b_0^n = C_{ij}^{nm}/b_0^n$; and for transit-based modes, we can also show that $w_i^{nm}/b_0^n + \tilde{c}_{ij}^{nm}/b_0^n = C_{ij}^{nm}/b_0^n$. Substituting these expressions into equations (2.12) and (2.13) for road- and transit-based modes respectively, we have:

$$\frac{1}{b_0^n \beta_2^{ng}} \ln T_{ij}^{nm} + \frac{C_{ij}^{nm}}{b_0^n} - \lambda_{ij}^{ng} = 0,\ i \in I,\ j \in J,\ n \in N,\ m \in g \in G. \tag{2.17}$$

Therefore, since $T_{ij}^{nm} > 0$,

$$T_{ij}^{nm} = \exp(-\beta_2^{ng} C_{ij}^{nm}) \exp(b_0^n \beta_2^{ng} \lambda_{ij}^{ng}),\ i \in I,\ j \in J,\ n \in N,\ m \in g \in G. \tag{2.18}$$

From the definition of the conservation of flows (2.7d) and the summation of equation (2.18) over all $m \in g$, we can show that:

$$\frac{T_{ij}^{nm}}{T_{ij}^{ng}} = \frac{\exp(-\beta_2^{ng} C_{ij}^{nm})}{\sum_{m' \in G} \exp(-\beta_2^{ng} C_{ij}^{nm'})},\ i \in I,\ j \in J,\ n \in N,\ m \in g \in G, \tag{2.19}$$

which satisfies the lower-level logit choice proportion of trips by users of class n from origin zone i to destination zone j who choose mode class m within group g. Using equations (2.18) and (2.19), we have:

$$\exp(b_0^n \beta_2^{ng} \lambda_{ij}^{ng}) = \frac{T_{ij}^{ng}}{\sum_{m' \in G} \exp(-\beta_2^{ng} C_{ij}^{nm'})},\ i \in I,\ j \in J,\ n \in N,\ g \in G. \tag{2.20}$$

Now, we define the log sum of the generalized cost of the various mode classes in group g as

$$L_{ij}^{ng} = -\frac{1}{\beta_2^{ng}} \ln \sum_{m \in G} \exp(-\beta_2^{ng} C_{ij}^{nm}).$$

Then, equation (2.20) gives:

$$\lambda_{ij}^{ng} = \frac{1}{b_0^n} \left(L_{ij}^{ng} + \frac{1}{\beta_2^{ng}} \ln T_{ij}^{ng} \right),\ i \in I,\ j \in J,\ n \in N,\ g \in G. \tag{2.21}$$

From equations (2.11) and (2.21), we have:

$$\frac{1}{b_0^n \beta_1^n} \ln T_{ij}^{ng} - \frac{1}{b_0^n \beta_2^{ng}} \ln T_{ij}^{ng} + \frac{1}{b_0^n} \left(L_{ij}^{ng} + \frac{1}{\beta_2^{ng}} \ln T_{ij}^{ng} \right) - \lambda_{ij}^{n} = 0,\ i \in I,\ j \in J,\ n \in N,\ g \in G. \tag{2.22}$$

Therefore,

$$T_{ij}^{ng} = \exp(-\beta_1^n L_{ij}^{ng}) \exp(b_0^n \beta_1^n \lambda_{ij}^{n}),\ i \in I,\ j \in J,\ n \in N,\ g \in G. \tag{2.23}$$

Now, from the definition of the conservation of total flows over the network (2.7e), we have:

$$\frac{T_{ij}^{ng}}{T_{ij}^{n}}=\frac{\exp\left(-\beta_1^n L_{ij}^{ng}\right)}{\sum_{g\in G}\exp\left(-\beta_1^n L_{ij}^{ng}\right)},\ i\in I, j\in J, n\in N, g\in G, \tag{2.24}$$

which satisfies the upper-level logit choice proportion of group g modes for each user class n. If we define the log sum of the generalized cost of travel in the network as:

$$L_{ij}^{n}=-\frac{1}{\beta_1^n}\ln\sum_{g\in G}\exp\left(-\beta_1^n L_{ij}^{ng}\right),$$

combining equations (2.23) and (2.24) gives:

$$\lambda_{ij}^{n}=\frac{1}{b_0^n}\left(L_{ij}^{n}+\frac{1}{\beta_1^n}\ln T_{ij}^{n}\right),\ i\in I, j\in J, n\in N. \tag{2.25}$$

3.4 Trip Distribution

From equations (2.10) and (2.25), we have:

$$\frac{1}{b_0^n\theta^n}\ln T_{ij}^{n}-\frac{1}{b_0^n\beta_1^n}\ln T_{ij}^{n}+\frac{1}{b_0^n}\left(L_{ij}^{n}+\frac{1}{\beta_1^n}\ln T_{ij}^{n}\right)+\alpha_i^n+\beta_j^n=0,\ i\in I, j\in J, n\in N. \tag{2.26}$$

Therefore,

$$\frac{1}{b_0^n\theta^n}\ln T_{ij}^{n}+\frac{1}{b_0^n}L_{ij}^{n}+\alpha_i^n+\beta_j^n=0,\ i\in I, j\in J, n\in N. \tag{2.27}$$

After rearranging, we can show that:

$$T_{ij}^{n}=\exp\left(-\theta^n L_{ij}^{n}\right)\exp\left[-b_0^n\theta^n\left(\alpha_i^n+\beta_j^n\right)\right]=\exp\left[-\theta^n\left(L_{ij}^{n}+b_0^n\alpha_i^n+b_0^n\beta_j^n\right)\right],\ i\in I, j\in J, n\in N, \tag{2.28}$$

or alternatively,

$$T_{ij}^{n}=A_i^n B_j^n\exp\left(-\theta^n L_{ij}^{n}\right),\ i\in I, j\in J, n\in N, \tag{2.29}$$

where $A_i^n=\exp\left(-b_0^n\theta^n\alpha_i^n\right)$ and $B_j^n=\exp\left(-b_0^n\theta^n\beta_j^n\right)$, which are the balancing factors of the gravity-type distribution model of user class n.

4. SOLUTION ALGORITHM

The combined distribution, hierarchical mode choice, and assignment network model is a convex programming problem when we choose $\beta_2^{ng}\geq\beta_1^n$, $n\in N$, $g\in G$ in the hierarchical logit structure. Here we adopt the partial linearization approach for the solution algorithm.

Step 1: Initialization The iteration number $r=1$. Select an initial feasible solution $T_{ij}^{nm(r)}$, $T_{ij}^{ng(r)}$, $T_{ij}^{n(r)}$, $v_a^{nm(r)}$, $\forall n,m,g,i,j$. Generally, we can select the initial solution that corresponds to the free flow network.

Step 2: Computation of generalized costs Compute the minimum travel costs $u_{ij}^{nm(r)}$ between O–D pairs for both the private network and the transit network, and then the corresponding generalized costs of $C_{ij}^{nm(r)}$, $L_{ij}^{ng(r)}$, $L_{ij}^{n(r)}$. For the transit-based network with preset paths, the minimum travel costs are constant and need computing only in the first iteration.

Step 3: Computation of demand flows in the combined model Compute the demand flows $Z_{ij}^{nm(r)}$, $Z_{ij}^{ng(r)}$, $Z_{ij}^{n(r)}$ based on the generalized costs that were obtained in Step 2. The partial linearization approach is equivalent to the following steps:

(a) Compute $Z_{ij}^{n(r)}$ by solving the doubly constrained gravity sub-model using the Furness procedure or the row and column balancing method as:

$$Z_{ij}^{n(r)} = A_i^n B_j^n \exp\left(-\theta^n L_{ij}^n\right),$$

subject to:

$$\sum_{i\in I} Z_{ij}^{n(r)} = D_j^n,\ j\in J,\ n\in N,$$

$$\sum_{j\in J} Z_{ij}^{n(r)} = O_i^n,\ i\in I,\ n\in N,$$

$$Z_{ij}^{n(r)} \geq 0,\ i\in I, j\in J, n\in N.$$

(b) Compute $Z_{ij}^{ng(r)}$ and $Z_{ij}^{nm(r)}$ using the following equations:

$$\frac{Z_{ij}^{ng(r)}}{Z_{ij}^{n(r)}} = \frac{\exp\left(-\beta_1^n L_{ij}^{ng}\right)}{\sum_{g'\in G}\exp\left(-\beta_1^n L_{ij}^{ng'}\right)},\ i\in I, j\in J, n\in N, g\in G,$$

$$\frac{Z_{ij}^{nm(r)}}{Z_{ij}^{ng(r)}} = \frac{\exp\left(-\beta_2^{ng} C_{ij}^{nm}\right)}{\sum_{m'\in G}\exp\left(-\beta_2^{ng} C_{ij}^{nm'}\right)},\ i\in I, j\in J, n\in N, m\in g\in G.$$

Step 4: Assignment of demand flows on the network For the road-based mode classes, assign the demand matrix ($Z_{ij}^{nm(r)}$) in Step 3 to the minimum cost paths that were computed in Step 2 (when computing the minimum travel costs) to obtain the link flow pattern $y_a^{nm(r)}$.

Step 5: Check convergence The following gap function is adopted for the solution algorithm as a convergence criterion (Lam and Huang 1992):

$$\Delta = \Big|\sum_{a\in A}\sum_{n\in N}\sum_{m\in g\in G_1}\sum_{l\in L}\left(t_a + \frac{b_l^m}{b_0^n} d_a^l\right)\left[y_a^{nm(r)} - v_a^{nm(r)}\right]$$

$$+\sum_{n\in N}\sum_{m\in g\in G_2}\sum_{j\in J}\sum_{i\in I}\left(\frac{1}{b_0^n}\hat{c}_{ij}^{nm}\right)\left[Z_{ij}^{nm(r)} - T_{ij}^{nm(r)}\right]$$

$$+\sum_{n\in N}\sum_{j\in J}\sum_{i\in I}\left(\frac{1}{b_0^n\theta^n}\ln T_{ij}^n-\frac{1}{b_0^n\beta_1^n}\ln T_{ij}^n\right)\left[Z_{ij}^{n(r)}-T_{ij}^{n(r)}\right]$$

$$+\sum_{n\in N}\sum_{g\in G}\sum_{j\in J}\sum_{i\in I}\left(\frac{1}{b_0^n\beta_1^n}\ln T_{ij}^{ng}-\frac{1}{b_0^n\beta_2^{ng}}\ln T_{ij}^{ng}\right)\left[Z_{ij}^{ng(r)}-T_{ij}^{ng(r)}\right]$$

$$+\sum_{n\in N}\sum_{m\in M}\sum_{j\in J}\sum_{i\in I}\left(\frac{1}{b_0^n\beta_2^{ng}}\ln T_{ij}^{nm}+\frac{1}{b_0^n}w_i^{nm}\right)\left[Z_{ij}^{nm(r)}-T_{ij}^{nm(r)}\right]\Big|$$

where Δ is the absolute value of the gap between the current and the optimum value of the objective function; that is, $\nabla F\cdot D$, where ∇F is the gradient of objective function at current iteration and D is the corresponding descent direction. If $\Delta\leq\varepsilon$, an acceptable tolerance, then stop; otherwise go to Step 6.

Step 6: Line search With the descent direction vectors $Z_{ij}^{nm(r)}$, $Z_{ij}^{ng(r)}$, $Z_{ij}^{n(r)}$, and $y_a^{nm(r)}$ obtained in Steps 3 and 4, find the optimal step size $0\leq\omega\leq 1$, which minimizes the objective function (2.7a) by replacing the variables T_{ij}^{nm}, T_{ij}^{ng}, T_{ij}^{n}, v_a^{nm} with:

$$\bar{T}_{ij}^{nm(r)}=T_{ij}^{nm(r)}+\omega\left[Z_{ij}^{nm(r)}-T_{ij}^{nm(r)}\right],\ \forall n,m,i,j,$$

$$\bar{T}_{ij}^{ng(r)}=T_{ij}^{ng(r)}+\omega\left[Z_{ij}^{ng(r)}-T_{ij}^{ng(r)}\right],\ \forall n,g,i,j,$$

$$\bar{T}_{ij}^{n(r)}=T_{ij}^{n(r)}+\omega\left[Z_{ij}^{n(r)}-T_{ij}^{n(r)}\right],\ \forall n,i,j,$$

$$\bar{v}_a^{nm(r)}=v_a^{nm(r)}+\omega\left[y_a^{nm(r)}-v_a^{nm(r)}\right],\ \forall n,m,a.$$

Step 7: Update the solution Set:

$$T_{ij}^{nm(r+1)}=T_{ij}^{nm(r)}+\bar{\omega}\left[Z_{ij}^{nm(r)}-T_{ij}^{nm(r)}\right],\ \forall n,m,i,j,$$

$$T_{ij}^{ng(r+1)}=T_{ij}^{ng(r)}+\bar{\omega}\left[Z_{ij}^{ng(r)}-T_{ij}^{ng(r)}\right],\ \forall n,g,i,j,$$

$$T_{ij}^{n(r+1)}=T_{ij}^{n(r)}+\bar{\omega}\left[Z_{ij}^{n(r)}-T_{ij}^{n(r)}\right],\ \forall n,i,j,$$

$$v_a^{nm(r+1)}=v_a^{nm(r)}+\bar{\omega}\left[y_a^{nm(r)}-y_a^{nm(r)}\right],\ \forall n,m,a,$$

where $\bar{\omega}$ is the optimal step size obtained from Step 6. Set $r=r+1$ and go to Step 2.

5. A CASE STUDY OF HONG KONG

5.1 Hong Kong Comprehensive Transport Study

Since the First Hong Kong Comprehensive Transport Study (Transport Department 1976), the four-step modeling approach has continuously been adopted in all territorial transpor-

tation studies (Transport Department 1989). The Third Hong Kong Comprehensive Transport Study (CTS-3) was conducted in 1999 (Transport Department 1999), and also adopted the traditional four-step model to forecast travel demand according to trip generation, trip distribution, modal split, and trip assignment. In addition, the CTS-3 model incorporated a freight transport model (FTM), which was developed from a Freight Transport Study (Transport Department 1991). In the CTS-3, the distribution and modal-split processes were described in the distribution/modal-split model, in which the distribution process formed the trip matrix based on generation and attraction totals, whereas the main modal split divided trips into public and private modes. The proportion of trips that were chosen on public transport was calculated with a binary logit model formulation (based on the private and public transport times and costs that were produced by the cost model). Based on the above, the model produced both private and public trip matrices.

The model then estimated the split between private cars and taxis through a car/taxi model. This model was also a binary logit model which took the daily private matrices produced by the distribution/modal-split model as input to produce private car and taxi matrices, based on the difference in private car and taxi travel cost matrices.

The sub-modal split model was also adopted in the CTS-3, which estimated the proportions of public transport demand among the various modes (such as mass transit railway (MTR), Kowloon-Canton railway (KCR), light rail transit (LRT), tram, ferry, public light bus (PLB), franchised bus, and special purpose bus (SPB)) that made up the public transport system. The sub-modal split model took the form of a multinomial logit structure, and the total public transport matrix was then input into the sub-modal split model to produce a set of trip matrices, one for each public transport mode. With the modal split and demand estimation in the above steps, the demand matrices were assigned to the road and transit networks. The public transport assignment model assigned the matrices of public transport trips to the public transport paths to estimate the demand for each line and link in the public transport network. Non-road-based public transport trips (MTR, KCR, ferry and so on) were assigned to their own paths, while road-based public transport trips (bus, PLB and so on) were preloaded to the road assignment model. It was assumed that the assignment results would not affect the road users' choice in the modal split models.

The goods vehicle model was used to forecast commercial vehicular traffic, and was based on the FTM, which comprised the road network, cost, trip end, and distribution models. The goods vehicle category was divided into light van, light goods, medium goods, heavy goods, tractor unit, and service vehicles.

The road-based assignment model considered all road-based modes, including private cars, taxis, and goods vehicles. A multi-class user-equilibrium assignment was then performed and the paths through the network were built according to the generalized cost, which comprised travel distance, travel speed, and toll.

A simplified version of the flowchart for the CTS-3 modeling approach is shown in Figure 2.2.

5.2 A Numerical Example

A numerical example will demonstrate the application of the combined model to the Hong Kong strategic network. The problem is formulated as close as possible to that

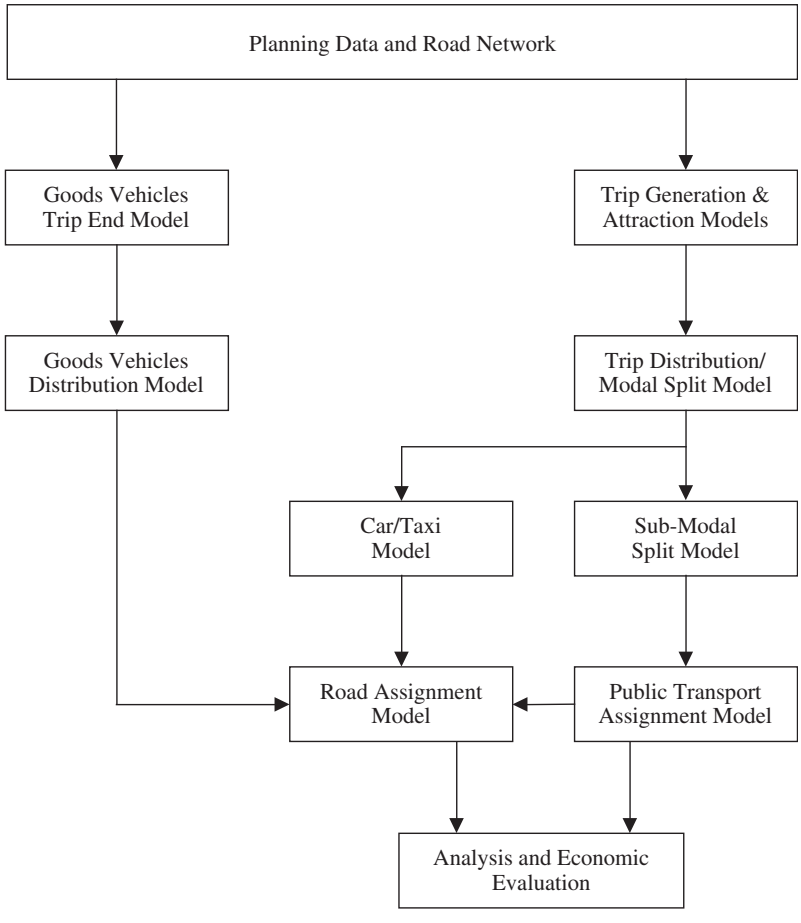


Figure 2.2 Flow chart of the CTS-3 transport demand model

adopted in the CTS-3 model. The strategic network consists of 20 zones, 77 nodes, and 216 links including 78 centroid connectors, and has road lengths ranging from 1 to 5 km. In the analysis, we assume that there are 18 user classes consisting of two-car availability users – car-owning and non-car-owning; combined with six trip purposes – home-based work high income (HBWH), home-based work low income (HBWL), home-based school (HBS), home-based other (HBO), non-home-based (NHB) and employer's business (EB); and six goods vehicle users – light van (LV), light goods (LG), medium goods (MG), heavy goods (HG), tractor unit (TU), and service (SR). There are nine mode classes, including private cars, taxis, public transport, and the six goods vehicle classes of LV, LG, MG, HG, TU, and SR. These mode classes are grouped into eight mode groups, which comprise private transport (private cars and taxis), public transport, and the six goods vehicle classes.

It is assumed that public transport users are insensitive to the road network condition (congestion), similar to the case of the CTS-3 model. Therefore, the generalized cost,

which can be obtained as the log sum of the respective costs of all public transport modes, is considered as fixed. For goods vehicles, six modes are considered, with each mode representing both the user class and the mode class of the particular type of goods vehicle. For example, LV represents both light van users as well as light van vehicles. However, travelers using a particular goods vehicle mode will not choose other goods vehicles.

The generations and attractions for each user class were estimated from previous studies, and the parameters for the distribution and modal split were chosen to be as close to the values used in the CTS-3 model as possible. For the assignment model in the network, the following travel impedance functions for link a is used,

$$t_a = t_a^0 \left[1 + 0.15 \left(\frac{v_a}{s_a} \right)^4 \right]$$

where t_a^0 is the free-flow travel time of link a, and v_a and s_a are the traffic flow and the practical capacity of link a, respectively.

With the input parameters and data, the mathematical program is solved with the proposed algorithm. The results from the model include the demand and generalized cost matrices of each user class, as well as the link loading conditions. As the data that are input into the model are largely hypothetical, we concentrate on computational issues in the following discussion. The objective is to show how this advanced combined modeling approach can be applied to the Hong Kong situation.

The aforementioned partial linearization algorithm was used to solve the multi-class problem. The level of congestion in the network affects the computing time because when the network is more congested it generally requires more iterations to converge. To investigate the influence of the congestion level in the network on the computational effort required for the solution algorithm, the demand level of each user class is factored by a common multiplier to simulate different levels of congestion.

Figure 2.3 shows the convergence characteristics of the model with different scaling factors on the demand flow in the network. The larger the scaling factor, the larger is the travel demand in the network, and hence it generates a higher level of congestion. The error for each iteration is measured by the gap between the current and the optimum value of the objective function. Here the solution is converged when the gap value is small than 0.05. For all cases, the solution converges satisfactorily to the equilibrium point. However, when the scaling factor increases, there are oscillations around the equilibrium solution, and hence the algorithm requires a large number of iterations to converge. This is because of the nature of the BPR (Bureau of Public Roads) type link impedance function, by which the travel time becomes more sensitive for higher degree of saturation. Hence, the solution algorithm shifts the flows back and forth around the equilibrium solution, which causes the oscillations in the convergence curves for scaling factors of 2.0 and 2.5.

Figure 2.4 shows the computing time required to solve the problem against the scaling factor on travel demand. The maximum degree of saturation, which is the maximum value of the volume/capacity (v/c) ratios among all links, is also shown in the figure. The computing time generally increases with the maximum degree of saturation in the network. Moreover, when the scaling factor increases, the maximum degree of saturation also increases, but the curve yields a factor above 2.0. This is because, as congestion increases, travelers tend to disperse in the network to move either to a closer destination or to a less congested path. The computing time increases linearly with the degree of saturation in the

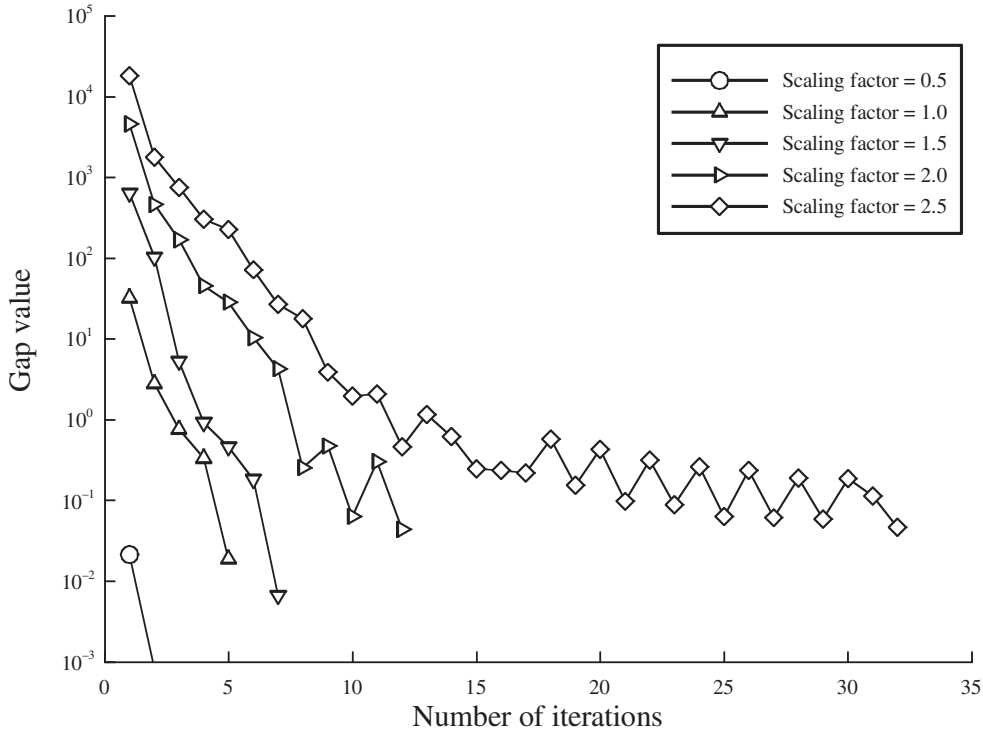


Figure 2.3 Convergence characteristics of different scaling factors on travel demand

network when the scaling factor is below 2.0, a range that the maximum *v*/*c* ratio is less than 2.5. However, when the scaling factor increases above 2.0, the computing time required increases substantially, because the algorithm needs more iterations to converge to the equilibrium solution as depicted in Figure 2.3.

6. CONCLUSIONS

This chapter presented a combined distribution, hierarchical mode choice, and assignment network model with multiple user and mode classes. The problem has been formulated as an equivalent mathematical program, which has been proved to satisfy all modeling requirements for the gravity-type distribution, hierarchical mode choice, and user equilibrium conditions of the multi-class problem. Evans's partial linearization algorithm has been proposed to solve the resultant problem. The strategic transportation network in Hong Kong was used as a case study to illustrate the potential applicability of the proposed methodology for solving complex transportation planning problems. The computing effort required for the proposed methodology was also studied.

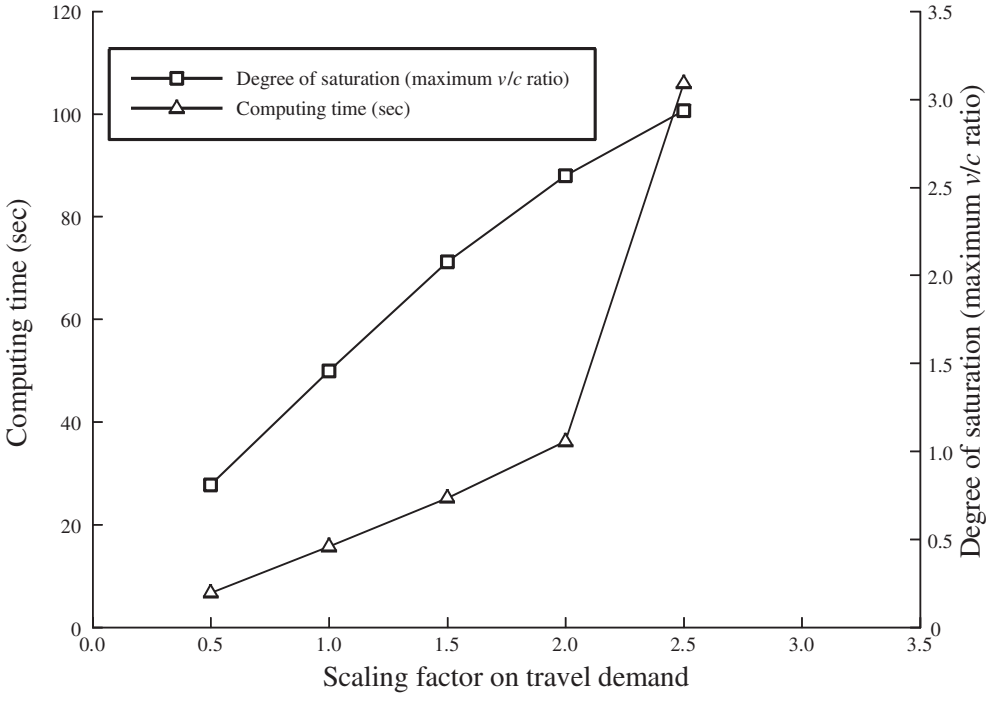


Figure 2.4 Computing time and degree of saturation against scaling factors on travel demand

NOTE

* We should like to gratefully acknowledge the support given to this project by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU7019/99E). This research was also supported by an Outstanding Young Researcher Award 2000 and a William Mong Young Researcher Award in Engineering 2002–2003 from the University of Hong Kong.

REFERENCES

Abrahamsson, T. and L. Lundqvist (1999), 'Formulation and estimation of combined network equilibrium models with applications to Stockholm', *Transportation Science*, **33**, 80–100.

Anderstig, C. and L.G. Mattsson (1991), 'An integrated model of residential and employment location in a metropolitan region', *Papers in Regional Science*, **70**, 167–84.

Anderstig, C. and L.G. Mattsson (1994), 'Modeling land-use and transport interaction: evaluations and policy analyses', Working Paper TRITA-IP 94-15, Department of Infrastructure and Planning, Royal Institute of Technology, Stockholm.

Boyce, D.E. (1984), 'Urban transportation network-equilibrium and design models: recent achievements and future prospects', *Environment and Planning*, **16A**, 1445–74.

Boyce, D.E. and H. Bar-Gera (2001a), 'Network equilibrium models of travel choices with multiple classes', in M.L. Lahr and R.E. Miller (eds), *Regional Science Perspectives in Economic Analysis*, Oxford: Elsevier Science, pp. 85–98.

Boyce, D.E. and H. Bar-Gera (2001b), 'Validation of urban travel forecasting models combining origin–destination, mode and route choices', Working Paper, Department of Civil and Materials Engineering, University of Illinois at Chicago.
Boyce, D.E., K.S. Chon, Y.J. Lee and K.T. Lin (1983), 'Implementation and computational issues for combined models of location, destination, mode and route choice', *Environment and Planning A*, **15**, 1219–30.
Boyce, D.E. and M.S. Daskin (1997), 'Urban transportation', in C. Revelle and A. McGarity (eds), *Design and Operation of Civil and Environmental Engineering Systems*, New York: Wiley, pp. 277–341.
Boyce, D.E., L.J. LeBlanc and K.S. Chon (1988), 'Network equilibrium models of urban location and travel choices: a retrospective survey', *Journal of Regional Science*, **28**, 159–83.
Evans, S.P. (1976), 'Derivation and analysis of some models for combining trip distribution and assignment', *Transportation Research*, **10**, 37–57.
Fernandez, J.E., J. De Cea, M. Florian and E. Cabrera (1994), 'Network equilibrium models with combined modes', *Transportation Science*, **28**, 182–92.
Fernandez, J.E. and T.L. Friesz (1983), 'Equilibrium predictions in transportation markets: the state of the art', *Transportation Research*, **17B**, 155–72.
Florian, M. (1977), 'A traffic equilibrium model of travel by car and public transit modes', *Transportation Science*, **11**, 166–79.
Florian, M. and S. Nguyen (1978), 'A combined trip distribution modal split and assignment model', *Transportation Research*, **12**, 241–6.
Florian, M., S. Nguyen and J. Ferland (1975), 'On the combined distribution assignment of traffic', *Transportation Science*, **9**, 43–53.
Florian, M., J.H. Wu and S.G. He (2002), 'A multi-class multi-mode variable demand network equilibrium model with hierarchical logit structures', in M. Gendreau and P. Marcotte (eds), *Transportation and Network Analysis: Current Trends: Miscellanea in Honor of Michael Florian*, London: Kluwer Academic, pp. 119–34.
Lam, W.H.K. and H.J. Huang (1992), 'A combined trip distribution and assignment model for multiple user classes', *Transportation Research*, **26B**, 275–87.
Meng, Q., H. Yang and S.C. Wong (2000), 'A combined land-use and transportation model for work trips', *Environment and Planning B*, **27**, 93–103.
Oppenheim, N. (1993), 'Equilibrium trip distribution/assignment with variable destination costs', *Transportation Research*, **27B**, 207–17.
Oppenheim, N. (1995), *Urban Travel Demand Modeling: From Individual Choices to General Equilibrium*, New York: John Wiley & Sons.
Ortuzar, J. de D. and L.G. Willumsen (1994), *Modelling Transport*, New York: John Wiley & Sons.
Safwat, N.K. and L.T. Magnanti (1988), 'A combined trip generation, trip distribution, modal split, and trip assignment model', *Transportation Science*, **18**, 14–30.
Transport Department (1976), *Hong Kong Comprehensive Transport Study – Final Report*, Hong Kong: Transport Department, Hong Kong Government.
Transport Department (1989), *Hong Kong Second Comprehensive Transport Study – Final Report*, Hong Kong: Transport Department, Hong Kong Government.
Transport Department (1991), *Freight Transport Study – Final Report*, Hong Kong: Transport Department, Hong Kong Government.
Transport Department (1999), *Hong Kong Third Comprehensive Transport Study – Final Report*, Hong Kong: Transport Department, Hong Kong Government.
Van Vliet, D., T. Bergman and W.H. Scheltes (1986), 'Equilibrium traffic assignment with multiple user classes', PTRC Summer Annual Meeting, Planning and Transport Research and Computation Co., 14–17 July, Brighton, England, pp. 111–21.
Yang, H. and Q. Meng (1998), 'An integrated network equilibrium model of urban location and travel choices', *Journal of Regional Science*, **38**, 575–98.

3. Combined travel forecasting models: formulations and algorithms

Hillel Bar-Gera and David Boyce*

1. INTRODUCTION

When planning improvements to transportation systems, various alternatives are considered. Careful evaluation requires forecasts of travel patterns for each alternative. Travel patterns are the result of many choices. Traditional modeling practice considers these choices as a sequential process with unique order: activity location choice (trip generation), joint choice of origin and destination (trip distribution), mode choice and finally route choice (assignment). Despite its intuitive appeal, justification for this order is not as trivial as it may seem.

Travelers usually do not think about modes and routes until they have chosen a destination. In many cases this is simply because they have a fairly good idea about the route of choice and its properties, and even more so about the mode of choice and its properties, for most origins and destinations under consideration, prior to choosing their activities. To a certain extent this is true even for choices of workplace or residential location, whether made simultaneously or sequentially in one order or another. In view of these observations it seems odd and perhaps even inappropriate to ask which choice comes first.

If all of the conditions that could affect travel choices are known in advance, the order of modeling the different choices should not matter. However, a basic assumption in most forecasting models is that travelers' choices are affected by the level of service of the transportation system. On the other hand, this level of service, and particularly travel times on the roadway network, depend upon the prevailing travel pattern and the associated congestion. The fact that the travel pattern depends on the level of service, which in turn depends on the travel pattern, is one of the main challenges of transportation modeling.

The need to consider congestion effects on route choice became apparent fairly early in the development of travel forecasting models. Early attempts included various computational procedures such as quantal loading, origin by origin loading and so on. In recent years user-equilibrium models have gradually replaced previous computational procedures. In these models behavioral assumptions are translated into mathematical conditions that need to be satisfied by the model solution. These well-defined conditions allow one to evaluate approximate solutions, and to examine the convergence of various algorithms.

In contrast to the development and penetration of user-equilibrium route choice models, travel forecasts are still based by and large on sequential procedures. Sequential procedures, even if they are based on user-equilibrium route choice model, still suffer from inconsistent consideration of travel times and congestion effects in the various steps of

the procedure. The inconsistent consideration of congestion is a well-known and often debated flaw of traditional sequential computational procedures. This flaw was a key issue in the San Francisco Bay Area lawsuit (Garrett and Wachs 1996). A common remedy for this flaw is to introduce a 'feedback' mechanism into the computational procedure, much like quantal loading in its different forms provides a 'feedback' mechanism in computational procedures for network assignment. An alternative approach is to state the behavioral assumptions, translate them into mathematical conditions, and seek solutions that satisfy these conditions. Such models are referred to as *combined* or *integrated* models.

The authors believe that whenever possible, models must be formulated mathematically. The first goal of this chapter is to demonstrate that for most models used in practice a mathematical formulation requires less effort than generally believed, and that there are important benefits to mathematical formulations which are not always appreciated.

Models that combine several travel choices together are far from new. The first mathematical formulation of user-equilibrium assignment by Beckmann et al. (1956) assumed in addition that the flow of travelers between every origin–destination (O–D) pair of is a function of the level of service for that O–D pair. Their convex optimization formulation was later extended to take substitution effects into account by the introduction of origin and/or destination constraints (Evans 1976). Evans was also the first to present an efficient convergent algorithm for solving this model. Other convex optimization models include the multi-mode model of Boyce et al. (1983); the multi-mode, multi-class model of Lam and Huang (1992); the multi-mode, multi-class model of Boyce and Bar-Gera (2001) and the origin-based algorithm for solving combined models formulated as convex optimization problems of Bar-Gera and Boyce (2003). Convex optimization formulations have the advantage of unique solutions and algorithms that are proven to converge. More general combined models were formulated as variational inequalities (VIs) by Dafermos (1982), Florian et al. (2002) and others. Algorithms for combined models are mostly link based, similar to Evans (1976), with the exception of the route-based algorithm of Lundgren and Patriksson (1998).

The remainder of the chapter is organized as follows. Section 2 presents the general fixed point formulation for combined models. Measures for solution accuracy are discussed in Section 3. Algorithms for combined models are presented in Section 4. Computational results are presented in Section 5. Conclusions and suggestions for future research are presented in Section 6.

2. FIXED POINT FORMULATIONS OF COMBINED MODELS

This section presents a mathematical formulation for the general combined model. Mathematical formulations are important tools for describing the goal of a computational process. Setting the goal is a crucial step that must come prior to any consideration of computational procedures, such as the popular feedback mechanism. Only with a clearly stated goal can anyone judge whether a certain procedure performs well or not. Fortunately, combined models can be formulated as fixed point problems in a way that is relatively intuitive and minimal in notation. Consider a *study area* which is divided into *zones*, during a certain time period of the day in a given year. Let Z denote the set of all zones. For every pair of origin $p \in Z$ and destination $q \in Z$ let d_{pq} denote the O–D flow

(persons/hour) from p to q. $\mathbf{d}$ is the array of O–D flows. Flows are averaged over the entire modeling period (for example, the morning peak) and over all work days during a specified year. The time period should not be too long, so that flows within it are fairly steady and reasonably represented by their average. Flows can be estimates for past years, or expected values for future years. In any case, it is important to note that as expected/average values, flows can be fractional and do not need to be integers.

The set of available routes from origin p to destination q is denoted by R_{pq}, and the set of these sets is $\mathbf{R} = \{R_{pq}\}_{p,q \in Z}$. The distribution of travelers from p to q among the routes in R_{pq} is described by a vector of non-negative route proportions (conditional probabilities) $\boldsymbol{\gamma}_{pq} = \{\boldsymbol{\gamma}_{pqr}\}_{r \in Rpq}$. $\boldsymbol{\gamma}$ is the array of route proportion vectors. Route proportions must add up to one for each O–D pair, hence the set of all feasible route proportion arrays is:

$$\Gamma(\mathbf{R}) = \left\{ \boldsymbol{\gamma} \in [0, 1]^{|\mathbf{R}|} : \sum_{r \in R_{pq}} \gamma_{pqr} = 1 \quad \forall p,q \in Z \right\}. \tag{3.1}$$

Given $\boldsymbol{\gamma}_{pq}$, the implied vector of route flows is $\mathbf{h}_{pq} = \{h_{pqr}\}_{r \in R_{pq}} = d_{pq} \cdot \boldsymbol{\gamma}_{pq}$. The array of route flow vectors is denoted by $\mathbf{h}$.

In working with these arrays of vectors it is convenient to consider two types of products. The dot product is interpreted as the sum of the product of the elements, similar to a vector dot product, that is $\mathbf{x} \cdot \mathbf{y} = \Sigma_{pqr} x_{pqr} \cdot y_{pqr}$. The cross-product is interpreted as a dot product of array elements, one by one. That is $\mathbf{z} = \mathbf{x} \times \mathbf{y}$ means $z_{pq} = x_{pq} \cdot y_{pq}$. As the algebraic product of matrices is not used in this chapter, there should not be any confusion with this notation. Using these conventions, the relationship between O–D flows, route proportions and route flows can be written in short form as $\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma}) = \mathbf{d} \times \boldsymbol{\gamma}$.

According to the user-equilibrium principle of Wardrop (1952), each traveler seeks to minimize the cost associated with his/her chosen route; therefore, at equilibrium the cost of every used route cannot be greater than the cost of any alternative route. The term cost is interpreted as a general measure of dis-utility, which incorporates travel time. Let $\mathbf{c} = \{\{c_{pqr}\}_{r \in R_{pq}}\}_{p,q \in Z}$ be the array of route cost vectors, which is a continuous function of the travel pattern, $\mathbf{c} = \mathbf{C}(\mathbf{h})$. The set of routes of minimum cost for a given O–D pair p, q is denoted by $R^*_{pq}(\mathbf{c}) = \text{argmin}\{c_{pqr} : r \in R_{pq}\}$. The array of such sets is denoted by $\mathbf{R}^*(\mathbf{c})$. For any non-empty subset of routes $\mathbf{R}'$; $\emptyset \subsetneq R'_{pq} \subseteq R_{pq}$, define the set of feasible route proportion arrays that are limited to $\mathbf{R}'$ as:

$$\Gamma(\mathbf{R}') = \{\boldsymbol{\gamma} \in \Gamma(\mathbf{R}) : \gamma_{pqr} = 0 \ \forall r \notin R'_{pq} \ \forall p,q \in Z\}. \tag{3.2}$$

In particular the set of minimum cost assignments is $\Gamma[\mathbf{R}^*(\mathbf{c})]$. It is obvious without any derivation that the travel pattern $\{\mathbf{d}, \boldsymbol{\gamma}\}$ satisfies the user-equilibrium requirements iff:

$$\boldsymbol{\gamma} \in F_1(\mathbf{d}, \boldsymbol{\gamma}) = \Gamma(\mathbf{R}^*\{\mathbf{C}[\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma})]\}). \tag{3.3}$$

In other words, user-equilibrium route proportions must belong to the set of feasible route proportions that are limited to the set of minimum cost routes, where route costs correspond to route flows that result from the chosen route proportions. We assume that the array of O–D flows is a continuous upper-bounded function of O–D costs, $\mathbf{d} = \Phi(\mathbf{u})$, where $\mathbf{u} = \{u_{pq}\}_{p,q \in Z}$ is the array of O–D costs. O–D costs equal average route costs,

weighted by flow, $\bar{U}_{pq}(\mathbf{c},\boldsymbol{\gamma}) = \boldsymbol{\gamma}_{pq} \cdot \mathbf{c}_{pq}$, or $\bar{\mathbf{U}}(\mathbf{c},\boldsymbol{\gamma}) = \boldsymbol{\gamma} \times \mathbf{c}$. The fixed point formulation of the combined model is:

$$\{\mathbf{d},\boldsymbol{\gamma}\} = \in F_2(\mathbf{d},\boldsymbol{\gamma}) = \left[\Phi\left(\bar{\mathbf{U}}\left[\mathbf{C}[\mathbf{h}(\mathbf{d},\boldsymbol{\gamma})],\boldsymbol{\gamma}\right]\right)\right] \times \Gamma\left(\mathbf{R}^*\left[\mathbf{C}[\mathbf{h}(\mathbf{d},\boldsymbol{\gamma})]\right]\right). \tag{3.4}$$

or equivalently:

$$\mathbf{d} = \Phi\left(\bar{\mathbf{U}}\left[\mathbf{C}[\mathbf{h}(\mathbf{d},\boldsymbol{\gamma})],\boldsymbol{\gamma}\right]\right) \tag{3.5}$$

$$\boldsymbol{\gamma} \in \Gamma\left(\mathbf{R}^*\left[\mathbf{C}[\mathbf{h}(\mathbf{d},\boldsymbol{\gamma})]\right]\right). \tag{3.6}$$

These equations state that at equilibrium O–D flows must correspond to prevailing O–D costs, and at the same time the user-equilibrium conditions must be satisfied. For user-equilibrium solutions $\bar{U}_{pq}(\mathbf{c},\boldsymbol{\gamma}) = U^*_{pq}(\mathbf{c}_{pq}) \equiv \min\{c_{pqr} : r \in R_{pq}\}$. Therefore, in the above formulation, we can replace $\bar{\mathbf{U}}$ with $\mathbf{U}^*$ and obtain an equivalent formulation.

This formulation can be easily extended to multi-mode and multi-class models, by adding a mode subscript m and a class superscript l to all variables. In other words if we let $\mathbf{d} = \{d^l_{mpq}\}$; $\mathbf{R} = \{R^l_{mpq}\}$; $\boldsymbol{\gamma} = \{\gamma^l_{mpqr}\}$; $\mathbf{h} = \{h^l_{mpqr}\}$; $\mathbf{c} = \{c^l_{mpqr}\}$, and adapt the interpretation of $\mathbf{R}^*$, Γ, Φ, $\mathbf{C}$, and $\bar{\mathbf{U}}$ accordingly, then equation (3.4) is a mathematical formulation of a generic multi-mode, multi-class combined model. Solution existence is demonstrated by Kakutani's extension to Brouwer's fixed point theorem (Kakutani 1941; Nikaido 1968, Theorem 4.4, p. 67). Nikaido defines a set-valued mapping $f: X \to 2^Y$, where 2^Y represents the set of subsets of Y, to be closed if $x^k \to x$; $y^k \to y$; $y^k \in f(x^k)$ implies $y \in f(x)$. Every continuous function is closed; hence $\boldsymbol{\Phi}$, $\bar{\mathbf{U}}$, $\mathbf{C}$ are closed. $\mathbf{R}^*$ is closed, with discrete topology on $\mathbf{R}$, and Γ is also closed. Therefore, F_2 is closed. $\Gamma(\mathbf{R}')$ is convex for every set of routes $\mathbf{R}'$, hence $F_2(\mathbf{d}, \boldsymbol{\gamma})$ is convex for every $(\mathbf{d}, \boldsymbol{\gamma})$. Due to the upper bound on O–D flows, M, the set of feasible solutions, $[0, M]^{|Z| \times |Z|} \times [0, 1]^{|\mathbf{R}|}$ is non-empty, compact and convex. Under these conditions, Kakutani's extension to Brouwer's fixed point theorem guarantees that the map F_2 has a fixed point. In other words, there is at least one solution for the combined model in equation (3.4).

3. ACCURACY MEASURES

One of the main advantages of mathematically formulated models is the ability to evaluate solutions by well-defined accuracy measures, and hence to determine whether a solution is sufficiently accurate for the specific analysis under consideration. In the case of O–D flows it is natural to compare O–D flows in the current solution $\mathbf{d}$, with those that result from the costs of travel under current conditions. For the latter we can choose either minimum O–D costs $\mathbf{d}' = \Phi\{\mathbf{U}^*[\mathbf{C}(\mathbf{h})]\}$ or average O–D costs $\mathbf{d}'' = \Phi\{\bar{\mathbf{U}}[\mathbf{C}(\mathbf{h}), \gamma]\}$. Both comparisons lead with similar results. We shall use $\mathbf{d}'$ simply because average O–D costs are not available for some algorithms. Possible aggregate measures of accuracy are the maximum positive difference, $\max\{d'_{pq} - d_{pq} : d'_{pq} \geq d_{pq}\}$, the maximum negative difference, $\max\{d_{pq} - d'_{pq} : d'_{pq} \leq d_{pq}\}$, and the total misplaced O–D flow, $\Sigma_{p,q \in Z} |d'_{pq} - d_{pq}|$. All are in units of person trips per hour.

The intuitive interpretation of these measures can be very helpful in setting conditions for sufficiently accurate solutions. For example, consider a study that examines the impact

of a new commercial facility, which is expected to attract 1000 trips per hour during the morning peak. It would be reassuring to know that the total misplaced O–D flow in the solution is less than 100 trips/hour. A total misplaced O–D flow of 1000 trips/hour may still be acceptable, assuming that it is spread over a wide region. But, a total misplaced O–D flow of 10000 trips/hour is probably not acceptable, as it is quite likely to have significant influence on the results of the study. Assignment accuracy measures can be based on the distribution of *excess cost*, $ec_r = c_{pqr} - U^*_{pq}(\mathbf{c}_{pq})$, among used routes. Possible aggregate measures based on access cost include the maximum excess cost, the 95th percentile, the portion of flow with excess cost above a certain value, say one minute and so on. The main aggregate measure used in this chapter is the *average excess cost*, $AEC_a = (1/d_{..}) \cdot \Sigma_{r \in R} h_r \cdot ec_r$, where $d_{..} = \Sigma_{p,q \in Z} d_{pq}$ is the total O–D flow (on the road network). In the context of fixed demand problems AEC is sometimes referred to as the 'normalized gap'.

Setting requirements for assignment accuracy is more challenging, since typically the goal is to make sure that link flows are sufficiently close to the true equilibrium solution. A case study (Boyce et al. 2003) examined the impact of adding a pair of freeway ramps in the Delaware Valley Region. The goal of the study was to estimate flow differences on links in the vicinity of the proposed improvement between the build and no-build scenarios. It was found that solutions should have AEC less than 0.001 vehicle-minutes, so that estimates for freeway links will be within 3 per cent from the true equilibrium solution, and estimates for arterials will be within 10 per cent from the true equilibrium solution. As additional case studies are conducted on different networks and for various levels of congestion, more definite recommendations will be available for practitioners.

A solution is considered to be sufficiently accurate only if it satisfies both conditions, that is if it has AEC less than, say, 0.001 vehicle-minutes, and total misplaced O–D flow less than, say, 1000 trips/hour.

4. ALGORITHMS

In this chapter we consider two algorithms; both are iterative. In the first algorithm, demand and route proportions are updated simultaneously in every iteration. It is similar to the algorithm proposed by Evans (1976) for a combined model formulated as a convex optimization problem. In every iteration of this algorithm, given the current solution $(\mathbf{d}^k, \boldsymbol{\gamma}^k)$; $\mathbf{h}^k = \mathbf{d}^k \times \boldsymbol{\gamma}^k$, a subproblem solution is found in the following way. First a minimum cost assignment is chosen, given the current costs, $\hat{\boldsymbol{\gamma}}^k \in \Gamma\{\mathbf{R}^*[\mathbf{C}(\mathbf{h}^k)]\}$; then, new O–D flows are found using the minimum costs found in the previous step, $\hat{\mathbf{d}}^k = \Phi\{\mathbf{U}^*[\mathbf{C}(\mathbf{h}^k)]\}$; finally, the new demand is assigned to the minimum cost routes found in the first step $\hat{\mathbf{h}}^k = \hat{\mathbf{d}}^k \times \hat{\boldsymbol{\gamma}}^k$.

Once a subproblem solution is found, a new solution is obtained by a weighted average of the current solution and the subproblem solution, $\mathbf{h}^{k+1} = (1-\lambda) \cdot \mathbf{h}^k + \lambda \cdot \hat{\mathbf{h}}^k$, where $0 \le \lambda \le 1$ is the *step size*, or the weight of the subproblem solution. Since total link flows are a linear function of route flows, averaging route flows and averaging total link flows lead to the same solution. Therefore, implementations of this algorithm typically store only total link flows, thus reducing memory requirements substantially. In the algorithm proposed by Evans (1976), the convex formulation was used to determine the step size. In

the general case, when a convex formulation is not available, different techniques must be used to determine the step size, as discussed below.

The second algorithm is similar to the origin-based algorithm proposed in Bar-Gera and Boyce (2003) for combined models with convex formulations, which is based on the origin-based assignment algorithm (Bar-Gera 1999, 2002). The general scheme of the algorithm is presented in Figure 3.1. Stopping conditions for the algorithm are based on total misplaced O–D flow and average excess cost, as discussed in Section 3. The key element in the proposed algorithm for combined models is the procedure for updating O–D flows, while retaining the route proportions of the current solution. Given a current

Initialization:
Let $\mathbf{u} = \mathbf{U}^*[C(0)]$
Let $\mathbf{d}^0 = \Phi(\mathbf{u})$
for p in Z do
  A_p = tree of minimum cost routes from p
  $\mathbf{f}_p$ = all or nothing assignment using A_p
end for

Main loop:
  for n = 1 to number of main iterations
    Update O–D flows, retain route proportions
    for p in Z do
      update restricting subnetwork A_p
      update origin-based approach proportions α_p
    end for
    for m = 1 to number of inner iterations
      for p in Z do
        update origin-based approach proportions α_p
      end for
    end for
  end for

Figure 3.1 An origin-based algorithm for combined models

solution, $\{\mathbf{d}^k, \boldsymbol{\gamma}^k\}$, subproblem O–D flows are determined according to average O–D costs $\hat{\mathbf{d}}^k = \Phi\,[\boldsymbol{\gamma}^k \times \mathbf{C}\,(\mathbf{d}^k \times \boldsymbol{\gamma}^k)]$. New O–D flows are obtained by a weighted average $\mathbf{d}_\lambda^{k+1} = (1-\lambda)\cdot\mathbf{d}^k + \lambda\cdot\hat{\mathbf{d}}^k$, where $0 \le \lambda \le 1$ is a chosen step size. In models that have a convex formulation, it can be used to determine the step size. A proof of convergence for the resulting algorithm is given in Bar-Gera and Boyce (2003). As shown there, the use of average O–D costs (rather than minimum O–D costs) to determine subproblem O–D flows is critical for convergence. As with the Evans-like algorithm, when a convex formulation is not available, different techniques must be used to determine the step size.

Once O–D flows are updated, route proportions are revised by an origin-based assignment iteration, while keeping O–D flows temporarily fixed. The main solution variables in the origin-based assignment algorithm are *origin-based approach proportions* $\alpha = \{\alpha_{ap}\}_{a\in A; p\in Z}$; $0 \le \alpha_{ap} \le 1$; $\Sigma_{a:a_h=j}\,\alpha_{ap} = 1\ \forall j \in N\ \forall p \in Z$, where N and A are the sets of nodes

and links on the road network, and a_t, a_h are the tail and the head of link $a=[a_t, a_h]$. For every origin an a-cyclic restricting subnetwork is chosen, $A_p \subseteq A$; $a \notin A_p \Rightarrow \alpha_{ap}=0$. Initial restricting subnetworks are trees of minimum cost routes. To update the restricting subnetwork, unused links are removed, v_i – the maximum cost to node i within the restricting subnetwork is computed, and all links $[i, j]$ such that $v_i < v_j$ are added to the restricting subnetwork. Approach proportions for origin p are updated by shifting flows within the restricting subnetwork A_p according to a boundary (piecewise linear) search in a direction determined by an approximate second-order method. Given the demand $\mathbf{d}$ and the current solution $\boldsymbol{\alpha}$, the set of restricting subnetworks for the next iteration is defined by a function $\mathbf{A}=\mathcal{A}(\mathbf{d}, \boldsymbol{\alpha})$, and the next iteration solution is defined by a map $\boldsymbol{\alpha}' \in \Theta^{\boldsymbol{\alpha}}(\mathbf{d}, \boldsymbol{\alpha})$.

Route proportions are determined by $\gamma_{pqr}=\Pi_{a\subseteq r}\alpha_{ap}$. It can be shown that origin-based link flows $f_{ap}(\mathbf{h})=\Sigma_{q\in Z}\ \Sigma_{r\in R_{pq};a\subseteq r}\ h_{pqr}$ and origin-based node flows $f_{jp}(\mathbf{h})=\Sigma_{q\in Z}\ \Sigma_{r\in R_{pq};j\in r}\ h_{pqr}=\Sigma_{a:a_h=j} f_{ap}$ maintain the relationship $f_{ap}=\alpha_{ap}\cdot g_{jp}$, demonstrating that α_{ap} is indeed the proportion of flow on approach a to node a_h for origin p. The availability of route proportions allows one to compute average O–D costs, as well as the assignment of new O–D flows by current route proportions. Due to the restriction to a-cyclic subnetworks, these computations can be done efficiently without route enumeration, in a time that is a linear function of the number of links times the number of origins. These properties are essential for the demand update procedure described above.

In both the Evans-like algorithm and the origin-based algorithm the main obstacle towards a general implementation for non-convex models is the determination of the step size. The best-known technique to determine step sizes in general problems is the method of successive averages (MSA), introduced in the seminal work of Robbins and Monro (1951). In this technique, step sizes are predetermined as $\lambda_k=1/k$. (Where k is the iteration index.) Polyak (1990) argues that in the context of stochastic approximation techniques, under certain conditions, it is better to use either a constant step size, or step sizes of $\lambda_k=k^{-\beta}$ where $0<\beta<1$.

Basic intuition for the behavior of different algorithms with various choices of step sizes can be developed by considering a very simple example of a single dimensional problem with a given feasible range $[0, M]$, and an unknown optimal solution x^*. Consider an averaging algorithm, $x^{k+1}=(1=\lambda_k)\cdot x^k+\lambda_k\cdot y^k$ that is based on subproblem solution $y^k=f_a(x^k)$, where:

$$f_a(x)=\begin{cases} M & x\le b_1 \\ x^*-a(x-x^*) & b_1\le x\le b_2 \\ 0 & b_2\le x \end{cases} \tag{3.7}$$

$$b_1=x^*-\frac{M-x^*}{a};\ b_2=x^*+\frac{x^*}{a}. \tag{3.8}$$

The value of a controls the accuracy of the subproblem. $a=0$ indicates perfectly accurate subproblem solutions, since $f_0(x)=x^*\ \forall x$. On the other hand, as $a\to\infty$ subproblem solutions are less accurate, and at the limit a semi-continuous point to set function is obtained:

$$f_\infty(x)=\begin{cases} M & x<x^* \\ [0,M] & x=x^* \\ 0 & x>x^*. \end{cases} \tag{3.9}$$

When subproblem solutions are accurate, that is, $a \to 0$, larger step sizes lead to faster convergence. MSA does not take advantage of accurate subproblem solutions, however, since even if $a = 0$, the convergence of MSA is given by $x^k - x^* = (x^0 - x^*)/k$, which is quite slow. The progress under constant step size, λ, while in the linear range $[b_1, b_2]$, is given by:

$$x^{k+1} = (1-\lambda) \cdot x^k + \lambda \cdot [x^* - a \cdot (x^k - x^*)] = x^* + (1 - \lambda - \lambda \cdot a) \cdot (x^k - x^*) \quad (3.10)$$

$$x^{k+1} - x^* = (x^k - x^*) \cdot [1 - \lambda \cdot (1+a)] = (x^0 - x^*) \cdot [1 - \lambda \cdot (1+a)]^k \quad (3.11)$$

The sequence x^k converges to x^* for every step size $0 < \lambda < 2/(1+a)$. Oscillations are avoided by any step size $0 < \lambda < 1/(1+a)$. Further reducing step sizes only slows convergence. In the case of f_∞, any constant step size leads to oscillations around x^*. Smaller step size leads to oscillations closer to x^*, but also to slower progress towards x^*.

In monitoring the progress of the algorithm, typically, the value of x^* is not known, hence $x^k - x^*$ cannot be evaluated. Instead, we can monitor the value of $gap^k = x^k - y^k$. This is similar to the accuracy measure for O–D flows proposed in Section 3. For an algorithm that is based on f_a, when the slope a is finite, the relative reduction in the gap within the linear range $[b_1, b_2]$ is given by:

$$\frac{gap^k - gap^{k+1}}{gap^k} = 1 - \frac{x^{k+1} - y^{k+1}}{x^k - y^k} = \lambda \cdot (1 + a). \quad (3.12)$$

With perfectly accurate subproblem solutions ($a = 0$) the relative gap reduction is equal to the step size λ. Note that in the first iterations of an algorithm based on f_∞, that is until oscillations start, the relative gap reduction is also equal to the step size. It is therefore interesting to monitor the relative gap reduction in computational experiments.

It is clear from the above discussion that the optimal step-size strategy depends on the algorithm. If subproblem solutions are continuous as a function of the current solution, as is the case with f_a when a is finite as well as for the origin-based algorithm described above, a constant step size, if sufficiently small, should not cause oscillations. Our goal in that case is to use the largest constant step size that does not lead to oscillations. On the other hand, if subproblem solutions are not continuous, as is the case with f_∞ and with the Evans-like algorithm, any constant step size will eventually lead to oscillations; therefore, we must use a decreasing sequence of step sizes.

5. EXPERIMENTAL RESULTS

This section presents computation results comparing the convergence of the proposed origin-based algorithm and the Evans-like algorithm for different step-size strategies. All experiments were conducted on the same Compaq Alpha Unix Server model DS20E, with CPU speed of 666 MHz, and 256MB RAM.

The algorithms were applied to a multi-modal model, which is similar to the model presented in Boyce and Daskin (1997). The network of the model, referred to as the 'Chicago Sketch Network', consists of 387 zones, 933 nodes, 2950 road links, and total O–D flow of about 1.25 million person trips per hour. The main inputs to this model are: the flow of person trips per hour from each origin $\bar{d}_{p\bullet}$; the flow of person trips per hour to each

destination $\bar{d}_{\bullet q}$; free flow travel times tt_a^0, capacities k_a, and lengths l_a for each link on the roadway network; parking costs pc_z and walking times to or from the parking place wt_z for each zone; in vehicle travel times c_{tpq}^{ivtt}, fares c_{tpq}^{fare}, and out of vehicle times c_{tpq}^{ovt}, when traveling by transit from origin p to destination q (these are fixed regardless of flows); and truck flows d_{pq}^{truck} by O–D in passenger cars equivalents.

Link travel time functions are of the BPR form:

$$tt_a(f_a) = tt_a^0 \cdot [1 + 0.15 \cdot (f_a/k_a)^4]. \tag{3.13}$$

Auto operating costs, including gasoline consumption, are a linear function of link length and travel time,

$$oc_a = \eta_1 \cdot tt_a(f_a) + \eta_2 \cdot l_a. \tag{3.14}$$

Link generalized costs are:

$$t_a(f_a) = \beta_{time}^{au} \cdot tt_a(f_a) + \beta_{cost}^{au} \cdot oc_a(f_a), \tag{3.15}$$

where the βs are calibration parameters. Parking costs and walking times are components of the additional auto costs, defined as:

$$ac_{apq} = \beta_{park}^{au} \cdot \frac{pc_p + pc_q}{2} + \beta_{walk}^{au} \cdot (wt_p + wt_q). \tag{3.16}$$

Route generalized costs by auto are $c_{apqr} = ac_{apq} + \Sigma_{a \in r} t_a$. O–D generalized costs by auto are $\bar{U}_{apq}(\mathbf{c}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}_{apq} \cdot \mathbf{c}_{apq}$, as before. O–D generalized costs by transit are:

$$u_{tpq} = \beta_{bias}^{tr} + \beta_{ivtt}^{tr} \cdot c_{tpq}^{ivtt} + \beta_{fare}^{tr} \cdot c_{tpq}^{fare} + \beta_{ovt}^{tr} \cdot c_{tpq}^{ovt}. \tag{3.17}$$

O–D flows are of the compound logit form:

$$d_{apq} = A_p \cdot B_q \cdot \exp(-\mu \cdot u_{apq}) \cdot u_{apq}^{-\rho} \tag{3.18}$$

$$d_{tpq} = A_p \cdot B_q \cdot \exp(-\mu \cdot u_{tpq}) \cdot u_{tpq}^{-\rho} \tag{3.19}$$

Flows of person trips by auto are converted to vehicle flows by a predetermined auto occupancy factor, *aof*. The same route proportions are used for auto flows and for truck flows, hence:

$$\mathbf{h}_{pq}(\mathbf{d}, \boldsymbol{\gamma}) = (d_{apq}/aof + d_{pq}^{truck}) \cdot \boldsymbol{\gamma}_{pq}. \tag{3.20}$$

The fixed point formulation in equation (3.4) applies to this model almost directly, except that the definition of $\mathbf{h}_{pq}(\mathbf{d}, \boldsymbol{\gamma})$ mentioned above is slightly different from before. The only difference between this model and the model considered in Bar-Gera and Boyce (2003) is the power term $u^{-\rho}$ in (3.18) and (3.19). For $\rho = 0$, we get the original model, that can be formulated as a convex optimization problem. In that case, the performance of the algorithms based on step sizes determined using the convex objective function can be used as a reference for the performance of any other step-size strategy.

Practitioners have reported that in some cases observed data is better explained by models with $\rho \approx 1$. Figures 3.2–7 show results of the two algorithms for $\rho = 0, 1, 2$. In the case of $\rho = 0$, the results of the algorithm described in Bar-Gera and Boyce (2003), which is based on a line search over the convex objective function, are included as a reference.

Equilibrium solutions are quite different for different values of ρ. For example, total intra-zonal flow values are about 105 900, 88 708 and 63 203 person trips per hour for $\rho = 0$, 1, 2, respectively. On the other hand, the behavior of the different algorithms is almost identical in all three cases.

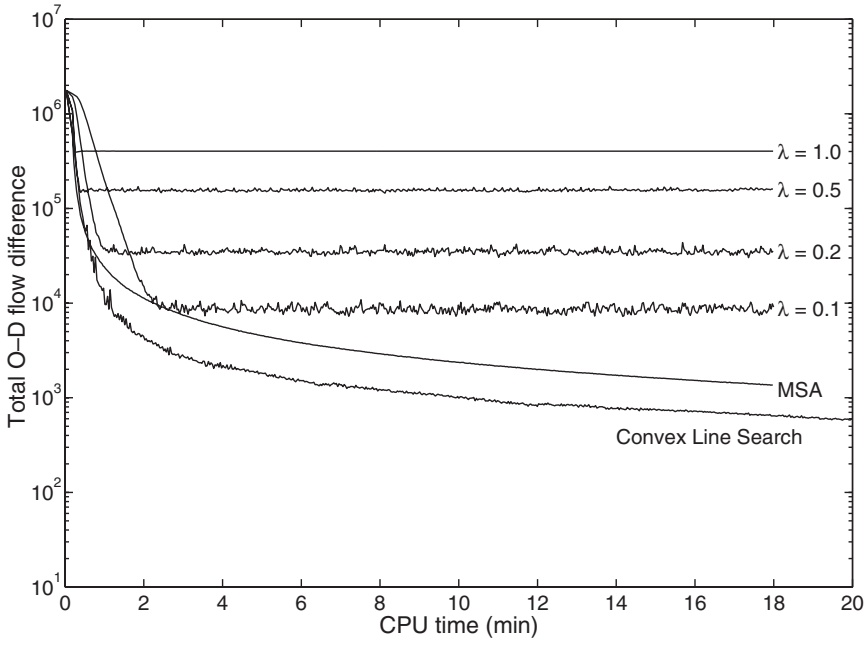


Figure 3.2 Convergence of Evans-like algorithms, $\rho = 0$

In the case of Evans-like algorithms, we find that any constant step size at some point leads to oscillations, as expected. In this case MSA performs reasonably well, although it is possible that a more sophisticated strategy will provide better results. On the other hand, in the case of origin-based algorithms, any step size less than 0.5 leads to convergence much faster than with MSA step sizes.

Comparing the convergence of the two algorithms with MSA step sizes with respect to CPU time, shows substantial advantage for the Evans-like algorithm. This is mainly because each origin-based iteration takes more CPU time. For example, 50 origin-based iterations take about 15 minutes of CPU time, and lead to total O–D flow difference of about 30 000 person trips/hour (with MSA step sizes); 50 Evans-like iterations take about 1.3 minutes of CPU time, and lead to total O–D flow difference of about 15 000 person trips/hour (with MSA step sizes); in 15 minutes of CPU time about 600 Evans-like iterations are performed, leading to total O–D flow difference of about 1500 person trips/hour (with MSA step sizes).

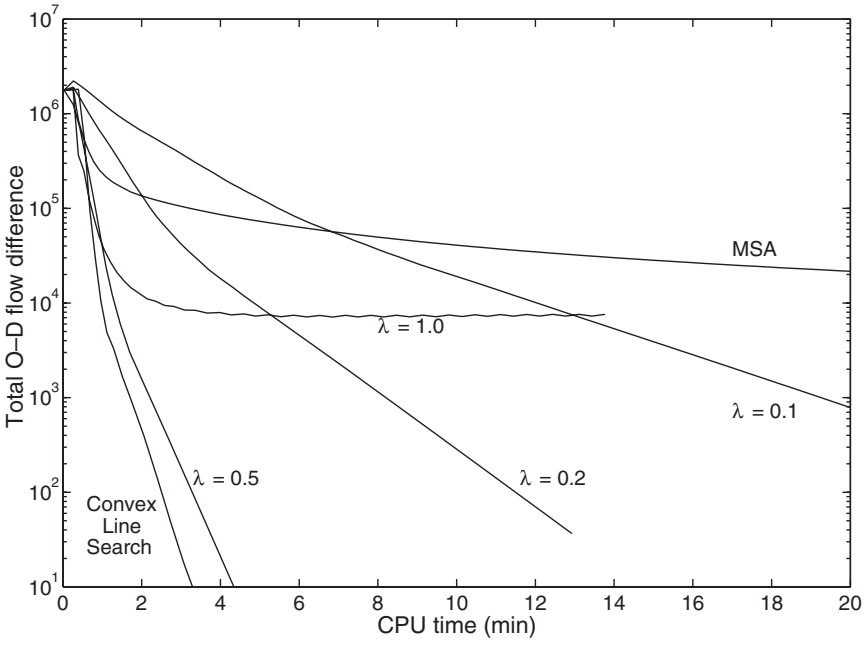


Figure 3.3 Convergence of origin-based algorithms, $\rho = 0$

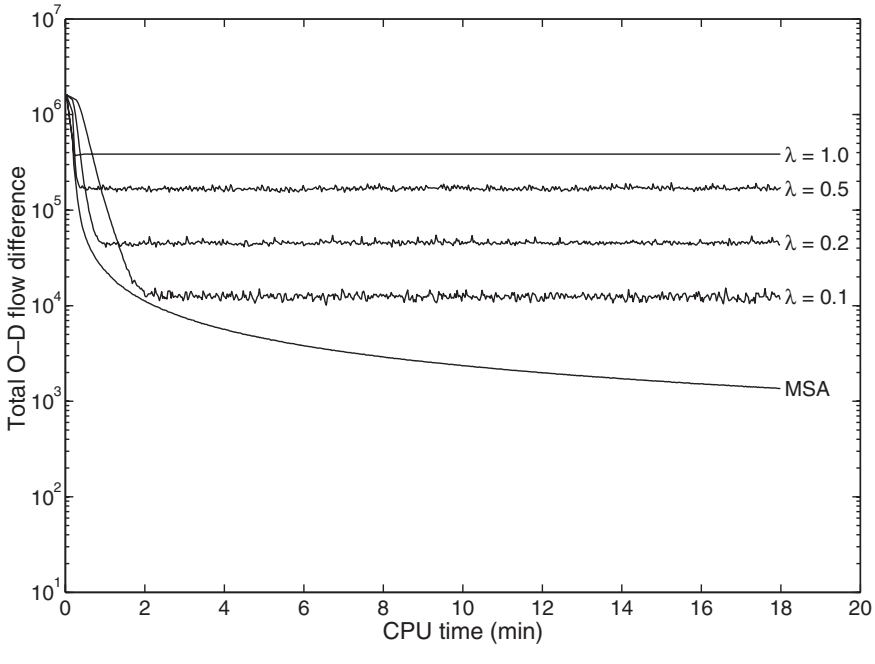


Figure 3.4 Convergence of Evans-like algorithms, $\rho = 1$

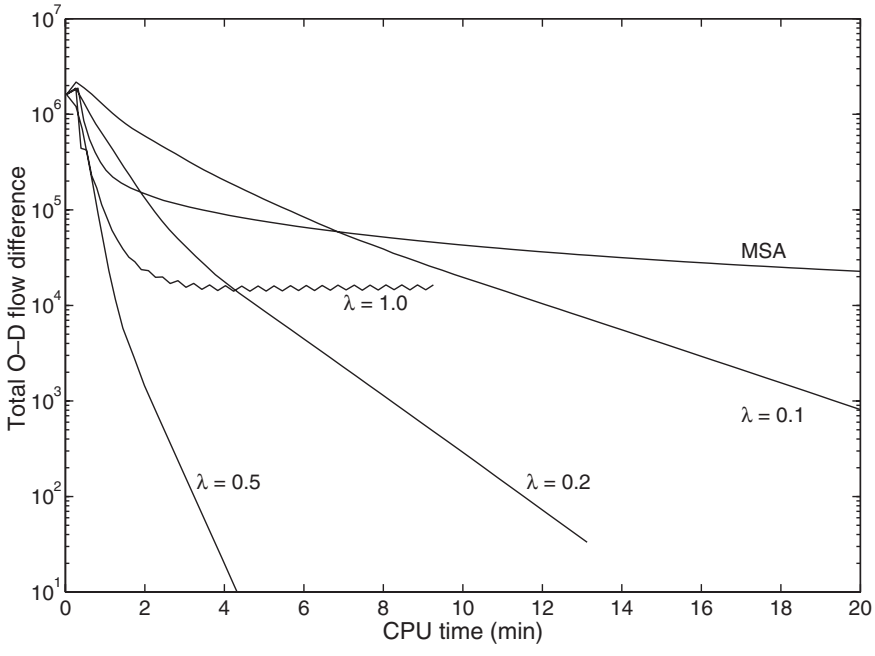


Figure 3.5 Convergence of origin-based algorithms, $\rho = 1$

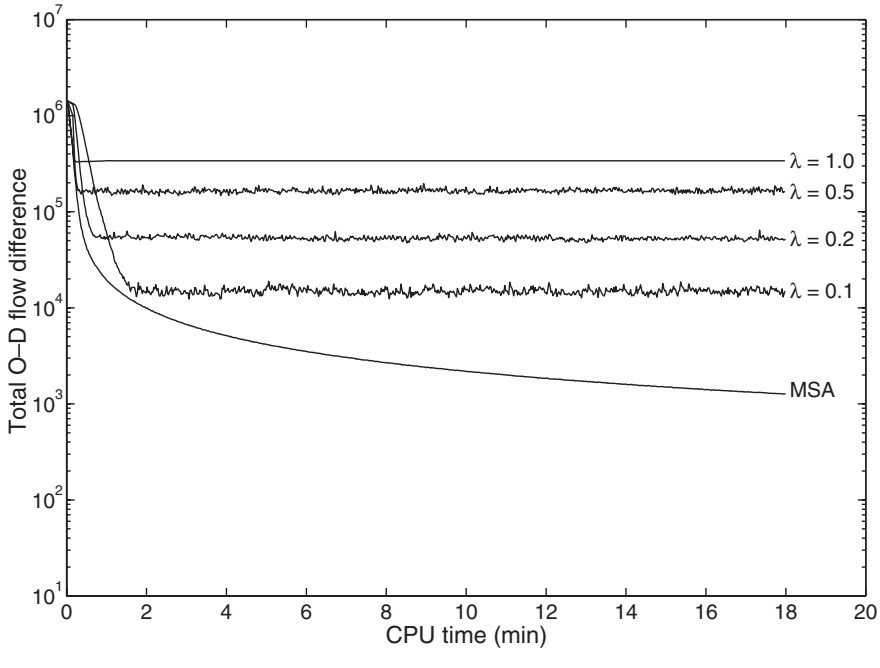


Figure 3.6 Convergence of Evans-like algorithms, $\rho = 2$

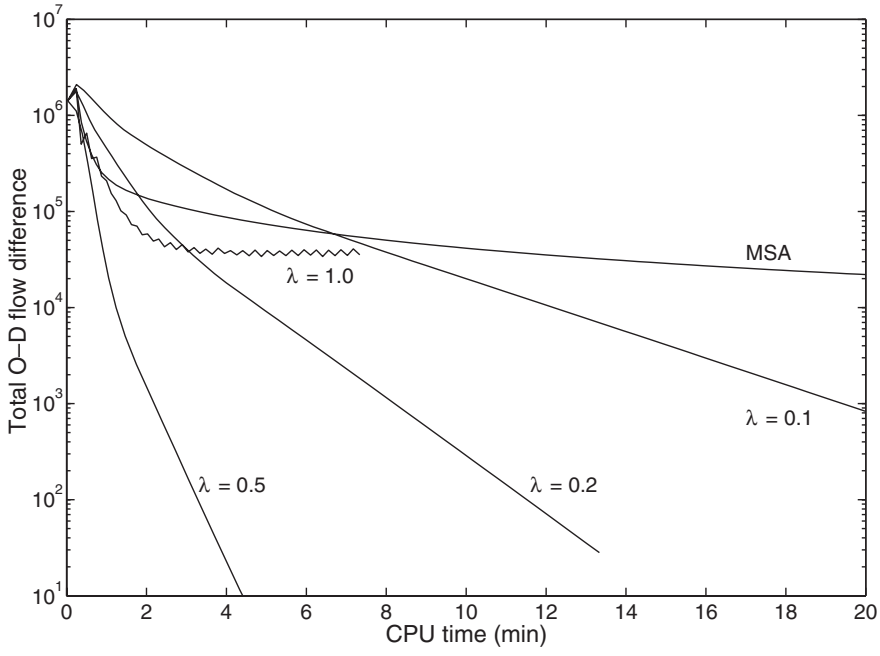


Figure 3.7 Convergence of origin-based algorithms, $\rho = 2$

Assignment convergence of solutions obtained by origin-based algorithms is substantially superior to that obtained by Evans-like algorithms. Solutions after 50 origin-based iterations with $\lambda = 0.5$, obtained in 4–5 minutes of CPU time, have average excess cost of less than 1E-10 vehicle minute equivalents. Solutions after 1000 Evans-like iterations with MSA step sizes, obtained in 20–30 minutes of CPU time, have average excess cost of 0.005 to 0.01 vehicle-minute equivalents.

It is interesting to point out that in most cases, in all algorithms while not in oscillation, the relative reduction in gap is fairly close to the step size. As discussed in the previous section, this is expected to happen when the subproblem solution does not change much as a function of the current solution, at least for most dimensions. Given the enormous number of dimensions in this problem, there may be other reasons for this observation as well. In any case, it may be possible to use this observation to develop an algorithm that adjusts the step size during the run. This remains a subject for future research.

6. CONCLUSIONS AND FUTURE RESEARCH

Traditional travel forecasting methods were based on sequential computational procedures. The need for integrated or combined models is becoming more and more convincing, in view of court decisions and legislative mandates in the United States in the last decade. The fixed point formulation presented in this chapter seems to be a natural tool

to formulate general combined models mathematically, including most models used in practice. The need for intuitive accuracy measures leads to separate consideration of assignment accuracy and the accuracy of O–D flows.

General combined models can be solved with either an Evans-like algorithm, or with an origin-based algorithm. In the first case, the sequence of step sizes used for averaging should be decreasing, as in the MSA. As a result, convergence is relatively slow. The proposed origin-based algorithm can provide much faster convergence, when a constant step size is used, as long as the step size is not too large. The results presented in this chapter should be examined and validated in other models, and particularly for different levels of congestion.

NOTE

* The authors wish to thank the Chicago Area Transportation Study, Chicago, IL for providing the Chicago network data.

REFERENCES

Bar-Gera, H. (1999), 'Origin-based algorithms for transportation network modeling', PhD Thesis, University of Illinois at Chicago, Chicago, IL.

Bar-Gera, H. (2002), 'Origin-based algorithm for the traffic assignment problem', *Transportation Science*, **36** (4), 398–417.

Bar-Gera, H. and D. Boyce (2003), 'Origin-based algorithms for combined travel forecasting models', *Transportation Research*, **37B** (5), 405–22.

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

Boyce, D. and H. Bar-Gera (2001), 'Network equilibrium models of travel choices with multiple classes', in M.L. Lahr and R.E. Miller (eds), *Regional Science in Economic Analysis*, Oxford: Elsevier Science, Ch. 6, pp. 85–98.

Boyce, D., K.S. Chon, Y.J. Lee, K.T. Lin and L.J. LeBlanc (1983), 'Implementation and evaluation of combined models of location, destination, mode and route choice', *Environment and Planning*, **A15**, 1219–30.

Boyce, D. and M.S. Daskin (1997), 'Urban transportation', in C. ReVelle and A.E. McGarity (eds), *Design and Operation of Civil and Environmental Engineering Systems*, New York: John Wiley & Sons, Ch. 7, pp. 277–341.

Boyce, D., B. Ralevic-Dekic and H. Bar-Gera (2003), 'Convergence of traffic assignments: how much is enough?', forthcoming in *ASCE Journal of Transportation Engineering*.

Dafermos, S. (1982), 'The general multimodal network equilibrium problem with elastic demand', *Networks*, **12**, 57–72.

Evans, S.P. (1976), 'Derivation and analysis of some models for combining trip distribution and assignment', *Transportation Research*, **10**, 37–57.

Florian, M., J.-H.Wu and S. He (2002), 'A multi-class multi-mode variable demand network equilibrium with hierarchical logit structures', in M. Gendreau and P. Marcotte (eds), *Transportation and Network Analysis: Current Trends*, Dordrecht, Netherlands: Kluwer Academic, Ch. 8, pp. 119–33.

Garrett, M. and M. Wachs (1996), *Transportation Planning on Trial: The Clean Air Act and Travel Forecasting*, Thousand Oaks, CA: Sage.

Kakutani, S. (1941), 'Generalization of Brouwer's fixed point theorem', *Duke Mathematical Journal*, **8** (3).

Lam, W.H.K. and H.J. Huang (1992), 'A combined trip distribution and assignment model for multiple user classes', *Transportation Research*, **26B**, 275–87.
Lundgren, J.T. and M. Patriksson (1998), 'An algorithm for the combined distribution and assignment model', in M.G.H. Bell (ed.), *Transportation Networks: Recent Methodological Advances*, Oxford: Elsevier, pp. 239–53.
Nikaido, H. (1968), *Convex Structures and Economic Theory*, New York: Academic Press.
Polyak, B.T. (1990), 'New method of stochastic approximation type', *Automation and Remote Control*, **51** (7), 937–46.
Robbins, H. and S. Monro (1951), 'A stochastic approximation method', *Mathematical Statistics*, **22**, 400–7.
Wardrop, J.G. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institution of Civil Engineers*, Part II, **1**, pp. 325–78.

4. Iteration-free microassignment

Michael Wegener*

1. INTRODUCTION

Integrated models of urban land use and transport capture the two-way interaction between location and mobility decisions of households and firms over time. Because of the slowness by which the physical stock of cities, such as residences and commercial and industrial buildings change, these models typically cover a 20- or 30-year period. To implement feedback between land use and transport, they have to run their land-use parts and their transport parts once in each simulation period.

This puts high demands on the speed by which the transport models embedded in land-use transport models are executed. Execution times of several hours, which may be acceptable if the transport model is applied only once, are prohibitive if it is to be executed once every year in a 30-year simulation. This constraint is in conflict with the current tendency to make urban travel models more disaggregate or even entirely microscopic down to the individual traveller, which typically leads to even longer execution times even with fast parallel computers. A significant part of the computing time requirements of highly disaggregate transport models is due to the large number of iterations required to achieve user-optimal equilibrium in trip assignment.

One way out of this dilemma is to review the rationale underlying these iterations. Obviously, reality does not iterate but produces a consistent sequence of trip patterns over the 24 hours of each day without trials. Why is it not possible to follow reality and produce consistent travel flows without iteration?

This chapter outlines a methodology to model activity patterns, trips and trip chains, destination, mode and route choice of individual travellers in urban regions by time of day, including within-day and period-to-period adjustment of behaviour, by microsimulation without iteration. The presentation is illustrated by a first simulation experiment using the urban region of Dortmund as a case study.

2. PROBLEM STATEMENT

Mathematical models for forecasting urban travel flows originated in the 1950s in the United States pioneered in the Chicago Area Transportation Study (CATS). The paradigmatic urban travel model consisted of four steps: (i) in the trip generation step the volume of trips originating in each travel analysis zone was estimated from socio-economic zonal data using statistically derived trip rates; (ii) in the trip distribution step these trips were allocated to possible trip destinations as a function of socio-economic characteristics of destination zones, or trip attractions, and the travel times or *generalized* cost between

them; (iii) in the modal split step these origin–destination flows were allocated to available travel modes as a function of the relative attractiveness of these modes, mostly expressed by their travel–time ratio; and (iv) in the trip assignment step these model flows were assigned to the links of the modal networks.

For the trip distribution step, the gravity model was used as the first spatial interaction (or in short SIA) model. Its straightforward physical analogy has later been replaced by better-founded formulations derived from statistical mechanics (Wilson 1967) or information theory (Snickars and Weibull 1976), yet even after these substitutions the SIA model did not provide an explanation for the spatial behaviour modelled. Only later did it become possible (Anas 1983) to link it via random utility theory (Domencich and McFadden 1975) to psychological models of human decision behaviour (Luce 1959).

It was soon becoming apparent that it was not sufficient to apply the four steps of the paradigmatic model sequentially. Depending on the flows assigned to the road network in the trip assignment step, travel times on congested links increased and became inconsistent with those used in the trip distribution and modal split steps. This inconsistency led to the definition of user-optimal network equilibrium (Beckmann et al. 1956), a state in which the pattern of flows in the network reflects the generalized costs on its links, which is equivalent to Wardrop's (1952) condition that each used route between each origin and each destination has the same generalized travel cost and no unused route has a lower cost.

There exist essentially three methods to achieve user-optimal network equilibrium through multiple iteration of trip distribution, modal split and trip assignment and averaging after each iteration (Boyce et al. 1994): the method of successive averages (MSA) over multiple all-or-nothing assignments, user-optimal assignment using Frank–Wolfe linearization and all-or-nothing assignment using partial linearization following Evans (1976). In all three methods travel times or generalized travel costs of each link are adjusted using speed–flow relationships (capacity restraint). The weights used for each iteration in the averaging are chosen to be the best in each iteration or pre-determined as $1/n$ in the nth iteration (Powell and Sheffi 1982). The iterations are largely responsible for the generally long computing times of state-of-the-art travel forecasting models. The problem prevails despite recent advances in assignment algorithms, such as the origin-based assignment by Bar-Gera (1999).

Things are getting worse with the current tendency to make urban travel models more disaggregate or even entirely microscopic down to the individual traveller in order to models multi-purpose uni-modal and inter-modal trip chains and time of day of trips, the interaction between activity and mobility patterns of household members, new lifestyles and work patterns, such as part-time work, telework and teleshopping, the interaction between travel demand, car ownership and residential and firm location, and environmental impacts of transport such as traffic noise and exposure to air pollution. Disaggregate travel models aim at a one-to-one reproduction of spatial behaviour by which individuals choose between mobility options in their pursuit of activities during a day (Axhausen and Gärling 1992; Ben-Akiva et al. 1996). Activity-based travel models start from interdependent 'activity programmes' of household members of a 'synthetic population' (Beckman et al. 1995) and translate these into home-based 'tours' consisting of one or more trips. Activity-based travel models do not model peak-hour or all-day travel but disaggregate travel behaviour by time of day, which permits the modelling of choice of departure time.

There are also disaggregate traffic assignment models based on queueing or cellular automata approaches, for example, in the TRANSIMS project (Barrett 1999; Nagel et al. 1999), which reproduce the movement of vehicles in the road network with a level of detail not known before.

However, microscopic disaggregation typically leads to even longer execution times even with fast parallel computers. As with aggregate models, a significant part of the computing time requirements of highly disaggregate transport models is due to the large number of iterations required to achieve user-optimal equilibrium in trip assignment. There are approaches to model within-day and day-to-day adjustment of behaviour by modelling dynamic network equilibrium (for example, Bernstein and Friesz 1998; Nagurney and Zhang 1998). These highly sophisticated methods, however, suffer from even larger computing time problems through iteration. Long computing times have also been a serious problem for TRANSIMS.

These long computing times become even more of a problem if the travel forecasting model is combined with an integrated urban land-use transport model. These models capture the two-way interaction between location and mobility decisions of households and firms over time. Because of the slowness by which the physical stock of cities, such as residences and commercial and industrial buildings change, they typically cover a 20- or 30-year period. To implement feedback between land use and transport, they have to run their land-use parts and their transport parts once in each simulation period. Moreover, these models are increasingly becoming more disaggregate to deal with aspects of urban form, travel demand management and environmental impacts (Spiekermann and Wegener 2000). This puts high demands on the speed by which the transport models are executed. Execution times of several hours, which may be acceptable if the transport model is applied only once, are prohibitive if it is to be executed once very year in a 30-year simulation.

One solution to this problem would be to develop a travel forecasting model that does not require iteration. Obviously, reality does not iterate but produces a consistent sequence of trip patterns over the 24 hours of each day without trials. Why is it not possible to follow reality and produce consistent travel flows without iteration?

A first step towards this goal is to review the rationale behind the concept of user-optimal network equilibrium. There are good reasons to doubt the half-century old proposition that user equilibrium is the best representation of travel behaviour. After all, the basic assumption underlying user equilibrium, complete rationality and complete information of all travellers, is highly unrealistic. Instead, many travellers often find themselves trapped in no-return situations, such as traffic jams, wrong lanes, no-turn intersections, train delays or missed connections that they would have avoided if they had had prior complete and timely information. Modelling travel behaviour then becomes the art of modelling decision making under uncertainty with incomplete information, short-term adjustment and trial and error with a significant proportion of routine and habitual behaviour. However, it can be assumed that travellers apply knowledge from previous experience. This can be exploited in a modelling environment in which the transport model is applied recursively in each simulation period.

3. MODEL FRAMEWORK

The urban travel model envisaged is part of long-term effort to develop a microsimulation model of urban land use, transport and environment (Wegener and Spiekermann 1996; Moeckel et al. 2002; Salomon et al. 2002) based on the existing land-use transport model of the urban region of Dortmund (Wegener 1998). Parts of the planned model are presently being implemented in the project ILUMASS (Integrated Land-Use Modelling and Transportation System Simulation) funded by the German Federal Ministry of Education and Research. The study area for tests and first applications of the model is the urban region of Dortmund.

The model consists of a number of microsimulation modules. A microsimulation module is a program unit that executes one elementary process (a choice, a transition or a policy) and stores the result in the common micro database. Each microsimulation module has defined input and output interfaces. Coordination between the modules is facilitated by a coordinator or scheduler program. The rows and columns of Figure 4.1 represent microsimulation modules ordered by increasing speed of change.

Transport infrastructure and buildings represent the slowest kind of change; their construction takes many years, and their life cycle is counted in decades. Firms and households also have life cycles of several years but are more easily established or dissolved. Firms and households change their location several times during their life yet even more frequently adjust their vehicle fleets to changing needs. Whereas all these changes are counted in years, logistics and household activities change from hour to hour during a single day. The fastest urban processes are goods transport and travel. They adjust in response to events in a matter of minutes. Environmental processes partly reflect the effects of human activities without delay but some have long-term consequences.

The microsimulation modules interact in various ways with one another. Figure 4.1 shows the direct interactions between microsimulation modules represented in the model. Some of these interactions are greatly delayed, that is, take their time to work their way through the system. For instance, increasing demand for office space or housing will result in new office space or new housing only after several years because of long planning and construction periods. Other impacts are much faster. For instance, dwellings vacated by households enter the supply of available housing after a few weeks. Still other impacts are almost immediate, such as driver response to congestion. This variety of response speeds requires that the exchange of information between the microsimulation modules is very efficient. This is achieved by the common micro database.

3.1 The Travel Microsimulation Module

The travel microsimulation module presented here is an attempt to combine several partially conflicting objectives:

- to model mobility decisions in a microscopic perspective in order to capture aspects of behaviour that are crucial for achieving sustainable urban transport, such as multi-purpose uni-modal and inter-modal trip chains and time of day of trips, the interaction between activity and mobility patterns of household members, new lifestyles and work patterns, such as part-time work, telework and teleshopping, the

Change of … causes \ … change of …	Road network	Public transport	Industrial buildings	Retail buildings	Office buildings	Residential buildings	Firm life-cycles	Household life-cycles	Person life-cycles	Industrial location	Retail location	Services location	Labour mobility	Residential mobility	Commercial vehicles	Car ownership	Logistics	Household activities	Goods transport	Travel	Energy, CO_2	Air pollution	Noise	Land take	Micro climate
Road network	●		●	●	●	●				●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Public transport		●											●	●		●		●		●	●	●	●	●	●
Industrial buildings			●							●											●	●	●	●	●
Retail buildings				●							●										●	●		●	●
Office buildings					●							●									●	●		●	●
Residential buildings						●								●							●	●		●	●
Firm life-cycles							●	●	●	●	●	●	●		●		●		●						
Household life-cycles						●		●	●					●		●		●		●					
Person life-cycles						●		●	●					●		●		●		●					
Industrial location			●							●		●	●	●	●		●	●	●	●					
Retail location				●							●	●	●	●	●		●	●	●	●					
Services location					●							●	●	●	●		●	●	●	●					
Labour mobility													●			●		●		●					
Residential mobility						●						●	●	●		●		●		●					
Commercial vehicles															●		●		●		●	●	●		
Car ownership						●							●	●		●		●		●	●	●	●		
Logistics															●		●		●						
Household activity													●	●		●		●		●					
Goods transport										●	●	●			●		●		●		●	●	●		●
Travel																●		●		●	●	●	●		●
Energy, CO_2			●	●	●	●															●				
Air pollution											●	●		●								●			
Noise											●	●		●									●		
Land take														●										●	
Micro climate														●											●

Figure 4.1 Interactions between microsimulation modules

interaction between travel demand, car ownership and residential and firm location, and environmental impacts of transport such as traffic noise and exposure to air pollution;

- to take into account that travellers have different travel preferences and perceptions of the transport system based on incomplete information about its current state, that they are uncertain about unexpected events, such as accidents, that they frequently find themselves trapped in situations they would have avoided if they had had prior information and that they base their travel decisions on previous experience that may be outdated or on habits and routines that are insensitive to current information;

- to reproduce the dispersion of mode and route choice that results from that diversity of preferences, incomplete information, uncertainty, trap situations and habits,
- to model both short-term adjustment, such as change of departure time or en route change of destination, mode or route as well as long-term learning based on prior experience;
- to develop efficient algorithms for activity generation, journey and trip generation, destination, mode and route choice and assignment that does not require extensive iterations.

To achieve these objectives, the stochastic microsimulation, already proposed by Burrell in 1968, is applied. However, unlike the procedure proposed by Burrell, congestion is taken into account by using generalized link travel costs based on the network link flows of the previous simulation period.

The travel microsimulation module models the selection of an activity programme for each member of each household and, following that, a departure time for each tour and a departure time, destination, mode and route for each trip (see Figure 4.2):

- *Select household* In the first step a household is selected for processing from the list of households. Each selected household is defined by its household attributes and the personal attributes of its members. The household attributes include its residential location. A location in the model is a micro location, that is, street address, geographical coordinates or raster cell of 100 × 100 metres.
- *Select person* Next, the first household member is selected. For each working person in the household the location of the workplace is known. For schoolchildren and university students the location of the school or university is known.
- *Select activity programme* Depending on the personal attributes of the household member, that is, age, sex and occupation, a daily activity pattern is selected from a catalogue of activity patterns. A daily activity pattern is defined as a schedule of tours.
- *Select car ownership and availability* Depending on household and personal attributes, it is determined whether the person has a car at his or her disposal.
- *Select tour departure time* The first tour of the activity programme is selected. The departure time is determined as a random variation of the scheduled departure time.
- *Select trip departure time* The first trip of the tour is selected. The departure time is determined as a random variation of the scheduled departure time.
- *Select destination* The destination of the trip is selected by logit choice. The locations of destinations are micro locations as above. Generalized costs of travel to the destinations are calculated as the log sum of the stochastic shortest routes (see below) of relevant modes. Relevant modes are by foot, bicycle, public transport and car (if available, see above). For work, school and university trips the destinations are already known.
- *Select mode* For the selected destination, mode choice is performed by logit choice based on the generalized costs of the stochastic shortest routes (see below).
- *Select route* For the selected mode the stochastic shortest route is selected as the route. The stochastic shortest route is the shortest route with a random disturbance

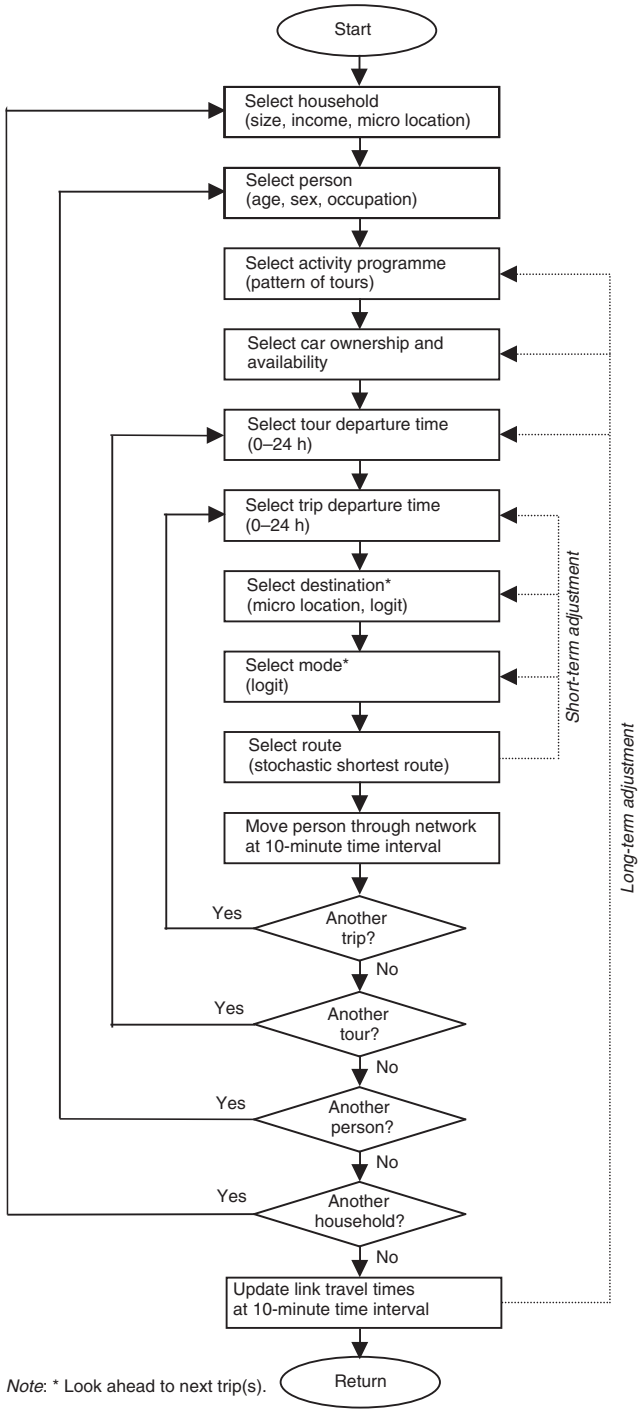


Figure 4.2 Microsimulation of travel behaviour

term added to each link generalized cost and each waiting/transfer time in the public transport network.

- *Move person through network* Each person travelling through the network is recorded on each traversed link at a 10-minute time interval.

After each trip the next trip of the route, if any, is selected. After each route, the next route, if any, is selected. After each person, the next person, if any, is selected. After each household, the next household, if any, is selected.

There are two ways of selecting the next trip. One intuitively appealing way is to start in the early morning hours with an empty network, process trips in the order of their departure time, that is spatially randomly, and after each trip update the generalized travel cost of all traversed links. In this way the gradual filling up of the network over the day is reproduced. This would be a microscopic version of the incremental loading assignment in use prior to the development of user-equilibrium assignment algorithms. If, however, it is assumed that travellers use their prior experience about network conditions when making travel decisions, an even simpler procedure can be applied. In this case the assignment does not start with free-flow generalized link costs but with the higher link travel costs of the loaded network of the previous simulation period. It is then not necessary to process trips in the order of their departure time. Only after all tours and trips have been executed, are the travel times of all traversed road links in each 10-minute interval of the day updated to account for congestion; however, this information will be used only in the next simulation period. Representative travel times and generalized costs between zones (required for accessibility calculations in the land-use model) are calculated on the basis of the shortest routes with updated travel times.

If during a trip a significant amount of congestion is encountered, short-term adjustment resulting in a postponement of the trip or a change of mode or route may occur. However, only changes of departure time and mode that can be made en route are implemented. Long-term adjustment of travel behaviour, such as going to work later or buying a monthly public transport pass, are based on the generalized costs of the network in the previous simulation period. Generalized costs are a combination of travel time and travel cost and can be different for each type of traveller to take account of the diversity of travel preferences.

Special provisions are necessary when no prior information is available, as in the first simulation period or in the case of large infrastructure changes. In the first simulation period either one aggregate user-equilibrium assignment using the Evans algorithm or one microassignment iteration starting from medium-flow generalized link travel costs may precede the actual assignment. Similarly, large infrastructure improvements, such as new road links, may be introduced with medium-flow generalized link travel costs representing the most likely expectation of travellers.

It is hoped that microassignment without iteration will produce a similar distribution of trips across destinations, modes and routes as user-optimal assignment with iteration. Total user benefit should be less due to the effects of uncertainty, incomplete information, trial and error and habitual behaviour. It will be an interesting task to examine the degree of sub-optimality and how the simulated travel behaviour compares to observed behaviour and the results of travel models based on user-optimal network equilibrium.

3.2 A First Test

As a first test of the proposed method, iteration-free microassignment was applied to peak-hour car trips in the Dortmund urban region and compared with the results of a user-optimal equilibrium assignment of the aggregate transport component of the existing land-use transport model of the region.

The existing transport model applies the Evans algorithm to arrive at a user equilibrium of trip generation, car ownership, trip distribution, modal split (by foot, bicycle, public transport, car) and route choice (Wegener 1986). Normally eight iterations are performed, but for this exercise the number of iterations was increased to 20. A simplified version of the model with 36 zones and 1800 network links was used; in the final version, 200 000 microlocations (grid cells) and 8000 network links will be used.

The resulting origin–destination matrix of peak-hour interzonal car trips was then assigned to the links of the road network on a car-by-car basis. For each car trip, a stochastic shortest route to its destination was determined using a shortest-route algorithm. For this, the generalized link travel costs of the aggregate user-optimal equilibrium assignment produced by the Evans algorithm described above were used – in the final model the travel costs of the loaded network of the previous simulation period will be used.

During the shortest-route search, these generalized link travel costs were disturbed by a uniformly disturbed random increment or decrement of up to 10 per cent. A 'once-through' shortest-route algorithm, in which the nodes already reached but not further processed (the 'candidates') are temporarily preserved in the 'candidate list' in the order of their travel cost from the origin node (and are hence processed only once), ensured that each link cost was disturbed only once. Only route changes were taken into account; no other behavioural responses, such as change of departure time or change of mode, were implemented nor were time intervals or microlocations considered – this will be left to future experiments.

In Figure 4.3 the resulting link flows of the microassignment are compared with the link flows generated by the aggregate user-equilibrium assignment using the Evans algorithm. It can be seen that the most link flows produced by the two methods are very similar, with a few significant deviations that need further investigation by comparison with observed link flows. It should be noted that a perfect fit of the two link-flow distributions cannot be expected – and is not even desirable if the hypothesis holds that user equilibrium is not the best approximation of actual travel behaviour.

4. CONCLUSIONS

This chapter has outlined a methodology to assign individual trips generated in a microscopic activity-based travel forecasting model to a multi-modal transport network without iteration. The rationale of the method rests on the assumption that travellers have only limited information about the current state of the network and that they base their travel decisions on prior knowledge from earlier experience. This assumption challenges the common assertion that user-optimal equilibrium represents the best approximation of travel behaviour.

The iteration-free nature of the approach makes it particularly suitable for integrated

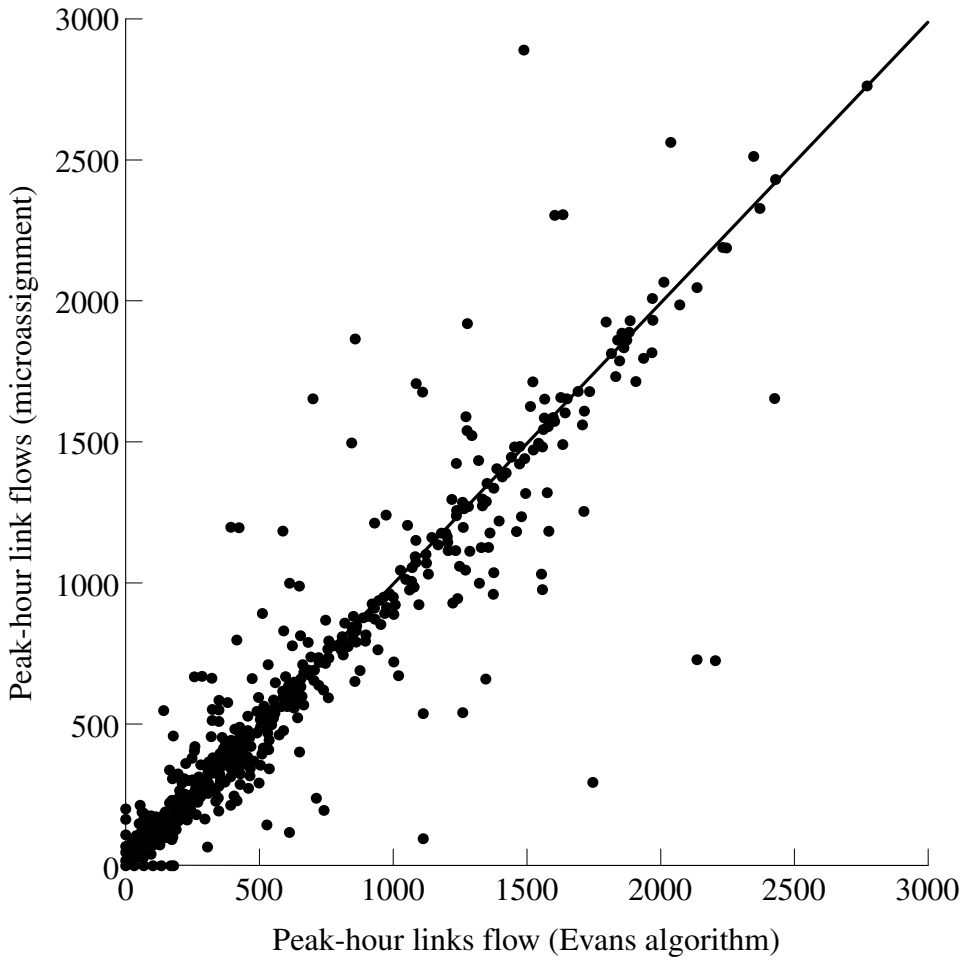


Figure 4.3 Microassignment versus assignment using the Evans algorithm

models of urban land use, transport and environment (LTE), which, as indicated, are increasingly becoming more disaggregate. Ideally, the transport component of a microscopic LTE model should also be microscopic but this conflicts with the need to have a very fast transport model to implement feedback between transport and land use. The proposed iteration-free microassignment method promises to be a solution to this conflict.

Nevertheless, before this becomes reality, several problems have to be solved. At a conceptual level, the question to what degree transport networks are in equilibrium needs to be investigated empirically. This will require new approaches of analysing travel choice behaviour from a cognitive-science perspective. In addition, a number of technical problems have to be addressed. Even iteration-free assignment is too slow unless efficient methods to calculate individual shortest routes between one origin and one destination

(not trees) are developed. Also the short-term adjustment conceptualized in the algorithm will have to be implemented in a theoretically sound and at the same time efficient manner. Finally, new methods of calibrating and validating the model against observed travel data will need to be developed.

NOTE

* The author is grateful to David Boyce for a highly useful discussion on the concepts underlying this chapter which led to a significant change in the original idea – bringing to mind a similarly important suggestion made by him when the author struggled with his first implementation of the Evans algorithm (Wegener 1986). The author also greatly benefited from several enlightening discussions with Britton Harris, who has developed innovative ideas about microassignment, although eventually an independent strategy was followed. Helpful comments by two anonymous referees are gratefully acknowledged.

REFERENCES

Anas, A. (1983), 'Discrete choice theory, information theory and the multinomial logit and gravity models', *Transportation Research*, **17B**, 13–23.

Axhausen, K.W. and T. Gärling (1992), 'Activity-based approaches to travel analysis: conceptual frameworks, models and research problems', *Transport Reviews*, **12**, 324–41.

Bar-Gera, H. (1999), 'Origin-based algorithms for transportation network modeling', PhD thesis, University of Illinois at Chicago, Chicago, IL.

Barrett, C.L. (1999), 'TRansportation ANalysis SIMulation System (TRANSIMS)', Version TRANSIMS-LANL-1.0, Volume 0 – Overview, LA-UR 99-1658, Los Alamos, NM: Los Alamos National Laboratory, http://transims.tsasa.lanl.gov/documents.html, 3 March 2003.

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

Beckman, R.J., K.A. Baggerly and M.D. McKay (1995), *Creating Synthetic Baseline Populations*, LA-UR-95-1985, Los Alamos, NM: Los Alamos National Laboratory.

Ben-Akiva, M.E., J.L. Bowman and D. Gopinath (1996), 'Travel demand model system for the information era', *Transportation*, **23**, 241–66.

Bernstein, D. and T.L. Friesz (1998), 'Infinite dimensional formulations of some dynamic traffic assignment models', in Lars Lundqvist, Lars-Göran Mattsson and Tschangho John Kim (eds), *Network Infrastructure and the Urban Environment*, Berlin/Heidelberg: Springer, pp. 112–24.

Boyce, D.E., Y. Zhang and M.R. Lupa (1994), 'Introducing "feedback" into four-step travel forecasting procedure vs. the equilibrium solution of a combined model', *Transportation Research Record*, **1443**, 65–74.

Burrell, J.E. (1968), 'Multipath route assignment and its application to capacity restraint', in *Proceedings of the 4th International Symposium on the Theory of Road Traffic Flows*, Karlsruhe: University of Karlsruhe, pp. 210–19.

Domencich, T.A. and D.L. McFadden (1975), *Urban Travel Demand: A Behavioral Analysis*, Amsterdam: North-Holland.

Evans, S.P. (1976), 'Derivation and analysis of some models for combining trip distribution and assignment', *Transportation Research*, **10**, 37–57.

Luce, R.D. (1959), *Individual Choice Behavior*, New York: Wiley.

Moeckel, R., C. Schürmann, K. Spiekermann and M. Wegener (2002), 'Microsimulation of urban land use', paper presented at the 42nd Congress of the European Regional Science Association, Dortmund, 27–31 August 2002.

Nagel, K., R.J. Beckman and C.L. Barrett (1999), *TRANSIMS for Transportation Planning*, LA-UR 98-4389, Los Alamos, NM: Los Alamos National Laboratory, http://transims.tsasa.lanl.gov/PDF_Files/LAUR98-4389.pdf, 3 March 2003.

Nagurney, A. and D. Zhang (1998), 'Introduction to projected dynamical systems for traffic network equilibrium problems', in Lars Lundqvist, Lars-Göran Mattsson and Tschangho John Kim (eds), *Network Infrastructure and the Urban Environment*, Berlin/Heidelberg: Springer, pp. 125–56.

Powell, W.B. and Y. Sheffi (1982), 'The convergence of equilibrium algorithms with predetermined step sizes', *Transportation Science*, **16**, 45–55.

Salomon, I., P. Waddell and M. Wegener (2002), 'Sustainable lifestyles? Microsimulation of household formation, housing choice, and travel behavior', in William R. Black and Peter Nijkamp (eds), *Social Change and Sustainable Transport*, Bloomington, IN: Indiana University Press, pp. 125–31.

Snickars, F. and J.W. Weibull (1976), 'A minimum information principle', *Regional Science and Urban Economics*, **7**, 137–68.

Spiekermann K. and M. Wegener (2000), 'Freedom from the tyranny of zones: towards new GIS-based models', in A. Stewart Fotheringham and Michael Wegener (eds), *Spatial Models and GIS: New Potential and New Models*, GISDATA 7, London: Taylor & Francis, pp. 45–61.

Wardrop, J.G. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institute of Civil Engineers*, Part II, **1**, 325–78.

Wegener, M. (1986), 'Transport network equilibrium and regional deconcentration', *Environment and Planning A*, 18, 437–56.

Wegener, M. (1998), 'The IRPUD model: overview', http://irpud.raumplanung.uni-dortmund.de/irpud/pro/mod/mod.htm, 3 March 2003.

Wegener, M. and K. Spiekermann (1996), 'The potential of microsimulation for urban models', in Graham P. Clarke (ed.), *Microsimulation for Urban and Regional Policy Analysis*, European Research in Regional Science 6, London: Pion, pp. 149–63.

Wilson, A.G. (1967), 'A statistical theory of spatial distribution models', *Transportation Research*, **1**, 253– 69.

5. Cost minimizing behavior in random discrete choice modeling

Sven Erlander and Jan T. Lundgren*

1. INTRODUCTION

In many planning situations the concern is to estimate the number of decision makers choosing each alternative among a set of available discrete alternatives. Models describing this type of choice are known as discrete choice models (McFadden 1974, 1978, 1981, 2001; Ben-Akiva and Lerman 1985; Boyce et al. 1988). Typical applications concern land-use and travel demand analysis, including residential and employment location and the choice of frequency, origin, destination and mode for travel (Boyce 1984; Boyce and Lundqvist 1987; Boyce and Mattsson 1999). In discrete choice models it is often postulated that decision makers exhibit a cost minimizing behavior in the choice process. Sometimes the behavior is formulated in terms of perceived utility, and then we talk about a utility maximization approach. In this chapter, however, we shall use cost instead of utility as the (negative) attractiveness measure.

Cost minimizing behavior can be characterized in many ways, and there are several ways to derive discrete choice models. A common element in the planning situation is a population of decision makers, each choosing exactly one alternative from a given choice set of alternatives. In a travel to work situation, for example, the trip makers choose residential and employment localities in a more or less rational way. The decisions are influenced by out-of-pocket money costs, time costs, parking costs and so on. Also, this influence may be different according to sex, income, family situation, car ownership and other properties of the trip maker. The notion of generalized cost is often used in order to extract information about this influence in the form of a linear function of the factors included, with the idea that the decision makers will tend to choose the alternative with the lowest generalized cost. Not all decision makers, however, will choose the least-cost alternative. There is always a variation in choice among the decision makers. This variation may be due to the fact that there are many factors not represented in the generalized cost; factors that nevertheless influence the decision making in unknown ways. Also, the individual characteristics of the decision makers may lead to different choices.

The variation in the decision making can be taken care of by using probabilistic models. The aim of the model derivation is then to obtain an expression for the probability of the decision maker to choose each alternative. From this, the proportion of all decision makers (or the total number) choosing each alternative can easily be computed.

One approach is to try to describe the behavior of one decision maker, who is assumed to behave rationally and who chooses the alternative which for him/her has the lowest cost. This cost is assumed to be composed of the generalized cost and an unknown

random cost component, which will make the decision makers choose different alternatives. The decision maker can be looked upon as a randomly chosen decision maker from the population of decision makers. By drawing a decision maker at random, the probability distribution of the random component describes the variation among the decision makers in the population.

The traditional way to derive the probability of choosing each alternative is to make an assumption about the form of the probability distribution of the random component of the cost. Usually, this probability distribution is assumed to be a (generalized) extreme value distribution or a normal distribution. In this chapter we shall proceed differently. Instead of considering the rational behavior for an *individual* decision maker, by making an assumption regarding the probability distribution for the random cost component, we shall define rational behavior for a *group* of decision makers. We shall introduce and define *cost minimizing behavior for a group of decision makers*. Cost minimizing behavior obtains if lower values of the total cost for the chosen alternatives in the group are more frequently observed than higher values of the total cost. Given this assumption about how a group of decision makers choose alternatives, we can derive the probability distribution for the choice between the alternatives without having to specify the probability distribution of the random cost component.

The purpose of introducing this new definition is threefold. First, we wish to show that there is an intuitively easily understandable way of describing cost minimizing behavior. In this new notion of cost minimizing behavior we avoid the need to introduce concepts and relations that cannot be observed (the probability distribution of the random cost component). Second, we wish to demonstrate that all discrete choice models that can be derived by the additive random utility maximizing (ARUM) approach with (generalized) extreme value distribution for the random component of the perceived utility (random cost component) can be derived directly from the appropriate version of the new definition of cost minimizing behavior. We shall discuss this in detail for the multinomial logit model. Third, we wish to bring out and discuss the underlying assumptions of the logit model. When formulated in terms of the new definition of cost minimizing behavior, we believe that the underlying assumptions are easy to understand and interpret. There is no need to introduce random components with non-observable properties.

The chapter presents a new form of derivation of discrete choice models, or maybe one could say a new explanation or interpretation of an old derivation. We do not propose a derivation leading to a different type of model and/or different predictions. Our aim is to propose a derivation of general discrete choice models which has a behavioral interpretation, different from the standard additive random perceived utility maximizing approach.

There will be no difference in the estimation or use of the model. Our contribution is to present an alternative derivation/interpretation which can add to all the others and in this way strengthen the discrete choice paradigm. We are convinced that many users of discrete choice models are uncomfortable with the assumption of a certain form of the stochastic component of the utility function. In our approach we do not need this assumption.

The new definition does not contradict other definitions. In particular, all models that can be derived by the additive random utility maximizing approach with generalized extreme value distribution for the non-observable random components satisfy our definition. In the standard derivation the expected achieved utility takes the form of the so-called log sum. Our approach removes some of the log-sum anomalies in limiting cases.

In the next section we describe and define the new notion of cost minimization behavior applied to the most simple case of a discrete choice situation. Then we show how the multi-attribute multinomial logit model and the gravity model for trip distribution can be derived from this new definition. We also present cost minimizing behavior in the general case of discrete choice modeling.

2. COST MINIMIZING BEHAVIOR

Consider a decision maker (or set of decision makers) confronted with the choice between K alternatives. For each alternative k we can specify a generalized cost, c_k, which incorporates all relevant characteristics of the alternative. This cost should be composed of known deterministic quantities for each alternative, the same for all decision makers under consideration. In a choice situation between different modes of travel the generalized cost can be the simple travel cost, or a cost composed of monetary costs, time costs, parking costs and so on. In this section we shall concentrate on the simple case where the generalized cost is given by c_k only. Later on we shall investigate the case where the generalized cost is specified as a linear function of the variables (attributes) used to characterize the alternatives.

The standard approach is the following. The individual decision maker is assumed to behave rationally and to choose the alternative which for him/her has the lowest cost. This cost is composed of the generalized cost c_k and a random cost component ε_k, reflecting variation in taste and availability of information among the decision makers, and the influence of unknown factors. Thus, the decision maker chooses the alternative which minimizes $c_k + \varepsilon_k$. While c_k is a deterministic component, we have to model ε_k as a random variable. To derive the probability p_k of choosing alternative k for the decision maker, we usually make an assumption about the form of the probability distribution of ε_k. Traditionally this probability distribution is assumed to be a (generalized) extreme value distribution or a normal distribution.

In this chapter we shall proceed differently to derive the probability p_k. Instead of making an assumption regarding the probability distribution for the cost component ε_k, and from this assumption define rational behavior for an *individual* leading to the probability distribution p_k, we shall define rational behavior for a *group* of decision makers. Given an assumption about how a group of decision makers choose alternatives we can derive the probability distribution p_k without having to specify the probability distribution of ε_k.

Consider a group of decision makers, where each individual in the group perceives the costs of the alternatives differently. The individuals will therefore choose different alternatives when according to their rational behavior they choose the cost minimizing alternative. In this way, alternatives with a higher generalized cost will also be chosen by some individuals in the group. However, the probability for a (randomly chosen) decision maker in the group to choose an alternative with higher generalized cost is lower. This means that we would expect an alternative with lower cost to be chosen by more decision makers, that is, such an alternative would be more frequently observed. Our assumption of cost minimizing behavior for the group then says that choices of the individuals of the group leading to lower value of *the sum of the generalized costs* would be more probable (fre-

quently observed). It turns out that by using this simple device we can derive the desired probability distribution. Thus, by assuming cost minimizing behavior in this way we can derive the choice probabilities without further assumptions about the specific behavior of the individual decision makers, that is, without specifying the form of the probability distribution of the random cost component ε_k.

Assume that there are N individuals in the group. Let $d=(d_1,\ldots, d_N)$ denote the decisions of all N individuals, which we shall call a *decision pattern*. Further, let z_k be the number of individuals choosing alternative k in the decision pattern. The total cost for the group, based on the generalized costs c_k, $k=1,\ldots, K$, can be expressed as:

$$c(d)=c(d_1)+\ldots+c(d_N)=\sum_{k=1}^{K} c_k z_k,$$

where $c(d_n)$ denotes the generalized cost of the chosen alternative for individual n.

The decision of each individual can be described by the unknown probability distribution $p=(p_1,\ldots,p_K)$. This is the probability distribution we want to derive and it is the same for all individuals in the group. Thus, the probability for individual n in the group to choose alternative k is $p_k=Pr(d_n=k)$, independent of which of the N individuals we consider. Therefore, the probability for decision pattern d can be expressed as:

$$p(d)=p(d_1)p(d_2)\ \ldots\ p(d_N)=\prod_{k=1}^{K} p_k^{z_k},$$

where $p(d_n)$ is the probability for the alternative chosen by individual n.

The decisions of the individuals in the group can be different, and alternatives with higher as well as lower generalized cost can be chosen. A reasonable condition on the probability distribution describing these choices is that an alternative with lower generalized cost should be more probable. This satisfies our intuitive feeling for the notion of cost minimizing behavior. We can also express this behavior by requiring that p_k should be a decreasing function of c_k. Since each member of the group behaves rationally and tries to minimize his/her individual cost, we can express the rationality in terms of the total generalized cost for the chosen alternatives of all members of the group. In the same way as for an individual choice, we can state for a group that *a decision pattern with lower total generalized cost should be more probable.* This latter statement is used to define cost minimizing behavior.

Assume that we have two groups of the same size N and that we compare the two decision patterns, d^1 and d^2, for the groups. Let z_k^1 and z_k^2 denote the corresponding number of times alternative k is chosen in decision patterns (groups) 1 and 2, respectively.

Definition 1: Cost minimizing behavior A probability distribution $p=(p_1,\ldots,p_K)$ represents cost minimizing behavior if and only if, for any group size N,

$$c(d^1)\leq c(d^2)\Rightarrow p(d^1)\geq p(d^2),$$

or equivalently,

$$\sum_{k=1}^{K} c_k z_k^1\leq \sum_{k=1}^{K} c_k z_k^2 \Rightarrow \prod_{k=1}^{K} p_k^{z_k^1}\geq \prod_{k=1}^{K} p_k^{z_k^2}. \tag{5.1}$$

A probability distribution p is a *cost minimizing probability distribution* if it satisfies condition (5.1) above.

Definition 1 is expressed in terms of properties of decision patterns involving simultaneous choices by a group of independent decision makers. It will, however, be used to derive the choice probability distribution for *one* decision maker. The right-hand side of (5.1) expresses the relation between the probabilities of the particular decision patterns, and a decision pattern with lower total cost is in Definition 1 assumed to be more probable than a decision pattern with higher total cost. The simultaneous probabilities in (5.1) are, because of independence, expressed in terms of the probability distribution p which is assumed to hold identically for each member of the group of decision makers.

Given that we assume this cost minimizing behavior to hold, the form of the probability distribution can be derived. A remarkable fact is that this simple definition (Definition 1) is enough to completely determine the probability distribution p.

Proposition 1 The probability distribution $p=(p_1,\ldots,p_K)$ represents cost minimizing behavior according to Definition 1 if and only if it is a log linear probability distribution

$$p_k = \exp(\mu - \beta c_k),$$

which can equivalently be written:

$$p_k = \frac{\exp(-\beta c_k)}{\Sigma_{k=1}^{K} \exp(-\beta c_k)}. \tag{5.2}$$

Remark Proposition 1 is a particular case of the general representation theorem which will be presented later.

Proof By substitution into (5.1) it is easy to verify that the probability distribution (5.2) satisfies the definition of cost minimizing behavior. To prove the reverse implication some more work is needed.

Assume that the probability distribution $p=(p_1,\ldots,p_K)$ is cost minimizing, hence satisfying (5.1). Let the generalized cost coefficients c_k, $k=1,\ldots,K$, be rational. Further, let $w=(w_1,\ldots,w_K)$ be arbitrarily given, where $w_k>0$ and integer, $k=1,\ldots,K$, such that $\Sigma_{k=1}^{K} w_k = W$.

Then, consider the linear program in the K variables $y=(y_1,\ldots,y_K)$:

$$\min \sum_{k=1}^{K} y_k \log p_k, \quad \text{subject to} \quad \sum_{k=1}^{K} c_k y_k \leq \sum_{k=1}^{K} c_k (w_k/W), \quad \sum_{k=1}^{K} y_k = 1,\ y_k \geq 0,\ k=1,\ldots,K.$$

Clearly, $(y_1,\ldots,y_K)=(w_1/W,\ldots,w_K/W)$ is a feasible solution, We shall show that this solution is also optimal, and the form of the probability distribution then follows from duality theory.

Since the linear program has rational coefficients and rational right-hand side, there is a rational optimal solution $\hat{y}=(m_1/n,\ldots,m_K/n)$ where m_k, $k=1,\ldots,K$, and n are integer numbers. Now, consider two decision patterns defined by $d^1=nW\hat{y}=(nW\hat{y}_1,\ldots,nW\hat{y}_K)$ and $d^2=n\mathrm{w}=(nw_1,\ldots,nw_K)$. These decision patterns are of the same group size, which is:

$$\sum_{k=1}^{K} nW\hat{y}_k = \sum_{k=1}^{K} nw_k = nW.$$

Since $\hat{y}$ by definition is feasible in the linear program we have that:

$$nW\sum_{k=1}^{K} c_k \hat{y}_k \leq nW\sum_{k=1}^{K} c_k w_k / W,$$

which can be expressed:

$$\sum_{k=1}^{K} c_k (nW\hat{y}_k) \leq \sum_{k=1}^{K} c_k (nw_k).$$

This inequality shows that the decision patterns $d^1 = nW\hat{y}$ and $d^2 = nw$ satisfy the inequality on the left-hand side of implication (5.1) in Definition 1. Since p is assumed to be cost minimizing, it follows that:

$$\prod_{k=1}^{K} p_k^{nW\hat{y}_k} \geq \prod_{k=1}^{K} p_k^{nw_k},$$

which can be written:

$$\sum_{k=1}^{K} \hat{y}_k \log p_k \geq \sum_{k=1}^{K} (w_k/W) \log p_k.$$

Hence, $y = w/W$ is also an optimal solution to the linear program (because $\hat{y}$ is optimal by assumption). Since $w_k > 0$, $k = 1,\ldots,K$, it follows from the complementary slackness theorem of linear programming that there are dual variables μ and β, such that the dual constraints are satisfied with equality. Hence, $\log p_k = \mu - \beta c_k$, which can be rewritten as the logit probability distribution (5.2).

We recognize the probability distribution (5.2) as the (multinomial) logit model. Hence, the logit model exhibits cost minimizing behavior according to Definition 1. Also, Proposition 1 says that the probability distribution (5.2) is the *only* model that has the cost minimization behavior according to Definition 1. Thus, we have demonstrated that any probability distribution that represents cost minimizing behavior in the sense of Definition 1 with rational cost coefficients has to be of the logit model form (5.2). In fact all discrete choice models that can be derived by the standard ARUM approach with generalized extreme value (GEV) distributions (McFadden 1974, 1978, 1981) satisfy an appropriate variant of our definition of cost minimizing behavior (see Definition 5). All such models can therefore be derived by our approach without having to assume the detailed structure of the decision making. Hence our new definition of cost minimizing behavior offers a new interpretation of the underlying structure of these models.

In the logit model the expected average cost can be expressed as:

$$c = E[c(d)/N] = \sum_{k=1}^{K} c_k p_k = \sum_{k=1}^{K} c_k \exp(-\beta c_k) / \sum_{k=1}^{K} \exp(-\beta c_k).$$

In the ARUM approach the expected achieved perceived disutility takes the form of the so-called log sum/composite cost (Williams 1977; Sheffi 1985)

$$-(1/\beta) \log \sum_{k=1}^{K} \exp(-\beta c_k). \tag{5.3}$$

One anomaly with the latter is the behavior for small values of β. In this case the probability distribution tends to the uniform distribution, that is, the cost values c_k are of little

importance. However, the expected achieved perceived disutility tends to minus infinity as $\beta \to 0$. This strange behavior is due to the fact that the log sum can be expressed as (Erlander 2001):

$$-(1/\beta)\log\sum_{k=1}^{K}\exp(-\beta c_k) = \sum_{k=1}^{K} c_k p_k + (1/\beta)\sum_{k=1}^{K} p_k \log p_k,$$

representing the sum of expected cost and the negative of freedom of choice. (Fisk and Boyce 1984) discuss modifications of the composite cost to avoid this problem.

Cost minimizing behavior as defined by Definition 1 is a property of a group of decision makers. We can alternatively interpret Definition 1 as a property of a population Ω of decision makers facing a set of alternatives. Each member of the set Ω chooses in some way an alternative with corresponding generalized cost, and the choice is unknown to us until we observe the decision maker. Since we cannot observe all members in the population (they are too many), we have to consider only a limited series (or sequence) of decision makers. Instead of using the concept of 'group', we can express Definition 1 in terms of 'samples' of members in the population Ω. Comparing samples of decision patterns (series of observations) we would expect that samples with lower values of the sum of the generalized cost would be more frequently observed, if cost minimizing behavior is at hand. In other words, series with lower value of the sum of the costs would be more probable. This is what is expressed in Definition 1. Using Proposition 1 we can then derive the probability distribution p of the decision taken by a decision maker drawn at random from the population Ω. Thus, by assuming cost minimizing behavior for the population Ω we can derive the choice probabilities without further assumptions about the specific behavior of the decision makers.

In summary, we can alternatively speak about cost minimizing behavior as a property of the individuals in a group of decision makers, as a property of members of the population Ω, as a property of the population itself or as a property of the choice probability distribution p.

3. DERIVATION OF DISCRETE CHOICE MODELS

In this section we shall describe how the definition of cost minimizing behavior can be used to derive two types of discrete choice models; the multi-attribute discrete choice model and the gravity model for trip distribution. We shall also consider the derivation of discrete choice models in general.

3.1 The Multi-attribute Discrete Choice Model

Assume that each alternative can be characterized by S attributes, and that we are given (deterministic) cost measures c_{sk}, $s=1,\ldots,S$, for each alternative k, $k=1,\ldots,K$. Given a decision pattern $d=(d_1,\ldots,d_N)$ and the corresponding values z_k, $k=1,\ldots,K$, of how many individuals that have chosen alternative k in this decision pattern, the total cost measure of attribute s for all N individuals in the group of decision makers is given by $\Sigma_{k=1}^{K} z_k c_{sk}$. We can now express rationality of the decision makers in the group by defining cost minimizing behavior in terms of this total cost measure: a decision pattern for a group that simultaneously has lower total cost measure for all S attributes should be more probable.

Assume that we have two groups of the same size N and that we compare the two decision patterns, d^1 and d^2 (with corresponding values z_k^1 and z_k^2, $k=1,\ldots,K$) for the groups.

Definition 2: Cost minimizing behavior for multi-attribute discrete choice A probability distribution $p=(p_1,\ldots,p_K)$ represents cost minimizing behavior if and only if, for any group size N,

$$\left(\sum_{k=1}^{K} c_{sk} z_k^1 \leq \sum_{k=1}^{K} c_{sk} z_k^2,\ s=1,\ldots,S\right) \Rightarrow \prod_{k=1}^{K} p_k^{z_k^1} \geq \prod_{k=1}^{K} p_k^{z_k^2}. \tag{5.4}$$

The implication in (5.4) says that *if* the total cost measure for each attribute is lower in one group compared to the other, the probability to observe the decisions of that group should be higher. Note that nothing is said about the relation between the probabilities of observing the decisions in the two groups, if the relations between the total cost measures do *not* hold.

By substitution into (5.4) we find that the logit model,

$$p_k = \frac{\exp\left(-\Sigma_{s=1}^{S} \beta_s c_{sk}\right)}{\Sigma_{k=1}^{K} \exp\left(-\Sigma_{s=1}^{S} \beta_s c_{sk}\right)},\ k=1,\ldots,K, \tag{5.5}$$

where $\beta_s \geq 0$, $s=1,\ldots,S$, satisfies the definition of cost minimizing probability distribution. This is the so-called linear-in-parameters multinomial logit model (see, for example, Ben-Akiva and Lerman 1985).

Also, the multi-attribute discrete choice model (5.5) is not just cost minimizing according to Definition 2, but it is in fact the *only* probability distribution with this property when there are S cost measures $[c_{sk}]$ available. This follows from the general representation theorem (see Proposition 2).

3.2 The Gravity Model for Trip Distribution

In our second example we shall consider a trip distribution problem. In this case we have a group of N trip makers (commuters) going from I origin zones to J destination zones. The trip makers correspond here to the decision makers in the previous examples. The decisions concern the simultaneous decisions about where to live and where to work, and the choice set consists of all pairs (i,j), $i=1,\ldots,I$, $j=1,\ldots,J$. Let c_{ij} denote the (generalized) cost of going from zone i to zone j. Hence, c_{ij} corresponds to the generalized cost c_k in the previous examples. We want to derive the probability distribution $p=(p_{ij})$, where p_{ij} is the probability that a trip maker chooses origin i and destination j.

Assume that each of the N trip makers makes a decision, that is, chooses one pair (i,j), independently of the other trip makers. Let $d_n=(i_n,j_n)$ be the decision taken by trip maker n, and let the *trip pattern* (decision pattern) $d=(d_1,\ldots,d_N)$ denote the decisions of all N trip makers. From this trip pattern we can compute the number of trip makers going from i to j, which we shall denote T_{ij}. Hence, T_{ij} corresponds to z_k in the previous examples.

The probability of trip pattern d can now be expressed as:

$$p(d) = \Pr[d_1=(i_1,j_1),\ldots,d_N=(i_N,j_N)] = p_{i_1j_1}\ldots p_{i_Nj_N} = \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{T_{ij}},$$

and the total cost of all the trips is given by:

$$c(d) = c_{i_1 j_1} + \ldots + c_{i_N j_N} = \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} T_{ij}. \tag{5.6}$$

To define cost minimizing behavior for a group we shall consider two trip patterns d^1 and d^2 of the same size N. The total number of trip makers going from i to j in the two trip patterns is given by the matrices $T^1 = [T^1_{ij}]$ and $T^2 = [T^2_{ij}]$.

The two trip patterns (groups) are *activity equivalent* if and only if:

$$\sum_{j=1}^{J} T^1_{ij} = \sum_{j=1}^{J} T^2_{ij},\ i = 1,\ldots,I \quad \sum_{i=1}^{I} T^1_{ij} = \sum_{i=1}^{I} T^2_{ij},\ j = 1,\ldots,J. \tag{5.7}$$

Activity equivalence means that the number of trip makers going to and from the zones $i = 1,\ldots,I$ and $j = 1,\ldots,J$, respectively, is equal for the trip patterns. If not activity equivalent, it makes no sense comparing them with respect to cost. Now, if the trip makers are rational and tend to minimize cost, activity-equivalent trip patterns (groups) with lower total cost would be more probable (frequently observed) than those with higher total cost.

Definition 3: Cost minimizing behavior in trip distribution A probability distribution $p = (p_{ij})$ represents cost minimizing behavior if and only if for any two independent activity-equivalent trip patterns, d^1 and d^2, of the same size N, we have:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} T^1_{ij} \leq \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} T^2_{ij} \Rightarrow \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{T^1_{ij}} \geq \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{T^2_{ij}}. \tag{5.8}$$

It is easy to see by substitution into (5.8) that the gravity model,

$$p_{ij} = \exp\,(\alpha_i + \beta_j - \gamma c_{ij}), \tag{5.9}$$

where $\gamma \geq 0$, satisfies the definition of cost minimizing behavior. (For a discussion of the gravity model, see, for example, Erlander and Stewart 1990.)

Also, it follows again from the representation theorem (Proposition 2) that the probability distribution p represents a cost minimizing probability distribution with respect to the marginal constraints (5.7) and the cost matrix $[c_{ij}]$ if and only if it has the gravity form (5.9). The gravity model is the only probability distribution expressing cost minimizing behavior (according to Definition 3) with respect to the trip costs $[c_{ij}]$ when the marginal trip numbers are identified as defining activity equivalence.

3.3 Discrete Choice Models in General

In discussing the gravity model we introduced the concept of activity equivalence, and in the trip distribution example this was defined by relations (5.7). Also, in treating the multi-attribute case we gave an example where cost minimizing behavior was defined with respect to multiple cost measures. In this section we shall consider the general case where we have several cost functions (cost measures) which influences the choice of the decision maker, and where activity equivalence is defined by a general set of linear equations. Cost minimizing behavior is then defined with respect to all cost functions, where we restrict

the comparison of decision patterns for groups to such patterns which have the same activity levels as defined by these linear equations. The main result is the representation theorem (Proposition 2), which shows that all cost minimizing probability distributions are of exponential (log-linear) type.

Assume, as before, that we have N decision makers choosing independently among K alternatives. Let $d=(d_1,\ldots,d_N)$ be the decision pattern, and let the vector $z=(z_1,\ldots,z_K)$ contain the numbers of decision makers choosing each alternative in this pattern.

Now, assume that we have S cost measures and that activity equivalence is defined with respect to M activity measures. Then, we can introduce a *cost matrix* $\mathrm{C} \in \mathrm{R}^{S\times K}$, where element c_{sk} denotes the value of the cost function (cost measure) s for alternative k. Similarly, we introduce an *activity matrix* $\mathrm{A} \in \mathrm{R}^{M\times K}$, where element a_{mk} indicates the level of activity measure m in alternative k.

The activity matrix A will be used to define M linear constraints on the vector z of the form Az, expressing the total level of the activity measures $1,\ldots,M$ in decision pattern d. We only want to compare decision patterns of the same size N and with identical levels of the activity measures. The cost matrix C will be used to define S cost measures which we are going to compare simultaneously.

Consider two independent decision patterns d^1 and d^2, with corresponding vectors z^1 and z^2, describing the choices of two groups of decision makers of the same size.

Definition 4: Activity equivalence For any given activity matrix $\mathrm{A} \in \mathrm{R}^{M\times K}$ two decision patterns of the same size N are activity equivalent if and only if $\mathrm{Az}^1 = \mathrm{Az}^2$.

Note that the levels of the activity measures are not specified. The definition only requires the levels in the two decision patterns to be equal. However, the values of the parameters to be derived in Proposition 2 below will be determined by specifying the levels of the activity measures.

In the trip distribution example we have $M=I+J$ activity measures, defining the total number of trip makers going from origin i and to destination j, respectively. Thus, each row corresponds to one origin *or* one destination. Element a_{mk} will take the value of one if alternative k (relation (i, j)) concerns origin (destination) m. The general activity equations in Definition 4 are then defined according to (5.7), and the levels of the activity are given by the marginal totals.

We are now ready to define *cost minimizing behavior* in the general case.

Definition 5: Cost minimizing behavior in general For any given activity matrix $\mathrm{A} \in \mathrm{R}^{M\times K}$ and cost matrix $\mathrm{C} \in \mathrm{R}^{S\times K}$, the probability distribution p is defined to be a cost minimizing probability distribution with respect to A and C if and only if for any independent decision patterns d^1 and d^2 of the same size N,

$$[\mathrm{Az}^1 = \mathrm{Az}^2,\ \mathrm{Cz}^1 \leq \mathrm{Cz}^2] \Rightarrow \prod_{k=1}^{K} p_k^{z_k^1} \geq \prod_{k=1}^{K} p_k^{z_k^2}, \tag{5.10}$$

holds for any N.

The definition contains M activity-equivalence constraints and S cost measures. The inequalities $\mathrm{Cz}^1 \leq \mathrm{Cz}^2$ must hold simultaneously for all S cost measures. Hence, a probability

distribution exhibits cost minimizing behavior if, when comparing activity-equivalent decision patterns, smaller values of all the cost measures imply higher probability of the decision pattern to be observed. Nothing is said about other possible decision patterns, or about the probability distribution if not all the cost relations hold.

Assuming that a probability distribution is cost minimizing is equivalent to making the assumption that decision patterns/samples with lower cost in all S cost components are at least as probable as decision patterns/samples with higher costs when activities are equivalent.

Definitions 4 and 5 were given by Smith for the trip distribution problem under the notion of efficiency (Smith 1978). The new element in our formulation is the cost minimizing behavioral interpretation. With this in mind we can say that our definition of cost minimizing behavior in general, Definition 5, is new. It is a natural generalization of the definitions used before with one cost condition (Definition 1), several cost conditions (Definition 2), and one cost condition and several activity-equivalence conditions (Definition 3).

From the assumption that a probability distribution is cost minimizing according to Definition 5 follows the specific form of the probability distribution. This is given in the general representation theorem presented next.

Proposition 2: Representation theorem for cost minimizing probability distributions For any given rational activity matrix $A \in R^{M \times K}$, and cost matrix $C \in R^{S \times K}$, the probability distribution p is a cost minimizing probability distribution with respect to A and C if and only if there exists a non-negative vector, $\gamma \in R^S$, such that for some vector, $\alpha \in R^M$, and some $\mu \in R$,

$$p_k = \exp(\mu + \alpha^T a_k - \gamma^T c_k),\ k = 1,\ldots,K, \tag{5.11}$$

where a_k, and c_k are column k of the matrices A and C, respectively.

Here μ represents a scaling parameter that guarantees that the probabilities sum up to one.

We do not give the proof. It follows from the corresponding theorems for the efficiency principle (Smith 1978; Erlander 1985, Theorem 2, and Erlander and Smith 1990, Theorem 2.1).

All standard discrete choice models of log-linear type can be written in the form of formula (5.11), and can therefore be derived from cost minimizing behavior according to Definition 5 (Erlander 1998).

The result of Proposition 2 is a remarkable fact, since very little is assumed in order to obtain the specific form of the probability distribution. The reverse implication is also true; any probability distribution of the specific form is cost minimizing. Proposition 1 is a special case of Proposition 2 and the probability distributions (5.5) and (5.9) of the multi-attribute discrete choice model and the gravity model are also special cases of this proposition.

The principal strength of the definition of cost minimizing behavior as we have defined it in Definition 5 is that very little has to be assumed about specific details of behavior in order to derive all cost minimizing probability distributions. Many different decision

mechanisms are in principle compatible with the definition. The definition is expressed in observable quantities, and the assumptions can therefore be tested and refuted by means of observations (Erlander 2000). No assumption has to be made about the form of the probability distributions. On the contrary, this form is derived by assuming cost minimizing behavior. Other derivations of the same form of the probability distribution are based on assumptions about, for example, maximizing perceived utility and extreme value distributed (Gumbel distributed) non-observable random elements as part of the perceived utility. These assumptions cannot be directly tested by observations, since they are expressed in terms of non-observable quantities.

NOTE

* This work was supported by the Swedish Transportation and Communication Research Board (KFB), Dnr 1998-0185.

REFERENCES

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

Boyce, D.E. (1984), 'Network models in transportation/land use planning', in M. Florian (ed.), *Transportation Planning Models*, Amsterdam: North-Holland, pp. 475–98.

Boyce, D.E., L.J. LeBlanc and K.S. Chon (1988), 'Network equilibrium models of urban location and travel choices: a retrospective survey', *Journal of Regional Science*, **28**, 159–83.

Boyce, D.E. and L. Lundqvist (1987), 'Network equilibrium models of urban location and travel choices: alternative formulations for the Stockholm region', *Papers of the Regional Science Association*, **61**, 93–104.

Boyce, D.E. and L.-G. Mattsson (1999), 'Modeling residential location choice in relation to housing location and road tolls on congested urban highway networks', *Transportation Research*, **B33**, 581–91.

Erlander, S. (1985), 'On the principle of monotone likelihood and loglinear models', *Mathematical Programming*, **21**, 137–51.

Erlander, S. (1998), 'Efficiency and the logit model', *Annals of Operations Research*, **82**, 203–18.

Erlander, S. (2000), 'A graphical test for utility maximizing behavior', Technical Paper LiTH-MAT-R-2000-11, Department of Mathematics, Linköping University, Linköping, Sweden.

Erlander, S. (2001), 'Benefit measures, freedom of choice and composite utility in the logit model', Technical Paper LiTH-MAT-R-2001-09, Department of Mathematics, Linköping University, Linköping, Sweden.

Erlander, S. and T.E. Smith (1990), 'General representation theorems for efficient population behavior', *Applied Mathematics and Computation*, **36**, 173–217.

Erlander, S. and N.F. Stewart (1990), *The Gravity Model in Transportation Analysis – Theory and Extensions*, Utrecht: VSP.

Fisk, C.S. and D.E. Boyce (1984), 'A modified composite cost measure for probabilistic choice modeling', *Environment and Planning A*, **16**, 241–8.

McFadden, D. (1974), 'Conditional logit analysis of qualitative choice behavior', in P. Zarembka (ed.), *Frontiers of Econometrics*, New York: Academic Press, pp. 105–42.

McFadden, D. (1978), 'Modeling the choice of residential location', in A. Karlqvist, L. Lundqvist, F. Snickars and J.W. Weibull (eds), *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland, pp. 75–96.

McFadden, D. (1981), 'Econometric models of probabilistic choice', in C. Manski and D.

McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, Ma: MIT Press, pp. 198–272.

McFadden, D. (2001), 'Disaggregate behavioral travel demand's RUM side – a 30-year retrospective', in D. Hensher (ed.), *Travel Behavior Research: The Leading Edge*, Amsterdam: Pergamon, pp. 17–64.

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

Smith, T.E. (1978), 'A general efficiency principle of spatial interaction', in A. Karlqvist, L. Lundqvist, F. Snickars and J.W. Weibull (eds), *Spatial Interaction Theory and Planning Models*, Amsterdam: North-Holland, pp. 97–118.

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

6. A modified iterative scheme for the equilibrium traffic signal setting problem

Claudio Meneguzzer

1. INTRODUCTION

Traffic-responsive signal control is considered by traffic engineers a powerful tool for making the most efficient use of existing intersection capacities by adjusting signal timing and phase sequencing so as to meet the time-varying demands of competing traffic streams. It is well known, however, that this form of control has the potential to influence significantly the network flow pattern due to the existence of mutual interactions between user route choices and signal setting actions. The idea that this influence can be used to the planner's advantage in order to achieve some network-wide objective (such as total travel time minimization), rather than being just an undesired side-effect, was first suggested by Allsop (1974) and has motivated the development of a class of models known as *equilibrium traffic signal setting* (ETSS) models.[1] In essence, ETSS models seek to determine a joint equilibrium of link flows and signal settings in a road network operating under traffic-responsive control. As discussed later in this chapter, the computation of such an equilibrium can be approached either from a normative standpoint, that is, treating it as a network design problem, or from a descriptive standpoint, that is, viewing it as the problem of simulating in a realistic fashion the evolution of network conditions induced by the mutual interaction between route choices and control decisions, *even though the eventual equilibrium may not be optimal in terms of network performance.*

The network effects of responsive control, which may materialize over relatively long time periods under manual adjustment of the signal settings, tend to be greatly accelerated and amplified in the presence of actuated control. As the latter has become the prevailing form of signal control in many urban road networks, the adoption of an ETSS framework appears to be highly warranted as a means of enhancing the realism and policy relevance of equilibrium assignment models used in the evaluation of traffic management schemes. However, while signal control may be designed and operated so as to adapt fully and promptly to any traffic-flow variation,[2] in real-world networks it is unrealistic to expect that drivers exhibit a fully reactive behavior in the adjustment of their route choices. For several reasons, such as the presence of indifference thresholds and habit effects, it seems more plausible to assume that only a fraction of the network users actually adjust their route choices following any change of signal settings. This can be expected to introduce a smoothing effect into the process of mutual interaction between traffic assignment and signal control. This basic consideration motivates the work presented in this chapter, whose main purpose is to evaluate the likely impact of the assumption of partial driver response upon key properties of ETSS.

Much emphasis has been placed in recent years on the role of driver information as a determinant of route choice behavior, and, consequently, much effort has been devoted to the development of driver information systems as a tool for alleviating congestion in road networks. This theme is clearly relevant to ETSS research, as the study of the potential benefits that may derive from the integration of route guidance (or, more generally, real-time information to network users) and traffic-responsive signal control is a key issue in the design of an ATIS/ATMS (advanced traffic management systems/advanced traveler information systems) environment (see, for example, Bell 1992; Van Vuren and Van Vliet 1992; Hu and Mahmassani 1997). It is well known that a rather simplified, but effective, way to account for the level of driver information in equilibrium traffic assignment is to allow for random variations in travel time perceptions, and hence in route choice, thus obtaining what is commonly known as stochastic user equilibrium (SUE) assignment (Daganzo and Sheffi 1977). The underlying rationale is that driver perceptions of travel times and other network features tend to be strongly dependent upon the available information. At the aggregate level, stochastic route choice results in 'dispersed' network equilibria, in which not all drivers follow minimum-time routes, and hence system travel time is higher than in the deterministic case, and tends to increase as driver information deteriorates. Based on these considerations, the modeling framework presented in this chapter assumes stochastic route choice, and the effect of the dispersion parameter of the SUE model upon the mutual interaction of signal control and traffic assignment is, in fact, one of the elements of our analysis.

This chapter is organized as follows. In Section 2 we provide the essential background on ETSS, by formally defining the problem and presenting a selective overview of the relevant literature. Section 3 discusses the motivation and the key features of the proposed approach to ETSS, while the details of the modeling framework and solution algorithm are described in Section 4. In Section 5 we present and discuss the results of numerical experiments in which the proposed approach is implemented on a small test network, but using realistic link delay functions as descriptors of intersection performance. Finally, Section 6 offers concluding remarks on the main findings of the study.

2. THE EQUILIBRIUM TRAFFIC SIGNAL SETTING PROBLEM

In order to formally state the ETSS problem, we consider a road network whose intersections operate under traffic-responsive signal control, and assume that control decisions are dictated by some *signal control policy*, that is a criterion used to determine a vector of signal settings (for example, cycle lengths and green time splits) for any given specification of a vector of relevant link flows (assumed known). The signal control policy may take on various specific forms, ranging from simple empirical rules to rigorous optimization procedures. Further, we assume that the travel demand is fixed and known and that, for given values of the signal settings, drivers' route choice behavior is described by the SUE paradigm, as discussed in the introductory remarks of Section 1. This means that, at the individual level, *no driver can reduce his or her perceived travel time by means of a unilateral change of route*, a statement which generalizes the well-known user equilibrium principle of Wardrop (1952) to the case of subjective, random perceptions of travel times.

Let: $\mathbf{f}^{SUE} | \mathbf{g}$ be the vector of SUE link flows given a vector $\mathbf{g}$ of signal settings, and $\mathbf{g}^P|$

f be the vector of signal settings determined through the control policy P given a vector **f** of link flows. Under these assumptions, the ETSS problem consists of finding a pair of vectors $(\mathbf{f}^*, \mathbf{g}^*)$ such that:

$$\mathbf{f}^* = \mathbf{f}^{SUE} \mid \mathbf{g}^* \tag{6.1}$$

$$\mathbf{g}^* = \mathbf{g}^P \mid \mathbf{f}^* \tag{6.2}$$

that is, $\mathbf{f}^*$ is a SUE when signals are set at $\mathbf{g}^*$, and $\mathbf{g}^*$ are precisely the signal settings corresponding to $\mathbf{f}^*$ under the specified control policy P. A pair of vectors satisfying (6.1) and (6.2), if it exists, is called a *mutually consistent flow-control equilibrium*, since under flows $\mathbf{f}^*$ and signal settings $\mathbf{g}^*$ none of the decision makers involved (the drivers and the signal setter) has an incentive to modify his or her current course of action.

Definitions (6.1) and (6.2) suggest a rather natural decomposition of the ETSS problem into two subproblems, namely a *traffic assignment subproblem*, in which SUE link flows are computed for fixed signal settings, and a *signal control subproblem*, in which signal settings are determined under fixed link flows through policy P. Essentially, the two subproblems are 'interfaced' by the delay functions of the signalized network links, which, acting as cost functions in the assignment model, convey the impact of signal settings upon route choice. Therefore, it is hardly surprising that the behavior of ETSS models may be strongly affected by the specific type of function adopted for modeling intersection-related delays. Building upon this concept in a deterministic framework, Smith and Van Vuren (1993) have shown how the ETSS problem can be viewed as a 'natural' extension of the ordinary equilibrium traffic assignment problem, and in pursuing this idea they have introduced the notion of *traffic pressure*, which, in a sense, is a 'cost' perceived by the signal setter when making his or her control decisions. They demonstrated how different signal control policies can be derived from appropriate assumptions about the traffic pressure, and hence that the choice of a specific control policy may influence the properties of the ETSS problem no less than the choice of a specific link delay function.

Since the inception of ETSS modeling in the 1970s, two main solution approaches have emerged. The first is an *iterative scheme* which does not build on any explicit model formulation, and simply seeks to mimic the real-world behavior of the traffic engineer and the network users by alternately updating signal settings for fixed link flows (*control step*) and solving a user equilibrium problem (either deterministic or stochastic) under fixed signal settings (*assignment step*). This approach, originally proposed by Allsop (1974) and later termed *iterative optimization and assignment* (IOA) by Tan et al. (1979), has an essentially heuristic and descriptive nature: there is no a priori guarantee that the mutually consistent flow-control equilibrium calculated by this method is even locally optimal in terms of some measure of network performance, such as, for example, total travel time. This is not surprising in light of the above noted lack of an underlying model formulation.

In fact, it has been argued in several papers (for example, Smith 1979; Dickson 1981) that application of IOA may lead to deterioration of network performance as compared to the initial conditions, unless the signal control policy is purposely designed to take explicit account of changes in driver route choices induced by its implementation. Notably, this is not the case with control policies commonly adopted by traffic engineers, such as equisaturation (Webster 1958) and delay minimization (Allsop 1971). To overcome this problem, Smith and colleagues (see, for example, Smith and Van Vuren 1993)

have proposed a family of alternative control policies which are capable of optimizing some system objective, such as network capacity, when persistently applied within an IOA framework. For instance, the so-called P_0 policy (Smith 1980, 1981) tends to encourage the use of routes having higher physical capacity by assigning them green splits that yield lower delays. This shows that good behavior of the iterative scheme may depend to a large extent upon an appropriate choice of the 'traffic engineering submodel' (the signal control policy). Perhaps, the main practical drawback of this idea is that it may be somewhat difficult for 'unconventional' policies such as P_0 to gain widespread acceptance within standard traffic engineering practice.

The second approach to the solution of ETSS problems consists of formulating a constrained optimization problem in which an upper-level decision maker (the traffic manager) is to choose a vector of signal settings so as to optimize some measure of system performance, under technical constraints on the signal settings plus the requirement that link flows are in user equilibrium. This type of formulation can be easily recognized as an instance of the more general *equilibrium network design* (END) problem.

There are two main difficulties with this approach: first, multiple local optima may exist, owing to the non-linearity of the user equilibrium constraint; second, exact solution algorithms for END problems are known to be computationally impractical for networks of realistic size. Efforts to overcome these difficulties have focused either on the use of suitable approximations to the user equilibrium constraint (for example, Heydecker and Khoo 1990) or on the adoption of alternative solution approaches, such as stochastic optimization (for example, Lee and Hazelton 1996). Promising results obtained by means of *ad hoc* heuristics have also been reported (for example, Chiou 1999). Recently, Maher et al. (2001) have proposed an algorithm for the solution of the bi-level formulation of the ETSS problem under logit-based SUE assignment. Their numerical tests show that the algorithm, though heuristic in nature, tends to exhibit convergent behavior and is able to identify the global optimal solution, at least on small artificial networks.

Further behavioral insights into ETSS may be gained through the concepts of game theory. Fisk (1984) pointed out that the END formulation of the problem has the structure of a *Stackelberg* (or leader–follower) *game*, as it is reasonable to assume that the upper-level player (the signal setter) is able to anticipate the reactions of the lower-level players (the drivers) to his or her decisions, but not vice versa (the drivers are usually unaware of the control strategy adopted by the signal setter). As noted by Friesz and Harker (1985), while the IOA procedure may be regarded only as a heuristic for solving such a hierarchical game, it is indeed an exact solution algorithm for a *Cournot-Nash game*, in which each player selects his or her strategy non-cooperatively and without taking into account the reactions of the other players.

The game-theoretic perspective may also be useful for interpreting bounds to the solution of the END formulation: Harker and Friesz (1984) have shown that the IOA solution provides an upper bound to the exact END solution, while a lower bound may be obtained by solving a fully system-optimal ETSS problem, in which a (rather unlikely) monopolistic traffic manager has control over both signal settings and link flows. The game-theoretic approach has also been extended to deal with the case in which both traffic control and traffic assignment are modeled in a dynamic framework (Chen and Ben-Akiva 1998).

Even though the above discussion does suggest a superiority of the END approach over IOA as far as the ability to achieve an optimal solution to the ETSS problem, it should be

emphasized that the END solution may not be a *global* optimum, and that IOA certainly represents a more appealing tool from the standpoint of both computational tractability and behavioral interpretation. As noted by Watling (1996), the iterative scheme accounts in a realistic manner for the effect that initial conditions (for example, current signal settings determined on the basis of local traffic engineering practice) may have on the evolution of network flow patterns and responsive signal control in subsequent time periods and on the eventual equilibrium, whereas the globally optimal flow-control equilibrium, whose calculation is the ultimate aim of the END approach, may correspond to a routing pattern too far from current route choice behavior, and therefore unlikely to evolve from current conditions.

The possible existence of multiple flow-control equilibria is perhaps the most critical issue arising in the context of ETSS modeling, regardless of the solution approach adopted. In practice, this means that the solution to the ETSS problem may depend on the starting values of the signal settings. While in ordinary equilibrium traffic assignment solution uniqueness is guaranteed a priori under relatively mild restrictions on the link cost functions, analytical conditions which are *sufficient* for uniqueness are violated in ETSS models.[3] Essentially, this happens because the cost–flow relationships need not be monotone due to the reactive nature of signal control, which allows link capacities to change in response to varying demand flows. It is, however, important to realize that, once again, alternative control policies may perform quite differently with respect to solution uniqueness (Van Vuren and Van Vliet 1992).

In practice, the possibility of non-unique equilibrium solutions may have rather disturbing implications when the model is used for policy assessment. This is because multiple equilibria are likely to result in different network flow patterns, and therefore it may be difficult to distinguish such differences from the effects of the options being evaluated. While the potential for multiple-equilibrium behavior of ETSS models has been demonstrated by several authors mainly on the basis of applications to simple networks (for example, Allsop and Charlesworth 1977; Van Vuren and Van Vliet 1992; Nihan et al. 1995), extensive computational experiments carried out by the author (Meneguzzer 1995, 1996) seem to suggest that unique solutions may be more likely to prevail in applications to realistic networks.

3. THE PROPOSED APPROACH

The main goal of this study is to propose and test a modified version of the iterative scheme for the solution of ETSS problems, which explicitly allows for *partial* driver rerouting in response to changes in the signal settings. Following Smith and Van Vuren (1993), we regard IOA as a highly idealized representation of the day-to-day (or, more generally, period-to-period) dynamics of the interaction between signal control and route choice, so that iteration k of the procedure represents period k of this dynamic process. We assume that, at any iteration of the process, only a fraction α of the network users adjust their route choices in reaction to the updated signal settings.

In the standard IOA procedure, it is implicitly assumed that route choices in period k are made on the basis of the signal settings experienced in period $k-1$; hence, the flow updating rule can be formally stated as:

$$\mathbf{f}^k = \mathbf{f}^{UE} \mid \mathbf{g}^{k-1}, \tag{6.3}$$

where $\mathbf{f}^k$ represents the vector of link flows obtained through user equilibrium assignment (either deterministic or stochastic) under signal settings $\mathbf{g}^{k-1}$. On the other hand, under the assumption of partial driver response, the flow updating rule becomes:

$$\mathbf{f}^{\alpha,k} = \alpha \mathbf{f}^k + (1-\alpha)\mathbf{f}^{\alpha,k-1} \quad 0 \leq \alpha \leq 1, \tag{6.4}$$

where $\mathbf{f}^k$ is computed as in (6.3) and $\mathbf{f}^{\alpha,k}$ represents the link-flow vector prevailing in period k under driver response of level α. Parameter α can be interpreted as an aggregate measure of reactiveness, or sensitivity, of the network users to signal control changes; clearly, the standard IOA procedure corresponds to the limiting case of $\alpha = 1$ (full driver response), while in the other limiting case ($\alpha = 0$), a completely rigid driver behavior would essentially preempt the function of responsive signal control and thus, in fact, yield a standard user equilibrium traffic assignment. We observe that expression (6.4) has a recursive character, and that, due to the convexity of the feasible region of the user equilibrium problem, $\mathbf{f}^{\alpha,k}$ is a feasible link-flow vector.

Moreover, if (for any given value of α) the modified iterative scheme converges to a mutually consistent equilibrium, this will also be a solution of the iterative scheme under full driver response. The following proposition, which holds for both deterministic and stochastic user equilibrium assignment, provides a formal statement of this result.

Proposition A mutually consistent flow-control equilibrium under partial driver response of level α ($0 < \alpha \leq 1$) is also a mutually consistent flow-control equilibrium for the standard IOA procedure.

Proof Assume the modified iterative scheme converges to ($\mathbf{f}^{\alpha,*}$, $\mathbf{g}^*$). Then, for sufficiently large k :

$$\mathbf{f}^{\alpha,*} = \mathbf{f}^{\alpha,k} = \mathbf{f}^{\alpha,k-1} \tag{6.5}$$

and

$$\mathbf{g}^* = \mathbf{g}^k = \mathbf{g}^{k-1}. \tag{6.6}$$

Substituting the second equality of (6.5) into (6.4) yields:

$$\mathbf{f}^{\alpha,k-1} = \alpha \mathbf{f}^k + (1-\alpha)\mathbf{f}^{\alpha,k-1}$$

or, equivalently:

$$\mathbf{f}^{\alpha,k-1} = \mathbf{f}^k. \tag{6.7}$$

Thus, by (6.3), (6.5), (6.6) and (6.7) one has:

$$\mathbf{f}^{\alpha,*} = \mathbf{f}^{UE} \mid \mathbf{g}^*. \tag{6.8}$$

Also, since:

$$\mathbf{g}^k = \mathbf{g}^P \mid \mathbf{f}^{\alpha,k}$$

using (6.5) and (6.6) we obtain:

$$\mathbf{g}^* = \mathbf{g}^P \mid \mathbf{f}^{\alpha,*}. \tag{6.9}$$

Equations (6.8) and (6.9) together imply that ($\mathbf{f}^{\alpha,*}$, $\mathbf{g}^*$) is a mutually consistent flow-control equilibrium for the standard IOA procedure. QED

Since the above result holds for any value of α ($0 < \alpha \leq 1$), its main implication is that the mutually consistent solution computed by the modified iterative scheme should turn out to be substantially insensitive to the driver reactiveness parameter. Note, however, that the same does *not* apply to the convergence properties of the iterative scheme, which can be expected to depend considerably on α. In fact, the findings of the numerical experiments reported in Section 5 appear to confirm these expectations.

As stated in the introductory remarks, it does not seem unreasonable to claim that the proposed modification of IOA may lead to a more realistic model of the interaction between signal control and route choice. In real-world networks, a fully reactive driver behavior is unlikely to prevail for several reasons. For example, some drivers may not realize that changes in the signal settings have taken place, perhaps because of differences in perception thresholds, or simply because they are occasional users of the network; others may not deem it worth switching to another route, even though they have become aware of the control changes, perhaps because of habit or inertia in route choice. In addition, there may be situations in which a traffic management agency provides information on the implementation of new signal settings, but such information reaches only a fraction of the network users (perhaps those whose vehicles are equipped with some kind of driver information system). Note that, by introducing parameter α, we are, in fact, separating the above effects from those embodied by the spread parameter of the SUE assignment submodel: the former captures, in a broad sense, *'perception errors' that relate to the changes in signal settings*, while the latter accounts for *'perception errors' that are inherent in route choice behavior under given signal settings* (such as, for example, errors due to the subjective estimation of travel times).

As a conclusion to this discussion, it is important to stress that the proposed approach suffers from the same limitation as the standard version of the iterative scheme, that is it does not *guarantee* that the calculated equilibrium of link flows and signal settings is optimal (not even locally) in terms of overall network performance. Unlike in the END approach to ETSS, here the emphasis is on the *descriptive* rather than on the *normative* side of the problem: we are mainly interested in describing realistically the evolution of network conditions induced by the mutual interaction of route choices and control decisions, rather than striving for the optimality of the eventual equilibrium.

4. MODELING FRAMEWORK AND SOLUTION ALGORITHM

4.1 Traffic Assignment Submodel

As pointed out in the introductory remarks, the SUE approach to traffic assignment provides a proper representation of route choice when both drivers' travel time misperceptions and congestion effects are to be allowed for. Moreover, the dispersion parameter of the probability distribution assumed for perceived travel times, which determines the spread of trips over non-optimal routes, may be regarded as a simple but meaningful proxy for driver information. Since in this study there is a major interest in investigating the effect of information upon the interaction of route choice and responsive signal control, a logit-based SUE model was chosen for the traffic assignment component of our modeling framework.

A formal definition of this model may be given as follows. Let:

- T_{ij} travel demand (trip rate) between origin i and destination j (assumed known);
- P_{ijr} probability of choosing route r between origin i and destination j;
- F_{ijr} flow on route r between origin i and destination j;
- f_l flow on link l;
- C_{ijr} cost of traveling from origin i to destination j along route r;
- c_l cost of using link l;
- δ_{lijr} typical element of the link-route incidence matrix (= 1 if link l is on route r between origin i and destination j, and 0 otherwise);
- β dispersion parameter of the logit route choice model;
- $\mathbf{g}$ vector of signal settings for the network;
- I set of trip origins;
- J set of trip destinations;
- R_{ij} set of routes between origin i and destination j;
- L set of network links;

Then, the following set of equations can be used to define a SUE flow pattern:

$$F_{ijr} = T_{ij} P_{ijr} \quad \forall i \in I, \forall j \in J, \forall r \in R_{ij} \tag{6.10}$$

$$P_{ijr} = \frac{\exp(-\beta C_{ijr})}{\sum_{k \in R_{ij}} \exp(-\beta C_{ijk})} \quad \forall i \in I, \forall j \in J, \forall r \in R_{ij} \tag{6.11}$$

$$C_{ijr} = \sum_{l \in L} \delta_{lijr} c_l(f_l | \mathbf{g}) \quad \forall i \in I, \forall j \in J, \forall r \in R_{ij} \tag{6.12}$$

$$f_l = \sum_{i \in I} \sum_{j \in J} \sum_{r \in R_{ij}} \delta_{lijr} F_{ijr} \quad \forall l \in L. \tag{6.13}$$

Note that equation (6.11) represents stochastic network loading under the logit assumption, and that in equation (6.12) the cost of using link l is evaluated *conditional upon a given vector of signal settings.* This is because, in the traffic assignment subproblem, signal settings are held fixed at their current values, and hence only the *direct* dependence of link costs on link flows is retained. For simplicity, we assume that at signalized intersections

only between-phase interactions among traffic movements are allowed,[4] so that, *within the traffic assignment subproblem*, cost functions are separable in the usual sense; see Meneguzzer (1997) for a discussion of this issue.

Equations (6.10) to (6.13) can be appropriately combined to yield a fixed point formulation of SUE in terms of link flows:

$$f_l = \sum_{i\in I}\sum_{j\in J}\sum_{r\in R_{ij}} \delta_{lijr} T_{ij} \frac{\exp\left[-\beta \sum_{l\in L} \delta_{lijr} c_l(f_l|\mathbf{g})\right]}{\sum_{k\in R_{ij}} \exp\left[-\beta \sum_{l\in L} \delta_{lijk} c_l(f_l|\mathbf{g})\right]} \quad \forall l \in L. \tag{6.14}$$

or, alternatively, in terms of route flows:

$$F_{ijr} = T_{ij} \frac{\exp\left[-\beta \sum_{l\in L} \delta_{lijr} c_l\left(\sum_{l\in I}\sum_{j\in J}\sum_{r\in R_{ij}} \delta_{lijr} F_{ijr}|\mathbf{g}\right)\right]}{\sum_{k\in R_{ij}} \exp\left[-\beta \sum_{l\in L} \delta_{lijk} c_l\left(\sum_{i\in I}\sum_{j\in J}\sum_{r\in R_{ij}} \delta_{lijr} F_{ijr}|\mathbf{g}\right)\right]} \quad \forall i \in I, \forall j \in J, \forall r \in R_{ij}. \tag{6.15}$$

From an algorithmic standpoint, the link-flow pattern that solves (6.14) can be determined through the method of successive averages (MSA); see, for example, Sheffi (1985).

4.2 Signal Control Submodel

The signal control submodel embodies, in a sense, the 'supply side' of ETSS, and its workings are governed by a signal control policy, as defined in Section 2. In order to investigate the effect that this important element may have on the behavior of ETSS, two control policies were considered for testing their interaction with route choice under the proposed approach. The first is the well-known equisaturation policy (Webster 1958), which yields approximately delay-minimizing signal settings by allocating green times to phases so as to equalize the degrees of saturation of critical movements across phases (unless any minimum green time constraint is active). The second is a capacity-maximizing policy due to Smith (1980), known as P_0, which tends to induce an efficient use of network capacity by diverting traffic toward routes that have a higher saturation flow; this is accomplished by assigning these routes green splits that result in lower delays. Even though P_0 does not, in general, minimize total travel time, it does satisfy sufficient conditions for the existence of a mutually consistent flow-control equilibrium, as shown by Smith (1981). In this study, both policies are implemented at the local level (that is, independently for each intersection), as we do not consider signal coordination or network-wide signal setting strategies.

The properties of the two control policies are best illustrated by an example. Consider a simple two-phase signal controlling the intersection of two links ($l = 1, 2$). Assume that cycle length is fixed, and there are no lost times. Let:

s_l	saturation flow of link l;
f_l	flow on link l;
$y_l = f_l/s_l$	flow ratio for link l;

λ_l green time split assigned to link l;
$d_l(f_l, \lambda_l)$ delay on link l as a function of flow and green split on the same link.

Then, according to the equisaturation control policy the green splits are:

$$\lambda_l = y_l / (y_1 + y_2) \quad l = 1, 2 \tag{6.16}$$

whereas under P_0 we compute λ_l ($l = 1, 2$) such that:

$$s_1 d_1(f_1, \lambda_1) = s_2 d_2(f_2, \lambda_2). \tag{6.17}$$

Even though the statement of both policies can be generalized by considering intersections of more than two links and signal plans with more than two phases, the simple case presented here is sufficient to understand intuitively why the two policies exhibit a quite different behavior when implemented in a traffic-responsive control framework. Equisaturation tends to favor more congested links/routes in the allocation of green times, and this, in turn, attracts even more traffic onto those links/routes; P_0, on the other hand, encourages the use of links/routes having a higher saturation flow, thus exploiting more fully the network's physical capacity. In terms of the resulting signal settings, this means that, *ceteris paribus*, equisaturation is more likely to yield extreme (that is, maximal or minimal) values of green splits as compared to P_0.

One of the purposes of our study is to investigate the behavior of both policies under SUE for varying levels of the spread parameter of the perceived travel time distribution. This is of interest because, as intuitively expected, the route choice dispersion inherent in stochastic traffic equilibria may significantly affect the flow-redistributional properties of the control policies under consideration.

4.3 Link Delay Function

Delay functions for signalized links are traffic-engineering-based mathematical relationships expressing the dependence of average vehicular delays on signal settings (typically cycle length and green split) and link flows for given values of the saturation flows. The latter represent, essentially, the physical capacities of the road segments approaching a signal-controlled intersection. A well-known formula due to Webster (1958) is usually regarded as a suitable representation of signalized intersection delay for steady-state, undersaturated conditions. In an equilibrium traffic assignment framework, however, it is often more appropriate to employ delay functions that allow for temporary oversaturation, not only because they provide a more effective model of peak-hour intersection operation, but also because traffic flows well in excess of capacity may prevail on certain links at some stage of the equilibration process, even though those links may turn out to operate below capacity *at equilibrium*. See Hurdle (1984) for a general discussion of this topic, and Meneguzzer (1997) for related remarks in the context of ETSS modeling.

In keeping with the above observation, a function due to Akçelik (1988) is employed in this study to model delay for signalized links. Like other so-called 'sheared' delay formulae, the Akçelik function covers oversaturated conditions through a time-dependent overflow delay term, thus overcoming the major limitation of Webster-like formulae. The

Akçelik function takes one of the following expressions, depending on the value of the degree of saturation x:

$$\text{for } x \leq 0.5: \quad d = \frac{0.5C(1-\lambda)^2}{1-\lambda x} \tag{6.18}$$

$$\text{for } x > 0.5: \quad d = \frac{0.5C(1-\lambda)^2}{1-\lambda \cdot \min(x,1)} + 900T\left\{x - 1 + \sqrt{\left[(x-1)^2 + \frac{8(x-0.5)}{KT}\right]}\right\}, \tag{6.19}$$

where:

d average delay incurred by vehicles on subject link (sec);
C length of signal cycle facing subject link (sec);
λ green time split facing subject link;
T duration of demand flow period on subject link (hrs);
K capacity of subject link (vehicles/hr);
x degree of saturation (flow-to-capacity ratio) of subject link.

Note that for flows well below capacity (equation (6.18)) the formula consists of a single component, termed *uniform delay* since it is derived under the assumption of a uniform vehicle arrival pattern, whereas for higher values of x (equation (6.19)) an *incremental delay* term is added to account for short-term overflows caused by random vehicle arrivals. This overflow term is made explicitly time-dependent through T, a parameter which essentially controls the slope of the function in the region of volume-to-capacity ratios greater than unity, so as to reflect the duration of the peak period of the demand flow.

4.4 Modified Iterative Optimization and Assignment Algorithm

The proposed modified version of the iterative scheme is obtained by embedding the MSA, employed for the solution of the logit-based SUE subproblem, into the IOA procedure, and introducing an additional step, devoted to the updating of link flows under the assumption of partial driver response (see equation (6.4)). The resulting algorithm consists of the following steps:

STEP 0: Initialization
$\mathbf{g} = \mathbf{g}^0$ (initial signal settings)
$k = 0$ (iteration counter).

STEP 1: Traffic assignment subproblem
$k = k + 1$
Compute link flows $\mathbf{f}^k$ solving SUE by MSA under signal settings $\mathbf{g}^{k-1}$.

STEP 2: Link flow updating
If $k = 1$ $\quad \mathbf{f}^{\alpha,k} = \mathbf{f}^k$
Else $\quad \mathbf{f}^{\alpha,k} = \alpha \mathbf{f}^k + (1-\alpha)\mathbf{f}^{\alpha,k-1} \quad 0 \leq \alpha \leq 1.$

STEP 3: **Signal control subproblem**
Compute signal settings $\mathbf{g}^k$ via control policy P under link flows $\mathbf{f}^{\alpha,k}$.

STEP 4: **Stopping rule**
If $k=1$ go to Step 1
Else
If $\delta(k-1, k) \leq \varepsilon$ Stop and set $\mathbf{f}^* = \mathbf{f}^{\alpha,k}$, $\mathbf{g}^* = \mathbf{g}^k$
Else go to Step 1,

where $\delta(k-1, k)$ is a measure of distance between two successive solutions (expressed in terms of either link flows or signal settings), ε is a prespecified tolerance, and $(\mathbf{f}^*, \mathbf{g}^*)$ represents a mutually consistent flow-control equilibrium which solves the ETSS problem. Note that the algorithm has a nested structure: assuming we run $k=1, ..., K$ iterations of IOA (also called outer iterations), and solve each SUE subproblem by means of N_k iterations of MSA (also called inner iterations), each consisting of a stochastic network loading, the overall computational effort will amount to $L=\Sigma_k N_k$ loadings.

In the implementation of the solution algorithm, the stopping criterion for IOA is based on the maximum relative change of link flows:

$$\max_l \frac{|f_l^{\alpha,k} - f_l^{\alpha,k+1}|}{f_l^{\alpha,k}} \leq \varepsilon \quad \forall l: f_l^{a,k} \neq 0, \tag{6.20}$$

where the maximum is taken over all links. A similar form is adopted for the convergence test of each SUE subproblem:

$$\max_l \frac{|f_l^n - f_l^{n+1}|}{f_l^n} \leq \sigma \quad \forall l: f_l^n \neq 0, \tag{6.21}$$

where index n represents the counter of the MSA iterations. Note that different orders of magnitude were selected for the tolerances of the two tests ($\varepsilon=0.001$, $\sigma=0.01$), so as to ensure a certain degree of 'streamlining' of the algorithm, which was found to enhance computational efficiency by previous studies (for example, Meneguzzer 1996).

5. NUMERICAL EXPERIMENTS

The proposed approach to ETSS is tested on the small network shown in Figure 6.1, which consists of nine nodes, 12 links and four origin–destination pairs (nodes O_1 and O_2 act as origins, while D_1 and D_2 act as destinations). The network includes three intersections, denoted by S_1, S_2, and S_3: in each, two approaching links are controlled by a traffic signal operating on a two-phase plan. For simplicity, it is assumed that cycle lengths are fixed and equal to 90 seconds for all junctions, and lost times are not considered. Thus, for each intersection, the signal control subproblem consists of determining the green time splits for the two approaching links, according to either the equisaturation policy (equation (6.16)) or Smith's P_0 policy (equation (6.17)). In order to keep the test case realistic from a traffic engineering standpoint, a minimum value of 0.1 is assumed for all green splits. Also, in Step 0 of the modified iterative scheme (see Section 4.4) all green splits are *initially* assigned a value of 0.5.

The test network data are displayed in Table 6.1. For signal-controlled links, travel time is taken to be the sum of a free-flow component and a delay term calculated according to the Akçelik formula (equations (6.18) and (6.19)) with $T = 0.25$ hours; for such links, capacity is endogenous and may be obtained as the product of saturation flow, shown in the third column of the table, and green time split, iteratively updated during the execution of IOA. On the other hand, travel time along unsignalized links is modeled according to the classical BPR function (Bureau of Public Roads 1964), and using the fixed capacities shown in column four. Table 6.2 shows, for each of the 10 origin-destination routes, the list of the constituent links and the sequence of traffic signals encountered by drivers using that route; note that each route traverses at least one signalized intersection.

The origin–destination trip rates for the test network are shown in Table 6.3. Since the effect of the congestion level on the mutual interaction of signal control and traffic assignment is one of the subjects of the subsequent investigations, two travel demand scenarios, denoted as 'low' and 'high', are specified for each O–D pair; note that the high values are obtained by multiplying the low ones by a scaling factor of 1.5.

The aim of the numerical experiments is to investigate the joint effect of four different factors upon the convergence behavior of the modified iterative scheme and the ensuing consistent flow-control equilibria. These factors are:

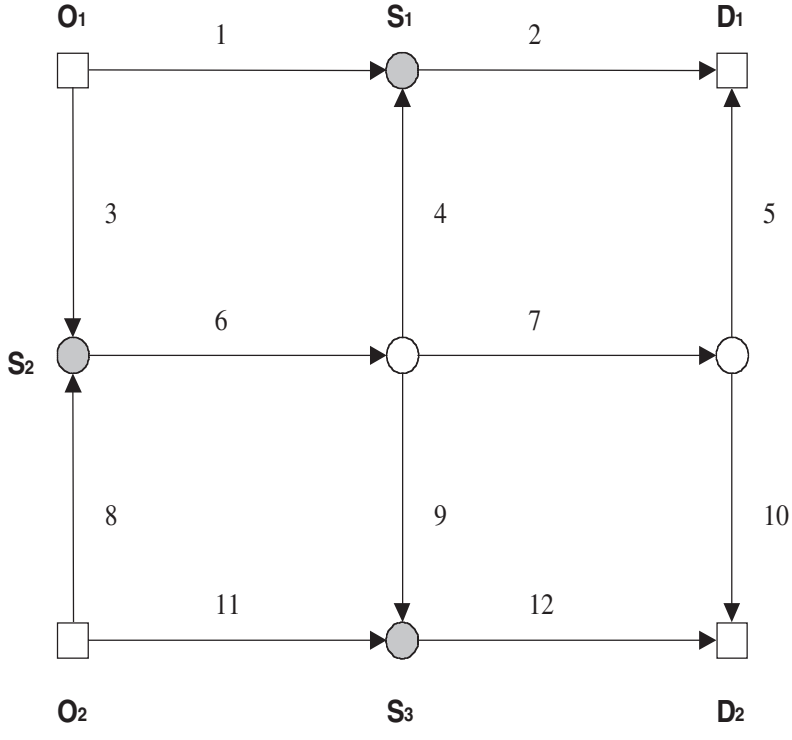


Figure 6.1 Test network

Table 6.1 Test network data

Link	FFTT (min)	SAT (vphg)	CAP (vph)
1	10	2000	–
2	6	–	2000
3	4	3000	–
4	4	2000	–
5	4	–	1500
6	6	–	3000
7	6	–	2500
8	4	3000	–
9	4	2000	–
10	4	–	1500
11	10	2000	–
12	6	–	2000

Note: FFTT: Free-flow travel time; SAT: Saturation flow; CAP: Capacity.

Table 6.2 Relation of traffic signals to test network routes

Route	Link sequence	Signals
1	1 – 2	S_1
2	3 – 6 – 4 – 2	$S_2 - S_1$
3	3 – 6 – 7 – 5	S_2
4	3 – 6 – 7 – 10	S_2
5	3 – 6 – 9 – 12	$S_2 - S_3$
6	8 – 6 – 4 – 2	$S_2 - S_1$
7	8 – 6 – 7 – 5	S_2
8	8 – 6 – 7 – 10	S_2
9	8 – 6 – 9 – 12	$S_2 - S_3$
10	11 – 12	S_3

Table 6.3 Origin–destination trip rates for test network under low and high demand (vph)

		D_1	D_2
O_1	Low	1600	1200
	High	2400	1800
O_2	Low	1360	1040
	High	2040	1560

1. α, the driver reactiveness parameter (see equation (6.4));
2. β, the dispersion parameter of the logit route choice model;
3. the signal control policy (equisaturation or P_0); and
4. the travel demand level (low or high).

The results of the numerical experiments are presented in Tables 6.4 to 6.7. Each of these tables refers to a specific combination of signal control policy and travel demand level, and shows, for pairs of selected values of α and β, the convergence and equilibrium properties of ETSS as described by the number of IOA (or outer) iterations needed for convergence, the *average* number of MSA (or inner) iterations per outer iteration, the mean system travel time (total travel time divided by total number of trips) at the resulting flow-control equilibrium, and the equilibrium values of the green splits for links approaching signalized intersections. In light of the previous discussion regarding the relationship between driver information and the dispersion parameter of the perceived travel time distribution (see Section 1), the values of β displayed in the tables are intended to represent situations of poor ($\beta = 0.15$), average ($\beta = 0.3$), and accurate ($\beta = 1.5$) information.

It should be pointed out that, since each SUE assignment subproblem is solved to convergence, the number of MSA iterations needed to achieve such convergence may vary as the IOA scheme progresses; therefore, it seems appropriate to use the *average* number of inner iterations as a representative measure of the speed of convergence of the embedded MSA procedure. Of course, the overall computational effort required to solve the ETSS problem may be meaningfully expressed in terms of the total number of stochastic network loadings performed, and this number is easily obtained as the product of the number of outer iterations and the average number of inner iterations; see Table 6.8, below, and related comments. It should also be noted that only three green time splits, namely those of links 1, 3 and 9, appear in the tables; this is because, under our assumptions, the values of the splits of the two links approaching any given intersection must add up to one, and thus a single value suffices to describe the signal control plan of each intersection.

5.1 Convergence of the Modified IOA Scheme

The speed of convergence of the modified IOA procedure as a function of α and β is illustrated by the histograms of Figures 6.2 to 6.5, each corresponding to a different policy-demand scenario. Under low demand, an increasing relationship between the number of IOA iterations and the value of β seems to prevail, suggesting that the dispersion of trips over alternative routes, induced by driver misperceptions, tends to accelerate the redistributional effect of traffic-responsive signal control. Thus, when route choices are dominated by stochastic effects, a smaller number of signal adjustments is needed for the network to settle into a consistent flow-control equilibrium. This pattern appears to be rather more pronounced under equisaturation than under P_0, a result which can be explained in light of the peculiar properties of the two policies: the former encourages the use of more congested routes, and thus yields more 'concentrated' flow patterns, that are prone to flip-flopping of trips among alternative routes, and exhibit slower convergence; the latter, on the other hand, is designed to exploit as fully as possible the network's

Table 6.4 Results of the numerical experiments (P_0 control policy and low demand)

		$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 1.00$
$\beta = 0.15$	IOA	2	3	4	4	3
	MSA	5	5	5	5	5
	MTT	22.305	22.305	22.306	22.307	22.307
	G(1)	0.506	0.506	0.506	0.505	0.505
	G(3)	0.508	0.508	0.509	0.509	0.509
	G(9)	0.527	0.527	0.527	0.527	0.527
$\beta = 0.3$	IOA	37	18	12	9	7
	MSA	4	4	4	4	4
	MTT	20.881	20.829	20.822	20.821	20.820
	G(1)	0.589	0.596	0.598	0.598	0.598
	G(3)	0.502	0.499	0.499	0.499	0.499
	G(9)	0.486	0.486	0.486	0.487	0.487
$\beta = 1.5$	IOA	84	24	13	9	6
	MSA	4.7	5	5.2	5.4	5.8
	MTT	19.214	19.207	19.206	19.206	19.206
	G(1)	0.784	0.785	0.785	0.786	0.786
	G(3)	0.481	0.480	0.480	0.480	0.480
	G(9)	0.398	0.397	0.397	0.397	0.397

Note: IOA: No. of IOA iterations; MSA: Average no. of MSA iterations per IOA iteration; MTT: Mean travel time (min); G(*i*): Green time split for link *i* ($i = 1, 3, 9$).

physical capacity by favoring flow redistribution, a feature which tends to induce faster convergence. Both policies, however, perform very similarly in the case of poor information ($\beta = 0.15$), where the random component of route choice is so pervasive that it effectively obscures any differences between the two types of supply action. A rather different behavior can be observed in the high-demand cases (Figures 6.4 and 6.5), in which convergence of IOA turns out to be slowest at the intermediate value of the dispersion parameter ($\beta = 0.3$), and the differences between the two control policies are generally less marked as compared to the scenario of low congestion.

Turning, then, to the effect of the driver reactiveness parameter, it is easy to identify a clear decreasing relationship between the number of IOA iterations at convergence and α. This confirms the intuitive expectation that the number of signal adjustments necessary to reach a consistent flow-control equilibrium tends to become smaller as the fraction of drivers reacting to signal re-setting increases. It should also be noted that the largest reduction is achieved for very low values of the parameter (from $\alpha = 0.05$ to $\alpha = 0.25$), suggesting that the greatest gains in terms of convergence speed are to be expected when the starting level of driver sensitivity to signal control changes is limited. The only exceptions to the above trend are the cases of poor information ($\beta = 0.15$) under low demand, where, again, the prevalence of stochastic effects in route choice prevents the emergence of any clear relationship.

Table 6.5 Results of the numerical experiments (equisaturation control policy and low demand)

		$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 1.00$
$\beta = 0.15$	IOA	2	5	4	4	3
	MSA	5	5	5	5	5
	MTT	22.304	22.309	22.311	22.313	22.313
	G(1)	0.507	0.507	0.507	0.507	0.507
	G(3)	0.509	0.509	0.509	0.509	0.509
	G(9)	0.559	0.560	0.561	0.561	0.561
$\beta = 0.3$	IOA	50	22	14	10	8
	MSA	4	4	4	4	4
	MTT	20.806	20.755	20.749	20.748	20.747
	G(1)	0.616	0.625	0.627	0.627	0.627
	G(3)	0.500	0.497	0.497	0.497	0.496
	G(9)	0.471	0.472	0.473	0.473	0.473
$\beta = 1.5$	IOA	177	60	35	25	19
	MSA	4.8	5.3	5.5	5.6	5.8
	MTT	19.039	19.033	19.032	19.032	19.032
	G(1)	0.891	0.897	0.898	0.898	0.898
	G(3)	0.469	0.469	0.469	0.469	0.469
	G(9)	0.283	0.284	0.285	0.285	0.285

Note: IOA: No. of IOA iterations; MSA: Average no. of MSA iterations per IOA iteration; MTT: Mean travel time (min); G(i): Green time split for link i ($i = 1, 3, 9$).

5.2 Convergence of the MSA

The convergence behavior of the SUE assignment submodel can be evaluated on the basis of the average number of MSA iterations per IOA iteration, which is shown in Tables 6.4 to 6.7 for the various test cases. A rather limited number of such iterations (four to six in most cases) is needed to achieve convergence, a result which follows from the choice of the value of the tolerance in the stopping criterion, intentionally aimed at streamlining the overall solution procedure; see equation (6.21), above, and related comments. Exceptions to this behavior are observed only in the cases of accurate information ($\beta = 1.5$) and high demand, where the average number of MSA iterations at convergence is seen to increase to 9–10, a result which appears to be consistent with the general convergence properties of MSA for logit-based SUE models (Sheffi 1985, pp. 330–31).

5.3 Network Performance at Consistent Flow-control Equilibria

The equilibrium performance of ETSS attained through the proposed version of the iterative scheme is measured in terms of mean system travel time at convergence (MTT), whose values are also displayed in Tables 6.4 to 6.7, and illustrated in Figure 6.6 for the intermediate case of $\alpha = 0.5$. First, we observe that, *ceteris paribus*, MTT is always a

Table 6.6 Results of the numerical experiments (P_0 control policy and high demand)

		$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 1.00$
$\beta = 0.15$	IOA	58	23	14	10	8
	MSA	5	5	5	5	5
	MTT	28.886	28.750	28.735	28.731	28.728
	G(1)	0.644	0.653	0.654	0.654	0.654
	G(3)	0.503	0.501	0.501	0.500	0.500
	G(9)	0.450	0.450	0.450	0.450	0.450
$\beta = 0.3$	IOA	106	35	20	14	11
	MSA	6	6	6	5.9	5.9
	MTT	26.089	26.015	26.007	26.004	26.003
	G(1)	0.761	0.767	0.768	0.768	0.769
	G(3)	0.489	0.488	0.488	0.488	0.488
	G(9)	0.377	0.377	0.377	0.377	0.377
$\beta = 1.5$	IOA	126	31	16	11	8
	MSA	9.3	9.6	9.8	10.1	10.3
	MTT	23.974	23.970	23.969	23.969	23.969
	G(1)	0.900	0.900	0.900	0.900	0.900
	G(3)	0.471	0.470	0.470	0.470	0.470
	G(9)	0.301	0.301	0.301	0.301	0.301

Note: IOA: No. of IOA iterations; MSA: Average no. of MSA iterations per IOA iteration; MTT: Mean travel time (min); G(*i*): Green time split for link *i* (*i* = 1, 3, 9).

decreasing function of β, a result which is fully consistent with the expected network inefficiency normally ascribed to stochastic route choice behavior.

Second, the effect of driver reactiveness upon equilibrium system performance appears to be practically negligible, as the values of MTT achieved for the various values of α tend to be extremely close to one another, perhaps with a few modest exceptions (corresponding to $\alpha = 0.05$ and low to medium values of β). Together with the results reported in Section 5.1, this finding seems to support the important conclusion that the fraction of drivers responding to signal re-setting may affect significantly the dynamics of the interaction between signal control and route choice, *but not the resulting equilibrium*. This conclusion, which is in agreement with the discussion of Section 3, appears to be corroborated by a detailed examination of equilibrium link flows (not shown here), from which a corresponding similarity across the values of α emerges.

Another factor that seems to have only a minor influence on the equilibrium performance of the model is the signal control policy. As suggested by the representative case of $\alpha = 0.5$, depicted in Figure 6.6, and confirmed by the results obtained in the remaining cases, the values of MTT achieved under P_0 and equisaturation are remarkably similar; in the few cases where noticeable differences exist, these are systematically in favor of equisaturation. This finding is somewhat at odds with the results of a study conducted by Smith et al. (1987), who concluded that equisaturation outperformed P_0 under low demand, but the reverse was true at high congestion levels. Note, however, that their model

Table 6.7 Results of the numerical experiments (equisaturation control policy and high demand)

		$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$	$\alpha = 1.00$
$\beta = 0.15$	IOA	63	25	15	11	8
	MSA	5	5	5	5	5
	MTT	28.797	28.662	28.648	28.644	28.643
	G(1)	0.656	0.665	0.666	0.666	0.666
	G(3)	0.502	0.500	0.500	0.500	0.500
	G(9)	0.444	0.445	0.445	0.445	0.445
$\beta = 0.3$	IOA	117	38	22	16	12
	MSA	6	6	6	5.9	5.9
	MIT	25.931	25.861	25.853	25.850	25.850
	G(1)	0.783	0.790	0.791	0.791	0.791
	G(3)	0.487	0.486	0.485	0.485	0.485
	G(9)	0.363	0.364	0.364	0.364	0.364
$\beta = 1.5$	IOA	108	27	14	9	6
	MSA	8.6	8.7	8.7	9.2	9.8
	MTT	23.956	23.949	23.948	23.948	23.948
	G(1)	0.900	0.900	0.900	0.900	0.900
	G(3)	0.469	0.469	0.469	0.469	0.469
	G(9)	0.274	0.273	0.273	0.273	0.273

Note: IOA: No. of IOA iterations; MSA: Average no. of MSA iterations per IOA iteration; MTT: Mean travel time (min); G(*i*): Green time split for link *i* (*i* = 1, 3, 9).

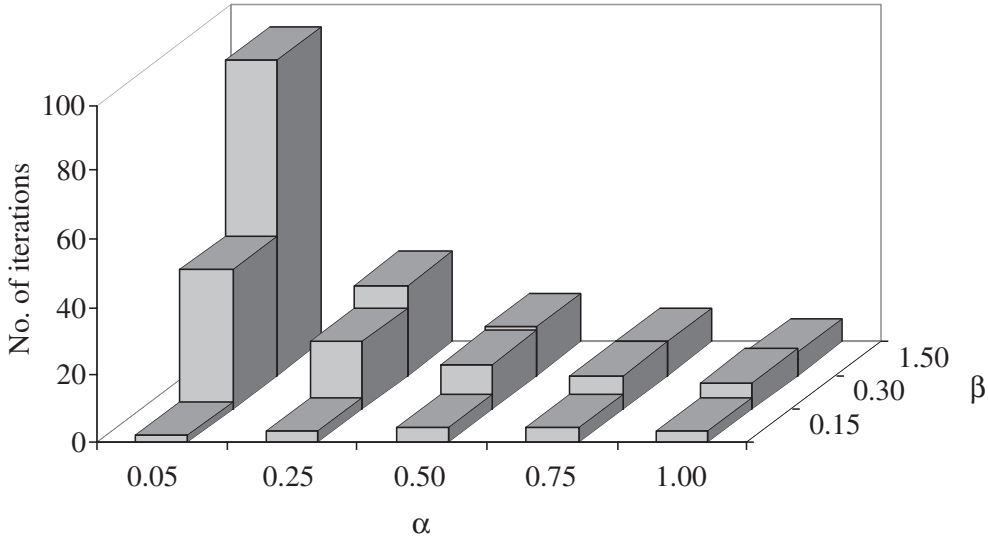


Figure 6.2 Number of IOA iterations at convergence for various values of α and β (P_0 control policy and low demand)

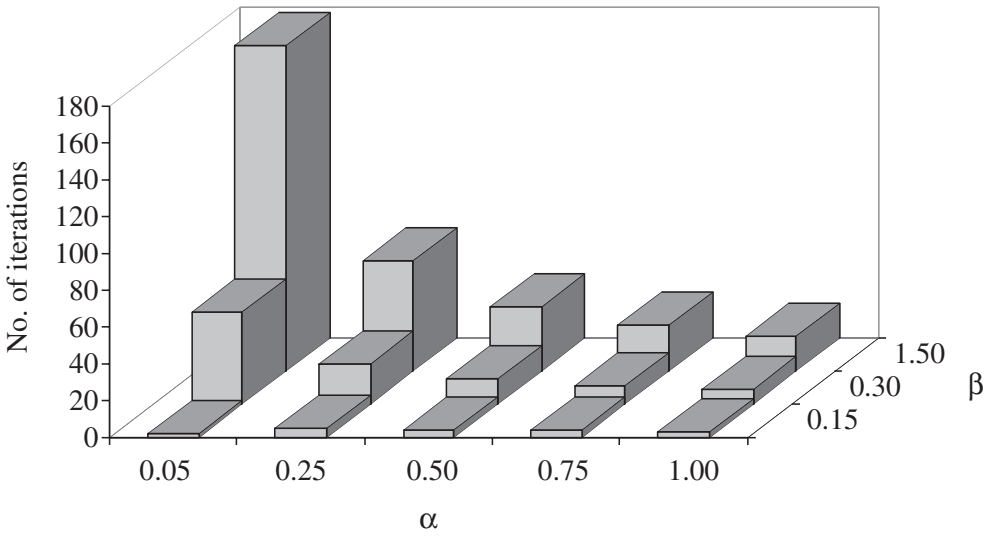


Figure 6.3 Number of IOA iterations at convergence for various values of α and β (equisaturation control policy and low demand)

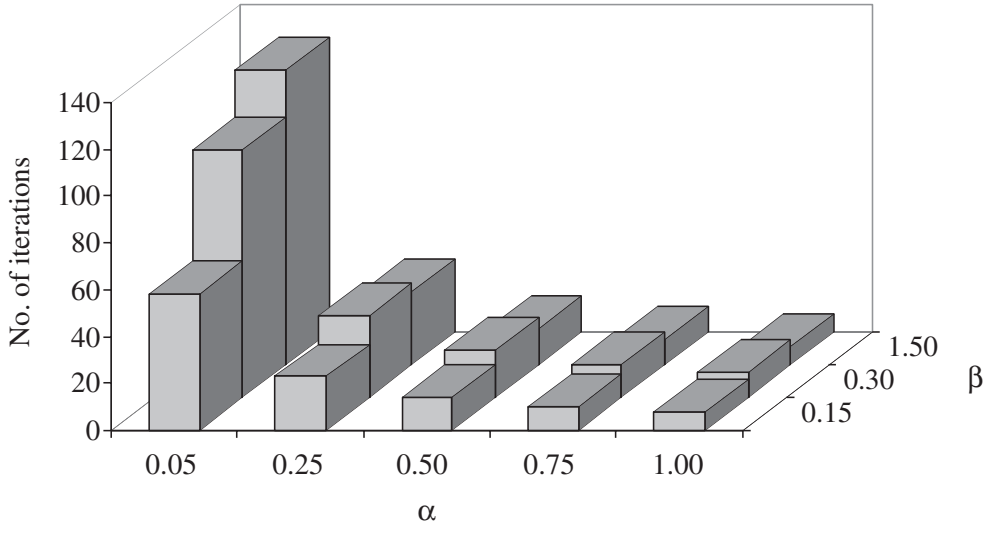


Figure 6.4 Number of IOA iterations at convergence for various values of α and β (P_0 control policy and high demand)

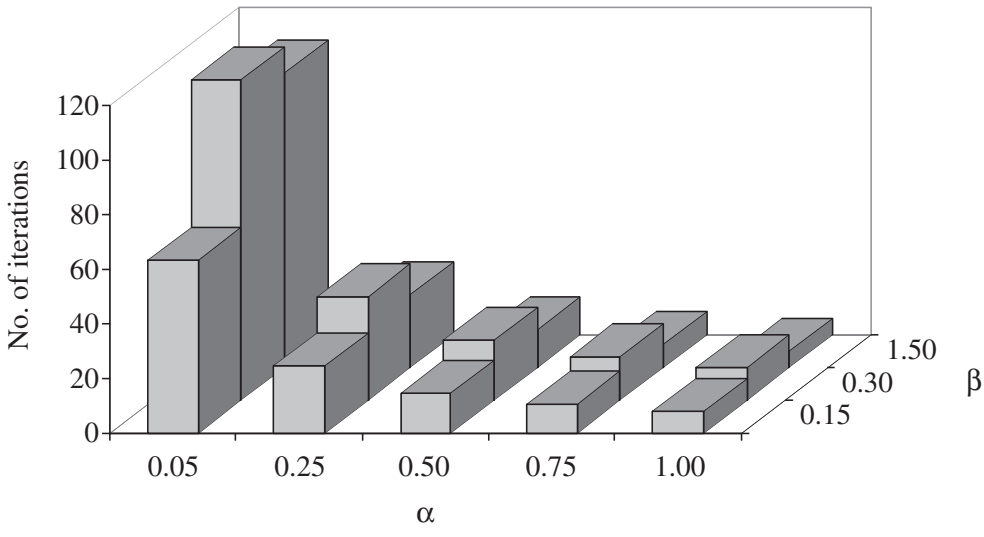
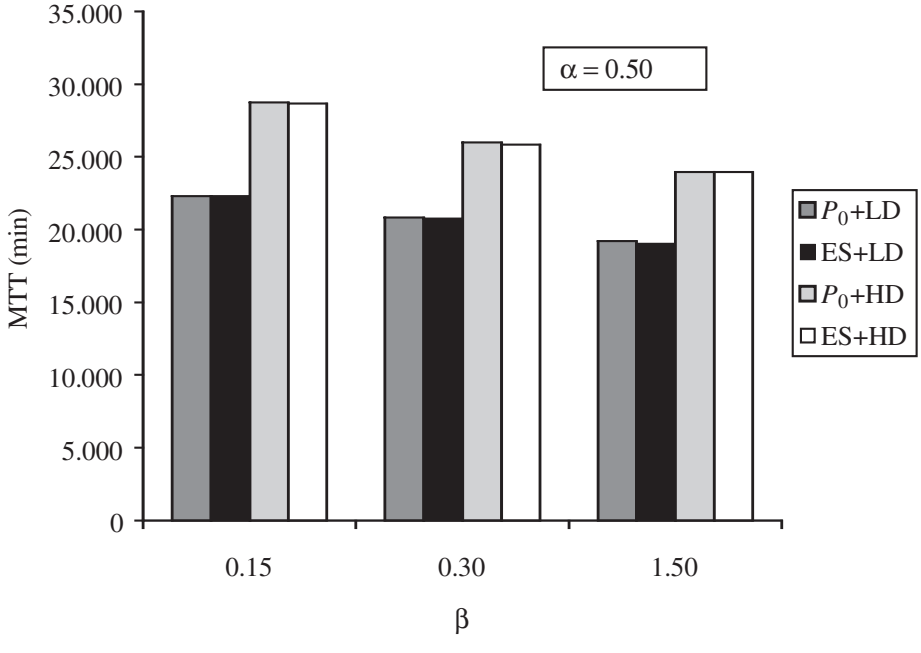


Figure 6.5 Number of IOA iterations at convergence for various values of α and β (equisaturation control policy and high demand)



Note: ES: Equisaturation; LD: Low demand; HD: High demand.

Figure 6.6 Mean system travel time at equilibrium versus β for α =0.5 and various policy-demand scenarios

assumed *deterministic* user equilibrium assignment, and thus their findings are not directly comparable to ours, inasmuch as the flow-redistributional capabilities of the two control policies may be significantly affected by the route choice paradigm. Intuitively, one would expect that the trip-spreading effect of stochastic route choice should counteract the tendency of equisaturation to yield concentrated routing patterns, which is precisely the reason for P_0's superiority under severe congestion in the deterministic case.

Last, the direct impact of the intensity of demand upon equilibrium network performance is, as expected, significant under both control policies, even though the differences in MTT between the low and high congestion scenarios tend to become proportionally smaller as β increases.

5.4 Equilibrium Green Time Splits

The operation of the three signalized intersections under mutually consistent link flows and control settings can be described by the values of the respective green time splits at equilibrium. Tables 6.4 to 6.7 present such values for links number 1, 3 and 9 under the headings G(1), G(3), G(9). As explained in the introductory remarks of Section 5, in our test network the signal control plan of each intersection can be adequately characterized by a single green split. As was the case with mean system travel time, a substantial rigidity of green splits with respect to parameter α emerges from the results, thus reinforcing our previous conclusion that, *ceteris paribus*, variations in the level of driver reactiveness are unlikely to result in significantly different flow-control equilibria.

The results of the numerical experiments are less conclusive regarding the effects of the remaining factors under consideration upon equilibrium green time splits. The reason may be that signal settings, unlike aggregate descriptors of network performance such as mean travel time, are local, intersection-specific variables, and therefore it is more difficult to relate them to network-wide conditions like the route choice dispersion and congestion levels. With this cautionary premise in mind, the following observations appear to be reasonable interpretations of the results.

An increasing level of route choice dispersion tends to reduce the differences between the equilibrium green splits obtained under P_0 and equisaturation, and to push the signal settings toward an even allocation of green time (0.5 to each approaching link), especially in the low-demand scenario. This seems to suggest that the impact of adopting alternative control strategies on equilibrium signal settings may be negligible when travel time misperceptions dominate route choice behavior. This result is intuitively plausible, as one would expect that in the limiting case of purely random route choice the nature of the supply action should become immaterial in determining the equilibrium outcome of the combined traffic assignment and signal setting process.

In the case of accurate driver information ($\beta = 1.5$) and low demand, the differences between the signal settings produced by the two control policies become significant, with more unbalanced green splits corresponding to equisaturation. This is consistent with the observation that P_0 is designed to encourage route flow redistribution even under light congestion, while more concentrated flow patterns tend to prevail with equisaturation; see Section 5.1. Increasing demand, on the other hand, seems to bring about more uneven green splits with both policies, indicating that the flow-redistributional capability of P_0 may be substantially impaired under severe congestion.

5.5 Patterns of Convergence of the Modified IOA Scheme

In order to gain further insights into the behavior of ETSS under partial driver response, it is useful to consider typical convergence patterns of the modified iterative scheme. Figures 6.7 and 6.9 illustrate such convergence in terms of δ, the left-hand side of equation (6.20), in two representative cases and for three levels of driver reactiveness, while Figures 6.8 and 6.10 apply to the same cases, but display MTT as a descriptor of convergence. First, we observe that all patterns are monotone, indicating good convergence behavior of the modified iterative scheme on the test network under consideration. Similar, monotonically decreasing convergence patterns were obtained in the remaining test cases, that are not shown due to space limitations.

Second, the level of driver reactiveness appears to have a rather substantial impact on convergence: consistent with intuitive expectation, a decrease in the fraction of drivers

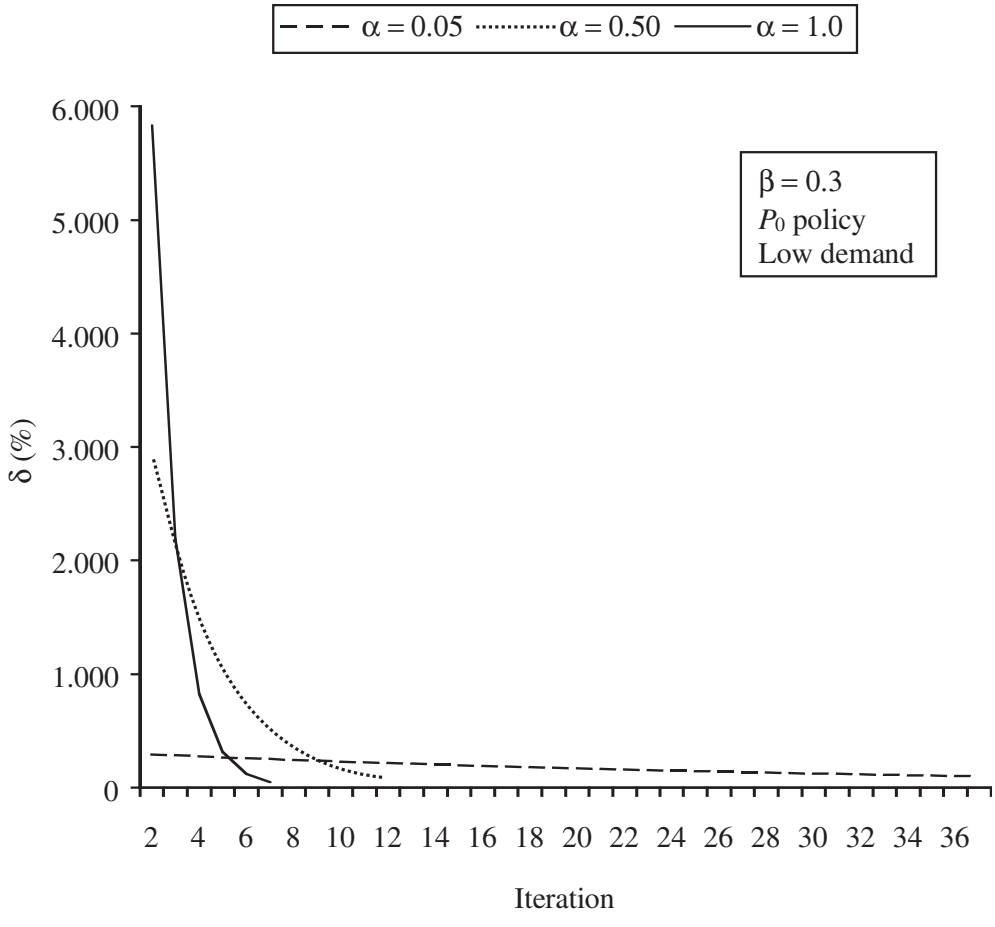


Figure 6.7 Convergence of modified IOA in terms of δ for selected values of α (β = 0.3, P_0 control policy and low demand)

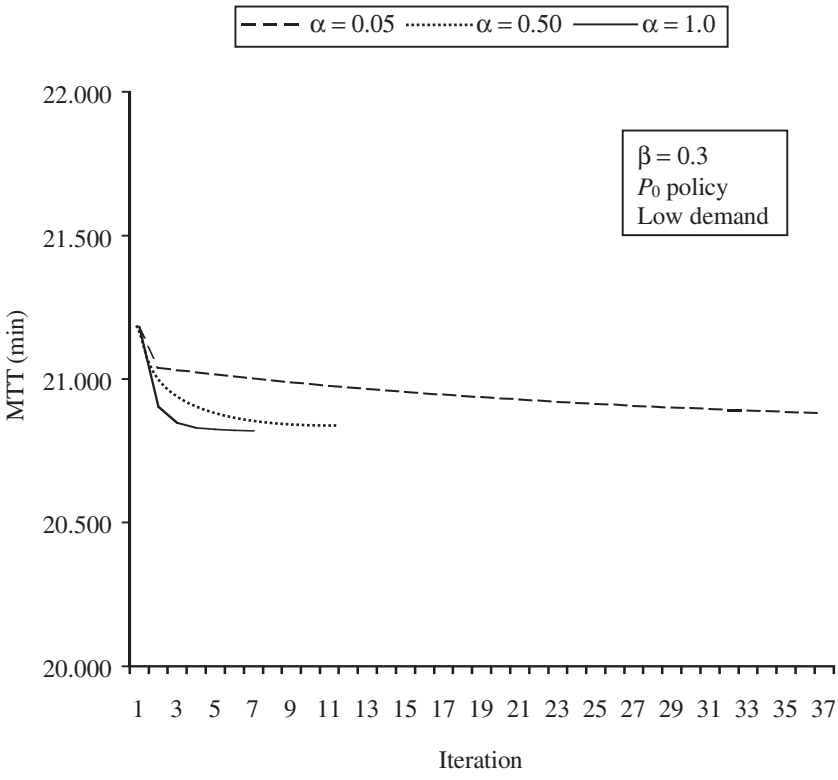


Figure 6.8 Convergence of modified IOA in terms of MTT for selected values of α (β = 0.3, P_0 control policy and low demand)

responding to signal re-setting tends to delay convergence, by producing smoother link-flow variations and slower improvement of network performance. Less obvious, however, is the result that the sensitivity of the convergence behavior to driver reactiveness is much higher in the lower half of the range of α values (that is, while the shape of the curves varies dramatically from $\alpha = 0.05$ to $\alpha = 0.5$, differences are less conspicuous between $\alpha = 0.5$ and $\alpha = 1.0$).

Last, a comparison between the patterns obtained under P_0 and equisaturation seems to suggest that, while the control policy may determine the number of iterations needed for convergence, the shape of the curves describing convergence behavior is essentially unaffected by the choice of policy.

5.6 Computational Requirements

Since the iterative scheme for the solution of the ETSS problem has a nested structure which involves the execution of outer (IOA) and inner (MSA) iterations, it is interesting to examine how the overall computational effort may vary depending on the factors analysed in the numerical experiments. As explained in the introductory remarks of Section

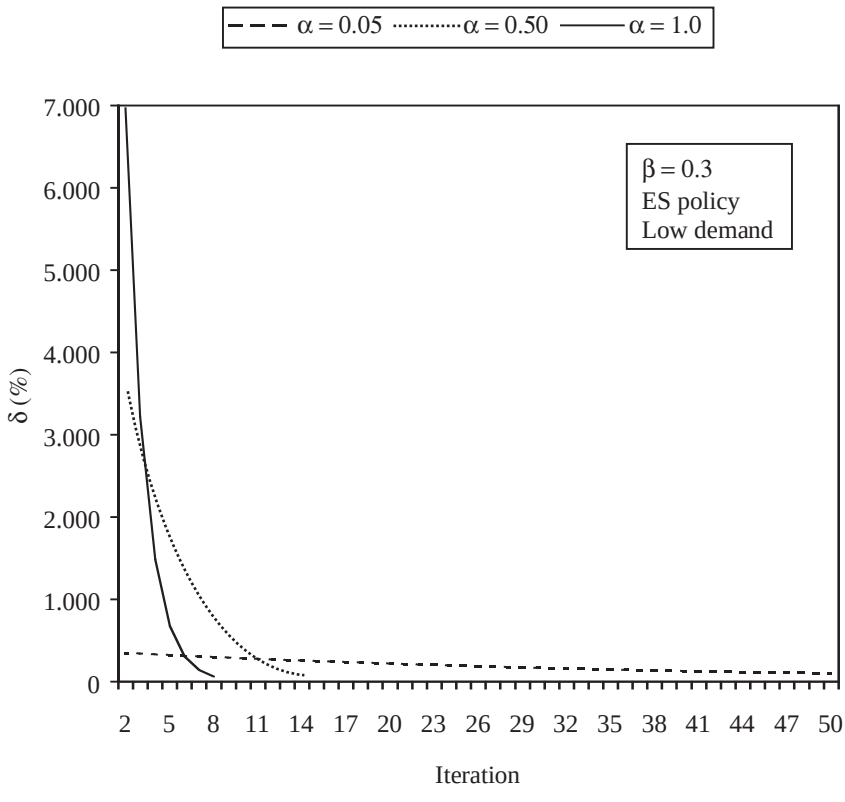


Figure 6.9 Convergence of modified IOA in terms of δ for selected values of α (β=0.3, equisaturation control policy and low demand)

6.5, such computational effort may be measured by the total number of stochastic network loadings performed, which, in turn, equals the product of the number of IOA iterations and the average number of MSA iterations. Note that in our implementation both procedures are solved to convergence, and there is no attempt to explore computational tradeoffs between the numbers of outer and inner iterations; for a systematic analysis of this issue in a deterministic ETSS context, see Meneguzzer (1996).

Table 6.8 shows, for each combination of α, β, policy and demand considered in the previous analyses, the total number of stochastic network loadings required to solve the ETSS problem. It is seen that this number exhibits a very large variation, ranging from a minimum of 10 to a maximum of 1177. With a few exceptions, the results are intuitively appealing, as they suggest that the computational burden increases with increasing congestion, and declines as driver reactiveness and route choice dispersion increase. As far as the effect of the control policy is concerned, P_0 is seen to outperform equisaturation, except for the case of accurate driver information and high demand.

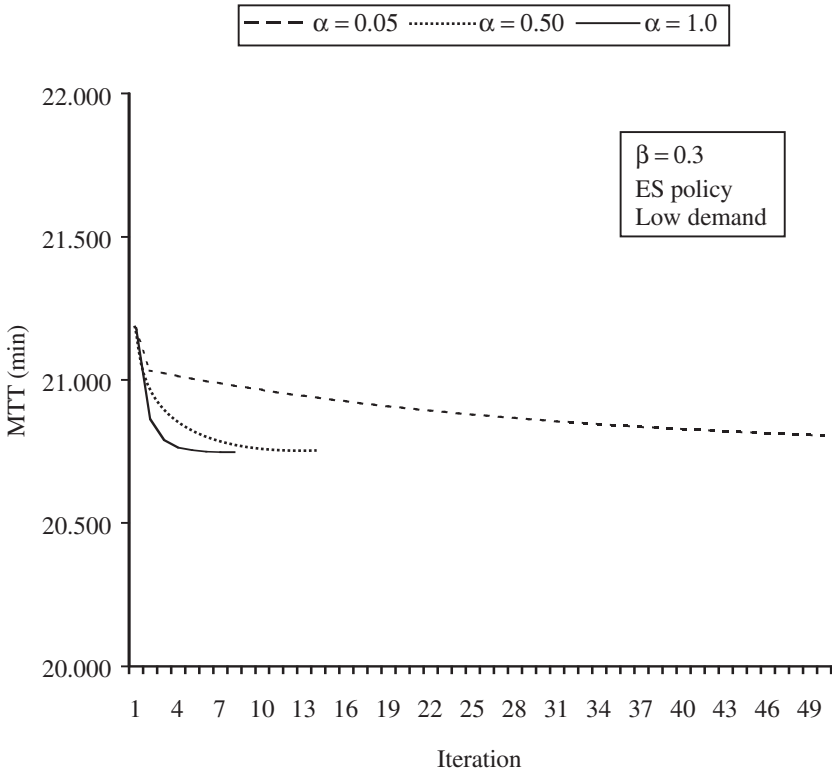


Figure 6.10 Convergence of modified IOA in terms of MTT for selected values of α (β = 0.3, equisaturation control policy and low demand)

Table 6.8 Computational requirements of the modified iterative scheme in terms of total number of stochastic network loadings performed

			α = 0.05	α = 0.25	α = 0.50	α = 0.75	α = 1.00
β = 0.15	P_0	LD	10	15	20	20	15
		HD	290	115	70	50	40
	ES	LD	10	25	20	20	15
		HD	315	125	75	55	40
β = 0.3	P_0	LD	148	72	48	36	28
		HD	635	209	119	83	65
	ES	LD	200	88	56	40	32
		HD	701	227	131	95	71
β = 1.5	P_0	LD	396	120	68	49	35
		HD	1177	297	157	111	82
	ES	LD	841	319	194	140	110
		HD	932	234	122	83	59

Note: ES: Equisaturation; LD: Low demand; HD: High demand.

6. CONCLUSIONS

A new approach to the solution of ETSS problems has been proposed and tested in this chapter. The main idea underlying the proposed approach is that the classical iterative scheme known as IOA can be appropriately modified so as to explicitly recognize that, in real-world networks, the rerouting of trips induced by traffic-responsive signal control may involve only a fraction of the network users. This partial driver response feature is introduced into IOA through an additional step, in which the updating of link flows in reaction to signal re-setting is smoothed by means of a 'driver reactiveness' parameter. The assumption of stochastic user equilibrium, adopted in the traffic assignment submodel, allows the explicit consideration of the effect that drivers' misperception of travel times (and hence driver information) may have on the interaction between signal control and route choice under conditions of partial user response.

Extensive numerical experiments have been conducted on a simple test network, in order to assess how the convergence behavior of the iterative scheme and the properties of the resulting flow-control equilibria may be affected by the driver reactiveness parameter and by the level of route choice dispersion. In addition, the performance of two alternative signal control policies, equisaturation and Smith's P_0, has been evaluated in interaction with the above parameters and under different demand levels.

Good behavior of the iterative scheme has been observed on the test network, as a smooth, monotone convergence to a mutually consistent flow-control equilibrium was achieved in all test cases. The main findings of the numerical experiments suggest that the fraction of drivers reacting to the updating of signal settings may have a considerable effect on the dynamics of the interaction between signal control and route choice, but not on the resulting equilibrium. This conclusion is substantiated by the observation that the link flows, green splits and aggregate network performance obtained for various values of the reactiveness parameter are essentially similar, whereas the speed and pattern of convergence of the modified IOA appear to be strongly dependent upon the parameter itself. In addition, there is evidence that this behavior may be characterized by 'decreasing returns', in the sense that improvements in convergence performance as a result of increased reactiveness tend to be more significant for lower initial values of the parameter.

The level of route choice dispersion has been found to affect significantly both the equilibrium and convergence properties of the proposed iterative scheme. Our results indicate that a greater dispersion of route choice leads to deterioration of network performance, thus confirming a well-known feature of ordinary stochastic user equilibrium assignment, and that the spread of trips over alternative routes, induced by driver misperception of travel times, may accelerate the redistributional effect of traffic-responsive signal control, especially under conditions of low congestion.

Other interesting findings pertain to the comparative behavior of the two signal control policies. While in our tests equisaturation and P_0 have produced remarkably similar results in terms of aggregate network performance, some noticeable differences in the allocation of green times to conflicting intersection approaches have emerged. These differences appear to follow from the different flow-redistributional capabilities that the two policies exhibit when applied in interaction with route choice. Interestingly, such differences tend to vanish as route choice dispersion increases, suggesting that the impact of

alternative control strategies on equilibrium signal settings may be negligible when the traffic assignment process is dominated by stochastic effects.

Overall, the results obtained from the numerical tests appear to be encouraging, and suggest that the proposed approach to ETSS may represent a worthwhile step toward an improved description of the real-world evolution of network conditions ensuing from the interaction of route choice and traffic-responsive signal control. Further developments of this line of research should address the question of how to estimate realistic values of the driver reactiveness parameter in specific applications, and should pursue the implementation of the proposed approach on larger and more realistic networks. It is envisaged that the former task could be accomplished through the observation of the actual day-to-day route choices of an appropriate sample of drivers in a network operating under responsive control, perhaps by means of GPS (global positioning systems)-based data collection techniques.

Finally, a key issue that needs to be thoroughly investigated is the behavior of the proposed iterative scheme with respect to uniqueness of the mutually consistent flow-control equilibria (see the discussion at the end of Section 2). A computational search for possible multiple solutions could be carried out by systematically changing the initial values assigned to the green splits of signalized links, under various combinations of demand level, control policy and model parameters.

NOTES

1. Alternatively, such models are sometimes referred to as *combined traffic assignment and control* models.
2. In fact, control decisions for actuated signals may even be based on *anticipated* traffic conditions, that is, flow estimates resulting from short-term forecasts; in this study, however, such a *proactive* approach is not considered, and control policies are assumed to be purely reactive.
3. However, it should be noted that the conditions being violated in the presence of responsive control are not *necessary* for solution uniqueness, so that single-equilibrium behavior cannot be ruled out a priori.
4. This is the same as saying that signal phasing is designed so that conflicting flows are never given way simultaneously.

REFERENCES

Akçelik, R. (1988), 'The Highway Capacity Manual delay formula for signalized intersections', *ITE Journal*, **58**, 23–7.

Allsop, R.E. (1971), 'Delay-minimising settings for fixed-time traffic signals at a single road junction', *Journal of the Institute of Mathematics and its Applications*, **8**, 164–85.

Allsop, R.E. (1974), 'Some possibilities for using traffic control to influence trip distribution and route choice', in D.J. Buckley (ed.), *Proceedings of the 6th International Symposium on Transportation and Traffic Theory*, New York: Elsevier, pp. 345–73.

Allsop, R.E. and J.A. Charlesworth (1977), 'Traffic in a signal controlled road network: an example of different signal timings inducing different routeings', *Traffic Engineering and Control*, **18**, 262–4.

Bell, M.G.H. (1992), 'Future directions in traffic signal control', *Transportation Research*, **26A**, 303–13.

Bureau of Public Roads (1964), *Traffic Assignment Manual*, Washington, DC: US Department of Commerce.

Chen, O.J. and M.E. Ben-Akiva (1998), 'Game-theoretic formulations of interaction between

dynamic traffic control and dynamic traffic assignment', *Transportation Research Record*, **1617**, 179–88.
Chiou, S.-W. (1999), 'Optimization of area traffic control for equilibrium network flows', *Transportation Science*, **33**, 279–89.
Daganzo, C. and Y. Sheffi (1977), 'On stochastic models of traffic assignment', *Transportation Science*, **11**, 253–74.
Dickson, T.J. (1981), 'A note on traffic assignment and signal timings in a signal-controlled road network', *Transportation Research*, **15B**, 267–71.
Fisk, C.S. (1984), 'Game theory and transportation system modeling', *Transportation Research*, **18B**, 301–13.
Friesz, T.L. and P.T. Harker (1985), 'Properties of the iterative optimization-equilibrium algorithm', *Civil Engineering Systems*, **2**, 142–54.
Harker, P.T. and T.L. Friesz (1984), 'Bounding the solution of the continuous equilibrium network design problem', in J. Volmuller and R. Hamerslag (eds), *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, Utrecht: VNU Science Press, pp. 233–52.
Heydecker, B.G. and T.K. Khoo (1990), 'The equilibrium network design problem', *Proceedings of the AIRO 90 Conference on Models and Methods for Decision Support*, Sorrento, Italy, pp. 587–602.
Hu, T.Y. and H.S. Mahmassani (1997), 'Day-to-day evolution of network flows under real-time information and reactive signal control', *Transportation Research*, **5C**, 51–69.
Hurdle, V.F. (1984), 'Signalized intersection delay models – a primer for the uninitiated', *Transportation Research Record*, **971**, 96–105.
Lee, S. and M. Hazelton (1996), 'Stochastic optimisation of combined traffic assignment and signal control junction modelling', in J.B. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, New York: Elsevier, pp. 713–35.
Maher, M.J., X. Zhang and D. Van Vliet (2001), 'A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows', *Transportation Research*, **35B**, 23–40.
Meneguzzer, C. (1995), 'An equilibrium route choice model with explicit treatment of the effect of intersections', *Transportation Research*, **29B**, 329–56.
Meneguzzer, C. (1996), 'Computational experiments with a combined traffic assignment and control model with asymmetric cost functions', in Y.J. Stephanedes and F. Filippi (eds), *Proceedings of the 4th International Conference on Applications of Advanced Technologies in Transportation Engineering*, New York: ASCE, pp. 609–14.
Meneguzzer, C. (1997), 'Review of models combining traffic assignment and signal control', *Journal of Transportation Engineering*, **123**, 148–55.
Nihan, N., M. Hamed and G. Davis (1995), 'Interactions between driver information, route choice, and optimal signal timing on a simple network', *Journal of Advanced Transportation*, **29**, 163–82.
Sheffi, Y. (1985), *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Englewood Cliffs, NJ: Prentice-Hall.
Smith, M.J. (1979), 'Traffic control and route-choice; a simple example', *Transportation Research*, **13B**, 289–94.
Smith, M.J. (1980), 'A local traffic control policy which automatically maximises the overall travel capacity of an urban road network', *Traffic Engineering and Control*, **21**, 298–302.
Smith, M.J. (1981), 'Properties of a traffic control policy which ensure the existence of a traffic equilibrium consistent with the policy', *Transportation Research*, **15B**, 453–62.
Smith, M.J. and T. Van Vuren (1993), 'Traffic equilibrium with responsive traffic control', *Transportation Science*, **27**, 118–32.
Smith, M.J., T. Van Vuren, B.G. Heydecker and D. Van Vliet (1987), 'The interactions between signal control policies and route choice', in N.H. Gartner and N.H.M. Wilson (eds), *Proceedings of the 10th International Symposium on Transportation and Traffic Theory*, New York: Elsevier, pp. 319–38.
Tan, H.N., S.B. Gershwin and M. Athans (1979), 'Hybrid optimization in urban traffic networks', Technical Report no. DOT-TSC-RSPA-797, Cambridge, MA: Laboratory for Information and Decision Systems, Massachusetts Institute of Technology.

Van Vuren, T. and D. Van Vliet (1992), *Route Choice and Signal Control: The Potential for Integrated Route Guidance*, Aldershot, UK: Avebury.
Wardrop, J.G. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institution of Civil Engineers*, Part II, **1**, pp. 325–78.
Watling, D. (1996), 'Modelling responsive signal control and route choice', in D. Hensher, J.K. and T.H. Oum (eds), *Proceedings of the 7th World Conference on Transport Research*, vol. 2, Oxford: Pergamon, pp. 123–37.
Webster, F.V. (1958), 'Traffic signal settings', Road Research Technical Paper No. 39, London: HMSO.

7. Transport and location effects of a ring road in a city with or without road pricing

Lars-Göran Mattsson and Lina Sjölin*

1. INTRODUCTION

Sustainability is one of the key issues in urban policy making of today. This does not mean that there is general agreement on its meaning or definition. In relation to urban transport policies, however, sustainability is very much a question about how to achieve a reasonable balance between actions that will increase mobility of people and goods and actions that will reduce transport demand and, in particular, the use of motorized vehicles.

Many cities grow more or less rapidly. For this and other reasons they often face increasing congestion problems on the roads. Citizens and trade and industry often express their claims for investments in the road network. However, in a situation when demand has been suppressed because of congestion, it is not clear to what extent congestion actually would be relieved by simply increasing the supply of road facilities. Many transport analysts argue that such a policy should be accompanied by a congestion pricing scheme to reduce the use of the enhanced road network to an efficient level.

Cities are very complex systems. To be able to evaluate transport policies in an appropriate way, decision support systems are necessary. By now many urban and traffic planners have access to sophisticated network-based travel demand modelling tools by which they can analyse the effects of different policies on, for example, trip generation, distribution, mode and route choice. However, road network investments, or the introduction of congestion pricing, will not only affect the demand for transport in a direct and immediate way. Such policies will also, in the long run, change the attractiveness of different places for location of activities and hence the land-use structure of the city. To be able to evaluate such policies appropriately, urban and traffic planners need tools that could help them clarify transport as well as land-use effects of different actions. For strategic planning they need to be able to analyse the interaction between the transport and land-use markets. As an example, will the effects of a policy instrument in the transport market be counteracted or amplified by the relocation of households and workplaces in the land market?

There is considerable interest in developing combined transport and land-use models (Wegener 1994; Wilson 1998) and there are also some commercial systems available. One of the authors has been involved in developing a residential and employment location model which, combined with a transport demand model, has been applied in Stockholm regional planning and to some extent also in other places in the Nordic countries (Anderstig and Mattsson 1991, 1998). A number of interesting modelling and policy studies have recently

been carried out in many EU projects such as PROSPECTS[1] and other projects within the LUTR cluster[2] (May et al. 2003; Minken et al. 2003). However, there is no doubt that the actual application of combined land-use/transport modelling tools in urban policy analysis is still fairly limited compared with stand-alone transport demand models. One obvious reason for this is the fact that the calibration, implementation and application of a land-use/transport interaction model is quite a demanding task.

This kind of experience motivates our present interest in simplified urban simulation models based on spatial symmetry assumptions. The idea of exploring symmetry in urban modelling is not new. In an early study, Lam and Newell (1967) developed an analytical model of flow-dependent traffic assignment in a circular city, but it was difficult to provide a general solution to the system of equations that would generate the equilibrium flows. Some interesting special cases could be solved numerically, however. Roy et al. (1998) developed a more comprehensive urban simulation model that in its first version also makes use of symmetry assumptions in much the same spirit as our approach. The model locates different kinds of households and jobs in a symmetric city and models the interaction between job and housing choices. The resulting commuting trips are assigned to symmetric road and public transport networks.

Of course, the symmetry assumptions in these and our study would exclude many details of a particular transport and land-use system. For example, the existence of an airport or other major transport facilities would typically be of a non-symmetrical character. Many real cities are located on the shoreline, which is another obvious example when the symmetry assumption is violated. On the other hand, in many policy studies one is interested in the principal or typical effects of a course of action. The abstract view that a completely symmetric city implies could then rather be an advantage. The effects of a policy will be simulated without distorting the results by the spatial peculiarities of a specific city. The most important advantage of the symmetry assumption is, as we shall see, that it allows us to build an operational urban simulation model that is much more realistic and comprehensive than would be possible within the tradition of analytic urban economics models, and still computationally much simpler than a fully-fledged operational land-use/transport interaction model.

We shall apply and slightly extend a stylized model of a 'generic' symmetric city developed by Eliasson and Mattsson (2001). The numerical solution of the model relies heavily on the assumed symmetry of the city. The city is star-shaped with a radial transport system including a viable public transport alternative, connecting discrete and homogeneous zones. There are four groups of actors: households, employers, shops and service establishments. The households commute to the workplaces and make shopping and service trips by car, public transport and a slow mode (cycling and/or walking). The trips may take place during the morning peak, office hours or the afternoon peak. In addition to personal travel, there are also road-based deliveries from the workplaces to the shops and service establishments. There is congestion on the roads. The car travel time for any link is an increasing function of the level of car traffic. The public transport system, on the other hand, is subject to increasing returns to scale. The public transport travel time between any two zones is assumed to go down as the demand goes up. The different actors locate in the city in response to accessibility factors and land prices in a way that is specific to each group of actors.

Eliasson and Mattsson (2001) used this model to simulate transport and land-use

effects of a congestion pricing scheme as well as of simplified pricing policies in the form of 'toll rings'. In this chapter, we first extend the modelling framework by allowing for a ring road connecting the innermost suburbs. With this modified model we can extend the previous analysis to the impact of a ring road in itself as well as to how it would function in combination with optimal (that is marginal cost-based) congestion pricing or a toll ring. The analysis includes the effects on travel time and travel distance by mode of transport and the effects on the location of households, workplaces, shops and service establishments.

The rest of the chapter is organized as follows. First, the policy context with special reference to the Stockholm region will be discussed further in Section 2. Then the model is described briefly in Section 3 and the scenarios to be analysed are defined in Section 4. The results from the simulation of the scenarios are presented and discussed in Section 5, followed by a summary and conclusions in Section 6.

2. RING ROAD AND ROAD PRICING: WHAT ARE THE LIKELY EFFECTS?

Many cities suffer from severe congestion problems (see Schneider et al. 2002, for an overview and discussion of possible actions). Stockholm is no exception. After a period of reduced traffic volume in connection with the economic recession during the mid-1990s, traffic volume and congestion have again increased to even higher levels. There has been considerable political turbulence about how to handle these problems. Over the years different strategic plans for the improvement of the transport system have been put forward but not, or only to a limited extent, implemented. The most notable example was the so-called 'Dennis Package' with its extensive proposal for investments in the road network and in the public transport system (Johansson and Mattsson 1995). This package also included a road-pricing scheme in the form of a toll ring. The pricing scheme was meant to fulfil the dual purpose of reducing traffic and hence congestion in the inner city and of raising necessary funds for the road investments. The political accord behind the package eventually broke down because of disagreement about certain controversial road links and the road-pricing scheme.

In the year 2001 a governmental committee comprising representatives from all parties in the national parliament was commissioned to propose actions to improve the functioning and capacity of both the public transport system and the road network in the extended Stockholm region. The improved transport system should be environmentally, socially and economically sustainable. Among road investments that are being considered is the completion of a ring road around the inner city, while congestion pricing is one of the economic policy measures that is being investigated. Following the 2002 general election, this review was fuelled by political agreement at the national level, according to which a full-scale test of congestion pricing is to be carried out for the city of Stockholm before the end of the present term of office, that is by 2006.

Much of the controversy around the transport policy in the Stockholm region has been related to the effectiveness of road investment and road pricing to reduce congestion. On the one hand, some argue that many large cities of the size of Stockholm have introduced some kind of ring road to improve accessibility and to alleviate congestion in the city

centre and so Stockholm, one of the few exceptions, should do likewise. The environmentalists, on the other hand, claim that new roads simply induce more traffic, leaving congestion at the same level as before. The argument is that economic incentives such as congestion pricing would be a more effective and cheaper policy to mitigate congestion. Both sides have very firm views and they disagree on what the effects of these two specific policies would be.

Although the effects of a ring road or a road-pricing scheme on transport demand are fairly clear, according to the research literature, the magnitude of these effects, and to what extent they counterbalance each other, is not evident. But when it comes to the long-term location effects, it is not even clear whether these policies will lead to centralisation or decentralisation of activities as it is reviewed in Eliasson and Mattsson (2001).

This policy context motivates our present study of transport and location effects of a ring road with or without some kind of road pricing. Congestion pricing is a policy that would be complicated to implement – a toll ring would be much simpler. A special question, therefore, is how closely congestion pricing can be approximated by a toll ring.

3. AN URBAN SIMULATION MODEL

3.1 The Original Version of the Model without a Ring Road

We shall briefly present the model that will be used. The original version is presented in Eliasson and Mattsson (2001), to which we refer for a full description of the mathematical structure and the choice of parameter values. The model has been calibrated to replicate location and transport patterns of a generic symmetric city. However, a decision had to be made concerning what is meant by a generic city, and in this chapter, behavioural parameters and size variables have, as far as possible, been chosen so that the model will resemble the situation in Stockholm.

The city consists of discrete, homogeneous zones connected to one another by a symmetric radial network, with four zones on each of eight rays, as illustrated in Figure 7.1. The links connecting the zones are all 5 kilometres in length, and there is one link representing two lanes in each direction.

Since the city is completely symmetric in all respects, it is sufficient to regard one of the rays when analysing the results of the simulations. The notation to be used is illustrated in Figure 7.2. The zones are denoted by the numbers 1 to 5, from the city centre to the outermost suburbs, and the links are denoted by the letters A to D, from the innermost links to the outermost ones.

There are four groups of activities in the city:

- *households* represented by one wage earner (interactions within a household are not modelled),
- *workplaces* of only one kind,
- *shops* that can be thought of as grocery stores and where different shops are perceived as rather similar when the households make their shopping destination choices, and
- *service establishments* that can be thought of as banks, hospitals, travel agencies and

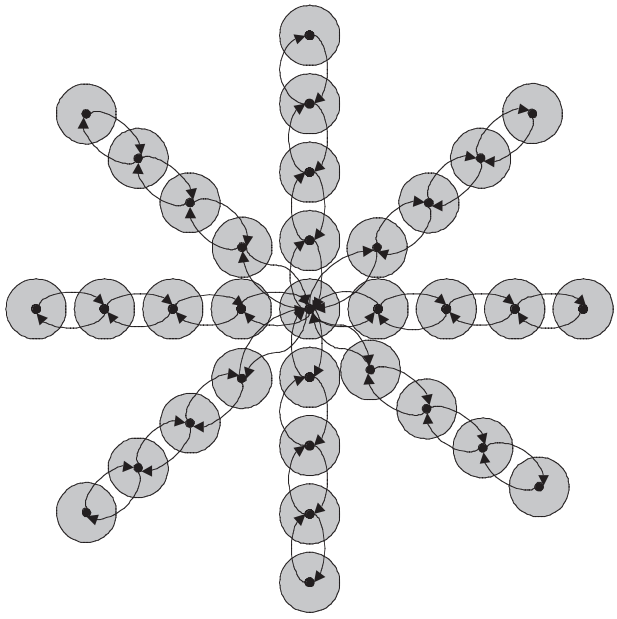


Figure 7.1 The star-shaped network of the original model

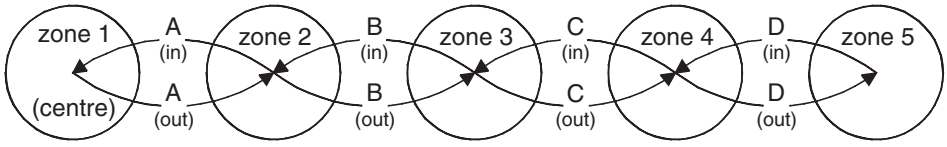


Figure 7.2 Notation along one representative ray of the symmetric original model

so on, and where different service establishments are perceived as fairly heterogeneous when the households make their service destination choices.

There are three travel modes available in the city:

- *car*, for which the (link) travel times increase with increasing car flows,
- *public transport*, which is subject to economies of scale – the more people that choose to travel by public transport, the higher the frequency of service, and the denser the network within the zones, leading to shorter waiting and access times, and
- *slow mode*, which is an aggregate of cycling and/or walking and for which the speed is assumed to be constant, and hence all link travel times will be equal.

The car and public transport modes are also subject to monetary costs that depend on the distance travelled, while the slow mode is free of charge. For the car mode, parking fees may be charged and some kind of road pricing may be levied on the users.

In this original version of the model there is no need to model the route choice explicitly, since there is a unique cheapest route between each pair of zones. When going from one zone to another, the travel distance as well as the travel time is the sum of the distances or travel times of the links that are used for the trip. The travel distance for intrazonal trips is assumed to be 2.5 kilometres, and the travel time is assumed to be half the average travel time of the links connected to the zone. This means that also the intrazonal travel times will be sensitive to congestion.

The different groups of activities interact with one another through four kinds of trips: work, shopping and service trips and deliveries. The trips can take place during three different time periods of different length: morning peak (2.5 hours), office hours (6 hours) and afternoon peak (4 hours). Each household makes a work trip when it travels to work during the morning peak and when it returns home during the afternoon peak. Shopping and service trips take place during office hours and the afternoon peak. If they are made by car, a parking fee is charged (parking is most expensive in the city centre and free in the outermost suburbs). The demand for shopping and service trips is elastic both with respect to the frequency and the choice of time period (office hours or afternoon peak). Deliveries are only made during office hours and always by car. Each shop and service establishment requires a fixed number of deliveries each day.

The household travel demand and the deliveries are modelled by nested logit models (see Ben-Akiva and Lerman 1985, for a thorough introduction to this modelling tradition and Eliasson and Mattsson 2001, for a detailed account of how it has been done in the present model). Transport demand is modelled conditional on the location of activities. For work trips we consider mode and destination choices, whereas for shopping and service trips we also model, as mentioned above, the choice of frequency and time period. The destination choices for work trips are modelled so that the total number of individuals working in a zone agrees with the total number of workplaces in that zone. Deliveries are only modelled with respect to destination choice, since, by assumption, they have made by car with fixed frequencies. In all these models, demand depends on travel times and travel costs with respect to the involved modes of transport. The different trip types have different sensitivities to time and cost in a way that is consistent with empirical data. As a consequence, shopping trips, for example, have a stronger preference for car than service trips have. It should be remembered that the car travel time on a link is an increasing function, and the public transport travel time a decreasing function, of the number of people choosing that specific mode. In this way it is possible to model how congestion affects the transport pattern in the city. It is also possible to model how different road pricing schemes would affect the transport pattern by changing the monetary car travel costs accordingly.

Finally, it remains to explain how the activities are located in the city. The city is 'closed', that is, the total number of members of each group in the city is constant. There is a fixed amount of land reserved in each zone for households, shops and service establishments, respectively. Therefore these different groups act on different land markets. For workplaces it is assumed that there is no scarcity of land in any zone. One interpretation of this is that workplaces are located at the edge of the zones where there is always enough land available.

The location of activities is also modelled by nested logit models, where:

- *households* value lot size (inversely proportional to the number of households in the zone), accessibility to shops and service establishments and to potential workplaces,

- *workplaces* value accessibility to households (as workforce), and
- *shops and service establishments* value accessibility to households (expressed in terms of the number of customers attracted to the zone), accessibility to workplaces (from which they get a certain number of deliveries each day) and lot size (shops to a greater extent than service establishments).

The location of activities is linked to the transport system through accessibility measures. These measures are operationalized as log sums from the travel demand model that is relevant for each activity (for a detailed account of how this has been done, see Eliasson and Mattsson 2001). Through these accessibility measures, travel times and travel costs by different modes will affect the location pattern, and hence also the origin of the demand for the different trip types and for the deliveries.

Thus, the travel demand for different trip types depends on where different activities are located and what the travel times and travel costs are on the different links in the transport network. The travel times (and in the case of congestion pricing also the travel costs) depend on the number of people choosing the different modes for the different links, that is, on the travel demand. Finally, the location of activities depends on the accessibility in the different zones, which in turn depends on the location of other activities and the travel times and travel costs between the zones. All these relationships result in a number of equations that are solved for a spatially symmetric equilibrium.

3.2 The Present Version with a Ring Road Added

The model has been extended to allow for a ring road connecting the innermost suburbs (zone 2), as shown in Figure 7.3. This ring road is open only for car use.

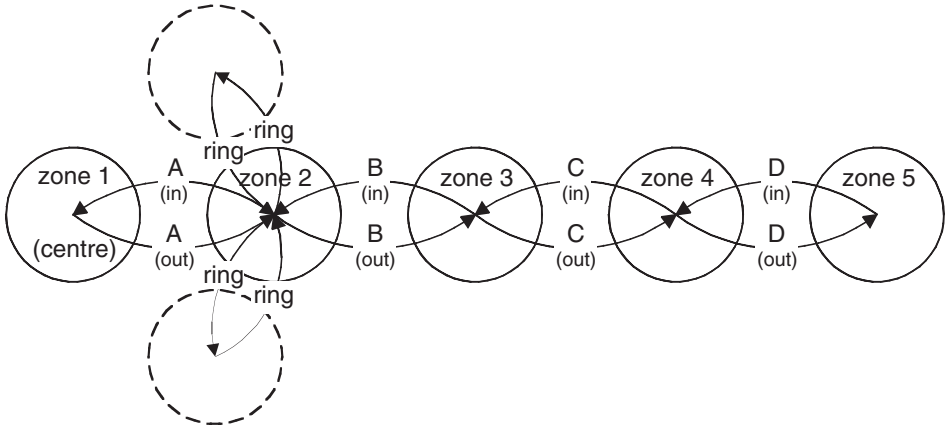


Figure 7.3 A representative ray of the star-shaped model, with the ring road added

With this ring road added to the network, there may be alternative routes for car trips between a pair of zones. We assume that the route choices follow Wardrop's principle of user equilibrium, that is, all used routes between a pair of zones have the same generalized cost and no unused route has a lower generalized cost. This generalized cost is

computed as the monetary cost for the route plus the travel time multiplied by an average time value.

Due to the convenient symmetry of the city it is possible to design a simple algorithm for the calculation of equilibrium routes without applying the weightier machinery of solving an equivalent non-linear optimization problem, which is the typical approach in large-scale operational models (see Boyce and Daskin 1997, for an introduction to the latter approach).

This simple algorithm follows from a couple of observations. If a trip is made between any two zones on the same ray (the city centre may be one of these zones), the unique user equilibrium route is to follow the radial route joining the two zones. Hence, it is only when a trip is made between a suburb on one ray and a suburb on another ray, thus avoiding the city centre, that there may exist multiple equilibrium routes. In these cases, it is only for the part of the route that goes between the innermost suburbs of the two rays that there are alternatives. For this part there are at the most two options in equilibrium, either to use the ring road or to go via the city centre. We now subdivide the trips between different rays according to 'how many rays away' they are. It follows from simple geometrical considerations that there may exist multiple routes in equilibrium for one of these categories at the most. Trips between rays that are closer together will all use the ring road, whereas trips between rays that are farther apart will all go via the city centre. Thus the idea of the algorithm is to find this category of mixed route choice if it exists, and if so how the trips are then divided between the two route options. Sjölin (2001) provides a detailed account of the algorithm.

4. SCENARIOS

To study transport and location impacts of a ring road, and to what extent the impacts are modified by different kinds of road pricing, eight scenarios in total have been constructed. These scenarios have been obtained by combining the original base scenario as well as the ring road scenario obtained by connecting the innermost suburbs, as illustrated in Figure 7.3, with four different road pricing schemes: no road pricing, (optimal) congestion pricing, an inner toll ring and an outer toll ring.

Optimal (first-best) congestion pricing, or congestion pricing for short, is achieved by charging a fee on each link so that the total marginal cost for the individual car driver will be equal to the social marginal cost. If all car users had the same value of time, τ, the optimal charge level for a specific link, as a function of the car flow f on that link, would be:

$$toll(f) = \tau \frac{dt(f)}{df} f,$$

where $t(f)$ is the travel time that is assumed to be an increasing function of the car flow on the link. This level of charge has an intuitive interpretation. To achieve optimal congestion pricing, each car user on a link should be charged a fee that is equivalent to the additional costs his/her presence on the link imposes on all other users of the same link, which is the additional travel time caused by him/her, $dt(f)/df$, times the number of car users affected, f, times their value of time, τ. Since this charge level depends on the actual

car flow, it will be different for different links and different time periods. One additional difficulty should be noted. The value of time varies between different trip types, and hence the average time value on a link depends on the mixture of trip types. For computational reasons we ignore this difficulty and apply consistently a common average value of time of $\tau = 42.60$ SEK/h.

The road pricing scheme of an inner toll ring is implemented by levying a fixed toll of 10 SEK for all time periods on all car trips on inbound links between the innermost suburbs and the city centre (that is, on all inbound A-links). Similarly, an outer toll ring is implemented by levying a fixed fee of 10 SEK on all car trips on inbound links immediately outside the innermost suburbs (that is, on all inbound B-links).

5. RESULTS AND ANALYSIS

In applying the urban simulation model to the scenarios of the previous section, all parameter values have been kept at the same level as in the previous study by Eliasson and Mattsson (2001). As for the size of the city, there are 1 million households, 1 million workplaces, 10 thousand shops and 10 thousand service establishments. Due to the symmetry of the city it is only necessary to present the results for one ray. Links in different directions between two zones do not generally have the same flows and travel times, although they are fairly similar. We shall suppress these differences and display only the average value for the two directions.

The application of the model to all eight scenarios generates a considerable number of results. We shall present only a selection of these to illustrate some interesting tendencies. For a full account of the results, see Sjölin (2001).

5.1 Impacts of a Ring Road with No Road Pricing

One immediate effect of introducing a ring road to connect the innermost suburbs (zone 2) is a drastic increase in the accessibility of activities that may be located in these suburbs. Figure 7.4 shows how this affects the location pattern in the city. The increase in accessibility evidently has a strong centralizing effect on location, particularly for the innermost suburbs, where the increase is 20 per cent or more for all activities. In the city centre (zone 1) the number of households increases by 7 per cent and the number of workplaces by 9 per cent. The number of shops and service establishments decreases in the centre. This is particularly pronounced for shops, which decrease by 8 per cent.

The way shops are modelled makes them more sensitive to generalized car costs than service establishments are, since shop customers have a stronger preference for car use. Apart from depending on distance, the generalized car costs depend on parking fees and travel times. Introducing the ring road makes travel between the inner suburbs less expensive and hence their attractiveness is increased (especially for shops). Moreover, shops and service establishments take into account the expected number of customers attracted to the zone they consider for location. If a zone has a large number of shops or/and service establishments, and consequently a large number of customers, more shops and service establishments will find it attractive to locate there. When they do, the zone attracts even more customers, making it even more attractive for location. The increase in the number

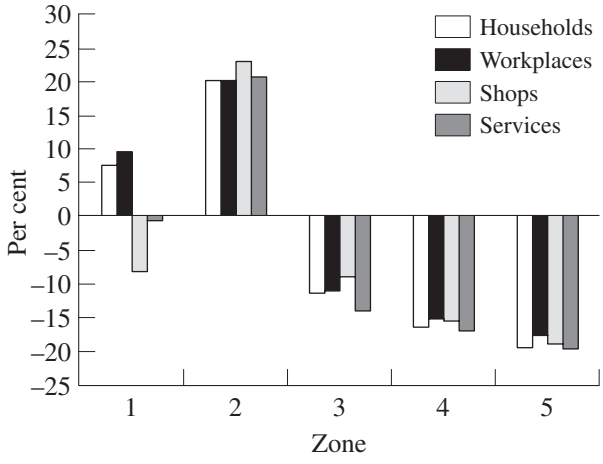


Figure 7.4 Percentage change in the location of activities by zone (ring road scenario with no road pricing compared with the base scenario with no road pricing)

of shops and service establishments in the innermost suburbs shows how the model is capable of capturing this kind of *clustering* phenomenon.

The ring road causes the car traffic to and through the city centre to decrease substantially. In the base scenario, car travel times on the innermost links (A-links) are about 30 minutes during the morning and afternoon peaks, and 22 minutes during office hours. With the ring road, morning travel time is 22 minutes, afternoon travel time is 23 minutes, and during office hours it is 13 minutes. Although the A-links are still quite congested after the introduction of the ring road (travel time at free flow is 6 minutes), the situation has obviously improved. Figure 7.5 shows the relative change in car travel flow (vehicles/hour), and Figure 7.6 in car travel time, on each link by time period for the ring road scenario compared with the base scenario. Evidently the ring road fulfils its purpose of relieving congestion in the city centre.

In contrast to this relieving effect in the city centre, the ring road will increase congestion quite severely on the links just outside of the ring road (the B-links), and consequently also the car travel times. The largest relative increase of car traffic on these links occurs during office hours, when it is 26 per cent and the resulting increase of the travel time is 55 per cent (from 11 to 17 minutes).

The introduction of the ring road makes it more advantageous to go by car. This is not only because car travel times decrease on many of the links as a consequence of less congestion. It is also an effect of the fact that the ring road is open only for cars. People going by the slow mode or by public transport from one inner suburb to another have to travel via the city centre. This increases the competitiveness of the car mode, since, for instance, to travel by public transport during office hours from one inner suburb to another on a neighbouring ray takes 39 minutes, while it takes only 13 minutes by car (using the ring road). In the base scenario the corresponding travel times are 39 minutes by public transport and 44 minutes by car.

Because of these improved conditions for car use in the ring road scenario, the share of the car mode increases, and the share of public transport decreases, on all links and during

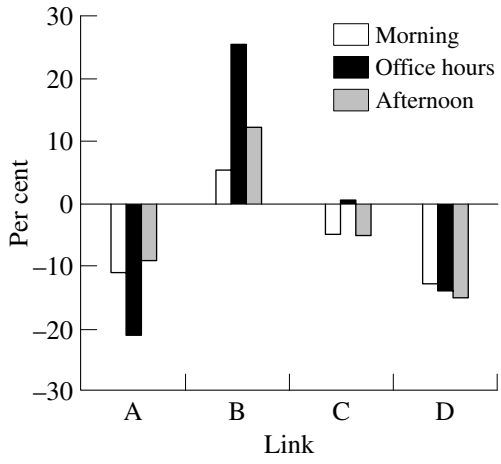


Figure 7.5 Percentage change in car travel flow by link and time period (ring road scenario with no road pricing compared with the base scenario with no road pricing)

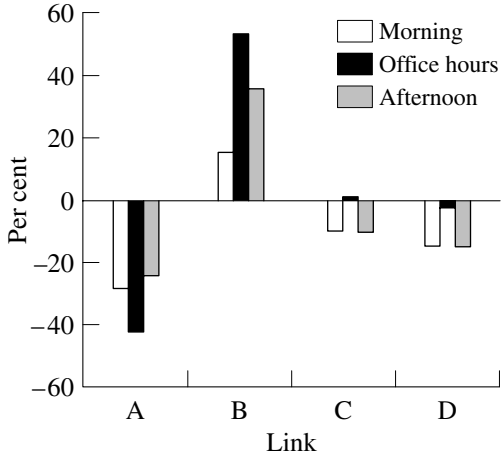


Figure 7.6 Percentage change in car travel time by link and time period (ring road scenario with no road pricing compared with the base scenario with no road pricing)

all time periods compared with the base scenario. The biggest changes are seen on the B-links, where the car share during office hours increases by 14 percentage units. In total, the car share increases by 9 percentage units on the B-links. The total increase on the C-links is 5 percentage units and on the D-links it is 3 percentage units. The largest increases of the car shares occur during office hours on all links, which is a consequence of the low service level of public transport during that time. The smallest changes in the mode shares occur on the A-links. These changes are all in the range of 1 percentage unit up or down.

Looking at the mode shares by residential zone, Eliasson and Mattsson (2001) concluded for the base scenario that the highest public transport shares and the lowest car shares occur in zone 2 rather than in the city centre (which might be expected). Their explanation for this was that households who are living in the centre and are using a car need pass only one of the highly congested A-links to get to the activities in any inner suburb, while households located in one of the inner suburbs have to pass two such links to reach another inner suburb. Hence in the base scenario, households living in the innermost suburbs are the ones that are least inclined to travel by car.

With the introduction of the ring road, residents in all zones increase their car use for all trip types, and decrease their public transport use, compared with the base scenario (see Figures 7.7 and 7.8). The ring road enhances the accessibility by car particularly for the inner suburbs, since people living there no longer have to pass the centre to go to other inner suburbs. Then it is no longer residents in zone 2 (or those in the city centre) who exhibit the highest share of public transport use and the lowest share of car use, but resi-

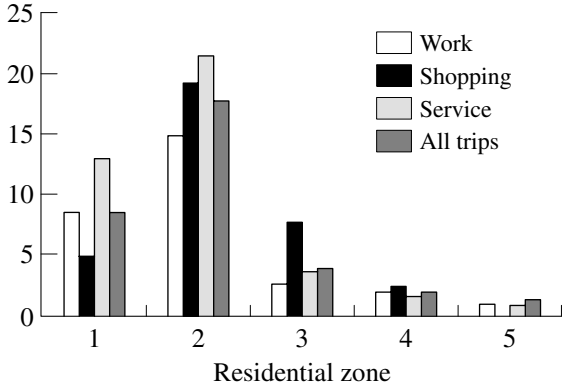


Figure 7.7 *Percentage change in car share by trip type and residential zone (ring road scenario with no road pricing compared with the base scenario with no road pricing)*

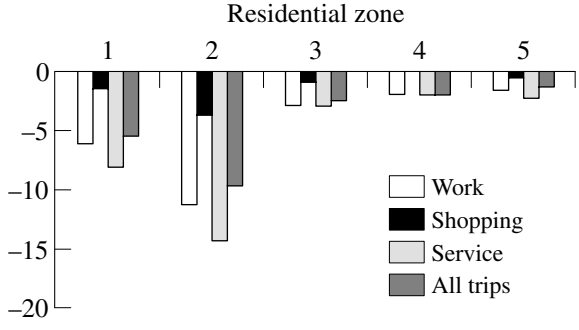


Figure 7.8 *Percentage change in public transport share by trip type and residential zone (ring road scenario with no road pricing compared with the base scenario with no road pricing)*

Figure 7.9 Percentage change in slow mode share by trip type and residential zone (ring road scenario with no road pricing compared with the base scenario with no road pricing)

dents in zone 3. To get to the activities on another ray or to zone 1 or 2 on their own ray, the households in zone 3 have to pass at least one now heavily congested B-link, and therefore they have a higher tendency than other households to use modes other than car. The introduction of the ring road also affects the use of the slow mode (see Figure 7.9). The largest decreases of the slow mode shares occur in zone 2, especially for shopping trips.

Because of the lesser use of public transport in combination with its assumed economies of scale, the public transport travel times increase on all links. For the A-links they increase by slightly more than 1 minute during morning and afternoon peaks, from 14 to 15 minutes and from 12 to 14 minutes, respectively. The increases are quite small on the two outermost links, where the public transport travel times in both scenarios are about 19 to 20 minutes.

All in all, the total time and distance travelled by public transport decrease by 17 and 21 per cent, respectively, compared with the base scenario. For the slow mode, both total time and distance travelled decrease by 18 per cent (they are equal because of the assumed constant speed). The total time and distance travelled by car increase by 25 and 9 per cent, respectively. This indicates that the ring road induces new traffic, both on the ring road and on the B-links, increasing congestion on the latter. As for the route choice, the ring road is used by almost everyone going by car between suburbs that are up to two ring road links apart, during all time periods.

5.2 Impacts of a Ring Road Combined with Congestion Pricing

How will the transport and location impacts change, if the ring road is combined with congestion pricing? Let us first see how the optimal charge level varies by time period and link. The charge levels in Table 7.1 are expressed as per distance unit and hence reflect the congestion situation. Obviously, the A-, B- and ring road links are the most congested ones, particularly during the morning and afternoon peaks.

Combining the ring road with congestion pricing does not affect the location pattern very much except for the shops in the city centre. For most of the other zones and activities

Table 7.1 Optimal congestion charges by time period and link (SEK/km)

Time period	A	B	C	D	Ring
Morning peak	3.1	3.5	1.9	0.8	4.3
Office hours	1.1	1.7	0.4	0.1	1.7
Afternoon peak	2.1	2.2	1.2	0.5	3.2

the changes caused by this pricing scheme are within 1 per cent. This is consistent with the conclusions for the base scenario according to Eliasson and Mattsson (2001). There is, however, a slight damping effect on the strong tendency to locate in the innermost suburbs that was observed for the ring road without road pricing. This is particularly evident for the shops. The previously observed out-moving of the shops from the city centre to the innermost suburbs is almost neutralized by the congestion pricing scheme. Congestion pricing implies that car travel becomes more expensive, and it is therefore quite natural that the location effects of the ring road will become slightly damped. Since the shops are particularly sensitive to car costs, this tendency is more pronounced for this activity than for others.

The most obvious effects of congestion pricing are those on the transport pattern. The morning car travel times on the A-links, for example, are reduced from 22 to 13 minutes. More generally, morning and afternoon peak travel times by car are reduced by 42 to 62 per cent for the A-, B- and the ring road links with the largest decrease for the afternoon peak on the ring road. This strong effect of congestion pricing is also consistent with the results in Eliasson and Mattsson (2001). As noticed in the previous section, the ring road without road pricing reduces car flows and travel times on all but the B-links. When the ring road is combined with congestion pricing, the car flows and travel times on the B-links are also smaller than in the base scenario (the flows are further reduced on all other links, too). As can be expected, the ring road has less impact on travel times when combined with congestion pricing. The morning car travel times on the A-links, for example, decrease from 31 to 22 minutes when the ring road is added to the base scenario without road pricing, whereas the reduction is only from 15 to 13 minutes when comparing the same scenarios combined with congestion pricing.

When the ring road is combined with congestion pricing, the link car shares are 7 to 11 percentage units lower compared with the ring road scenario without road pricing, and 3 to 7 percentage units lower compared with the base scenario without road pricing. Compared with the base scenario with congestion pricing, however, the shares are 3 to 5 percentage units higher for all but the A-links, for which they are about the same.

Table 7.2 shows the impact of congestion pricing on total time and distance travelled by car and public transport with and without a ring road. One interesting observation is that congestion pricing has about the same effect irrespective of whether it is combined with a ring road. Congestion pricing reduces total distance travelled by car by about 25 per cent and total time travelled by car by just below 60 per cent. If we instead compare the ring road scenario with congestion pricing with the base scenario without congestion pricing, then the reduction is about 18 per cent in total distance travelled by car and about 48 per cent in total time travelled by car. For public transport, congestion pricing leads to

Table 7.2 Effect of congestion pricing on total time and distance travelled by car and public transport in the base and the ring road scenarios (percentage change)

	Car		Public transport	
Scenario	Distance	Time	Distance	Time
No ring road	−24	−57	15	10
Ring road	−25	−59	17	12

an increase in total distance travelled by about 15 per cent and in total time travelled by about 10 per cent, both for the base and for the ring road scenarios.

5.3 Impacts of a Ring Road Combined with a Toll Ring

To implement optimal congestion pricing would be a complicated procedure – not the least, technically. As an alternative to congestion pricing, a toll ring is often considered and also sometimes implemented. This is the case for the three largest Norwegian cities (see Small and Gomez-Ibañez 1997). Here we study two alternative locations for a toll ring, either an inner toll ring on the A-links or an outer toll ring on the B-links. In both cases a fixed toll fee of 10 SEK is charged on all car trips in an inbound direction on the tolled links.

Eliasson and Mattsson (2001) concluded for the base scenario that an inner toll ring has a decentralizing, and an outer toll ring a centralizing effect on location. This is still true in the presence of the ring road. Compared with the ring road scenario with no road pricing, an inner toll ring relocates 1 to 3 per cent of the activities in the city centre and about 1 per cent of the activities in the innermost suburbs to the suburbs farther out. An outer toll ring, on the other hand, relocates 0 to 4 per cent of the activities in zone 3 to the city centre and the innermost suburbs (zone 2), while the outer zones are scarcely affected. Compared with the ring road scenario combined with congestion pricing, an inner toll ring leads to fewer activities in the city centre whereas an outer toll ring leads to more activities in the city centre and the innermost suburbs, even if the picture is not uniform for all activities.

When it comes to the transport effects of a toll ring, the first conclusion is that they are essentially local and limited to the car mode. Adding an inner toll ring to the ring road scenario leads to a decrease in the car travel times on the then tolled A-links by 3 to 7 minutes depending on the time period. Also, an outer toll ring leads to the same decreases but for the then tolled B-links. The effects for all other links are usually well below 1 minute. The effects on the public transport travel times are negligible.

Also the effects on the car shares are primarily local. The car shares are 3 percentage units lower on the A-links for an inner toll ring, and 3 percentage units lower on the B-links for an outer toll ring. The public transport and slow mode shares are 1 to 3 percentage units higher for corresponding links, whereas the changes for the rest of the links are small for all modes. Compared with the base scenario with no road pricing, the car shares are still higher on all links for both an inner and an outer toll ring except for the A-links.

Table 7.3 Effect of an inner and an outer toll ring on total time and distance travelled by car and public transport in the base and the ring road scenarios (percentage change)

Scenario	Car		Public transport	
	Distance	Time	Distance	Time
Base scenario and an inner toll ring	−4	−16	0	0
Base scenario and an outer toll ring	−5	−11	1	1
Ring road and an inner toll ring	−4	−9	1	1
Ring road and an outer toll ring	−4	−12	0	0

Table 7.3 shows the impacts of an inner and an outer toll ring on total distance and time travelled by car and public transport for the base scenario and the ring road scenario, respectively. Again, the presence of the ring road does not seem to change the effects of a toll ring very much. Moreover, an inner and an outer toll ring have about the same overall effects on total distance as well as on total time travelled. Compared with the effects of congestion pricing (see Table 7.2), the effects of a toll ring with the assumed toll level of 10 SEK is much more modest (4 per cent reduction in total distance travelled compared with 25 per cent for congestion pricing). The effects of a toll ring on public transport are negligible both with respect to total distance and total time travelled.

5.4 Comparison of the Road Pricing Schemes

It is evident from the previous analysis that congestion pricing reduces car traffic more effectively than a toll ring. One obvious reason is that the costs that are imposed on the car users under congestion pricing are much higher than in either of the two toll ring alternatives. The total daily revenue for the ring road combined with congestion pricing is 18.3 million SEK, whereas it is 2.7 million SEK for an inner toll ring (when the A-links are tolled) and 3.1 million SEK for an outer toll ring (when the B-links are tolled). Would it be possible to attain the same level of car traffic reduction for a toll ring by simply increasing the fee level? Figures 7.10 to 7.12 show, for the ring road scenario, how total toll revenue, total time travelled and total distance travelled by car per day vary with the fee level for an inner and outer toll ring.

Figure 7.10 indicates that the maximum total revenue that it is possible to attain for a toll ring is much lower than for congestion pricing. For an inner toll ring, total revenue reaches a peak of slightly more than 5 million SEK per day at a fee of about 35 SEK. For higher fees the traffic on the A-links decreases to such a degree that the total revenue decreases. The same thing happens for an outer toll ring, but then the total revenue reaches a peak of 8.5 million SEK per day for a fee of about 65 SEK.

The relationship between total revenue and fee level for an inner toll ring may look somewhat peculiar, compared with the smooth curve for an outer toll ring. This is because a toll placed on the A-links affects the division of traffic between the A-links and the ring road in an entirely different manner compared with a toll placed on the B-links.

Figure 7.11 shows the total car travel time per day as a function of the toll fee for the

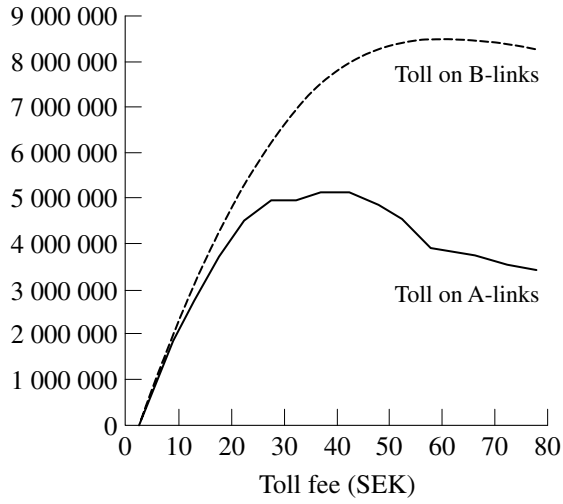


Figure 7.10 Total toll revenue per day (SEK) by level of toll fee for the ring road scenario

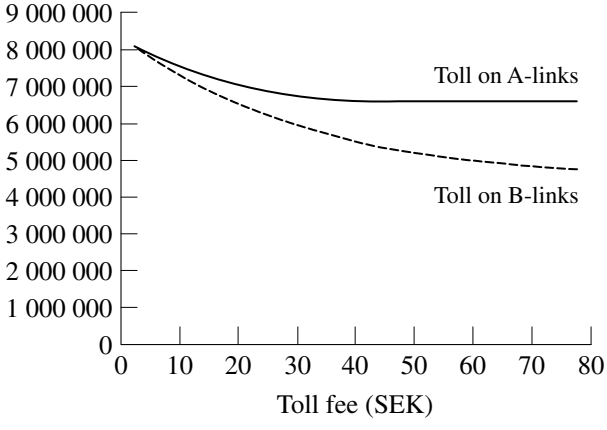


Figure 7.11 Total time travelled by car per day (hr) by level of toll fee for the ring road scenario

two alternative toll rings. At a toll fee of 10 SEK the total travel time is around 740000 hours for an inner toll ring, and 710000 hours for an outer toll ring. The corresponding figure for congestion pricing is as low as 330000 hours. The total time decreases faster for an outer than for an inner toll ring. In the latter case, the total time seems to converge to about 650000 hours. The total time for an outer toll ring seems to converge to about 450000 hours. Therefore, compared with congestion pricing, it seems unlikely that the same figure will be reached by means of a toll ring, irrespective of the chosen fee level.

The total distance travelled by car, on the other hand, can reach the same level as that achieved by congestion pricing. Figure 7.12 shows how the total distance varies for increasing toll fees. At a fee level of about 60 SEK for an outer toll ring, the total distance

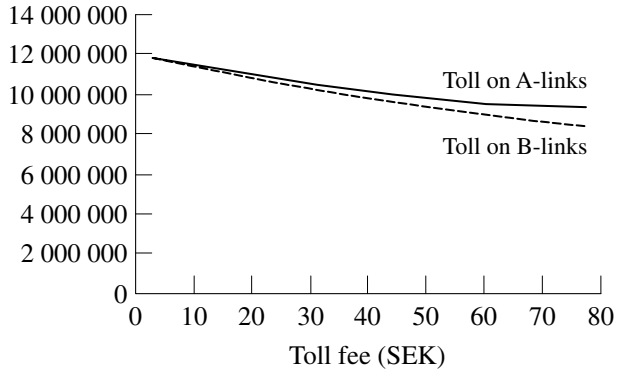


Figure 7.12 Total distance travelled by car (km) per day by level of toll fee for the ring road scenario

travelled reaches 9 000 000 kilometres, which was achieved for congestion pricing. It seems unlikely that the total distance can ever be reduced to that level if a toll is levied on the A-links. At a fee of 75 SEK for an inner toll ring, the total distance travelled is about 9 500 000 kilometres. However, with these very high toll fees, congestion would be severe on many of the other links. For an inner toll ring, the congestion would be worst on the ring road, because of all the traffic that would be redirected from the tolled A-links, but it would also be severe on the B-links. For an outer toll ring, the congestion would be severe on all links except for the tolled B-links.

A conclusion from these experiments is that we cannot attain the same order of reduction of time and distance travelled by car by means of an inner toll ring as is possible by congestion pricing. It is possible to achieve larger reductions by means of an outer toll ring, but the fee levels that would be necessary would probably be unrealistically high. However, if such tolls were imposed, they would change not only the travel pattern but also the location pattern.

6. SUMMARY AND CONCLUSIONS

Congestion is a serious problem in many cities. Road investment and road pricing are two possible policy options to relieve the problem. Cities are very complex systems and the impacts of different policies are difficult to predict. Road investment and road pricing will not only affect the demand for transport in various respects but may also, in the long run, change the location of activities. To be able to evaluate transport policies appropriately, tools that could clarify transport as well as land-use effects of different policies are needed. Eliasson and Mattsson (2001) developed a stylized model of a generic symmetric city for the simulation of such policies. The model was used to evaluate transport and land-use effects of congestion pricing and of an inner and outer toll ring in the road network. In the present study we extend the model and the analysis to the effects of a ring road connecting the innermost suburbs. Different scenarios are constructed in which the ring road is either combined or not with congestion pricing or an inner or outer toll ring.

A *ring road* through the innermost suburbs *without road pricing* leads to a significant relocation of activities to the inner suburbs from those farther out. The increase is 20 per cent or more. Households and workplaces are also attracted to the city centre, whereas shops move out to the inner suburbs. The ring road relieves the congestion in the city centre by enabling new route choice alternatives. On the other hand, the ring road increases in a substantial way the congestion on the links immediately outside the ring road. The car shares increase for trips from all residential zones and especially from the inner suburbs. These increases are at the expense of lower shares for public transport and even more so for the slow mode. In total, the ring road induces new car traffic to the amount of almost 10 per cent.

The effects of *the ring road combined with congestion pricing*, compared with *the ring road scenario with no road pricing*, can be summarized as follows. Congestion pricing *per se* has little effect on location. However, it dampens the attraction of the inner suburbs somewhat. In particular, it dampens the relocation of shops from the city centre. The most notable impact of congestion pricing is its substantial effect on relieving congestion. Car travel flows and travel times are reduced considerably on all links. The decreases are largest for the A- and B-links and for the ring road, for which the travel times go down by 40 to 60 per cent. Car shares are lower on all links, compared both with the ring road scenario without road pricing and with the base scenario. They are, however, not as low as when congestion pricing was applied to the base scenario. The total amount of car travel is reduced considerably: by 25 per cent in distance and almost 60 per cent in time. This leads to a considerable increase in total travel by both public transport and the slow mode. These effects are about the same as when congestion pricing was introduced to the base scenario without a ring road.

Next we turn to the effects of a *ring road combined with a toll ring* with a fixed fee of 10 SEK on inbound traffic, compared with a *ring road scenario with no road pricing*. An inner toll ring placed on the A-links has a decentralizing effect on location, whereas an outer toll ring placed on the B-links has a centralizing effect. This is intuitively reasonable. If the area inside the toll ring is large enough, it is better to be located inside the toll ring to avoid the effects of the tolls, whereas the opposite is true if the area inside the toll ring is too small. The effects of a toll ring on car travel flows and travel times are mainly local, with traffic volumes being reduced on the tolled links. This is also the case for the car shares. An outer toll ring placed on the B-links has, compared with an inner toll ring placed on the A-links, about the same decreasing effect on total car travel time (around 10 per cent), and the same decreasing effect on total car distance (4 per cent). This holds also for the base scenario without a ring road. The effects on public transport are negligible with respect to both total distance and total time travelled.

The total revenue obtained by a toll ring can never reach the same level as that obtained by congestion pricing. The maximum revenue for an outer toll ring is about half as big as for congestion pricing and for an inner toll ring about one-third as big. The total time travelled by car cannot, for any level of toll fee, get as low as that obtained by congestion pricing. The total distance travelled by car can, for an outer toll ring, ultimately be reduced to the same level as for congestion pricing but only at a very high toll fee (60 SEK).

According to our analysis, a ring road through the inner suburbs has a strong effect on the location pattern. Activities from the outer suburbs will to a significant extent be

attracted to the inner suburbs. A ring road will also relieve congestion in the city centre, however, at the cost of increased congestion on the links immediately outside the ring road. Increasing the capacity of these links might reduce the latter effect. The problem is that such an action will induce even more traffic than the ring road in itself would generate. Another alternative that has been studied would be to combine the ring road with some form of road pricing. This offers a possibility of increasing accessibility without also inducing new car traffic. The ring road could in that way lead to a substantial reduction of congestion all over the city. When it comes to the design of a road pricing system, the analysis suggests that a system that approximates congestion pricing is more effective in achieving the desirable effects than a simpler system in the form of a toll ring.

The analysis is built on the application of a model that is consciously kept very simple in many respects. This is to some extent the point of the approach. We have been interested in qualitative relationships rather than the exact magnitude of various effects. However, it would be interesting to develop the model in different ways. One obvious such extension would be to allow public transport on the ring road. Another extension would be to allow the location of the ring road to be flexible. We could then test whether a ring road farther out would relieve congestion in the city centre without increasing congestion on the radial links in the same way as in the present case. Neither of these extensions is straightforward to implement, however, because they will both make the route choice far more complicated than in the present model.

NOTES

* Mattsson's research was given financial support from the Swedish Agency for Innovation Systems. The authors are grateful for helpful comments from two anonymous referees. They would also like to thank Jonas Eliasson, who wrote the computer code of the original version of the simulation model. An earlier version of this chapter was presented at the 42nd European Regional Science Association Congress, Dortmund, 2002.

1. Procedures for Recommending Optimal Sustainable Planning of European City Transport Systems: www-ivv.tuwien.ac.at/projects/prospects.html.
2. The Land Use and Transport Research Cluster: www.ess.co.at/LUTR/.

REFERENCES

Anderstig, C. and L.-G. Mattsson (1991), 'An integrated model of residential and employment location in a metropolitan region', *Regional Science*, **70**, 167–84.

Anderstig, C. and L.-G. Mattsson (1998), 'Modelling land-use and transport interaction: policy analyses using the IMREL model', in L. Lundqvist, L.-G. Mattsson and T.J. Kim (eds), *Network Infrastructure and the Urban Environment: Advances in Spatial Systems Modelling*, Berlin: Springer-Verlag, pp. 308–28.

Ben-Akiva, M. and S.R. Lerman (1985), *Discrete Choice Analysis*, Cambridge, MA: MIT Press.

Boyce, D.E. and M.S. Daskin (1997), 'Urban transportation', in C. Revelle and A.E. McGarity (eds), *Design and Operation of Civil and Environmental Engineering Systems*, New York: Wiley, pp. 277–341.

Eliasson, J. and L.-G. Mattsson (2001), 'Transport and location effects of road pricing: A simulation approach', *Journal of Transport Economics and Policy*, **35**, 417–56.

Johansson, B. and L.-G. Mattsson (1995), 'From theory and policy analysis to the implementation of road pricing: the Stockholm region in the 1990s', in B. Johansson and L.-G. Mattsson (eds),

Road Pricing: Theory, Empirical Assessment and Policy, Boston: Kluwer Academic, pp. 181–204.

Lam, T.N. and G.F. Newell (1967), 'Flow dependent traffic assignment on a circular city', *Transportation Science*, **1**, 318–61.

May, A.D., A. Karlstrom, N. Marler, B. Matthews, H. Minken, A. Monzon, M. Page, P. Pfaffenbichler and S. Shepherd (2003), *Decision Makers' Guidebook*, Deliverable no. 15, PROSPECTS (Procedures for Recommending Optimal Sustainable Planning of European City Transport Systems), Funded by the European Commission, 5th Framework – Energy, Environment and Sustainable Development (EESD).

Minken, H., D. Jonsson, S. Shepherd, T. Järvi, T. May, M. Page, A. Pearman, P. Pfaffenbichler, P. Timms and A. Vold (2003), *A Methodological Guidebook*, Deliverable no. 14, PROSPECTS (Procedures for Recommending Optimal Sustainable Planning of European City Transport Systems), Funded by the European Commission, 5th Framework – Energy, Environment and Sustainable Development (EESD).

Roy, J.R., M.A.P. Taylor, T. Ueda and L.O. Marquez (1998), 'Development of a compact urban simulation model', in L. Lundqvist, L.-G. Mattsson and T.J. Kim (eds), *Network Infrastructure and the Urban Environment: Advances in Spatial Systems Modelling*, Berlin: Springer-Verlag, pp. 344–60.

Schneider, F., A. Nordmann and F. Hinterberger (2002), 'Road traffic congestion: the extent of the problem', *World Transport Policy and Practice*, **8**, 34–41.

Sjölin, L. (2001), 'Introduction of a ring road to a symmetric urban model: analysis of the impacts on transport and location for different road pricing schemes', MSc thesis No. 01-175, Department of Infrastructure and Planning, Royal Institute of Technology, Stockholm.

Small, K.A. and J.A. Gomez-Ibañez (1997), 'Road pricing for congestion management: the transition from theory to policy', in T.H. Oum, J.S. Dodgson, D.A. Hensher, S.A. Morrison, C.A. Nash, K.A. Small and W.G. Waters II (eds), *Transport Economics: Selected Reading*, Amsterdam: Harwood Academic, pp. 373–403.

Wegener, M. (1994), 'Operational urban models: state of the art', *Journal of the American Planning Association*, **60**, 17–29.

Wilson, A.G. (1998), 'Land-use/transport interaction models: past and future', *Journal of Transport Economics and Policy*, **32**, 3–26.

8. Optimal integrated pricing in a bi-modal transportation network

Hai Yang, Qiang Meng and Timothy D. Hau

1. INTRODUCTION

The number of private automobiles on the roads is increasing at a rate faster than our infrastructure program and traffic management schemes can ever hope to accommodate. The change in modal split against public transport causes further deterioration in the level of service for both private and public transport users. Clearly this presently observed modal mix is at least partially the result of a misallocation of resources in transportation, and there is a drastic need to break out of this vicious cycle. Currently, road pricing is being considered as a measure for restraining auto use and providing revenue for public transport improvement, and thus reduces the costs of urban road traffic congestion to achieve optimal modal split. This novel idea relies on the fact that the private automobile user does not pay the full costs of his/her choice: he/she is, in fact, subsidized. In fact, road pricing has gained significant support in recent years, and various intelligent charging technologies such as transponder and smartcard for automatic non-stop payment have been developed (Blythe and Hills 1994). In the case of Hong Kong, the Transport Department has begun a technical feasibility study of an advanced charging system as a novel means of managing traffic and reducing the environmental impact of road traffic. A main goal of the demonstration is to link road pricing and public transport provision via the use of congestion charge revenue to subsidize or improve bus and transit services for optimal modal split (HKTD 1995). In Singapore, the electronic road pricing system has been implemented for several years (www.lta.gov.sg/index.htm).

It is thus conceivable that a new generation of road-use pricing technologies will eventually be introduced into the road network. All these emerging technologies offer efficient tools and challenges to implement integrated transport pricing for traffic management, and also generate a great demand for the development of efficient transport pricing models. The greatest need perhaps is the development of a trans-modal integrated transport pricing system that enables coordination of road tolls and transit fares for optimum modal split. Particularly for congested urban areas like Hong Kong, an efficient trans-modal transport pricing scheme should be developed to capture the spatial and temporal distribution of travel demand.

Road pricing and transit subsidy policy may be attractive to a great majority of users and politically palatable. Their combination can be viewed as offering a carrot for transport users to switch to more socially desirable modes with lower social marginal cost. Apart from this efficiency, transport pricing and subsidy policy can also be justified on the ground of the equity issue, which is often the primary focus of road pricing arguments.

First, this policy puts money directly back into the hands of commuters, while giving them the flexibility to avoid some or all of the higher fees by shifting modes, routes or times of day. Second, suppose the commuters have only two options: travel by auto or travel by transit, the equity in the system can be improved by subsidizing the transit system using the congestion toll revenue. The reason is that the poor will mostly refrain from traveling by auto and will have to make use of transit services when the travel cost to auto users is increased by the optional congestion toll. As a result, the price of transit travel should be lower so as to offset inequities. Consequently, a joint pricing and subsidy policy could possibly lead to a more equitable income redistribution and even possibly benefit every commuter.

Transport pricing has been a subject of many research studies and has been covered in most transport economics monographs or textbooks (for example, Hau 1992a, 1992b; Button 1993; O'Sullivan 1993; Johansson and Mattsson 1994; to name but a few). Most studies considered the determination of both congestion delay and toll independent of the modal split. A few studies for inter-modal transport pricing have been conducted, largely relying upon the economic theory of deterministic utility-maximizing consumers. Optimal prices of private cars and/or transit in the absence of any restrictions on the pricing instruments or the second-best prices for various special cases such as in the presence of imposition of budgetary restrictions are derived, and the welfare distributional effects of different toll regimes are examined (Sherman 1971; Braeutigam 1979; Viton 1983; Jara-Diza and Tudela 1993; Nowlan 1993; Borger et al. 1996, among many others).

On the other hand, although the subsidy issues in transport are often alluded to in the literature, most studies assume that the subsidy is financed by some tax system independent of the transportation activity, the congestion cost and toll are determined independently of the modal split, and the objective is to examine the effects of government subsidies on the public transit output, productivity and income redistribution (Sherman 1972; Frankena 1983; Obeng et al. 1997). Glaister and Lewis (1978) developed a comprehensive modal split model that attempted to combine both congestion costs and subsidies in determining optimal pricing policy. Guria (1987) considered the interdependence of demands for auto and transit in determining the combination of subsidy, fare and service quality of public transit services for optimal allocation of the commuters between the two modes. Some authors explored the combination of pricing and rationing (or strategies similar to rationing, such as special lanes for buses and high-occupancy vehicles), and the conditions under which they might become attractive (Mohring 1972; Small 1983; Daganzo 1995; Yang and Huang 1999). Recently, congestion pricing in general road networks is receiving attention (Yang and Lam 1996; Bergendorff et al. 1997; Yang and Bell 1997; Yang and Huang 1997a,b, 1999; Dial 1999a,b; Oberholzer-Gee and Weck-Hannemann 2002; Paulley 2002). A variety of new problems related to the road-pricing issue are created (Labbe et al. 1998; De Palma and Lindsey 2000; May and Milne 2000; Yang and Meng 2000, 2002; Yang et al. 2002; Yang and Zhang 2002). In addition, solution methods for these road-pricing problems have gained considerable success (Hearn and Ramana 1998; Dial 1999c, 2000; Li 1999, 2002; Brotcorne et al. 2001; Verhoef 2002). However, road congestion cost and toll are still determined independently of the modal split. Miyagi and Suziki (1996) examined a Ramsey price equilibrium model, and its computational procedure in a bi-modal network, rigorous quantitative modeling of integrated trans-modal transport pricing and subsidy policy is not yet fully developed.

This chapter investigates the relationship between trans-modal transport pricing and subsidy policy for optimal modal split and presents optimization models on a bi-modal transportation network. We consider the morning commuters who travel from their origin (home) to their destination (workplace) in a network. Suppose there exist two parallel modes of travel, auto vehicle and transit, between which commuters can switch. Each mode for travel is characterized with a random utility function and at the equilibrium the modal split at aggregate demand level is governed by a logit formula. Then the first question of interest is: what is the optimal modal split for a system optimization (social welfare maximization) and how can the system optimization be achieved?

We would seek trans-modal integrated transport pricing to achieve socially optimal choices of the travel modes. Economic efficiency would require pricing at marginal cost for both transportation services. However, in this case the transit sector will operate at a deficit since marginal cost pricing cannot cover its fixed and variable operating cost. It is, therefore, rational for the government to utilize the revenue generated from road toll charge to subsidize transit. To the extent that subsidization induces travelers to switch from roads to transit, road congestion will be alleviated.

Given the possibility of subsidization, the next question of interest is: what magnitude of the subsidy is required for an optimal shift from road to transit travel, and can this amount be fully covered by the road toll revenue?

In the case of impossibility of perfect transport pricing and subsidy program such as that due to the existence of a transit budget constraint, we then seek the second-best policy options using a bi-level programming approach. In particular, we investigate how the second-best prices and benefits obtained may deviate from social optimum.

2. DEMAND AND EQUILIBRIUM

2.1 The Assumptions

We suppose that the transportation network comprises two types of modes to provide transportation service from origins (homes) to destinations (workplaces). Mode 1 represents a public transit (that is, on-street buses or off-street subways) and mode 2 represents private auto such as car. It is assumed that a certain number of homogeneous commuters could either travel on the highway by car (auto mode) or travel by transit (transit mode) with pedestrian access to reach their workplace. We assume everyone has to go to work every morning, and therefore the total number of commuters during the peak period is fixed irrespective of the modal split. Clearly, auto mode has limited capacity due to its physical conditions, while transit capacity is virtually unlimited by assuming the transit service (frequency and so on) will be provided in response to demand. Furthermore, to keep the model simple, we deal with the situation in which there is no congestion interaction between the two modes, such as transit vehicle travel on dedicated guideways (for example, subway, elevated guideways, reserved bus lanes and so on). These assumptions mean that travel time or level of service by transit mode (including access and egress time) is constant (independent of either the automobile flow or the transit patronage), while travel time by auto mode will monotonically increase with the auto volume on the road.

Let $G=(N,A)$ be an expanded and directed bi-modal transportation network defined

by a set N of nodes and a set A of directed links, $A = A_A \cup A_T$ where A_A and A_T are two subsets of auto mode links and transit mode links. Each auto mode link $a \in A_A$ has an associated flow-dependent cost, $t_a(v_a)$, that denotes the cost per unit flow or average cost on each link (the automobile occupancy factor is assumed to be one). The auto flow cost function, $t_a(v_a)$, $a \in A_A$ is assumed to be differentiable and monotonically increasing with the amount of flow v_a. A transit route connecting an origin–destination (O–D) pair is represented by a separate link $a \in A_T$. Let W denote the set of O–D pairs and R_A^w denote a set of all routes by auto mode between O–D pair $w \in W$.

2.2 The Demand Model

Based on the above assumptions, we now specify the demand for both travel modes. Suppose each mode between each O–D pair is characterized by a random utility function.[1] The utility that a transit mode user receives from his/her trip between O–D pair $w \in W$ can be defined as:

$$U_T^w = U_0^w - \beta T_T^w - P_T^w + \varepsilon_T^w, w \in W \tag{8.1}$$

where U_0^w is a constant term of utility received through trip-making, P_T^w is a flat fare for use of the transit mode, β is the value of time and βT_T^w is the time cost of travel, being a constant specific to the transit mode (including in-vehicle time, access time from home to transit station and egress time from transit station to workplace), ε_T^w represents the uncertainty in specifying the transit mode's utility because of the individual perception of the characteristics of the variants to be heterogeneous or not completely observable due to taste variations across commuters. For simplicity, the total travel and waiting time for a transit user is assumed to be constant (T_T^w is fixed) even if transit frequency is changed slightly.

On the other hand, the utility between O–D pair $w \in W$ that an auto user receives will consist of time cost of travel and auto toll:

$$U_A^w = U_0^w - \beta T_A^w - P_A^w + \varepsilon_A^w, w \in W \tag{8.2}$$

where βT_A^w is the time cost of travel by auto, P_A^w is the auto toll cost, and again ε_A^w is a random variable.

Suppose the random terms (ε_T^w, ε_A^w) with mean zero are identically and independently distributed with a Gumbel probability distribution function, then at equilibrium, the mode split at aggregate demand level is governed by a logit formula specified below (Ben-Akiva and Lerman 1985):

$$Q_T^w = Q^w \frac{\exp(\theta \bar{U}_T^w)}{\exp(\theta \bar{U}_T^w) + \exp(\theta \bar{U}_A^w)}, w \in W \tag{8.3}$$

$$Q_A^w = Q^w \frac{\exp(\theta \bar{U}_A^w)}{\exp(\theta \bar{U}_T^w) + \exp(\theta \bar{U}_A^w)} (= Q_w - Q_T^w), w \in W, \tag{8.4}$$

where:

$$\bar{U}_T^w = U_0^w - \beta T_T^w - P_T^w, w \in W \tag{8.5}$$

$$\bar{U}_A^w = U_0^w - \beta T_A^w - P_A^w,\ w \in W \tag{8.6}$$

represent the systematic utility of travel by transit and auto, respectively, between O–D pair $w \in W$. The positive value of parameter θ (1/$) is related to the standard deviation of the terms ε_T^w and ε_A^w, Q^w is the total demand for travel between O–D pair $w \in W$, Q_T^w is the number of transit commuters, and Q_A^w is the auto volume. As mentioned before, the total demand Q^w, $w \in W$ for the combined transport services is assumed to be perfectly inelastic.

2.3 Transit Fares

Transit systems in most urban areas are operated by private firms and typically supervised by a board of elected public officials. It is natural to assume that the government is the controlling agency that determines the transit fare based on certain criteria (Berechman 1993). Suppose that a lump-sum subsidy, $S_w, w \in W$, is provided to the transit firm, and a flat transit fare for each O–D pair is set so that the fare revenue together with the subsidy is sufficient to cover the fixed and variable operating cost of transit:

$$P_T^w Q_T^w + S_w - \left(I_w + \gamma_w Q_T^w\right) = B_w,\ w \in W, \tag{8.7}$$

where the first term of the left-hand side represents the total revenue generated from fare collection, the first term in the brackets represents the fixed operating costs of transit whereas the second term represents the variable cost that is assumed to be proportional to the number of passengers (for example, as the number of passengers grows, more transit units, fuel, and labor hours are needed).[2] Constant B_w, $w \in W$ represents a predetermined, allowable surplus budget. If $B_w = 0, w \in W$, then the transit firm must at least break even with the fare revenue and government subsidy. Note that the budget constraint (8.7) is given with respect to an individual O–D pair. This could correspond to the situation where the transit service between each O–D pair is provided by different firms. Otherwise, equation (8.7) would be a strong requirement.

From the budget constraint, the transit fare should be set at:

$$P_T^w + \gamma_w + \frac{B_w + I_w - S_w}{Q_T^w},\ w \in W. \tag{8.8}$$

2.4 Transportation Equilibrium

With the mutual exclusive transit network, the network equilibrium problem is simplified as a binary mode choice/traffic assignment problem for a given toll charge and transit fare, and can be formulated as the following convex optimization problem (Florian and Spiess 1983):

$$\min \mathbf{Z}(\mathbf{v},\mathbf{Q}) = \sum_{a \in A} \int_0^{v_a} \left[\beta t_a(\omega) + P_A^a\right] d\omega + \sum_{w \in W} \int_0^{Q_T^w} \left(\frac{1}{\theta} \ln \frac{\omega}{Q^w - \omega} + \beta T_T^w + P_T^w\right) d\omega \tag{8.9}$$

subject to:

$$\sum_{r \in R_A^w} f_r^w = Q_A^w,\ w \in W \tag{8.10}$$

$$Q_A^w + Q_T^w = Q^w,\ w \in W \tag{8.11}$$

$$f_r^w \geq 0,\ Q_T^w \geq 0,\ w \in W,\ r \in R_w, \tag{8.12}$$

where $v_a = \Sigma_{w \in W} \Sigma_{r \in R_w} f_r^w \delta_{ak}^w$, $a \in A_A$ and $\mathbf{v}$ is a vector of all road link flows, f_r^w denotes auto flow on path r between O–D pair $w \in W$, and $\delta_{ar}^w = 1$ if link a is in path r between O–D pair $w \in W$, and $\delta_{ar}^w = 0$ otherwise. P_A^a is the auto toll charge on road link a. If the congestion toll is set at marginal external cost, then $P_A^a = \beta v_a \left[\partial t_a(v_a)/\partial v_a\right]$, $a \in A_A$. Also, if the transit fare is set on a self-supporting basis, then $S_w = 0$, $w \in W$ and P_T^w is given by equation (8.8). Therefore, the equilibrium modal share $\mathbf{Q} = (Q_A^w, Q_T^w)$, $w \in W$ can be determined for any given values of $(Q_w, \beta, \gamma_w, B_w, I_w, S_w, T_T^w)$, $w \in W$ under a specific pricing scheme.

Note that one would be able to deal with the general situation where there exists a general road and a general transit network with flow interaction between the two modes, and formulate the problem as a variational inequality. This more complex situation does not add much more insight into our network treatment of the subject, and is not further pursued here.

3. FIRST-BEST PRICING

We next consider how to determine the congestion toll and transit fare for a system optimum. The system optimum can be defined as the maximization of the social welfare that can be measured as the total user benefit minus total cost. Different from the previous studies, the logit representative travel behavior approach (Oppenheim 1995) with micro-economic justification is used to drive the optimal pricing.

Define the expected indirect utility received by an individual as $EU^w = E\left[\max\left(U_A^w, U_T^w\right)\right]$, $w \in W$, which is the expected indirect utility an individual traveler enumerated at random receives on average from his/her repeated utility maximizing selections. It is easy to show that, with the modal split governed by the logit models (8.3) and (8.4), EU^w can be written as (Williams 1977; Hau 1987):

$$EU^w = \frac{1}{\theta} \ln\left[\exp\left(\theta \bar{U}_A^w\right) + \exp\left(\theta \bar{U}_T^w\right)\right],\ w \in W, \tag{8.13}$$

where $\bar{U}_A^w$ and $\bar{U}_T^w$ are the systematic components of the utility functions defined by (8.5) and (8.6), respectively. The expected utility in (8.14) is a measure of consumer surplus attributable to the differentiated transport services since the quasi-linear utility specification yields marginal utility of income as unity. In other words, utility is synonymous with income or dollars.

First, we assume that the conditions for the existence of a 'representative traveler' are met (Anderson et al. 1988; Oppenheim 1995). Furthermore, in the absence of income effects the demand system generated by the logit model can be associated with a single consumer maximizing a deterministic utility, namely, the representative traveler's direct utility function or the utility at aggregate demand level. This direct utility function for transportation can be expressed as:

$$DU = \sum_{w \in W} \left[-\frac{1}{\theta} \left(Q_A^w \ln Q_A^w + Q_T^w \ln Q_T^w \right) + \frac{1}{\theta} Q^w \ln Q^w \right], \tag{8.14}$$

subject to $Q_A^w + Q_T^w = Q^w$, $w \in W$ (for a detailed derivation, the reader might refer to Anderson et al. 1988, or Oppenheim 1995). The representative traveler's direct utility can be regarded as a measure of the user benefit from travel at aggregate demand level. In fact the direct utility function (8.15) consistent with the logit model is an entropy-type function which has been used as a benefit measure in terms of interactivity in trip distribution (Erlander and Stewart 1990).

On the other hand, the total social cost can be simply calculated as:

$$TC = \left[\sum_{a \in A_A} \beta v_a t_a(v_a) \right] + \left[\sum_{w \in W} \left(\beta Q_T^w T_T^w + I_w + \gamma_w Q_T^w \right) \right]. \tag{8.15}$$

The first term represents the total time cost of travel by auto mode, and the second term represents the sum of the time cost of travel by transit mode and the total operating cost of the transit system.

Therefore, the social welfare, $W(\mathbf{v},\mathbf{Q})$, can be measured as:

$$W(\mathbf{v},\mathbf{Q}) = DU - TC. \tag{8.16}$$

By deleting the constant term and changing the sign of the objective function, the social welfare maximizing problem can be characterized by the following nonlinear minimization problem:

$$\min\left[-W(\mathbf{v},\mathbf{Q}) \right] = \sum_{w \in W} \frac{1}{\theta} \left(Q_A^w \ln Q_A^w + Q_T^w \ln Q_T^w \right) + \sum_{a \in A_A} \beta v_a t_a(v_a) + \sum_{w \in W} \left(\beta Q_T^w T_T^w + I_w + \gamma_w Q_T^w \right) \tag{8.17}$$

subject to constraints (8.10)–(8.12).

Now we show that maximizing net benefit leads to the well-known marginal-cost pricing rules. Note that the objective function (8.17) is strictly convex, thus the following Kuhn–Tucker conditions for any $Q_A^w > 0$, $Q_T^w > 0$ are also sufficient to obtain a unique optimal solution:

$$\beta T_T^w + \gamma_w + \frac{1}{\theta} \left(\ln Q_T^w + 1 \right) - \lambda_w = 0, \; w \in W \tag{8.18}$$

$$\beta \mu_A^w + \frac{1}{\theta} \left(\ln Q_A^w + 1 \right) - \lambda_w = 0, \; w \in W, \tag{8.19}$$

where μ_A^w and λ_w are the Lagrange multipliers associated with the constraints (8.10) and (8.11), and μ_A^w also represents the shortest social marginal cost between O–D pair $w \in W$, as given below:

$$\sum_{a \in A_A} \bar{t}_a(v_a) \delta_{ar}^w = \mu_A^w, \text{ if } f_r^w > 0 \tag{8.20}$$

$$\sum_{a\in A_A} \bar{t}_a(v_a)\delta_{ar}^w \geq \mu_A^w, \text{ if } f_r^w = 0, \tag{8.21}$$

where:

$$\bar{t}_a(v_a) = t_a(v_a) + v_a \frac{\partial t_a(v_a)}{\partial v_a}. \tag{8.22}$$

Evidently, the first term, βT_T^w, of the right-hand side of (8.18) is the actual time cost of travel incurred by a transit commuter and the second term, γ_w is the marginal transit operation cost for an additional commuter. Similarly, the first term in (8.22) is the actual link travel time incurred by a traveler and the second term is the additional travel time that a traveler imposes on all other travelers in the road. Equations (8.20) and (8.21) mean that user equilibrium is satisfied if each auto user is facing his/her social travel cost, and this equilibrium can be achieved if each link is charged at social marginal cost. Namely,

$$P_A^a = \beta v_a \frac{\partial t_a(v_a)}{\partial v_a}, \; a \in A. \tag{8.23}$$

In addition, from (8.18) and (8.19), we have:

$$\ln Q_T^w = \theta\left(\lambda_w - \beta T_T^w - \gamma_w\right) - 1, \; w \in W \tag{8.24}$$

$$\ln Q_A^w = \theta\left(\lambda_w - \beta \mu_A^w\right) - 1, \; w \in W. \tag{8.25}$$

Solving these equations together with $Q_A^w + Q_T^w = Q^w$, $w \in W$ will yield the following logit model:

$$Q_T^w = Q^w \frac{\exp\left[\theta\left(-\beta T_T^w\right) - \gamma_w\right]}{\exp\left[\theta\left(-\beta T_T^w\right) - \gamma_w\right] + \exp\left[\theta\left(-\beta \mu_A^w\right)\right]}, \; w \in W \tag{8.26}$$

$$Q_A^w = Q^w \frac{\exp\left[\theta\left(-\beta \mu_A^w\right)\right]}{\exp\left[\theta\left(-\beta T_T^w\right) - \gamma_w\right] + \exp\left[\theta\left(-\beta \mu_A^w\right)\right]}, \; w \in W. \tag{8.27}$$

By comparing equations (8.26) and (8.27) with their counterparts (8.3) and (8.4) with the deterministic utility components (8.5) and (8.6), we find that the exact correspondence between (8.26) and (8.3) or (8.27) and (8.4) requires that, in addition to the marginal cost-based toll charge on the road network, the transit fare is determined as:

$$P_T^w - \gamma_w, w \in W. \tag{8.28}$$

From the above derivation, we can easily observe that, without the budgetary constraint (8.7), the optimality conditions (8.19)–(8.23) for maximization of the social welfare function $W(\mathbf{v},\mathbf{Q})$ are also the conditions that govern the logit-based modal split models (8.3) and (8.4). This implies that the social welfare maximization problem can be supported as a logit-based mode choice/network equilibrium problem. The only requirement for this equivalence is that the transit fare and road link toll should be charged with an amount equal to the additional cost or user externalities, respectively. In other words, the conventional marginal-cost pricing (first-best pricing) should be applied. This would ensure that

the users' optimal private choices will also be optimal social choices that lead to the efficient utilization of the transportation services.

4. OPTIMAL TRANSIT SUBSIDY

As shown above, the system optimization requires transport pricing at social marginal cost for both auto and transit commuters. For each O–D pair, the optimal transit fare is equal to $\gamma_w, w \in W$. It is clear that under this marginal cost fare, the budget constraint (8.7) is not satisfied if $S_w = 0, w \in W$ (without subsidy). The fare revenue can just cover the variable transit operating cost and the transit sector will operate at a deficit which is equal to its fixed operating cost, $I_w, w \in W$. In this case an appropriate subsidy for transit is essential to maintain operation of the transit system at system optimum.

We now consider how to link the road toll revenue and the optimal transit subsidy via the transit budget constraint (8.7). Suppose that at the system optimum the optimal auto link volume is $v_a^*, a \in A$ and transit ridership between O–D pair $w \in W$ is Q_T^{w*}, the revenue in dollars generated from road-use charge is thus given by:

$$R = \sum_{a \in A} P_A^{a*} v_a^* = \sum_{a \in A_A} \beta v_a^{*2} \frac{\partial t_a(v_a)}{\partial v_a}\bigg|_{v_a = v_a^*}. \tag{8.29}$$

The required total amount of subsidy for transit at system optimum is determined by:

$$S = \sum_{w \in W} S_w = \sum_{w \in W} (I_w + \gamma_w Q_T^{w*}) + \sum_{w \in W} B_w - \sum_{w \in W} \gamma_w Q_T^{w*} = \sum_{w \in W} (B_w + I_w), \tag{8.30}$$

where $\Sigma_{w \in W} \gamma_w Q_T^{w*}$ is the total revenue generated at optimal transit fare $P_T^{w*} = \gamma_w, w \in W$. Thus the optimal subsidy is equal to the sum of the budget surplus and the fixed transit operating cost. Note that, at marginal-cost pricing, the optimal subsidy is a constant independent of modal split, while the road toll revenue will generally increase with the level of demand. If $R \geq S$, then the road toll revenue is sufficient to cover the transit subsidy, this will be true as travel demand grows beyond a certain level. In this case, a simple reallocation of income within the system would lead to a precisely optimal modal split and would even make all commuters better off without the requirement of input of any system-external resources. Note that oversubsidization for transit would lead to an excessive transit ridership, resulting in an increase in total social cost. On the other hand, if $R < S$, then the toll revenue cannot cover the transit deficit gap and the system optimum cannot be achieved if there is no subsidy from external sources. This renders the fact that transit is worth building only in large cities.

5. SECOND-BEST PRICING

In the presence of a perfect transport pricing instrument, the optimal auto toll and transit fare can be determined straightforwardly from the road use externality and marginal transit use cost that appears in the optimality conditions of system optimum. However, in reality transport pricing may be subject to various restrictions and the first-best pricing

environment may not be available. In this case the optimal auto toll and/or transit fare have to be determined under given constraints and subject to the commuters' behavior response in terms of route and mode changes. This second-best social optimum can be obtained by the following bi-level welfare maximization problem:

$$\min_{\mathbf{P}} \{-W[\mathbf{v}(\mathbf{P}), \mathbf{Q}(\mathbf{P})]\}, \tag{8.31}$$

subject to:

$$\mathbf{G}[\mathbf{v}(\mathbf{P}), \mathbf{Q}(\mathbf{P})] \leq 0, \tag{8.32}$$

where road link flow $\mathbf{v}(\mathbf{P})$ and mode share $\mathbf{Q}(\mathbf{P})$ are determined by solving the binary mode choice/traffic assignment problem (8.9)–(8.12) for a given road toll and transit fare pattern $\mathbf{P} = [\ldots, P_A^a, \ldots, \ldots, P_T^w, \ldots]$. Note that $\mathbf{G}$ is a vector of the upper-level constraints. It could include the break-even constraint (8.7), upper or lower bound of auto link tolls and/or transit fare if any, or zero value of auto link tolls if road congestion pricing is not introduced.

The bi-level programming problem has been applied extensively in transportation modeling and optimization by Yang and his co-authors (for example, Yang and Lam 1996; Yang and Bell 1997). Sensitivity analysis-based algorithms (Yang 1997) or some direct search algorithm without using derivatives such as the method of Hooke and Jeeves (Bazaraa et al. 1993) could be applied to solve the above bi-level optimization problem.

6. GRAPHICAL REPRESENTATION

In what follows we explain graphically the aforementioned transport pricing and subsidy problem using a simple network. The network consists of a parallel transit route and a highway route connecting a single O–D pair (Figure 8.1). Figure 8.2 presents such a graphical interpretation where for clarity linear demand functions are plotted.

Without transit budgetary constraint, the marginal-cost pricing will give rise to the optimal auto volume Q_A^2 (where the demand curve intersects the social trip cost curve) and the optimal transit ridership is Q_T^2 (where the demand curve with congestion toll intersects the marginal social cost curve). This optimal modal split will lead to the maximization of social welfare.

When there is no congestion toll or transit subsidy, the auto volume is Q_A^1. The underpricing of auto travel shifts the demand curve for transit downward. The equilibrium is shown by point H with a transit ridership Q_T^1 where the demand curve intersects the isobudget curve. Note that the two candidates of equilibrium points, T and F, are unstable against small perturbations, and thus are ignored here. Clearly, in the absence of both congestion toll and transit subsidy, the auto volume exceeds the optimum volume and the transit ridership is less than the optimum ridership.

In the presence of a congestion toll, but absence of a transit subsidy, the automobile volume is Q_A^3 and the transit ridership is Q_T^3. Therefore, auto pricing will cause some auto drivers to shift to transit, and both market shares move closer to their optimum points. It can be observed that auto pricing alone narrows, but does not eliminate, the gaps between

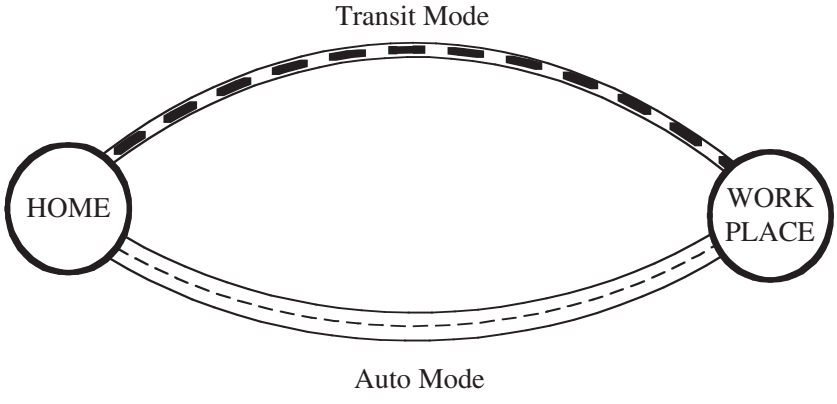


Figure 8.1 A schematic representation for the morning commute (a single O–D pair connected by two travel modes)

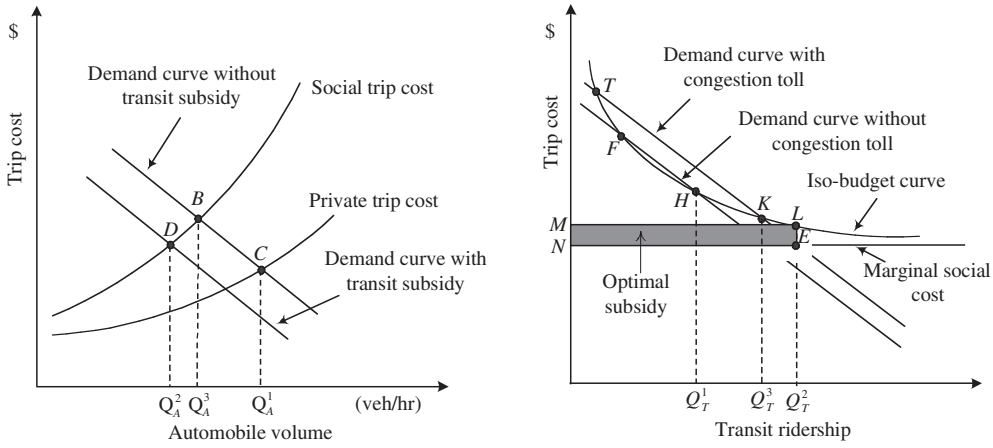


Figure 8.2 Transport pricing and subsidy policy for optimal modal split between two alternative travel modes

the system optimum and equilibrium volumes. As shown in the figures, Q_A^3 is still more than Q_A^2 and Q_T^3 is still less than Q_T^2.

As shown before, the system optimum does not satisfy the budget target constraint (8.7), and the operational subsidy for transit is essential. The required subsidy for transit is equal to the shaded area *MNEL*. Generally, any small transit subsidy up to the amount *MNEL* will drive the equilibrium points to move closer to the system optimum by diverting some users from automobile to transit. However, further subsidy may induce ridership above its optimum level, and the benefit of the transit subsidy generated from the diversion of drivers from congested roads may become smaller than the cost due to excessive transit ridership.

7. A NUMERICAL EXAMPLE

Now we present a numerical example to demonstrate the various transport pricing and subsidy schemes. We seek the first-best pricing that maximizes net social benefit and discuss how the toll/fare and the benefit obtained may deviate from the first-best solution when perfect pricing is not available and thus a system optimal modal split is not established.

Figure 8.3 shows the bi-modal transportation network used in the numerical example, which consists of auto and transit modes. Nodes 1∼5 are centroids (both origins and destinations), each O–D pair is connected directly by a transit link. Travel time on each auto link depends on the link flow $v_a, a \in A$ on that link only and is given by the BPR (Bureau of Public Road) formula:

$$t_a(v_a) = t_a^0 \left[1 + 0.15 \left(\frac{v_a}{c_a} \right)^4 \right] \tag{8.33}$$

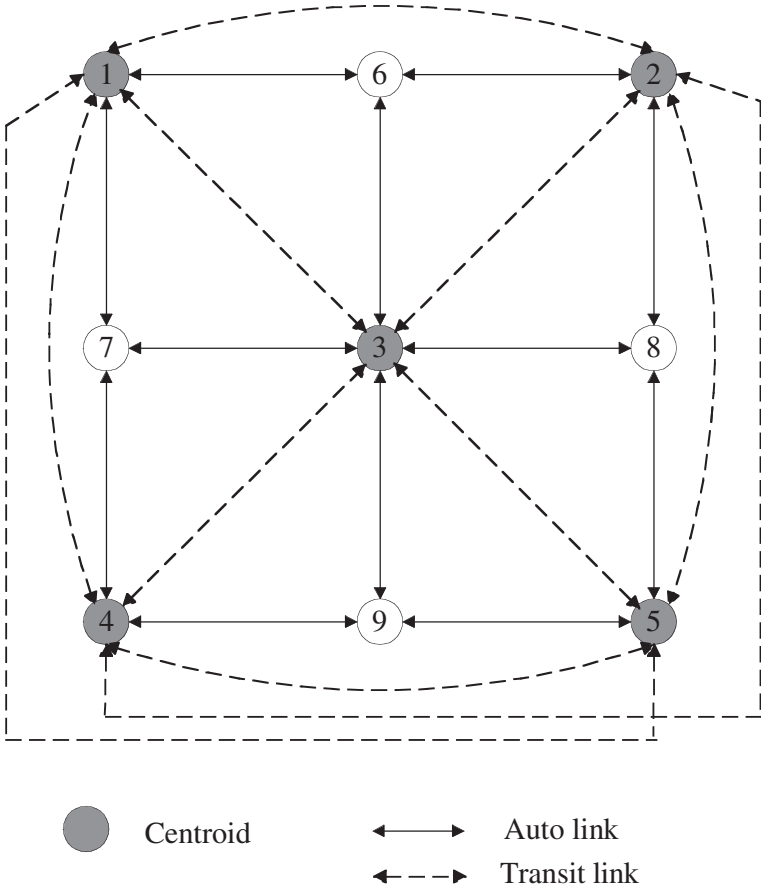


Figure 8.3 The bi-modal transportation network used for numerical calculation

Table 8.1 Input data for auto link cost function of the network shown in Figure 8.3

Link	Free-flow travel time (min)	Capacity (veh/hr)	Link	Free-flow travel time (min)	Capacity (veh/hr)
1–6	8.5	2500	6–1	8.5	2500
1–7	8.0	1800	6–2	9.0	2000
2–6	9.0	2000	6–3	7.0	1400
2–8	9.0	1800	7–1	8.0	1800
3–6	7.0	1400	7–3	10.0	1100
3–7	10.0	1100	7–4	6.0	2500
3–8	7.0	1500	8–2	9.0	1800
3–9	6.5	1600	8–3	7.0	1500
4–7	6.0	2500	8–5	11.0	1650
4–9	9.0	1700	9–3	6.5	1600
5–8	11.0	1650	9–4	9.0	1700
5–9	10.5	2000	9–5	10.5	2000

where link free-flow travel time t_a^0 and link capacity c_a are reported in Table 8.1. The basic transit data including transit travel time (assumed to be constant), fixed and unit variable operating cost between each O–D pair are shown in Table 8.2. For other parameters pertaining to transit, we set $B_w = 0$, $w \in W$ and $\theta = 1.00$ (1/\$), $\beta = 0.20$ (\$/min). Finally, the total travel demand in terms of number of trips per hour by both modes between each O–D pair is fixed and given in Table 8.3.

Here we consider the following six cases:

1. the benchmark do-nothing (auto-unpriced and transit-unsubsidized equilibrium) case, where transit service is charged at average cost so the transit firm will break even;
2. the optimum system under marginal-cost pricing on both road and transit (first-best case), in which we shall examine whether the road toll revenue can cover transit deficit;
3. marginal-cost based auto toll versus average-cost-based transit fare;
4. optimal auto toll versus average-cost-based transit fare;
5. marginal-cost pricing on transit alone, in which the transit will also operate at a deficit;
6. optimal pricing on transit alone, in which the transit firm will also operate at a deficit.

Note that Case 3 differs from Case 4 and Case 5 differs from Case 6, respectively. The reason is that the marginal-cost pricing criterion on a single mode as suggested by the conventional analysis is not necessarily optimal when the two modes are interdependent and integrated. Taking Case 3 (marginal-cost-based auto toll) and Case 4 (optimal auto toll) as examples, the conventional approach assumes a movement along the demand curve and flow distribution on alternative routes while determining the optimal auto congestion tolls. However, in the current case where the total number of peak period trips is assumed to be fixed, the commuters priced out of road services are expected to move to public transit services. Thus because of the integration of transport modes, the new equilibrium

Table 8.2 Input data for transit mode of the network shown in Figure 8.3

O–D pair w	Travel time by transit T_T^w (min)	Fixed transit operating cost I_w (\$)	Variable transit operating cost γ_w (\$/person)
1–2	27	1300	0.3
1–3	25	1200	0.3
1–4	22	1000	0.2
1–5	46	1800	0.6
2–1	27	1300	0.3
2–3	21	1000	0.3
2–4	45	1800	0.6
2–5	27	1400	0.3
3–1	25	1200	0.3
3–2	21	1000	0.3
3–4	23	1100	0.3
3–5	24	1100	0.3
4–1	22	1000	0.2
4–2	45	1800	0.6
4–3	23	1100	0.3
4–5	27	1300	0.3
5–1	46	1800	0.6
5–2	27	1300	0.3
5–3	24	1100	0.3
5–4	27	1300	0.3

Table 8.3 Total travel demand (trips/hr) between each O–D pair for the network of Figure 8.3

O/D pair	1	2	3	4	5
1	–	2500	2300	1900	2300
2	2500	–	1850	2400	2050
3	2300	1850	–	2000	1800
4	1900	2400	2000	–	2150
5	2300	2050	1800	2150	–

after road congestion toll is imposed is reached by both shift of the demand curve as well as movement along the demand curve. In this case a road toll pattern that considers both route and mode choices will result in higher social welfare than a marginal-cost-based road toll pattern does.

Table 8.4 presents the numerical results of the transit fares under each policy option. From this table we can observe that when no congestion toll is charged on the road and no subsidy is provided for transit (the do-nothing case), a high transit fare (generally higher than all other cases) must be charged in order to cover the transit operating cost (or break even) because a large percentage of commuters will rely on auto mode. This high

Table 8.4 Transit fares ($) under alternative pricing and subsidy schemes in the numerical example

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
O–D pair	The benchmark do-nothing case	System optimum under marginal-cost pricing	Marginal-cost-based auto toll vs average-cost-based transit fare	Optimal auto toll vs average-cost-based transit fare	Marginal-cost pricing on transit alone	Optimal pricing on transit alone
1–2	1.798	0.300	1.307	1.295	0.300	0.000
1–3	1.554	0.300	1.905	1.081	0.300	0.000
1–4	1.867	0.200	1.819	0.827	0.200	0.000
1–5	2.205	0.600	2.115	1.507	0.600	0.000
2–1	1.771	0.300	1.307	1.257	0.300	0.000
2–3	1.459	0.300	1.346	1.131	0.300	0.000
2–4	2.475	0.600	2.052	1.437	0.600	0.000
2–5	2.404	0.300	1.312	1.469	0.300	0.000
3–1	1.557	0.300	1.905	1.060	0.300	0.000
3–2	1.461	0.300	1.346	1.134	0.300	0.000
3–4	1.846	0.300	1.992	0.868	0.300	0.000
3–5	2.313	0.300	1.483	1.172	0.300	0.000
4–1	1.866	0.200	1.819	0.827	0.200	0.000
4–2	2.501	0.600	2.052	1.436	0.600	0.000
4–3	1.832	0.300	1.992	0.869	0.300	0.000
4–5	1.664	0.300	1.196	1.346	0.300	0.000
5–1	2.223	0.600	2.115	1.506	0.600	0.000
5–2	2.091	0.300	1.234	1.401	0.300	0.000
5–3	2.367	0.300	1.483	1.168	0.300	0.000
5–4	1.683	0.300	1.196	1.373	0.300	0.000

transit fare will lower the transit ridership, making the system even worse. Without a budget constraint, the socially optimal transit charge should be below the marginal social cost of the service provided in order to offset the overuse of roads, if no toll for auto is charged for road use. In fact the optimal transit-alone fare (Case 6) in the example should be negative but here a non-negativity condition of the transit fare is imposed. This implies that if the automobile does not pay its marginal social costs, then the correct policy is to adjust the fare on transit users downwards (prices of transit users should be lower than the marginal social cost and of course below the average cost, and thus call for a subsidy). In this case the split of travel demand between modes will be adjusted toward its optimal level.

Table 8.5 presents the numerical results of optimal road tolls under each policy option. It is noteworthy that the auto tolls are not necessarily higher than the marginal-cost pricing toll when transit use is charged at average cost so that the transit firm must break even in order to attain self-financial viability. The reason is that Case 4 (optimal auto toll versus average-cost-based transit fare) attempts to induce more travelers to switch from auto to transit mode (moving closer to social optimum) than Case 3 (marginal-cost-based auto toll versus average-cost-based transit fare) does. As more travelers move to transit mode, the required average-cost-based transit fare to satisfy break-even condition will decrease, and thus the auto tolls have to be adjusted downward correspondingly to maintain an optimal mode share. Indeed this can be seen from Table 8.4 where the average-cost-based transit fare in Case 4 is generally lower than that in Case 3.

The above observations show that in the case of the impossibility of perfect integrated trans-modal transport pricing, marginal-cost pricing on a single mode alone may not necessarily be an efficient solution due to the existence of the interdependence of the demands for auto use and transit ridership.

Now we turn to Table 8.6, which shows the percentage changes in modal split in comparison with the do-nothing case. As expected, each policy option will cause modal shift toward system-optimal modal split. It can be observed that Case 4 (optimal auto toll versus average-cost-based transit fare) leads to excessive shift from auto to transit or auto use is overpenalized in relation to the system optimum. In contrast, in Cases 5 and 6 auto mode is still overused (underpriced or subsidized due to free toll) in relation to the system optimum.

Table 8.7 shows the marginal improvements on the do-nothing case, as measured by the percentage changes in users' benefit, total social costs and total social welfare (users' benefit minus total social cost). Note that all policy options give rise to an increase in social welfare and a decrease in total social cost, and of course different options lead to different degrees of improvement. The user benefit measured in terms of the direct utility function either increases or decreases, depending on the specific scheme.

Now we shall discuss the system optimum or the first-best solution in more detail. The system optimum or the optimal modal split is achieved if both road and transit are charged at their respective marginal-cost levels. As stated above, the revenue generated from marginal-cost pricing of transit can only cover its variable operating cost. The transit firm will operate at a deficit that is equal to its fixed operating cost. It is thus interesting to see whether the road pricing revenue can sufficiently cover this deficit gap. Note that the relative magnitudes of auto toll revenue and transit deficit depend on the total travel demand. Figure 8.4 plots the change of the ratio (R/S) of the total road toll revenue

Table 8.5 Auto link tolls ($) under alternative pricing and subsidy schemes in the numerical example

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Link	The benchmark do-nothing case	System optimum under marginal-cost pricing	Marginal-cost-based auto toll vs average-cost-based transit fare	Optimal auto toll vs average-cost-based transit fare	Marginal-cost pricing on transit alone	Optimal pricing on transit alone
1–6	0.000	1.326	2.359	0.695	0.000	0.000
1–7	0.000	1.761	5.887	0.585	0.000	0.000
2–6	0.000	0.850	0.847	1.273	0.000	0.000
2–8	0.000	1.596	1.806	0.411	0.000	0.000
3–6	0.000	2.131	2.233	0.780	0.000	0.000
3–7	0.000	4.709	10.020	0.000	0.000	0.000
3–8	0.000	2.148	2.821	0.015	0.000	0.000
3–9	0.000	3.683	6.919	0.000	0.000	0.000
4–7	0.000	0.501	0.575	0.000	0.000	0.000
4–9	0.000	1.504	2.650	0.285	0.000	0.000
5–8	0.000	0.958	0.535	2.488	0.000	0.000
5–9	0.000	0.762	1.338	0.316	0.000	0.000
6–1	0.000	1.326	2.359	0.585	0.000	0.000
6–2	0.000	0.850	0.847	1.273	0.000	0.000
6–3	0.000	2.131	2.233	0.868	0.000	0.000
7–1	0.000	1.761	5.887	0.001	0.000	0.000
7–3	0.000	4.709	10.020	0.000	0.000	0.000
7–4	0.000	0.501	0.575	0.000	0.000	0.000
8–2	0.000	1.596	1.806	0.425	0.000	0.000
8–3	0.000	2.148	2.821	0.014	0.000	0.000
8–5	0.000	0.958	0.535	2.460	0.000	0.000
9–3	0.000	3.683	6.919	0.000	0.000	0.000
9–4	0.000	1.504	2.650	0.316	0.000	0.000
9–5	0.000	0.762	1.338	0.285	0.000	0.000

Table 8.6 The percentage change (comparison with do-nothing case) in mode split between each O–D pair under alternative pricing and subsidy schemes in the numerical example

	Case 2		Case 3		Case 4		Case 5		Case 6	
O–D pair	System optimum under marginal-cost pricing		Marginal-cost-based auto toll vs average-cost-based transit fare		Optimal auto toll vs average-cost-based transit fare		Marginal-cost pricing on transit alone		Optimal pricing on transit alone	
	Auto volume	Transit ridership	Auto volume	Transit ridership	Auto volume	Transit ridership	Auto volume	Transit ridership	Auto volume	Transit ridership
1–2	−34.61	65.07	−25.96	48.80	−26.86	50.49	−12.80	24.06	−13.20	24.82
1–3	−21.60	30.31	15.60	−21.89	−43.18	60.59	−8.26	11.59	−8.64	12.12
1–4	−16.24	35.20	−1.37	2.96	−76.64	166.09	−17.00	36.84	−16.05	34.79
1–5	−23.30	24.47	−5.64	5.92	−73.29	76.97	0.02	−0.02	−7.44	7.81
2–1	−33.96	62.09	−25.22	46.11	−29.39	53.74	−11.94	21.82	−12.34	22.57
2–3	−26.35	30.13	−9.39	10.74	−34.47	39.42	−8.51	9.73	−13.02	14.88
2–4	−11.89	17.84	−19.44	29.17	−82.73	124.11	−7.51	11.27	−9.07	13.61
2–5	−41.83	87.06	−51.87	107.95	−38.46	80.03	−19.29	40.14	−20.63	42.94
3–1	−21.76	30.68	15.37	−21.66	−46.41	65.43	−8.45	11.91	−8.82	12.44
3–2	−26.49	30.42	−9.56	10.98	−34.12	39.18	−8.68	9.97	−13.18	15.14
3–4	−11.04	20.00	4.76	−8.61	−95.07	172.19	−15.38	27.85	−17.97	32.54
3–5	−24.09	55.27	−30.60	70.19	−57.02	130.82	−21.27	48.79	−23.74	54.46
4–1	−16.23	35.14	−1.35	2.91	−76.63	165.97	−16.98	36.78	−16.04	34.73
4–2	−12.67	19.44	−20.16	30.93	−83.09	127.47	−8.33	12.78	−9.88	15.15
4–3	−10.60	18.92	5.29	−9.44	−94.89	169.45	−14.95	26.70	−17.55	31.35
4–5	−47.72	59.94	−41.65	52.32	−24.23	30.44	−8.97	11.26	−9.82	12.33
5–1	−24.11	25.88	−6.64	7.12	−73.69	79.10	−1.04	1.12	−8.42	9.04
5–2	−39.18	71.51	−49.68	90.66	−34.40	62.79	−15.61	28.49	−17.02	31.05
5–3	−24.94	59.42	−31.37	74.74	−58.03	138.23	−22.15	52.77	−24.60	58.59
5–4	−48.27	62.10	−42.27	54.38	−22.47	28.91	−9.92	12.77	−10.76	13.85

Table 8.7 System performance measures in terms of the percentage improvement on the do-nothing case under alternative pricing and subsidy schemes in the numerical example

	Case 2	Case 3	Case 4	Case 5	Case 6
O–System performance measure pair	System optimum under marginal-cost pricing	Marginal-cost-based auto toll vs average-cost-based transit fare	Optimal auto toll vs average-cost-based transit fare	Marginal-cost pricing on transit alone	Optimal pricing on transit alone
Total user benefit	1.92	0.24	−20.48	4.29	4.36
Total social cost	−12.05	−6.19	−10.75	−7.82	−9.12
Total net benefit	15.14	7.57	8.76	10.76	11.95

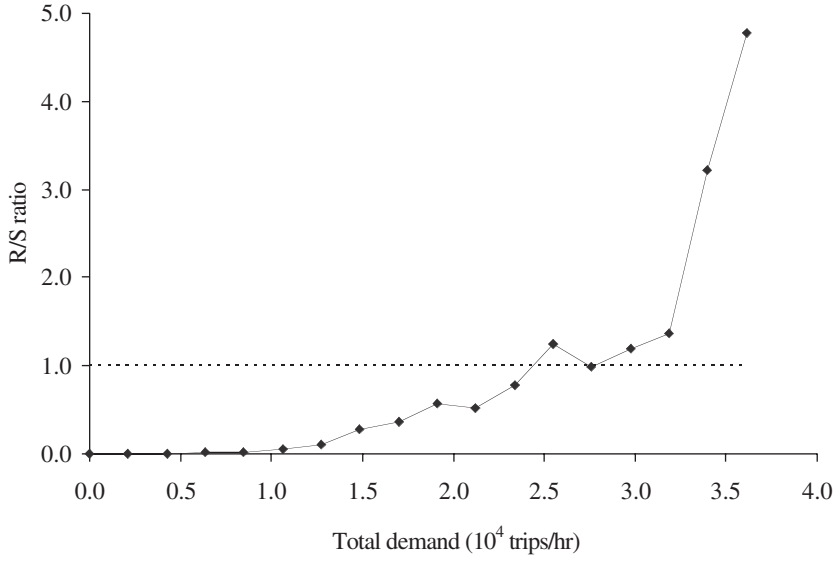


Figure 8.4 Ratio of the road pricing revenue to the required transit subsidy at various levels of total demand

to the total required transit subsidy with respect to the total travel demand ($Q = \Sigma_{w \in W} Q^w$); various levels of total demand are obtained by uniform scaling down or up of the present O–D matrix. Note that the ratio (R/S) does not necessarily have to increase monotonically with the level of demand due to their inherent nonlinear complex relationship. It is clear from the figure that when the total demand is small, this ratio is less than 1.0, which means that the road pricing revenue is not sufficient to cover the deficit of transit operation. However, the R/S ratio will increase quickly as total demand Q grows. These results illustrate the fact that the transit mode should be built only when Q exceeds a certain threshold, and be expanded as travel demand grows further. Therefore, in this case it would be interesting to investigate the optimal allocation of toll revenue between transport infrastructure investment and transit subsidy together with transport pricing.

8. CONCLUSIONS

This chapter investigated the relationship among the auto toll, transit fare and subsidy scheme at a bi-modal network level. Transport pricing and transit subsidy are sought for optimal modal split under the assumption of a transit budget constraint. Our introduction of idiosyncratic modal utility for optimal transport pricing differs from the classical consumer behavior theory. Our numerical results clearly demonstrate that the independent determination of optimal congestion toll or transit fare is underestimated compared to that under the integrated system; a system optimum in terms of social welfare maximization can generally be achieved under marginal-cost pricing and an appropriate transit subsidy program. It is thus desirable to link road pricing and public

transport provision via the use of congestion charge revenue to subsidize or improve transit services for optimal modal split. This scheme is also justifiable on the ground of equity. Future work would be toward the incorporation into our modeling framework of user heterogeneity, fixed costs in road capacity supply and overall demand elasticity as well as departure time choice (see, for example, Arnott et al. 1993; Tabuchi 1993; Yang and Huang 1997a and Yang and Meng 1998 for discussions of single and general bottleneck pricing problems).

NOTES

1. We assume that a consumer's representative utility is quasi-linear in non-transportation goods, following the literature (Hau 1987; Oppenheim 1995). The utility therefore appears as an additively separable component in the traveler-consumer's conditional indirect utility function. This simple quasi-linear utility specification yields zero income effects for transportation since our numerical results are based on changes with respect to the base case. The income term cancels out and does not affect our result. The additive income term drops out of the analysis, and this can be clearly seen from the numerator and denominator of the logit model equation to be introduced later. Without loss of generality, all the utility functions expressed here are based on the transportation attributes of travel cost and time only.
2. Note that with demand–response transit service, passenger waiting time will be reduced with increased transit frequency. This increasing returns from scheduling frequency or the so-called 'Mohring effect' (Mohring 1972) is not modeled here because incorporating this effect would considerably complicate our model at this stage.

REFERENCES

Anderson, S.P., A. De Palma and J.-F. Thisse (1988), 'A representative consumer theory of the logit model', *International Economic Review*, **29**, 461–6.

Arnott, R., A. de Palma and R. Lindsey (1993), 'A structural model of peak-period congestion: a traffic bottleneck with elastic demand', *American Economic Review*, **83**, 161–79.

Bazaraa, M.S., H.D. Sherali and C.M. Shetty (1993), *Nonlinear Programming: Theory and Algorithms*, New York: John Wiley & Sons.

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

Berechman, J. (1993), *Public Transit Economics and Deregulation Policy*, New York: North-Holland.

Bergendorff, P., D.W. Hearn and M.V. Ramana (1997), 'Congestion toll pricing of traffic networks', in Panos M. Pardalos, Donald W. Hearn and William W. Hager (eds), *Network Optimization*, Springer-Verlag Series: Lecture Notes in Economics and Mathematical Systems, New York: Springer, pp. 51–71.

Blythe, P.T. and P.J. Hills (1994), 'Automatic debiting and electronic payment for transport – the ADEPT project', *Traffic Engineering and Control*, **34**, 56–61.

Borger, B.D., I. Mayeres, S. Proost and S. Wouters (1996), 'Optimal pricing of urban passenger transport', *Journal of Transport Economics and Policy*, **30**, 31–54.

Braeutigam, R. (1979), 'Optimal pricing with intermodal competition', *American Economic Review*, **69**, 38–49.

Brotcorne, L., M. Labbe, P. Marcotte and G. Savard (2001), 'A bilevel model for toll optimization on a multicommodity transportation network', *Transportation Science*, **35**, 345–58.

Button, K. (1993), *Transport Economics*, (2nd edn), Aldershot, UK and Brookfield, US: Edward Elgar.

Daganzo, C. (1995), 'A Pareto optimum congestion reduction scheme', *Transportation Research*, **29B**, 139–54.

De Palma, A. and R. Lindsey (2000), 'Private toll roads: competition under various ownership regimes', *Annals of Regional Science*, **34**, 13–35.
Dial, R.B. (1999a), 'Network-optimized road pricing. Part I: A parable and a model', *Operations Research*, **47**, 54–64.
Dial, R.B. (1999b), 'Network-optimized road pricing. Part II: Algorithms and examples', *Operations Research*, **47**, 327–36.
Dial, R.B. (1999c), 'Minimal-revenue congestion pricing. Part I: A fast algorithm for the single-origin case', *Transportation Research*, **33B**, 189–202.
Dial, R.B. (2000), 'Minimal-revenue congestion pricing. Part II: An efficient algorithm for the general case', *Transportation Research*, **34B**, 645–65.
Erlander, S. and N.F. Stewart (1990), *The Gravity Model in Transportation Analysis: Theory and Extensions*, Utrecht: VSP.
Florian, M. and H. Spiess (1983), 'On binary mode choice/assignment models', *Transportation Science*, **17**, 32–47.
Frankena, M.W. (1983), 'The efficiency of public transport objectives and subsidy formulas', *Journal of Transport Economics and Policy*, **18**, 67–76.
Glaister, S. and D. Lewis (1978), 'An integrated fares policy for transport in London', *Journal of Public Economics*, **9**, 341–55.
Guria, J.C. (1987), 'Optimal pricing in an integrated transport system', *International Journal of Transport Economics*, **14**, 267–82.
Hau, T.D. (1987), 'Using a Hicksian approach to cost–benefit analysis in discrete choice: an empirical analysis of a transportation corridor simulation model', *Transportation Research*, **21B**, 339–57.
Hau, T.D. (1992a), 'Economic fundamentals of road pricing: a diagrammatic analysis', World Bank Policy Research Working Paper Series WPS 1070, World Bank, Washington, DC, December.
Hau, T.D. (1992b), 'Congestion charging mechanisms for roads: an evaluation of current practice', World Bank Policy Research Working Paper Series WPS 1071, World Bank, Washington, DC, December.
Hearn, D.W. and M.V. Ramana (1998), 'Solving congestion toll pricing models', in Patrice Marcotte and Sang Nguyen (eds), *Equilibrium and Advanced Transportation Modeling*, Boston, MA: Kluwer Academic, pp. 109–24.
Hong Kong Transport Department (HKTD) (1995), 'Measures to address traffic congestion', Public Consultation Report.
Jara-Diaz, S.R. and A.M. Tudela (1993), 'Multiobjective pricing of bus–subway services in Santiago, Chile', *Journal of Advanced Transportation*, **27**, 261–77.
Johansson, B. and L.G. Mattsson (1995), *Road Pricing: Theory, Empirical Assessment and Policy*, Boston, MA: Kluwer Academic.
Labbe, M., P. Marcotte and G. Savard (1998), 'A bilevel model of taxation and its application to optimal highway pricing', *Management Science*, **44**, 1608–22.
Li, M.Z.F. (1999), 'Estimating congestion toll by using traffic count data – Singapore's area licensing scheme', *Transportation Research*, **35E**, 1–10.
Li, M.Z.F. (2002), 'The role of speed–flow relationship in congestion pricing implementation with an application to Singapore', *Transportation Research*, **36B**, 731–54.
May, A.D. and D.S. Milne (2000), 'Effects of alternative road pricing systems on network performance', *Transportation Research*, **34A**, 407–36.
Miyagi, T. and T. Suzuki (1997), 'A Ramsey price equilibrium model and its computational procedure', *Journal of the Eastern Asia Society for Transportation Studies*, **2**, 1047–62.
Mohring, H. (1972), 'Optimization and scale economies in urban bus transportation', *American Economic Review*, **62**, 591–604.
Nowlan, D.M. (1993), 'Optimal pricing of urban trips with budget restrictions and distributional concerns', *Journal of Transport Economics and Policy*, **28**, 253–76.
Obeng, K., A.H.M. Golam Azam and R. Sakano (1997), *Modeling Economic Inefficiency Caused by Public Transit Subsidies*, Westport, CT and London: Praeger.
Oberholzer-Gee, F. and H. Weck-Hannemann (2002), 'Pricing road use: politico-economic and fairness considerations', *Transportation Research*, **7D**, 357–71.

Oppenheim, N. (1995), *Urban Travel Demand Modeling: From Individual Choices to General Equilibrium*, New York: John Wiley & Sons.
O'Sullivan, A. (1993), *Urban Economics*, Boston, MA: Richard D. Irwin.
Paulley, N. (2002), 'Recent studies on key issues in road pricing', *Transport Policy*, **9**, 175–7.
Sherman, R. (1971), 'Congestion interdependence and urban transit fares', *Econometrica*, **39**, 565–76.
Sherman, R. (1972), 'Subsidies to relieve urban traffic congestion', *Journal of Transport Economics and Policy*, **7**, 22–31.
Small, K.A. (1983), 'The incidence of congestion tolls on urban highways', *Journal of Urban Economics*, **13**, 90–110.
Tabuchi, T. (1993), 'Bottleneck congestion and modal split', *Journal of Urban Economics*, **34**, 414–31.
Verhoef, E.T. (2002), 'Second-best congestion pricing in general networks. Heuristic algorithms for finding second-bet optimal toll levels and toll points', *Transportation Research*, **36B**, 707–29.
Viton, P.A. (1983), 'Pareto-optimal urban transportation equilibrium', in T.E. Keeler (ed.), *Research in Transportation Economics*, vol. 1, Greenwich, CT: JAI Press, pp. 75–101.
Williams, H.C.W.L. (1977), 'On the formation of travel demand models and economic evaluation measures of user benefit', *Environment and Planning*, **9**, 285–344.
Yang, H. (1997), 'Sensitivity analysis for the elastic-demand network equilibrium problem with applications', *Transportation Research*, **31B**, 55–70.
Yang, H. and M.G.H. Bell (1997), 'Traffic restraint, road pricing and network equilibrium', *Transportation Research*, **31B**, 303–14.
Yang, H. and H.J. Huang (1997a), 'Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory', *Transportation Research*, **31B**, 425–40.
Yang, H. and H.J. Huang (1997b), 'Principle of marginal-cost pricing: how does it work in a general road network?', *Transportation Research*, **32A**, 45–54.
Yang, H. and H.J. Huang (1999), 'Carpooling and congestion pricing in a multilane highway with high-occupancy-vehicle-lane', *Transportation Research*, **33A**, 139–55.
Yang, H. and W.H.K. Lam (1996), 'Optimal road tolls under conditions of queuing and congestion', *Transportation Research*, **30A**, 319–32.
Yang, H. and Q. Meng (1998), 'Departure time, route choice and congestion toll in a queuing network with elastic demand', *Transportation Research*, **32B**, 247–60.
Yang, H. and Q. Meng (2000), 'Highway pricing and capacity choice in a road network under a build–operate–transfer scheme', *Transportation Research*, **34A**, 207–22.
Yang, H. and Q. Meng (2002), 'A note on highway pricing and capacity choice under a build–operate–transfer scheme', *Transportation Research*, **36A**, 659–63.
Yang, H., W.H. Tang, W.M. Cheung and Q. Meng (2002), 'Profitability and welfare gain of private toll roads with heterogeneous users', *Transportation Research*, **36A**, 537–54.
Yang, H. and X. Zhang (2002), 'The multi-class network toll design problem with social and spatial equity constraints', *Journal of Transportation Engineering*, **128**, 420–28.

9. Planning transport network improvements over time

Hong K. Lo and W.Y. Szeto

1. INTRODUCTION

Transport infrastructure is in an active phase of planning and development in many parts of Asia. In Hong Kong, for example, there have been heated debates on whether Route 10 should be built. On a regional scale, the question is how and when should Hong Kong be better linked with the Pearl River Delta? Specifically, should the Macau–Zhuhai bridge be built? These are expensive projects, involving billions of Hong Kong dollars. In times of constrained government expenditures, they must be carefully scrutinized for cost-effectiveness. In addition to the question of 'whether', the timing and scale of their implementation are also important considerations. To answer these questions, we cannot simply analyse the traffic around the highway link of interest, but must take a system approach to analyse the entire network, as network route changes may occur prior to the highway link. Traditionally, this analysis belongs to the discipline of transport network design.

Transport network design involves the interaction of supply and demand. The transport network can be considered as the supply, provided in a certain way for a certain objective, and the traveling public as the demand, who decide whether to travel, and if so, which route(s) to use. This interaction involves complex issues, such as: (i) the objective of the transport network design, (ii) the travel and route choice behavior of travelers, (iii) means of financing, and (iv) demand projections. Furthermore, as the costs and benefits of transport network improvements accrue over a long period of time, subject to the ever-changing demand patterns and gradual network upgrades, it is important to include the temporal dimension in the analysis.

Traditionally, transport network design is formulated as a bi-level programming problem. In this problem, the upper level is the network planner's problem, whereas the lower level is the network users' problem, which can be described by either the trip distribution/assignment problem or the traffic assignment problem. In general, this bi-level problem is referred to as the *network design problem* (NDP). A number of studies have been conducted on the NDP: Boyce (1984), Magnanti and Wong (1984) and Friesz (1985) have carried out comprehensive reviews in the past; Yang and Bell (1998) provide a recent review on models and solution approaches.

In fact, different combinations of the lower- and upper-level problems constitute different NDPs. Various traffic assignment problems have been considered in the lower-level problem, including deterministic user equilibrium (for example, Meng et al. 2001), stochastic user equilibrium (for example, Chen and Alfa 1991; Davis 1994), dynamic user equilibrium (for example, Heydecker 2002), and probabilistic user equilibrium (for example, Lo

2002; Lo and Tung 2003). Trip distribution/assignment problems have also been studied and incorporated (for example, Boyce and Janson 1980; Lam et al. 2002).

The upper-level problem has also been formulated with different decision variables and objective functions. The decision variables can be either discrete or continuous, depending on the problem of interest. When the decision variables are solely discrete (continuous), the resultant NDP is called the discrete (continuous) network design problem or the DNDP (CNDP); when the decision variables contain both discrete and continuous variables, the resultant NDP is called the mixed network design problem (MNDP). The DNDP is concerned with the network topology itself. Examples include the determination of new roadway links to be added to the network (for example, LeBlanc 1975; Chen and Alfa 1991), and the bus network design problem (for example, Wan et al. 2002). The CNDP takes the network topology as given and is concerned with optimizing the network parameters. Examples include determining the link widening problem (for example, Marcotte 1986; Meng et al. 2001), the combined traffic signal and assignment problem (for example, Fisk 1984; Smith and Van Vuren 1993; Cascetta et al. 1998; Chen and Ben-Akiva 1998; Gartner and Stamatiadis 2001), and the build–operate–transfer (BOT) problem – setting user charges and link capacities (for example, Yang and Meng 2000). The MNDP combines the DNDP and the CNDP. An example includes the simultaneous determination of new roads to be added and widening of existing roads (for example, Yang and Bell 1998).

In terms of objective functions, past NDP efforts have considered:

- total travel cost minimization (for example, Maher et al. 2001);
- reserve capacity maximization (for example, Wong and Yang 1997);
- consumer surplus maximization (for example, Lam et al. 2002); and
- multi-objective optimization including total user cost, construction cost and so on (for example, Friesz and Harker 1983; Current and Min 1986; Friesz et al. 1993; Tzeng and Tsaur 1997).

The above illustrates the variety of objectives that has been considered. Indeed, one may introduce other objectives as necessary. All of these efforts, however, optimize the network for a certain future time without clearly defining the time dimension within the formulation over the planning horizon. In this study, we introduce the time dimension to the continuous network design problem. For simplicity, we refer to this extension as the CNDP-T. With this extension, time-dependent travel demands, changing land-use patterns, and the gradually upgraded network during the planning horizon can be considered. The optimal project start time and phasing can also be outlined. In addition, it becomes possible to analyse the financial arrangements over the planning horizon, as the improvement budget may not be provided as a lump sum but rather as portions to be available once every few years.

The time scale considered in the CNDP-T is typically in years, as compared with the second-to-second scale of traffic dynamics, or the day-to-day scale of route choice dynamics. For ease of exposition, we consider the planning horizon to be discrete. For each discrete time interval, we consider the traffic assignment pattern as following a user-equilibrium pattern. And for simplicity in this first study, we follow the static equilibrium approach. On the other hand, travel demand is considered as elastic in each year, varying with time along the planning horizon. In this CNDP-T, we consider consumer surplus as the objective to be

maximized across the entire planning horizon by determining the network capacity enhancements over time subject to the budget, physical boundary, and potential demand constraints. The lower-level problem contains copies of the user-equilibrium sub-problems, one for each discrete time interval along the planning horizon, with respect to the time-dependent design variables specified in the upper-level problem. By expressing the lower-level problem as a set of constraints, we formulate the CNDP-T to a single-level mathematical program. This formulation allows the use of existing nonlinear programming techniques to solve the CNDP-T.

To illustrate the importance of capturing the time dimension, we set up two scenarios and compare the total consumer surplus obtained for the planning period through the traditional versus the proposed approach. The traditional design approach determines the network improvements by considering only the ultimate potential demands at the end of the planning horizon, whereas the proposed approach considers the time-dependent potential demands over the entire planning horizon. Both scenarios are solved by the generalized reduced gradient (GRG) method (Abadie and Carpentier 1969). The numerical studies demonstrate that this extended formulation produces superior results as compared with the traditional formulation.

The outline of this chapter is as follows. Section 2 formulates the CNDP-T. Section 3 depicts the GRG method. Section 4 contains the numerical studies. Finally, Section 5 provides some concluding remarks.

2. FORMULATION

We consider a general transportation network with multiple origin–destination (O–D) flows over the planning horizon $[0,T]$. The horizon is divided into N intervals, with each interval representing one year.

The following notations are adopted throughout this chapter:

1. *Set notations*
   - RS set of O–D pairs
   - P^{rs} set of routes for travelers connecting O–D pair rs
   - A set of links.
2. *Indices*
   - rs O–D pair, $rs \in RS$
   - p route between O–D pair rs, $p \in P^{rs}$
   - a link, $a \in A$
   - τ year, $\tau \in \{1,...,N\}$.
3. *Variables to be determined*
   - $f_{p,\tau}^{rs}$ representative hourly flow on route p between O–D pair rs in year τ
   - $y_{a,\tau}$ capacity enhancement in year τ
   - $\mathbf{f}$ column vector of $[f_{p,\tau}^{rs}]$

y column vector of $[y_{a,\tau}]$
s column vector of slack variables
x column vector of $[\mathbf{f}^T,\mathbf{y}^T,\mathbf{s}^T]^T$.

4. *Parameters given*
   - α parameter in link performance function
   - β parameter in link performance function
   - k proportionality parameter in construction cost function
   - n factor converting the consumer surplus from an hourly to an annual basis
   - B_τ budget provided by the government at the beginning of year τ
   - m^{rs} slope of the travel demand function of O–D pair *rs*
   - h^{rs} growth rate of potential demand between O–D pair *rs*
   - c_a initial capacity of link *a*
   - u_a upper bound of the capacity of link *a*
   - t_a^0 free-flow travel time of link *a*
   - δ_a^p link-path incidence indicator, $\delta_a^p = 1$ if *a* is on *p*, $\delta_a^p = 0$ otherwise
   - $\tilde{q}_1^{rs}$ potential demand between O–D pair *rs* in the first year.
5. *Functions of capacity enhancements and equilibrium route flows*
   - $v_{a,\tau}$ representative hourly flow on link *a* in year τ
   - $t_{a,\tau}$ travel time on link *a* in year τ
   - q_τ^{rs} representative hourly travel demand between O–D pair *rs* in year τ
   - $\eta_{p,\tau}^{rs}$ route time for travelers taking route *p* between O–D pair *rs* in year τ
   - π_τ^{rs} the shortest travel time between O–D pair *rs* in year τ
   - $D_\tau^{rs}(.)$ continuous travel demand function for O–D pair *rs* in year τ
   - $\mathbf{h}(.)$ vector function.
6. *Functions of capacity enhancements*
   - $R\tau$ unspent cumulative allocated budgets saved for the future year after τ
   - $g_a(.)$ improvement cost function of link *a*.
7. *Function of potential demand*
   - $G^{rs}(.)$ potential demand function for O–D pair *rs*.

The following assumptions are made in this study:

1. The potential demand growth over time is predicted perfectly.
2. Traffic assignment follows the user-equilibrium criterion.
3. Only link widening is considered.
4. The link cost and travel demand functions are separable.
5. The functional form of the travel demand function maintains over time.
6. The interest, inflation, and benefit discount rates are taken as zero for simplicity.

Most of these assumptions can be relaxed in future studies. Therefore, they are not restrictive from a modeling perspective, but are adopted merely to simplify the analysis.

The CNDP-T includes N annual-based traffic assignment sub-problems. For each sub-problem, travelers are assumed to follow the Wardrop principle (1952). This principle requires that route p between O–D pair rs will not be used if its travel time is longer than the shortest travel time between O–D pair rs. Conversely, any used route p must have its travel time equal to the shortest travel time between O–D pair rs. Mathematically, this principle for each sub-problem in year τ can be expressed as:

$$f_{p,\tau}^{rs}\left(\eta_{p,\tau}^{rs}-\pi_{\tau}^{rs}\right)=0,\ \forall rs,p,\tau, \tag{9.1}$$

$$\eta_{p,\tau}^{rs}-\pi_{\tau}^{rs}\geq 0,\ \forall rs,p,\tau, \tag{9.2}$$

where $f_{p,\tau}^{rs}$ and $\eta_{p,\tau}^{rs}$ are, respectively, the representative hourly flow and the route travel time for route p between O–D pair rs in year τ; π_{l}^{rs} is the shortest travel time between O–D pair rs in year τ.

Route travel times and the shortest travel times can be determined once the equilibrium route flows **f** and capacity enhancements **y** are known, expressed as:

$$v_{a,\tau}=\sum_{rs}\sum_{p}f_{p,\tau}^{rs}\delta_{a}^{p},\ \forall a,\tau, \tag{9.3}$$

$$t_{a,\tau}=t_{a}^{0}\left[1+\alpha\left(\frac{v_{a,\tau}}{c_{a}+\sum_{i=1}^{\tau}y_{a,i}}\right)^{\beta}\right],\ \forall a,\tau, \tag{9.4}$$

$$\eta_{p,\tau}^{rs}=\sum_{a\in A}t_{a,\tau}\cdot\delta_{a}^{p},\ \forall rs,\ p,\ \tau, \tag{9.5}$$

where $v_{a,\tau}$ and $t_{a,\tau}$ are the representative hourly flow and the travel time on link a in year τ, respectively; $y_{a,\tau}$ is the capacity enhancement on link a in year τ, meaning that the capacity of link a is increased by $y_{a,\tau}$ units at the beginning of year τ; δ_{a}^{p} is a link-path incidence indicator, which equals one if link a is on route p, zero otherwise; t_{a}^{0} is the free-flow travel time of link a; c_{a} is the capacity of link a before the capacity improvement; α, β are parameters of the link performance function.

Equation (9.3) states that link flow is obtained by summing the corresponding route flows on the link. Equation (9.4) is a typical link performance function. The summation term in (9.4) represents the total capacity enhancements of link a up to year τ. Therefore, the denominator inside the bracket denotes the link capacity in year τ after implementing the enhancements before and inclusive of year τ. Equation (9.5) computes the route travel time based on the corresponding link travel times.

Each traffic assignment sub-problem also includes the flow conservation and non-negativity conditions, expressed as:

$$\sum_{p}f_{p,\tau}^{rs}=q_{\tau}^{rs},\ \forall rs,\tau, \tag{9.6}$$

$$f_{p,\tau}^{rs} \geq 0, \forall rs, p, \tau, \tag{9.7}$$

where q_τ^{rs} is the travel demand of O–D pair rs in year τ.

The travel demand of O–D pair rs in year τ, q_τ^{rs}, is not fixed but is a function of the potential demand of that year $\tilde{q}_\tau^{rs}$ and its shortest travel time π_τ^{rs}:

$$q_\tau^{rs} = D_\tau^{rs}(\pi_\tau^{rs}, \tilde{q}_\tau^{rs}), \forall rs, \tau, \tag{9.8}$$

where $D_\tau^{rs}(.)$ is the continuous travel demand function for O–D pair rs in year τ. The travel demand function is generally decreasing, implying that higher shortest travel times lead to lower travel demands, and vice versa.

The potential demand per O–D pair of each year represents the potential travel growth due to population growth and/or changes in land-use patterns over time, expressed as:

$$\tilde{q}_\tau^{rs} = G^{rs}, (\tilde{q}_{\tau-1}^{rs}), \tau > 1, \forall rs, \tag{9.9}$$

where $G^{rs}(.)$ is the potential demand function for O–D pair rs. The potential demand in year τ is modeled to depend on the potential demand of year $\tau - 1$ but not to depend on the traffic conditions.

This formulation shares some similarity with the dynamic traffic assignment (DTA) problem, as both involve variable interactions across time in an intertwined manner. The difference is that in the CNDP-T, the equilibrium conditions hold at each discrete time interval in the planning horizon, typically years apart; whereas in DTA, the equilibrium conditions hold for traffic at the same departure time, typically seconds or minutes apart (Lo and Szeto 2002). In a sense, the CNDP-T is simpler. Traffic at one discrete time interval in the planning horizon does not interact with traffic at a different one in the planning horizon, whereas in DTA, one must consider such interactions.

In this study, we choose to use consumer surplus (CS) as the performance measure, which internalizes the effect of network congestion and the public's propensity to travel. Consumer surplus measures the difference between what consumers would be willing to pay for travel and what they actually pay. For the same network and demand characteristics, a higher consumer surplus implies a better performing system. Mathematically, it is expressed as:

$$CS = \sum_\tau \sum_{rs} n \left[\int_0^{q_\tau^{rs}} D_\tau^{rs-1}(v)\, dv - \pi_\tau^{rs} q_\tau^{rs} \right], \tag{9.10}$$

where n is a factor converting the consumer surplus from an hourly basis to an annual basis. CS in (9.10) is obtained from summing the consumer surplus measure over the planning horizon for all O–D pairs, representing the overall network consumer surplus over the planning horizon.

The CNDP-T can be formulated as the following constrained nonlinear maximization program:

$$\max_{\mathbf{f},\mathbf{y}} CS = n \sum_\tau \sum_{rs} \left[\int_0^{q_\tau^{rs}} D_\tau^{rs-1}(v)\, dv - \pi_t^{rs} q_\tau^{rs} \right],$$

subject to these link improvement and budgetary constraints:

$$\sum_a g_a(y_{a,1}) + R_1 = B_1, \tag{9.11}$$

$$\sum_a g_a(y_{a,\tau}) + R_\tau = R_{\tau-1} + B_\tau, \forall \tau > 1, \tag{9.12}$$

$$c_a + \sum_\tau y_{a,\tau} \leq u_a, \forall a, \tag{9.13}$$

$$y_{a,\tau} \geq 0, \forall a, \tau, \tag{9.14}$$

together with the annual-based traffic assignment constraints (9.1)–(9.9). In this formulation, B_τ is the network improvement budget allocated by the government at the beginning of year τ. The variable R_τ represents the cumulative allocated funds that have not yet been spent in year τ and are available for use at the beginning of year $\tau+1$. The variable $g_a(y_{a,\tau})$ is the cost of increasing the capacity of link a by $y_{a,\tau}$; u_a is the maximum allowable capacity of link a.

In the above nonlinear constrained mathematical program, (9.11) and (9.12) are the budgetary constraints. Equation (9.12) states that the sum of the total cost of improvements for year τ and the funds unspent or saved for the future years after τ is equal to the budget allocated by the government for year τ plus the funds saved from year 1 to $\tau-1$. Equation (9.11) describes the beginning of the planning horizon, assuming that no funds were saved before the planning horizon. This set of recursive equations permits the modeling of various budget and expenditure scenarios. For example, if the government provides a lump sum only at the beginning of the planning horizon, then $B_1 > 0$ and $B_\tau = 0$, $\tau > 1$. Then (9.11) and (9.12) can be reduced to $\sum_\tau \sum_a g_a(y_{a,\tau}) \leq B_1$. Equations (9.13) and (9.14) define the physical boundary conditions on capacity enhancements. The maximum allowable capacity constraint (9.13) limits the total capacity $c_a + \sum_\tau y_{a,\tau}$ of link a to be less than the maximum allowable capacity of that link u_a. Equation (9.14) is the non-negativity condition of capacity enhancements.

The CNDP-T (9.1)–(9.14) is a generalization of the traditional CNDP and can be reduced to the traditional CNDP by setting $T = 1$ (that is, the design period is one year) and removing the potential demand constraint. Without loss of generality, the problem (9.1)–(9.14) can be written as follows:

$$\max CS(\mathbf{x}), \tag{9.15}$$

subject to:

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}, \tag{9.16}$$

$$\mathbf{x} \geq \mathbf{0}, \tag{9.17}$$

where $\mathbf{x} = [\mathbf{f}^T, \mathbf{y}^T, \mathbf{s}^T]^T$; $\mathbf{h}(\mathbf{x})$ is a vector function of $\mathbf{x}$ representing the relationships (9.1)–(9.9) and (9.11)–(9.13); $\mathbf{s}$ is a column vector of slack variables. Slack variables $\mathbf{s}$ are introduced in the problem (9.15)–(9.17) to convert all inequality constraints in the problem (9.1)–(9.14) into equality constraints. The problem (9.15)–(9.17) constitutes a constrained nonlinear mathematical program that can be solved by existing optimization algorithms. Nevertheless, we note that it contains non-convex constraints due to the annual-based user-equilibrium conditions. Therefore, general-purpose nonlinear optimization codes will

stop at local minima. How to take advantage of the special structure of this problem to establish stable, efficient solution algorithms is an important research topic. It is beyond the scope of this chapter and is left for future studies.

In summary, the CNDP-T is to determine optimal capacity enhancements and the equilibrium flow patterns to maximize the consumer surplus over the planning horizon subject to the budgetary, physical boundary, potential demand, and annual-based traffic assignment constraints.

3. SOLUTION METHOD

This study adopts the generalized reduced gradient algorithm (Abadie and Carpentier 1969) to solve (9.15)–(9.17). The GRG method can work with mathematical programs with nonlinear objectives and nonlinear equality constraints. It combines the ideas of linearization and Wolfe's reduced gradient method. The latter is extended from the simplex method to solve nonlinear programming problems with linear equality constraints.

Like the simplex method, the GRG method classifies variables as basics (basic or dependent variables) and non-basics, and requires the determination of bases. In each iteration, the GRG method forms a reduced problem by using the first-order Taylor approximation to the nonlinear equality constraints at the current best point. A set of generalized reduced gradients and the displacement direction are then determined based on the current basis and the gradient of the objective function with respect to basics and non-basics at the current point $\mathbf{x}^k$. Having specified the displacement direction, the GRG method performs a minimum ratio check to determine the largest step size and avoid variables reaching negative values. It then conducts a one-dimensional search to determine the best step size for the reduced problem. Due to the non-linearity of the equality constraints, the point $\mathbf{x}'^k$ generated by the one-dimensional search does not always satisfy the equality constraints. A correction step is needed after each one-dimensional search to restore feasibility. This step can be done by the Newton–Raphson method. Starting with $\mathbf{x}'^k$ and keeping all the basic components of $\mathbf{x}'^k$ fixed, the Newton–Raphson method solves the equality constraints. However, this method does not take any notice of the non-negativity constraints. If during the iterative process, one of the basic variables vanishes or becomes negative, the GRG method requires updating the basis, re-computing a new reduced gradient relative to the new basis at $\mathbf{x}^k$, determining $\mathbf{x}'^k$, and employing the Newton–Raphson method again. Otherwise, a new point $\mathbf{x}^{k+1}$ that satisfies both the non-negativity and equality constraints is obtained, and a new iteration starts without updating the basis until a local optimal solution is found.

The detailed algorithmic steps are the following:

Step 0: Initialization Choose stopping tolerance $\varepsilon > 0$ and any starting feasible solution $\mathbf{x}^0$. Partition vector $\mathbf{x}$ into basic variables $\mathbf{x}_b$ and non-basic variables $\mathbf{x}_n$. Therefore, the initial point $\mathbf{x}^0$ can be represented by:

$$\begin{bmatrix} \mathbf{x}_b^0 \\ \mathbf{x}_n^0 \end{bmatrix}.$$

Construct the basis matrix $\mathbf{B} = (\partial \mathbf{h}/\partial \mathbf{x}_b)(\mathbf{x}^0)$ and the non-basis matrix $\mathbf{N} = (\partial \mathbf{h}/\partial \mathbf{x}_n)(\mathbf{x}^0)$. Set $k = 0$.

Step 1: Direction finding Compute the generalized reduced gradients $\mathbf{r}_n^k$ at $\mathbf{x}^k$ by:

$$\mathbf{r}_n^k = \frac{\partial CS}{\partial \mathbf{x}_n}(\mathbf{x}^k) - \frac{\partial CS}{\partial \mathbf{x}_b}(\mathbf{x}^k)\mathbf{B}^{-1}\mathbf{N}.$$

Determine the displacement direction

$$\mathbf{d}^k = \begin{bmatrix} \mathbf{d}_b^k \\ \mathbf{d}_n^k \end{bmatrix}$$

by: $\mathbf{d}_b^k = -\mathbf{B}^{-1}\mathbf{N}\mathbf{d}_n^k$, $\mathbf{d}_n^k = [d_n^k]$, where

$$d_n^k = \begin{cases} r_n^k & \text{if } r_n^k > 0 \text{ or } x_n^k > 0 \\ 0 & \text{otherwise} \end{cases},$$

$r_n^k(x_n^k)$ is the element of $\mathbf{r}_n^k(\mathbf{x}_n^k)$.

Step 2: Convergence test If $\|\mathbf{d}^k\| \leq \varepsilon$, stop.

Step 3: Feasibility limit Compute feasibility limiting step $\lambda_{\max}^k$ by:

$$\lambda_{\max}^k = \begin{cases} \infty & \text{if } \mathbf{d} > \mathbf{0} \\ \min \left(\dfrac{x_j^k}{-d_j^k}, \forall j \in \{n,b\} \mid d_j^k < 0 \right) & \text{otherwise} \end{cases}$$

in which d_b^k is the element of $\mathbf{d}_b^k$.

Step 4: Line search Perform a one-dimensional search to determine the step size λ^k solving:

$$\max CS(\mathbf{x}^k + \lambda^k \mathbf{d}^k)$$

subject to $0 \leq \lambda^k \leq \lambda_{\max}^k$.

Step 5: Move Set $\mathbf{x}'^k = \mathbf{x}^k + \lambda^k \mathbf{d}^k$.

Step 6: Correction step:

Step 6.1 Set $l = 0$ and $\mathbf{z}^0 = \mathbf{x}_b'^k$.

Step 6.2 Compute $\mathbf{z}^{l+1}$ by: $\mathbf{z}^{l+1} = \mathbf{z}^l - \mathbf{B}^{-1}\mathbf{h}(\mathbf{z}^l, \mathbf{x}_n'^k)$.

Step 6.3 If $\mathbf{z}^{l+1} \geq \mathbf{0}$ and $\|\mathbf{h}(\mathbf{z}^{l+1}, \mathbf{x}_n'^k)\| \leq \varepsilon$, Set $\mathbf{x}_b^{k+1} = \mathbf{z}^{l+1}$, $\mathbf{x}_n^{k+1} = \mathbf{x}_n'^k$, and $k = k+1$. Go to Step 1.

Step 6.4 If $\mathbf{z}^{l+1} \geq \mathbf{0}$, set $l = l+1$ and go to Step 6.2.

Step 6.5 Substitute the basic variable that becomes negative by a non-basic variable. Form a new basis $\mathbf{B} = (\partial \mathbf{h}/\partial \mathbf{x}_b)(\mathbf{x}^k)$ and $\mathbf{N} = (\partial \mathbf{h}/\partial \mathbf{x}_n)(\mathbf{x}^k)$. Go to Step 1.

4. NUMERICAL STUDIES

To show the effect of capturing the time dimension in planning network improvements, we set up two scenarios to compare the total consumer surplus obtained through the traditional versus the proposed approach. The traditional CNDP approach uses the ultimate potential demands in the planning horizon to determine the network improvements whereas the proposed CNDP-T approach considers the time-dependent potential demands throughout the planning horizon. In these two scenarios, the first one, without route choice considerations, can be simplified to a convex program to obtain the global maximum. The second scenario is set up to study the general situation with route choice considerations.

4.1 Scenario 1: Network without Route Choice

A degenerate case is considered in this scenario, in which travelers have no route choice. The network example, as shown in Figure 9.1, consists of three nodes, two links, and two

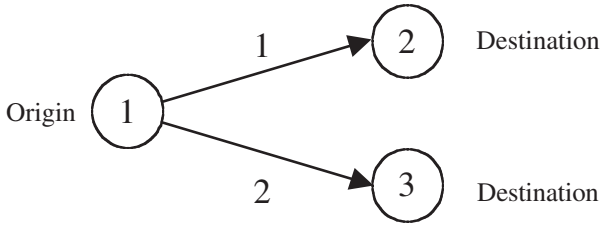


Figure 9.1 Example network for Scenario 1

O–D pairs. The two O–D pairs are from node 1 to node 2 and from node 1 to node 3. The following linear travel demand function is adopted:

$$\pi_\tau^{rs} = m^{rs}\,(q_\tau^{rs} - \tilde{q}_\tau^{rs}), \tag{9.18}$$

where m^{rs} is the slope of the travel demand function of O–D pair rs. The construction cost function takes the following linear form:

$$g(y_{a,\tau}) = kt_a^0 y_{a,\tau}, \tag{9.19}$$

where k is a proportionality parameter. The potential demand function is defined as:

$$\tilde{q}_\tau^{rs} = \tilde{q}_{\tau-1}^{rs}(1 + h^{rs}), \tag{9.20}$$

where h^{rs} is the growth rate of potential demand between O–D pair rs. The parameters include:

(a.) link capacities: $c_1 = 1$, $c_2 = 1$;
(b.) free-flow travel times: $t_1^0 = 15$, $t_2^0 = 10$;
(c.) maximum allowable link capacities: $u_1 = 2$, $u_2 = 2.5$;
(d.) potential demands after year 1: $\tilde{q}_1^{12} = 20$, $\tilde{q}_1^{13} = 42$;
(e.) growth rates: $h^{12} = 1.75$, $h^{13} = 1/14$;
(f.) link performance function parameters: $\alpha = 1$, $\beta = 1$;
(g.) slope of travel demand functions: $m^{12} = -1$, $m^{13} = -1$;
(h.) proportionality parameter: $k = 100$;
(i.) budgets: $B_1 = 1500$, $B_2 = 0$;
(j.) converting factor: $n = 8760$;
(k.) convergence tolerance: $\varepsilon = 0.0001$.

To formulate the traditional CNDP for this scenario, we use the same notations as the CNDP-T with $\tau = 1$ and the potential demands for the second year. To maintain readability, we put the detailed formulation and simplification procedures in the appendix. The CNDP can be simplified to the following:

$$\max CS = \frac{7\,008\,000}{\left(1 + \frac{1}{1 + y_{1,1}}\right)^2} + \frac{5\,365\,500}{\left(1 + \frac{1}{1 + y_{2,1}}\right)^2}, \tag{9.21}$$

subject to:

$$1.5y_{1,1} + y_{2,1} \leq 1.5 \tag{9.22}$$

$$y_{1,1}, y_{2,1} \geq 0. \tag{9.23}$$

By checking the Hessian of (9.21), the objective function is found to be concave. As (9.22) and (9.23) are linear, the program (9.21)–(9.23) itself is convex, which contains only one global maximum. By using the GRG method with any initial solutions, the global maximum is obtained.

In a similar manner, the CNDP-T can be simplified to the following convex program:

$$\begin{aligned} \max CS = & \frac{109\,500}{\left(1 + \frac{1}{1 + y_{1,1}}\right)^2} + \frac{4\,485\,120}{\left(1 + \frac{1}{1 + y_{2,1}}\right)^2} \\ & + \frac{7\,008\,000}{\left(1 + \frac{1}{1 + y_{1,1} + y_{1,2}}\right)^2} + \frac{5\,365\,500}{\left(1 + \frac{1}{1 + y_{2,1} + y_{2,2}}\right)^2}, \end{aligned} \tag{9.24}$$

subject to:

$$1.5y_{1,1} + y_{2,1}, + 1.5y_{1,2} + y_{2,2} \leq 1.5 \tag{9.25}$$

$$y_{1,1}, y_{2,1}, y_{1,2}, y_{2,2} \geq 0, \tag{9.26}$$

We solve the reduced programs (9.21)–(9.23) and (9.24)–(9.26) to determine the network improvements **y** and the corresponding travel demand and annual consumer surplus. Tables 9.1–3 summarize the results.

Table 9.1 Hourly travel demand and annual consumer surplus

	O–D pair (1,2) according to the CNDP		O–D pair (1,2) according to the CNDP-T	
Year	Travel demand	Annual CS (10^6)	Travel demand	Annual CS (10^6)
1	3.00	0.04	2.53	0.03
2	24.00	2.52	20.27	1.80

Table 9.2 Hourly travel demand and annual consumer surplus

	O–D pair (1,3) according to the CNDP		O–D pair (1,3) according to the CNDP-T	
Year	Travel demand	Annual CS (10^6)	Travel demand	Annual CS (10^6)
1	20.36	1.82	22.75	2.27
2	22.27	2.17	24.88	2.71

Table 9.3 Comparison between the CNDP and CNDP-T solutions for Scenario 1

	CNDP	CNDP-T
Link 1 improvement, 1st year ($y_{1,1}$)	0.50	0.03
Link 2 improvement, 1st year ($y_{2,1}$)	0.75	1.46
Link 1 improvement, 2nd year ($y_{1,2}$)	N/A	0.00
Link 2 improvement, 2nd year ($y_{2,2}$)	N/A	0.00
Total expenditure	1500.00	1500.00
Overall consumer surplus	6.55×10^6	6.81×10^6

The hourly travel demand and annual consumer surplus are shown in Tables 9.1 and 9.2. As expected, both the hourly travel demand and annual CS of each O–D pair increase with year. However, the two approaches induce different travel demands for each O–D pair. The CNDP approach induces more demand for O–D pair (1,2), whereas the CNDP-T approach induces more for O–D pair (1,3).

As far as the network improvements are concerned (Table 9.3), the two approaches produce markedly different strategies. The CNDP approach introduces similar levels of improvements to both links, whereas the CNDP-T approach invests almost entirely in

Link 2. Another observation is that all capacity enhancements for the CNDP-T are completed within the first year. In fact, this ought to be the best strategy as improvements completed within the first year can benefit both the first and second years. This result, of course, is subject to the assumption that the budget is available as a lump sum at the beginning of the planning horizon and the interest and discount rates are ignored. In a way, this scenario is deliberately set up to be limited so as to allow the problem to be solvable by both the CNDP and CNDP-T approaches. Had the improvement budget been distributed gradually over the planning horizon, or had the interest and discount rates varied over the horizon, the CNDP approach would not have been applicable.

According to Table 9.3, the whole budget is used up for both approaches. Comparing the CS measures between the CNDP and the CNDP-T approaches, the CNDP-T approach introduces an additional increase of 4 per cent in CS. This result is notable as the scenario considers only a horizon of two periods (or years). One would expect that for longer planning horizons, and more complex networks, the CNDP-T approach would produce even more significant additional improvements. Nevertheless, even in this very simple example, the two approaches determine very different improvement strategies.

4.2 Scenario 2: Network with Route Choice

The network for Scenario 2, as shown in Figure 9.2, consists of three nodes, three links, and two O–D pairs. The two O–D pairs are from node 1 to node 2 and from node 3 to node 2. O–D pair (1,2) has two routes whereas O–D pair (3,2) has only one route. Route

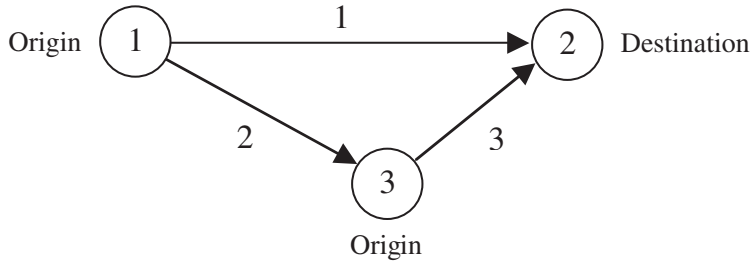


Figure 9.2 Example network for Scenario 2

1 constitutes link 1. Route 2 constitutes links 2 and 3. Route 3 constitutes link 3. The planning horizon is 5 years. The travel demand, potential demand, and construction cost functions assume a linear form as in Scenario 1. The proportionality parameter, convergence tolerance, and converting factor also follow the values as in Scenario 1. Other parameters for this scenario include:

(a.) link capacities: $c_1 = 10$, $c_2 = 10$, $c_3 = 15$;
(b.) free-flow travel times: $t_1^0 = 15$, $t_2^0 = 5$, $t_3^0 = 10$;
(c.) maximum allowable link capacities: $u_1 = 20$, $u_2 = 20$, $u_3 = 30$;
(d.) potential demands after year 1: $\tilde{q}_1^{12} = 40$, $\tilde{q}_1^{32} = 40$;
(e.) growth rates: $h^{12} = 0.22$, $h^{32} = 0.04$;

(f.) link performance function parameters: $\alpha = 0.15$, $\beta = 4$;

(g.) slopes of travel demand functions: $m^{12} = -1$, $m^{32} = -1$; and

(h.) budgets: $B_1 = 10\,000$, $B_2 = B_3 = B_4 = B_5 = 0$.

The mathematical programs for the CNDP and CNDP-T are both non-convex. In employing the GRG solution method (Section 3), different initial solutions are input so as to increase the chance of finding the global solutions. Table 9.4 shows the route flows according to the CNDP and CNDP-T over the five-year planning horizon. Table 9.5 shows the corresponding route travel times. One can verify that in each year, all the used routes attain the minimum route travel time whereas the unused one (example, route 2 in the first year according to the CNDP) has equal or higher route travel time. That is, in each year, the deterministic user-equilibrium conditions are satisfied for both approaches.

The hourly travel demand and annual consumer surplus for each O–D pair from the two approaches are shown in Tables 9.6 and 9.7. It can be verified that the hourly travel demands in Tables 9.6 and 9.7 are exactly equal to the corresponding sums of route flows in Table 9.4, indicating that the demand constraints are satisfied. However, unlike Scenario 1, the hourly travel demand and annual consumer surplus of each O–D pair do not always increase with time despite the network improvements. As route 2 overlaps with route 3, the travel demands of the two O–D pairs interact with each other. This interaction is nontrivial even in this simple example. From Table 9.7, one can see that both the travel demand and the consumer surplus for O–D pair (3,2) drop over time whereas those

Table 9.4 Route flows

	According to the CNDP			According to the CNDP-T		
	Route number			Route number		
Year	1	2	3	1	2	3
1	20.17	0.00	22.46	12.26	7.66	25.07
2	23.64	1.05	22.49	13.81	11.82	24.29
3	25.89	5.53	20.22	15.51	16.00	23.16
4	28.31	10.61	17.22	17.33	20.00	21.83
5	31.04	15.51	14.08	19.21	23.78	20.41

Table 9.5 Route travel times

	According to the CNDP			According to the CNDP-T		
	Route number			Route number		
Year	1	2	3	1	2	3
1	19.83	22.54	17.54	20.08	20.08	14.93
2	24.11	24.11	19.11	23.18	23.18	17.31
3	28.11	28.11	23.04	28.03	28.03	20.11
4	33.72	33.72	27.77	35.30	35.30	23.16
5	42.06	42.06	32.72	45.63	45.63	26.38

Table 9.6 Hourly travel demand and annual consumer surplus

	O–D pair (1,2) according to the CNDP		O–D pair (1,2) according to the CNDP-T	
Year	Travel demand	Annual CS (10^6)	Travel demand	Annual CS (10^6)
1	20.17	1.78	19.92	1.74
2	24.69	2.67	25.62	2.88
3	31.43	4.33	31.51	4.35
4	38.91	6.63	37.33	6.11
5	46.55	9.49	42.98	8.09

Table 9.7 Hourly travel demand and annual consumer surplus

	O–D pair (3,2) according to the CNDP		O–D pair (3,2) according to the CNDP-T	
Year	Travel demand	Annual CS (10^6)	Travel demand	Annual CS (10^6)
1	22.46	2.21	25.07	2.75
2	22.49	2.22	24.29	2.58
3	20.22	1.79	23.16	2.35
4	17.22	1.30	21.83	2.09
5	14.08	0.87	20.41	1.82

Table 9.8 Comparison between the CNDP and CNDP-T solutions for Scenario 2

	CNDP	CNDP-T
Link 1 improvement, 1st year ($y_{1,1}$)	6.67	0.00
Link 2 improvement, 1st year ($y_{2,1}$)	0.00	1.39
Link 3 improvement, 1st year ($y_{3,1}$)	0.00	9.31
Link improvements in subsequent years ($y_{a,\tau}$, $a = 1,2,3, \tau = 2,...,5$)	N/A	0.00
Total expenditure	10 000.00	10 000.00
Overall consumer surplus	3.33×10^7	3.48×10^7

of O–D pair (1,2) increase over time. This raises the issue of equity – network improvements that maximize the overall consumer surplus do not necessarily benefit all O–D pairs. In fact, some O–D pairs may be impaired by the network improvements, as seen in this example.

Table 9.8 presents the improvement strategies for the two approaches. Similar to the Scenario 1 result, the CNDP and CNDP-T approaches produce very different network improvement strategies, even though both approaches use up the available budget. Link 1 is selected for improvements in the CNDP approach whereas links 2 and 3 are selected

for improvements in the CNDP-T approach. The end result is that the CNDP-T approach can gain an additional 5 per cent increase in consumer surplus as compared with the CNDP approach.

Again similar to Scenario 1, all the improvements are completed within the first year. As argued earlier, this result is reasonable, as the improvements, once completed within the first year, will also be benefiting the subsequent years. Therefore, there is no reason to postpone the improvements to a future time. This result would not be expected, however, if the formulation also considered the annual available improvement budget, the maintenance costs for the widened networks, and the discount and interest rates. Again, this simplification is deliberate, so as to allow the CNDP to be capable of finding solutions for comparison purposes.

4.3 Summary

These two scenarios demonstrate that capturing the time dimension in the network design problem will improve performance. Other than these performance improvements, the extended formulation provides a more flexible framework to allow the optimal project initiation time, phasing, and financial arrangement to be determined over the planning horizon. Unfortunately, these other aspects are not demonstrated numerically in this first study, as they cannot be modeled with the traditional CNDP, rendering the comparison impossible. We leave further extensions for a future study.

5. CONCLUDING REMARKS

This study extended the traditional CNDP by incorporating the time dimension explicitly within the formulation. This more flexible framework permits the modeling of time-dependent travel demands, changing land-use patterns, and the gradually upgraded network over the planning horizon. With this extension, one can design for the optimal project initiation time, phasing, and financial arrangements over the planning horizon.

We formulated this extended CNDP-T as a single-level optimization program, solved it through the GRG solution algorithm, and set up two scenarios to compare its performance versus that of the traditional CNDP. The results showed that the network improvement strategies obtained by the CNDP-T were superior to those of the CNDP.

As a first study of this approach, the results presented here are somewhat preliminary. Nevertheless, we believe the discussions have raised a number of research extensions. First, network improvements are often subject to annual budgets, economies of scale, the associated maintenance costs, and the tradeoff between inflation and interest rates, and the discount rate of benefits. All of these factors should be duly considered for a complete analysis. Second, the CNDP-T is non-convex due to the user-equilibrium constraints. It is beneficial to develop effective sensitivity- or penalty-based algorithms to deal with these constraints for larger-scale implementations. Third, deterministic user equilibrium may not be a realistic assignment principle. Using the formulation developed herein as a platform, one may extend it for other types of assignment principles, such as the stochastic dynamic user optimal principle (Ran and Boyce 1996), the probabilistic user equilibrium principle (Lo and Tung 2003), or others.

APPENDIX

This section derives the program (9.21)–(9.23) for Scenario 1. The objective is:

$$\max\ CS = 8760\left[\int_0^{q_1^{12}} (55 - v)\,dv - q_1^{12}\pi_1^{12} + \int_0^{q_1^{13}} (45 - v)\,dv - q_1^{13}\pi_1^{13}\right]. \tag{9A.1}$$

We formulate the constraint step by step as follows:

a. From (9.1) and (9.2), we obtain:

$$f_{1,1}^{12}\left[\eta_{1,1}^{12} - \pi_1^{12}\right] = 0, \tag{9A.2}$$

$$f_{2,1}^{13}\left[\eta_{2,1}^{13} - \pi_1^{13}\right] = 0, \tag{9A.3}$$

$$\eta_{1,1}^{12} - \pi_1^{12} \geq 0, \tag{9A.4}$$

$$\eta_{2,1}^{13} - \pi_1^{13} \geq 0. \tag{9A.5}$$

b. As there is only one path passing through each link, according to (9.3), we get:

$$v_{1,1} = f_{1,1}^{12}, \tag{9A.6}$$

$$v_{2,1} = f_{2,1}^{13}. \tag{9A.7}$$

c. From (9.4), we obtain:

$$t_{1,1} = 15 + \frac{v_{1,1}}{1 + y_{1,1}}, \tag{9A.8}$$

$$t_{2,1} = 10 + \frac{v_{2,1}}{1 + y_{2,1}}. \tag{9A.9}.$$

d. As each path is made up of one link, from (9.5), we derive:

$$\eta_{1,1}^{12} = t_{1,1}, \tag{9A.10}$$

$$\eta_{2,1}^{13} = t_{2,1}. \tag{9A.11}$$

e. As there is only one route for each O–D pair, we have:

$$\pi_1^{12} = \eta_{1,1}^{12}, \tag{9A.12}$$

$$\pi_1^{13} = \eta_{2,1}^{13}, \tag{9A.13}$$

$$q_1^{12} = f_{1,1}^{12}, \tag{9A.14}$$

$$q_1^{13} = f_{2,1}^{13}. \tag{9A.15}$$

f. According to (9.7), we have:

$$f_{1,1}^{12}, f_{2,1}^{13} \geq 0. \tag{9A.16}$$

g. The demand functions for this scenario are:

$$\pi_1^{12} = 55 - q_1^{12}, \tag{9A.17}$$

$$\pi_1^{13} = 45 - q_1^{13}. \tag{9A.18}$$

h. According to (9.11)–(9.14) we have:

$$1 + y_{1,1} \leq 2, \tag{9A.19}$$

$$1 + y_{2,1} \leq 2.5, \tag{9A.20}$$

$$1500y_{1,1} + 1000y_{2,1} \leq 1500, \tag{9A.21}$$

$$y_{1,1}, y_{2,1} \geq 0. \tag{9A.22}$$

Since equations (9A.12) and (9A.13) are the special case of constraints (9A.4) and (9A.5), the latter are redundant and can be omitted. Putting (9A.12) and (9A.13) into (9A.2) and (9A.3), we obtain (9A.16). This means that (9A.2) and (9A.3) are also redundant.

To express the program (9A.1), (9A.6)–(9A.22) in terms of $f_{1,1}^{12}$, $f_{2,1}^{13}$, $y_{1,1}$, $y_{2,1}$ only, first, we substitute (9A.14) and (9A.15), (9A.17) and (9A.18) into (9A.1), integrate and simplify the resultant objective function to obtain the following equivalent program:

$$\max 4380\left[f_{1,1}^{\,2} + f_{2,1}^{\,2}\right], \tag{9A.23}$$

subject to (9A.6)–(9A.22).

Then we further simplify the constraint set (9A.6)–(9A.22) in terms of $f_{1,1}^{12}$, $f_{2,1}^{13}$, $y_{1,1}$, $y_{2,1}$ only. We substitute (9A.6), (9A.8), (9A.10), (9A.12) and (9A.14) into (9A.17) and rearrange the resultant equation to obtain the relationship between the route flow $f_{1,1}^{12}$ and capacity enhancement $y_{1,1}$ as follows:

$$f_{1,1}^{12} = \frac{40}{15 + \dfrac{1}{1 + y_{1,1}}}. \tag{9A.24}.$$

Similarly, we put (9A.7), (9A.9), (9A.11), (9A.13), and (9A.15) into (9A.18) to get:

$$f_{2,1}^{13} = \frac{35}{15 + \dfrac{1}{1 + y_{2,1}}}. \tag{9A.25}$$

Thus, the constraints (9A.6)–(9A.15), (9A.17) and (9A.18) can be replaced by (9A.24) and (9A.25) and the above program can be expressed as:

$$\max 4380\left[f_{1,1}{}^{2} + f_{2,1}{}^{2}\right], \tag{9A.26}$$

subject to (9A.16), (9A.19)–(9A.22), (9A.24) and (9A.25).

A final step is to express the program (9A.16), (9A.19)–(9A.22), (9A.24) and (9A.25), (9A.26) in terms of $y_{1,1}$, $y_{2,1}$ only. First, we put (9A.24) and (9A.25) into (9A.26) to obtain (9A.21) and do not include (9A.24) and (9A.25) in the constraint set. Second, (9A.16) is dropped because there are no route flow variables in the new objective function and constraints (9A.19)–(9A.22). Third, we rewrite (9A.19)–(9A.21) as follows:

$$y_{1,1} \leq 1, \tag{9A.27}$$

$$y_{2,1} \leq 1.5, \tag{9A.28}$$

$$1.5y_{1,1} + y_{2,1} \leq 1.5. \tag{9A.29}$$

By graphical means, we observe that equations (9A.27) and (9A.28) are redundant and can be deleted. The final constraint set consists of (9A.22) and (9A.29), which is equivalent to the constraint set (9A.22) and (9A.23).

REFERENCES

Abadie, J. and J. Carpentier (1969), 'Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints', in R. Fletcher (ed.), *Optimization*, New York: Academic Press, pp. 37–47.

Boyce, D.E. (1984), 'Urban transportation network equilibrium and design models: recent achievements and future prospects', *Environment and Planning*, **16A**, 1445–74.

Boyce, D.E. and B.N. Janson (1980), 'A discrete transportation network design problem with combined trip distribution and assignment', *Transportation Research*, **14B**, 147–54.

Cascetta, E., M. Gallo and B. Montella (1998), 'Models and algorithms for the optimization of signal setting on urban networks with stochastic assignment', Sixth Meeting of the EURO Working Group on Transportation, 9–11 September, Gothenburg, Sweden: School of Mathematics and Computing Science, Chalmers University of Technology.

Chen, M.Y. and A.S. Alfa (1991), 'A network design algorithm using a stochastic incremental traffic assignment approach', *Transportation Science*, **25** (3), 215–24.

Chen, O.J. and M.E. Ben-Akiva (1998), 'Game-theoretic formulations of interaction between dynamic traffic control and dynamic traffic assignment', *Transportation Research Record*, **1617**, 179–88.

Current, J. and H. Min (1986), 'Multiobjective design of transportation networks: taxonomy and annotation', *European Journal of Operational Research*, **26** (2), 187–201.

Davis, G.A. (1994), 'Exact local solution of the continuous network design problem via stochastic user equilibrium assignment', *Transportation Research*, **28B** (1), 61–75.

Fisk, C.S. (1984), 'Game theory and transportation systems modeling', *Transportation Research*, **18B** (4–5), 301–13.

Friesz, T.L. (1985), 'Transportation network equilibrium, design and aggregation: key developments and research opportunities', *Transportation Research*, **19A** (5–6), 413–27.

Friesz, T.L., G. Anandalingam, N.J. Mehta, K. Nam, S.J. Shah and R.L. Tobin (1993), 'The multiobjective equilibrium network design problem revisited: a simulated annealing approach', *European Journal of Operational Research*, **65** (1), 44–57.

Friesz, T.L. and P.T. Harker (1983), 'Multicriteria spatial price equilibrium network design: theory and computational results', *Transportation Research*, **17B** (5), 411–26.

Gartner, N.H. and C. Stamatiadis (2001), 'Combining traffic assignment and adaptive control in a dynamic traffic management system', in E. Schnieder and U. Becker (eds), *Control in Transportation Systems 2000: Proceedings of the 9th IFAC Symposium*, vol. 1, New York: Elsevier Science, pp. 281–6.

Heydecker, B.G. (2002), 'Dynamic equilibrium network design', in M.A.P Taylor (ed.),

Transportation and Traffic Theory in the 21st Century: Proceedings of the 15th International Symposium on Transportation and Traffic Theory, New York: Elsevier Science, pp. 349–70.
Lam, W.H.K., M.L. Tam and M.G.H. Bell (2002), 'Optimal road tolls and parking charges for balancing the demand and supply of road transport facilities', in M.A.P. Taylor (ed.), *Transportation and Traffic Theory in the 21st Century: Proceedings of the 15th International Symposium on Transportation and Traffic Theory*, New York: Elsevier Science, pp. 561–82.
LeBlanc, L.J. (1975), 'An algorithm for a discrete network design problem', *Transportation Science*, **9** (3), 183–99.
Lo, H. (2002), 'Trip travel time reliability in degradable transport networks', in M.A.P Taylor (ed.), *Transportation and Traffic Theory in the 21st Century: Proceedings of the 15th International Symposium on Transportation and Traffic Theory*, New York: Elsevier Science, pp. 541–60.
Lo, H. and W.Y. Szeto (2002), 'A cell-based variational inequality formulation of the dynamic user optimal assignment problem', *Transportation Research*, **36B** (5), 421–43.
Lo, H. and Y.K. Tung (2003), 'Network with degradable links: capacity analysis and design', *Transportation Research*, **37B**, 345–63.
Magnanti, T.L. and R.T. Wong (1984), 'Network design and transportation planning: model and algorithm', *Transportation Science*, **18** (1), 1–55.
Maher, M.J., X. Zhang and D. Van Vliet (2001), 'A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows', *Transportation Research*, **35B** (1), 23–40.
Marcotte, P. (1986), 'Network design problem with congestion effects: a case of bilevel programming', *Mathematical Programming*, **34** (2), 142–62.
Meng, Q., H. Yang and M.G.H. Bell (2001), 'An equivalent continuously differentiable model and a locally convergent algorithm for the continuous network design problem', *Transportation Research*, **35B** (1), 83–105.
Ran, B. and D. Boyce (1996), *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*, Berlin: Springer.
Smith, M.J. and T. Van Vuren (1993), 'Traffic equilibrium with responsive traffic control', *Transportation Science*, **27** (2), 118–32.
Tzeng, G.H. and S.H. Tsaur (1997), 'Application of multiple criteria decision making for network improvement', *Journal of Advanced Transportation*, **31** (1), 49–74.
Wan, K.H., H. Lo and C.W. Yip (2002), 'Optimal integrated transit network design', in Wang, C.P. et al. (eds), *Proceedings of the Seventh International Conference of the ASCE Applications of Advanced Technologies in Transportation*, Boston, MA: American Society of Civil Engineers, pp. 736–43.
Wardrop, J. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institute of Civil Engineers*, Part II, **1**, pp. 325–78.
Wong, S.C. and H. Yang (1997), 'Reserve capacity of a signal-controlled road network', *Transportation Research*, **31B** (5), 397–402.
Yang, H. and M.G.H. Bell (1998), 'Models and algorithms for road network design: a review and some new developments', *Transport Reviews*, **18** (3), 257–78.
Yang, H. and Q. Meng (2000), 'Highway pricing and capacity choice in a road network under a build–operate–transfer scheme', *Transportation Research*, **34A** (3), 207–22.

10. Estimating link delays for arterial streets

Elliott A. Torres, Peter C. Nelson, Nagui M. Rouphail and Joseph Raj*

1. INTRODUCTION

This chapter explores the use of artificial intelligence techniques to improve the efficiency and accuracy of delay estimates for arterial streets. Results from this process may be applied to critical functions of advanced traveler management systems (ATMS) and advanced traveler information systems (ATIS), including surface street incident detection. To demonstrate the feasibility and applicability of the proposed concepts in intelligent transportation systems (ITS), the technique of artificial neural networks was selected. These neural networks were calibrated (trained) and validated with actual field data sets. The use of field data is considered superior to traffic simulation models as the latter can only approximate field data. Simulation models require considerable input, as well as extensive verification, calibration and validation. They also do not reflect irregularities that can be observed in the field. The normal range of operating conditions in the field include inconsistent driver behavior, variation in traffic volumes (free traffic flow to saturated) and composition, variation in platoon dispersion and associated signal control, variation in on-line data (such as probe and detector frequency), parking/stopping, and other factors which produce inconsistent data.

An artificial neural network is an information processing system that is non-algorithmic, non-digital, and intensely parallel (Caudill and Butler 1992). Unlike traditional computing systems, neural networks are capable of learning how to classify input/output patterns that are both linear and nonlinear (Lippmann 1987). This capability makes neural networks suitable for solving complex problems like producing delay estimates at signals. In addition, neural networks are highly fault tolerant in that given an input pattern with a lot of noise or disturbance, neural networks are still capable of classifying that input pattern and producing a reasonable result (Wasserman 1987). Neural networks are also capable of learning how to classify input/output patterns, and to use outputs from previous time intervals as inputs to subsequent time intervals. This allows dynamic changes in traffic conditions to be reflected in the estimation of delays.

Data from inductive loop detectors were chosen as the inputs to the neural networks and a statistical alternative using regression models; delay was the selected output. Loop detector data are routinely available for many arterials equipped with modern signal systems. The selection of delay (Figure 10.1) over travel time was made with the following considerations in mind. Delay is not dependent on the length of the link, while other factors such as intervehicular and signal delays are accounted for. By subtracting the constant free-flow/cruise time, based on the lowest observed travel time, time-varying travel

$$delay = \begin{cases} 0 & \text{if } tt \le free \\ tt - free & \text{if } tt > free \end{cases}$$

tt – travel time

$free$ – free-flow travel (cruise) time

$delay$ – delay time

Figure 10.1 Calculation of delay

time is expressed in terms of delay alone. Conversion from detector output to travel time was done on the basis of link-specific formulae. Once the free-flow travel time is calibrated appropriately for a link, delay becomes the varying part within the total travel time. Hence signal delay, delays due to traffic interaction, pedestrian interaction, weather and so on are reflected in the second component of travel time, such as delay. Therefore, it is only necessary to establish this relationship between detector flow/occupancy and delay to produce meaningful output.

The research results reported in this chapter include the development of neural network architectures as well as regression models. Section 2 describes the motivation for this work. Section 3 discusses the issues related to collection and screening of field data. The regression models are introduced in Section 4. Neural network architectures are explored in Section 5. Section 6 provides a comparison of selected neural networks models with regression. Concluding remarks are discussed in Section 7.

2. MOTIVATION AND CONCEPT

The use of dynamic real-time data for travel times on arterial streets is applicable to a variety of ITS applications:

- short-term travel times,
- congestion-level predictions,
- modes of travel,
- route selections,
- traffic management systems,
- traffic control systems, and
- incident management systems.

These applications require the ability to predict traffic flow patterns, estimate link travel times, and detect incident occurrences.

Traditional approaches to travel time estimates include simple data selection, statistical averaging, and statistical regression methods. The last were chosen as the baseline for com-

parison to the neural networks in this research. The first two ATIS applications in the United States, Pathfinder and Travtek, utilized a simple selection and aging technique (Sisiopiku and Rouphail 1993). While this approach was successful, it lacked the ability to integrate data from all sources, as only the source with the maximum score was selected. The ADVANCE (Advanced Driver and Vehicle Advisory Navigation ConcEpt) Project located in the metropolitan Chicago area was an ambitious, multi-year demonstration program in Illinois, aimed at the design, implementation, and evaluation of an in-vehicle navigation and route guidance system with dynamically updated travel time information (Boyce et al. 1991; Dillenburg et al. 1995). ADVANCE converted arterial detector data into travel times using a broken-line regression approach. It subsequently used a simple weighted averaging technique for fusing this data with the probe vehicle data (Berka et al. 1995).

Studies conducted by Ritchie et al. (1992) have shown that artificial neural networks can be applied very effectively to incident detection on freeways using detector data. Work by Palacharla (1995) has further demonstrated that a technique based on both fuzzy logic and artificial neural networks can significantly improve the accuracy of arterial travel time estimation in comparison to other pattern recognition approaches that used simulation data only for training. Therefore, artificial neural networks were selected as the most suitable technique to demonstrate the feasibility and applicability of this class of methods to the challenging problem of travel time estimation on arterial streets.

Backpropagation, backpropagation through time, counterpropagation, cascade correlation and recurrent cascade correlation neural network architectures were developed into appropriate neural network models for further study. The neural network models (as well as the comparative regression model) were put through the process of calibration with the screened data. Once the neural networks were trained, their performance was assessed through validation experiments. This validation cycle involved presenting the trained networks with field data sets not involved in the calibration phase. The process of calibration and validation is performed iteratively and the decision to stop training is made on the basis of whether the validation error is acceptable (Figure 10.2). The results produced during the validation phase were then statistically evaluated.

3. DATA COLLECTION AND DATA SETS

Field data collected from the ADVANCE Project were used both for training and validation. Preliminary investigation indicated that a segment of Dundee Road was appropriate for the purposes of this research. The link section (Figure 10.3) of Dundee Road between Dundee/Portwine (a minor unsignalized T-intersection) and Dundee/Milwaukee (a major signalized semi-actuated intersection) was selected. The approximate length of the link was 925 metres. This link had higher occupancies and flows than other segments across the sub-network. In addition to detector data, ADVANCE-equipped vehicles also provided probe reports consisting of link travel times for this area. These travel times were subsequently converted into delay, based on the formula specified previously (see Figure 10.1). Supplemental field data were also collected to complement the ADVANCE probe reports, through a license plate matching technique described in Section 3.2. The travel times collected from ADVANCE and the supplemental field data collection consisted only of vehicles that traveled straight through the intersection.

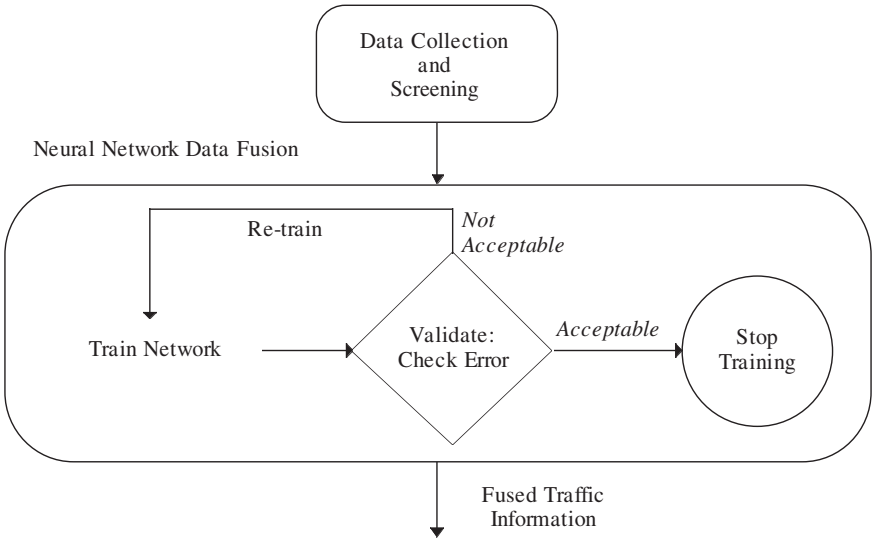


Figure 10.2 Neural network training process

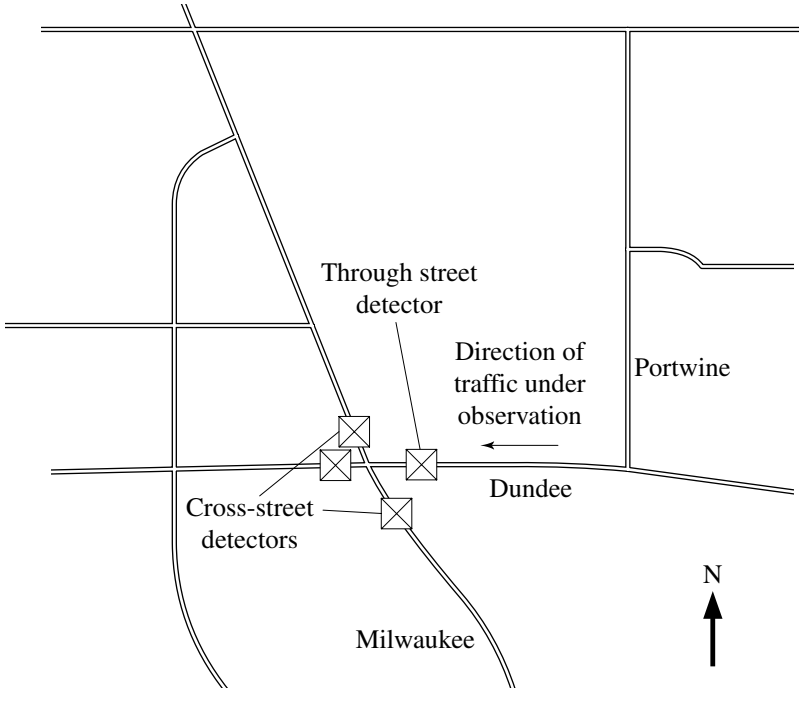


Figure 10.3 Arterial roadway segment Dundee/Portwine to Dundee/Milwaukee

The training of candidate neural network models and comparative regression models required the definition of a set of inputs and outputs. The inputs used included aggregated 5-minute flow and occupancy data from loop detectors. The outputs are the aggregated travel time reports from vehicles that traveled the segment. It was found that the average delays were more highly correlated to the mean of independent variables (flow and occupancy) than the median or the mode. Therefore, the average delay was obtained from all the probes traversing the study link over 5-minute aggregation periods. The outputs represent the expected current link delays, given the conditions described through the input data, and evaluated on the basis of errors, biases, good-fit and so on. Historical and anecdotal information were not used in the study. However, this information should be considered in future neural network models. All data was *link* and *time period* specific. The conversion of detector output to delay may be done on the basis of link-specific or generalized formulae. It is assumed that exogenous variables such as signal delay, free-flow speed, traffic intensity, weather and so on are reflected in the detector output in the link-specific formulae. Therefore, it is only necessary to establish the relationship between detector flow/occupancy and delay. However, the generalized formulae require most of the significant exogenous variables. Link-specific models, while easier to use and calibrate, are not transferable. A link-specific model was developed due to constraints of on-line data, restricting the development of a generalized model. When the data collection was completed an aggregation was then performed for each 5-minute time period to obtain a single data point for each period.

3.1 ADVANCE Data

Loop detector information was collected in 5-minute intervals. These data included the time stamp, the occupancy and the flow of traffic across the detector. ADVANCE probe reports were obtained every time a vehicle crossed the entire selected link. These data were then screened to remove any invalid loop detector and probe report data. Initial analysis of the data using both neural networks and comparative regression models revealed that the number of probe reports being observed was inadequate in order to derive a true statistical average for the probe travel times/delays. The data consisted of approximately 1–4 probe reports per 5-minute interval. Further investigation suggested that 5–10 probe reports per 5-minute interval were needed to represent the population mean. Based on this information it was decided that a supplemental data collection for the link was necessary to be used for training and validation in place of the probe reports.

3.2 Supplemental Data Collection

Field data were collected during a Tuesday and a Thursday afternoon and included off-peak and peak times. The data collection was performed using a license plate matching method. This approach was not cost-prohibitive, requiring a minimum number of observers. The field data were subsequently screened and processed to obtain travel times for vehicles traveling the link.

The method is briefly described here. Two observers were positioned at each entry and exit point of the link with laptop computers. The system clocks on the laptops were synchronized with that of the loop detectors. The last three digits of the license plates were

entered into a laptop computer held by each observer. The data collection software automatically time stamped each entry and wrote it into a flat file. Upon completion of the collection period, the entry and exit point data were merged. License plates were then matched and the travel time (the difference between the two recorded times) for each match vehicle was calculated. Based on the frequency of observation points, the upper and lower limits for the match was determined to be between 30 to 180 seconds (Figure 10.4). Thus an approximate cruise (free-flow) travel time of 30 seconds was observed for this link.

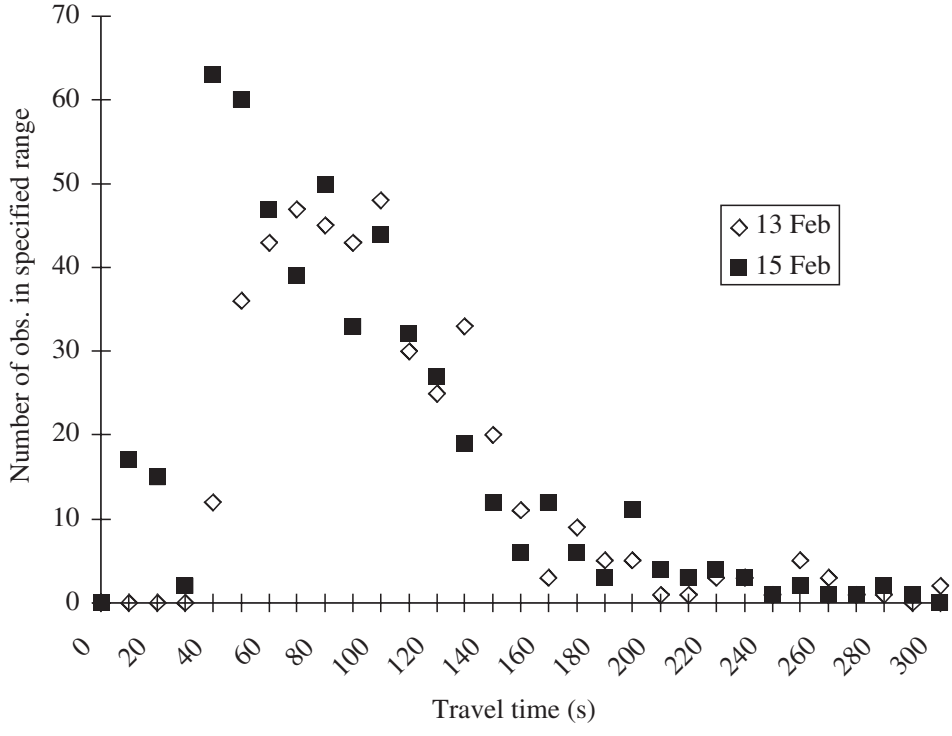


Figure 10.4 Travel time distributions on two different days

3.3 Data Sets

The generation of training and validation data sets is straightforward. As the complete database of travel data was generated and processed, a certain percentage of records were extracted for validation and the remaining data were selected for training. Using this method to separate the data we were assured that the network was being trained and validated with distinct data during each step. Apart from this, the use of recurrent backpropagation neural networks required the use of prior travel time data for subsequent training and validation. The entire data set for either of the days collected was used for training/validation where one day is declared the training day and the other the validation day. The typical training data set consisted of 34 patterns consisting of loop detector inputs,

both the through-street and cross-street, and 5-minute aggregated delay reports for the output. The validation test set consisted of 49 patterns.

4. REGRESSION ANALYSIS

The conversion of loop detector data into travel times using a conventional approach such as regression analysis is quite complex and the parameters involved are numerous, particularly under interrupted arterial traffic conditions (Sisiopiku and Rouphail 1994b). The link studied was a two-lane road section with a left-turn bay. By subtracting the constant free-flow/cruise time based on the lowest observed travel time, varying travel time could be expressed in terms of delay alone. Delay is preferred over travel time, as the effect of the length of the link is not directly considered, whereas other factors such as intervehicular and signal delays are accounted for.

4.1 Exploratory Analysis of Variables

Table 10.1 presents the variables that were explored for use in the initial statistical analysis, correlations, plots and the subsequent regression model analysis. The average delay in a 5-minute interval over the subject link had a mean value of approximately 48.84 seconds. The standard deviation, minimum and maximum observed values are also depicted in the table. The present occupancy (OCC) on the subject link is reported on each lane every 5 minutes. The detector cross-occupancy (COCC) is the maximum of the cross-occupancies on the two cross-legs at the intersection. The square component is $COCC^2$. Flow is represented in vehicles per hour per lane whereas cross-flow (CFLOW) is the maximum of either cross-leg-flow rate.

Multiple regression analysis techniques were used to derive the requisite regression models using the SAS 6.08 package. Models with both linear and polynomial terms were explored. The degree of the polynomial is usually determined by building the model. The model is built by sequentially fitting equations with higher-order terms until a satisfactory degree of fit has been accomplished. For further details on regression model development and application to arterial delay estimation, the reader is referred to Sen and Srivastava

Table 10.1 Simple statistics for a sample training data set

Variable	Mean	Std. Dev.	Minimum	Maximum
Delay	48.84	24.47	6.83	119.08
OCC	18.50	11.62	2.00	49.00
OCC^2	474.31	516.41	4.00	2401.00
FLOW	548.15	139.59	240.00	762.00
COCC	27.25	15.23	4.00	60.00
$COCC^2$	968.85	939.95	16.00	3600.00
CFLOW	484.62	91.61	276.00	660.00
$CFLOW^2$	243030	84817	76176	435600.00

Note: Number of reports: 39.

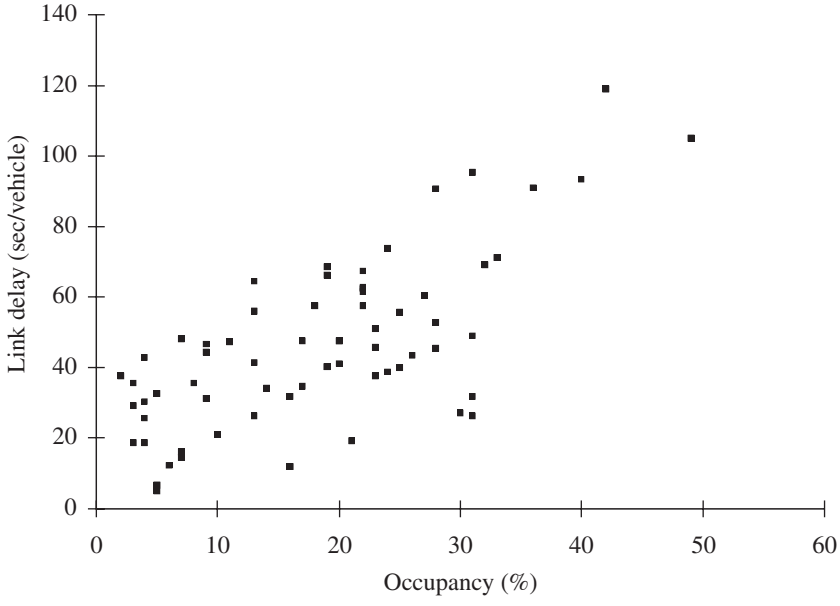


Figure 10.5 Plot of delay versus occupancy (OCC)

(1990), SAS System for Regression (Freund and Littell 1986) and Sisiopiku and Rouphail (1994b).

The initial step in model development was to generate a number of scatter plots (Figures 10.5–8) between delay and the candidate independent variables. Link delay versus link occupancy exhibit a linear trend, as depicted in Figure 10.5. In macroscopic traffic theory, however, link speed in most cases is considered to vary linearly with density (or occupancy). Implicitly therefore, travel time (or delay) will have a nonlinear relation to occupancy. Therefore it was decided to retain the square term of occupancy for further exploration. In Figure 10.6, delay appears to be insensitive to flow at low flow rates, and then increases in a nonlinear fashion at high flows. Again, this trend is consistent with most macroscopic flow models. The same pattern emerges in Figure 10.7 with cross-flow. Higher-cross flows imply a longer red time on the main street (as the signal is semi-actuated) and subsequently higher delays on the subject link. Finally, there is an evident trend between cross-occupancy and link delay, as depicted in Figure 10.8. High cross-occupancy is an indication of increased demand on the cross-street, and an ensuing higher demand on the subject link. One can also observe an increase in the variance of link delays as cross-occupancy increases.

Prior to model fitting, two additional modifications were made to the data set. All flows and occupancies were converted into equivalent values on a per lane basis, to enable a more general application of the model. Second, the intercept term was dropped on (lack of) statistical significance and practical considerations, that is, zero flow/occupancy should yield zero delays.

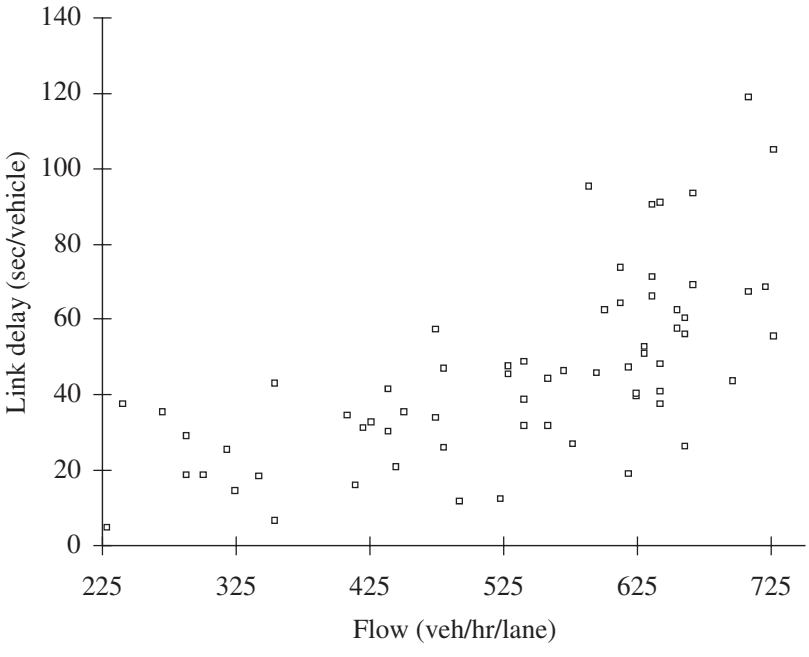


Figure 10.6 Plot of delay versus flow (FLOW)

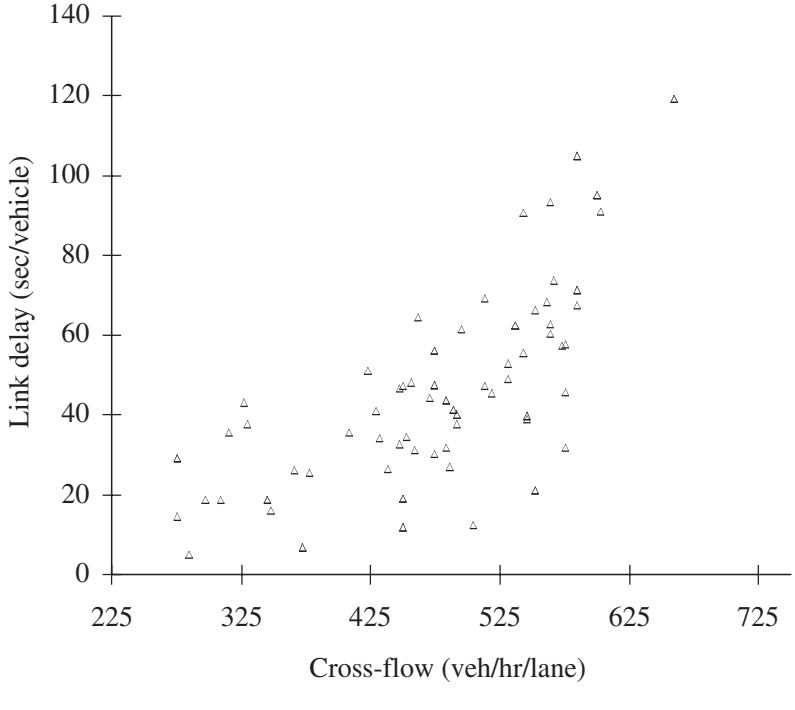


Figure 10.7 Plot of delay versus cross-flow (CFLOW)

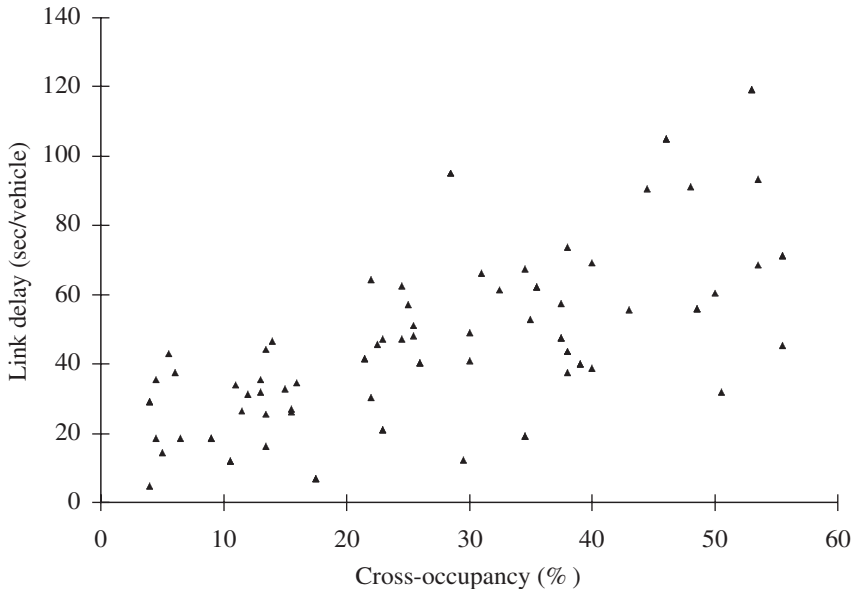


Figure 10.8 Plot of delay versus cross-occupancy (COCC)

Table 10.2 Correlation coefficients

	Delay	FLOW	OCC	OCC^2	CFLOW	COCC	$CFLOW^2$	$COCC^2$
Delay	1.00	0.63	0.76	0.78	0.65	0.60	0.69	0.58
FLOW	0.63	1.00	0.76	0.64	0.79	0.73	0.76	0.61
OCC	0.76	0.76	1.00	0.95	0.79	0.76	0.80	0.69
OCC^2	0.78	0.64	0.95	1.00	0.67	0.67	0.70	0.64
CFLOW	0.65	0.79	0.79	0.67	1.00	0.75	0.99	0.64
COCC	0.60	0.73	0.76	0.67	0.75	1.00	0.75	0.97
$CFLOW^2$	0.69	0.76	0.80	0.70	0.99	0.75	1.00	0.65
$COCC^2$	0.58	0.61	0.69	0.64	0.64	0.97	0.65	1.00

Note: All flows and occupancies expressed on a per lane basis.

4.2 Recommended Model

The exploratory runs with each of the selected variables included checking the model fit, the residuals and the cross-correlations. Two variables were found to be significant: occupancy^2 (OCC^2) and cross-flow (CFLOW). Note that flow was strongly related to occupancy and could not appear in the model with occupancy. Nor could it appear with cross-flow. Second, cross-occupancy (COCC) was also quite highly related to occupancy and thus rejected by the model. Upon checking the cross-correlation of the field delay to each of the variables as shown in Table 10.2, the selection of OCC^2 and cross-flow should

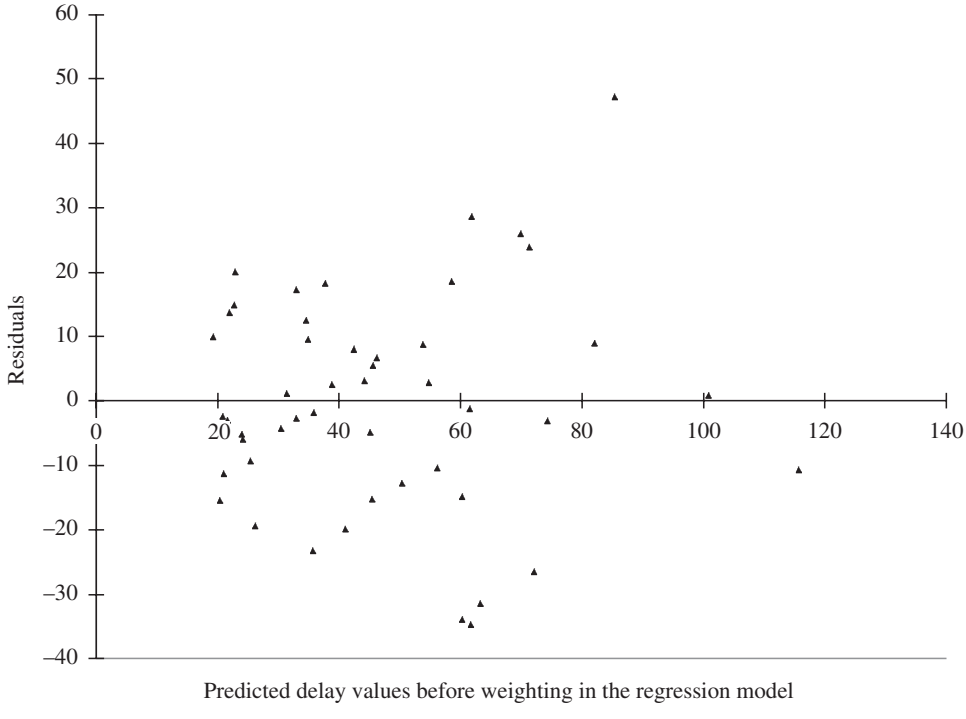


Figure 10.9 Residuals versus predicted values showing the effect of heteroscedasticity

not be a surprise. These two variables correlated less to each other and much higher with delay.

A test for multicollinearity between the variables selected was conducted by checking the variance inflation factors (VIFs) and the COLLINOINT option. It should be remembered that the existence of multicollinearity does not affect the estimation of the dependent variable but it is a contributor to the variances of the regression coefficients. Referring to Table 10.2 we also note that in comparison to the correlations of cross-flow and the other variables, the correlation between CFLOW and OCC^2 is smaller than with FLOW or OCC, reaffirming their significance in the model. The VIFs obtained were within the limits; the eigen values were far greater than 0 and the condition indices (square root of the ratio of the largest to the smallest eigen value) were $\ll 30$ (the rule of thumb based on Freund and Littell 1986). The selected variables thus appear to be excellent candidates in the model with negligible multicollinearity.

The proposed two-variable model was initially run and a plot, residuals versus predicted delays, was generated, as shown in Figure 10.9. Examination of the plots revealed the presence of heteroscedasticity (Sen and Srivastava 1990) with greater errors at the larger predicted values. Weighted least squares (WLS) was applied to reduce heteroscedasticity. WLS is a direct application of generalized least squares such that the estimated parameters obtained by this method minimize the weighted residual sum of squares:

$$\Sigma\, w_i\,(Y_i - B_0 - B_1 x_1 - \ldots\, B_m x_m)^2$$

where w_is are a set of non-negative weights assigned to the individual observations. Observations with small weights contribute less to the sum of squares and thus have less influence on the estimation of parameters, and vice versa for observations with larger weights. Based on the use of generalized least squares, the best linear unbiased estimates are obtained if the weights are proportional to the variances of the residuals. By using the iterative NLIN option in SAS (Freund and Littell 1986), a weight of 1/(model.delay)^x was applied and the heteroscedasticity removed (note that the value of x varied between 1 to 2 in the data sets). The resulting residual plot is shown in Figure 10.10.

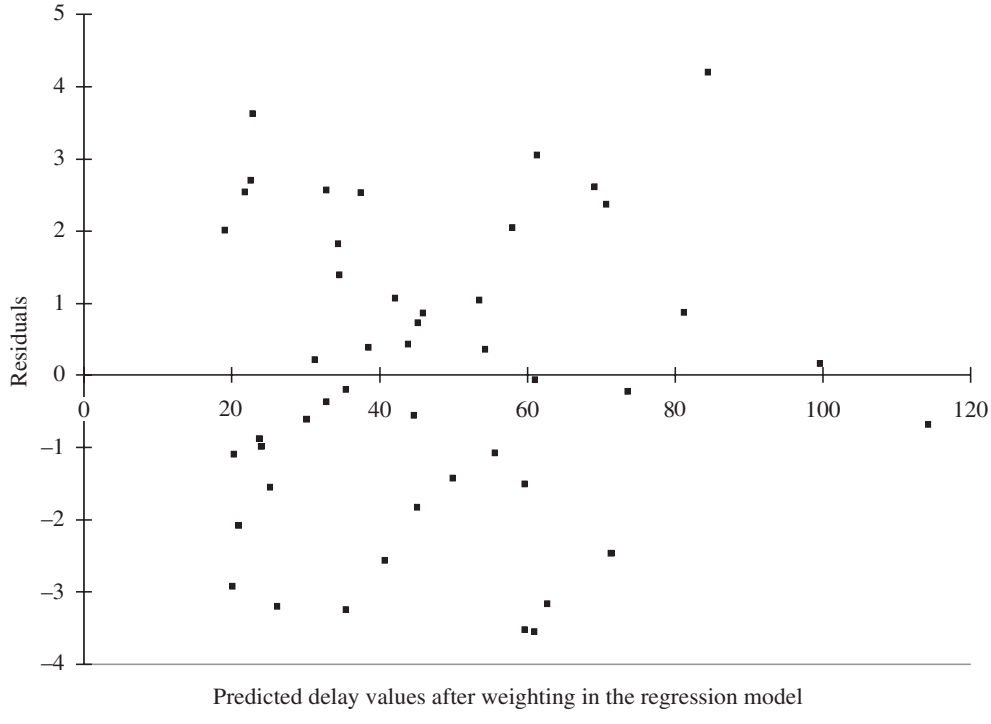


Figure 10.10 Residuals versus predicted values after removal of heteroscedasticity

The final regression model developed was of the form:

$$\text{Link Delay} = B_0\text{*OCC}^2 + B_1\text{*CFLOW.}$$

For the calibration/training set, the following model was obtained with an R-sq. of 0.90, a high F value and low mean squared errors:

$$\text{Delay} = 0.026\text{*OCC}^2 + 0.73\text{*CFLOW.}$$

Overall, all models appear to be well calibrated and all statistics indicate that the models are quite acceptable. When applied to the validation data set the model delays had high

correlations with the measured delays (an average of 0.794). Further details on the performance of the regression model appear in Section 6.

By way of comparison, Sisiopiku and Rouphail (1994a) found that there was a strong correlation between flow and occupancy, restricting their use in the same model. Our study also confirmed this. None of the earlier studies referenced in Sisiopiku and Rouphail (1994a) has implemented occupancy and cross-flow as independent variables to estimate link delay. However, occupancy and cross-flow were also incorporated in our regression model, stressing the importance of cross-flow as a contributor to delay. For example, occupancy alone had a 70 per cent correlation with the delays. By including cross-flow as an additional independent variable, the correlation with delays increased by another 10 per cent.

Most of the models developed previously have not been trained and validated with field data. They also do not incorporate the nonlinear relationship between flow and occupancy, thereby underpredicting delay at high occupancies. The proposed model incorporates occupancy as a polynomial term but ignores the actual signal settings. The rationale for this approach is that the effects of the settings are implicitly accounted for in the detector output information.

5. NEURAL NETWORK ARCHITECTURES

5.1 The Architectures

Neural network architectures were developed for feed-forward and recurrent neural network models. Initially, two types of simple feed-forward neural network models were selected based on work by Palacharla (1995): the backpropagation (Rumelhart et al. 1986) and the counterpropagation (Hecht-Nielsen 1988) network models. The backpropagation neural network model uses a supervised training algorithm to adjust weights in multi-layer neural networks. The neural network model shown in Figure 10.11 is a three-layered backpropagation neural network. The input layer has two inputs, one for flow and the other for occupancy. The hidden layer consists of three units and the output layer is a single unit

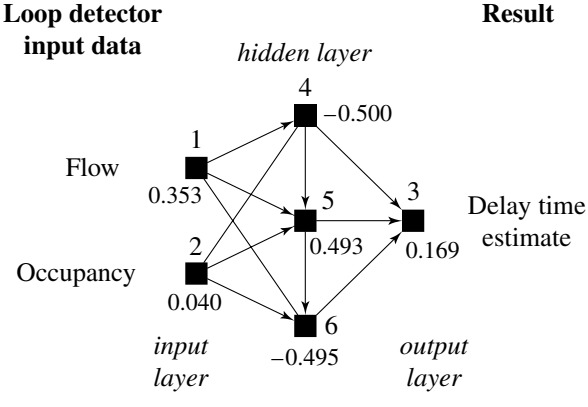


Figure 10.11 Backpropagation neural network

representing the delay estimate. The backpropagation algorithm teaches the network how to associate inputs with outputs by propagating the output error back through the network layer by layer, adjusting weights at each layer. The counterpropagation model (Hecht-Nielsen 1988) uses a combination of techniques that behave more like an adaptive look-up table capable of generalization. Generalization is an important characteristic for the neural network, enabling it to handle incomplete and noisy data, and giving it the ability to classify data which the neural network has not seen before. Additionally, the development of three more neural network architectures was considered: backpropagation through time (Hecht-Nielsen 1988), cascade correlation (Fahlman and Lebiere 1991) and recurrent cascade correlation (Fahlman 1991). The backpropagation-through-time neural network is a fully recurrent architecture that has the capability of taking the output from the previous time interval as input to the present time interval. The dynamic nature of this model has the potential to be very useful because it considers changes in current traffic flow when estimating travel time outputs. Cascade correlation was developed to overcome limitations present in ordinary feed-forward neural networks such as backpropagation. Recurrent cascade correlation, a recurrent implementation of the cascade correlation neural network, was also considered.

5.2 Backpropagation Neural Networks

The backpropagation neural network requires two sources of input flow and occupancy values, obtained from the loop detectors. These values are mapped into fuzzy sets (Zadeh 1965; Palacharla and Nelson 1995). The occupancy is divided into a fuzzy set vector consisting of 16 members and the flow into a fuzzy set of nine members. Two backpropagation models were used, with the primary difference being the number of units in the middle layer. The two models were chosen based on previous experience (Palacharla and Nelson 1995) with simulated data. The output layer corresponds to the delay value generated by the neural network.

The two parameters used in backpropagation (Werbos 1974; Parker 1985; and Rumelhart et al. 1986) are η, a learning parameter that specifies the step width of the gradient descent, and $d_{\max}$, the maximum tolerated difference between the training value and the output unit. The gradient descent will directly affect the speed at which the neural network converges and should be chosen carefully. A value which is too small may cause the network to get stuck in a local minima; a value too large may cause the network to overstep the minimum. The maximum difference is the value at which the network should propagate the error back through the network. For example, if the value is set to 0.1, the network will treat generated values from 0.9–1.0 to be represented as 1.0 and values from 0.0–0.1 as 0.0. This prevents the network from overtraining, which impairs generalization.

The initial backpropagation neural network was made up of 12 hidden units, and the training parameters were $\eta = 0.2$ and $d_{\max} = 0.1$. There were 15 output units which covered a range of delay values from 0–150 seconds. The first training session was set for 1000 epochs (training cycles), with a training data set of 34 aggregated delay reports. The validation test set was presented to the neural network every 100 cycles and consisted of 49 reports. The results of training the neural network are presented in Figures 10.12 and 10.13. The training data plot (Figure 10.12) graphs the estimated delay versus the time of day the delay report was collected, and the actual delay versus the report times of the

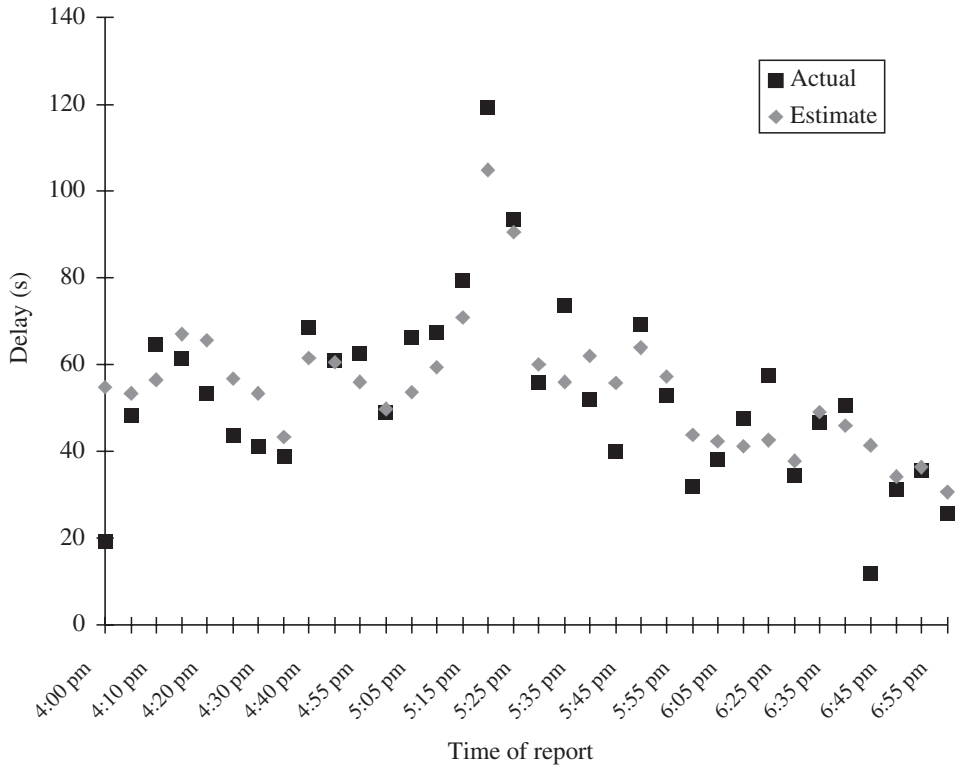


Figure 10.12 Training data plot of delays for bp_out *neural network*

training data set values. The validation data plot (Figure 10.13) graphs the same items for the validation data set. The delay values on the graphs are in seconds.

Observe that the estimated delay curve for the training data (Figure 10.12) fits the actual delay curve fairly consistently throughout the plot. As the number of training cycles increases for the training data set, the estimated delay curve will better fit the actual delay curve. The estimated delay curve for the validation data (Figure 10.13) is able to fit peaks (4:45–6:00 pm) in the actual delay curve. However, as the pattern becomes less pronounced (6:30–8:00 pm) and as traffic approaches free-flow conditions, the estimated delay curve begins to flatten out to produce an estimate delay curve which is centered and bisects portions of the actual delay curve on the plot. This appears to be caused by the network's inability to fully account for the variance generated by signal delay on the segment. At off-peak lower flows and occupancies, the sample from which the average delays are drawn is smaller, producing greater variations in actual delays. This effect is reduced when additional information is presented to the network during training, such as the cross-street detector information discussed in Section 5.4.

A summary of the results for several backpropagation neural networks is presented in Table 10.3. This table and the remaining neural network summary tables contain the following information:

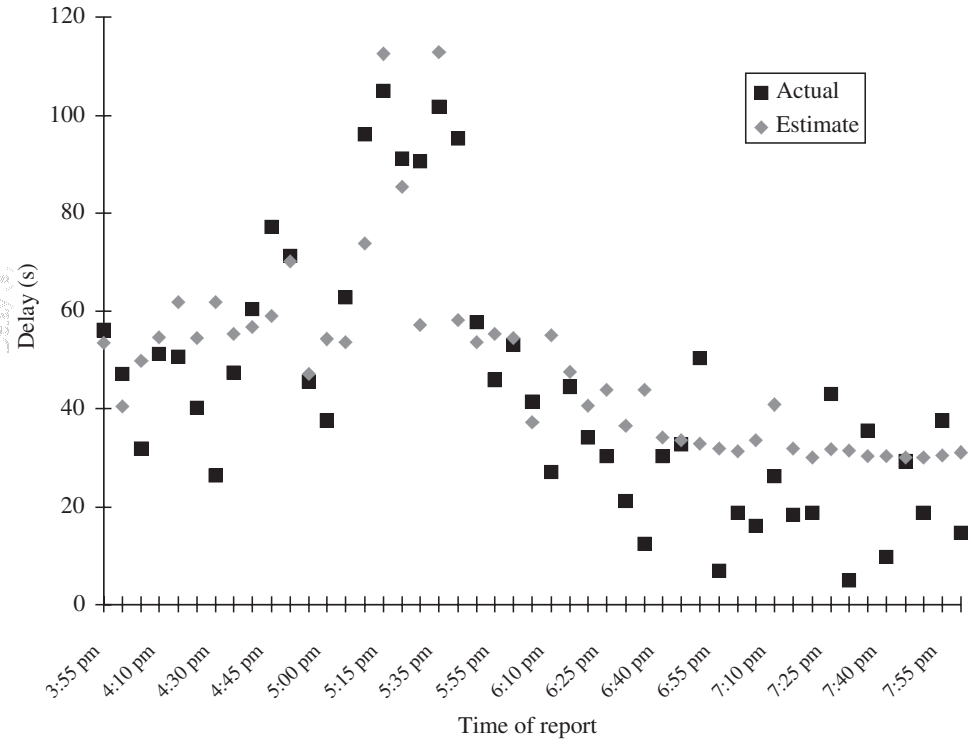


Figure 10.13 Validation data plot of delays for bp_out *neural network*

Table 10.3 Backpropagation results summary

Neural network name	Hidden units	η	d_{max}	Training cycles	Correlation validation	Correlation training	RMSE validation	RMSE training
Bp_out	12	0.2	0.1	500	0.813	0.843	15.452	11.437
Bp_out2	12	0.4	0.1	400	0.807	0.870	16.860	8.863
Bp_out3	15	0.2	0.1	500	0.804	0.861	15.765	10.828
Bp_out4	15	0.4	0.1	900	0.782	0.895	16.542	9.513

- *neural network name* – a label for each neural network being trained;
- *hidden units* – the number of hidden units in this particular network;
- *learning parameters* – these columns will vary based upon the architecture;
- *training cycles* – the number of epochs needed to train the network;
- *correlation validation* – the correlation between the actual delay and the estimated value given by the neural network for the validation data set;
- *correlation training* – the correlation between the actual delay and the estimated value given by the neural network for the training data set;

- *RMSE validation* – the root mean squared error between the actual delay and the estimated value given by the neural network for the validation data set is calculated by:

$$RMSE = \sqrt{\left[\sum_{r \in reports} (\text{actual}_r - \text{delay}_r)^2 / \#\text{ of } reports \right]}$$

- *RMSE training* – the root mean squared error between the actual delay and the estimated value given by the neural network for the training data set.

Although better recognition of the training data is observed for *bp_out2* and *bp_out4*, the other networks (*bp_out* and *bp_out3*) have better generalization capabilities. Additionally, the smaller networks (*bp_out* and *bp_out2*) have better generalization characteristics than the larger networks trained with the same learning parameters, as expected.

During training the network is tested periodically with validation data, typically every 100 cycles. This test is used to make the decision of when to stop training. The sum squared error (SSE) is given by:

$$SSE = \sum_{p \in patterns} \sum_{j \in output/units} (t_{pj} - o_{pj})^2,$$

and the mean squared error (MSE) by:

$$MSE = SSE/\#\ of\ patterns,$$

where t_{pj} is the value of the trained output unit j and o_{pj} represent the actual value of the output unit. A typical training session (Figure 10.14) will continually reduce the MSE value for both the training and validation data sets. However, as the network becomes more specialized the MSE value for the validation data set begins to rise. Near this point training should be halted, as the network will gradually begin to lose its ability to generalize. Note that the MSE tends to be more sensitive to the loss of the generalization than the RMSE. This is because the MSE is calculated at the output unit level for each member of the fuzzy set vector which has a value between 0 and 1.

5.3 Counterpropagation Neural Networks

The counterpropagation neural network model developed by Hecht-Nielsen (1988) is a combination of the self-organizing map of Kohonen (1988), and the outstar structure of Grossberg (1982), each of which correspond to a specific layer in the network. The Kohonen layer of a counterpropagation network uses competitive learning to select the closest node that corresponds to the input pattern presented. The winning node has its output set to 1 and all other nodes are set to 0. The signals from the Kohonen layer are then sent to the Grossberg layer. This third layer then proceeds to learn the averages of the output patterns presented during training based on the winner of the Kohonen layer competition. The resulting network functions as an adaptive look-up table capable of generalization (Hecht-Nielsen 1990). The generalization capability of the network allows it to produce a correct output pattern even when given an input pattern that is partially incomplete or partially incorrect. The networks train between 10 and 100 times faster than backpropagation neural networks (Hertz et al. 1991).

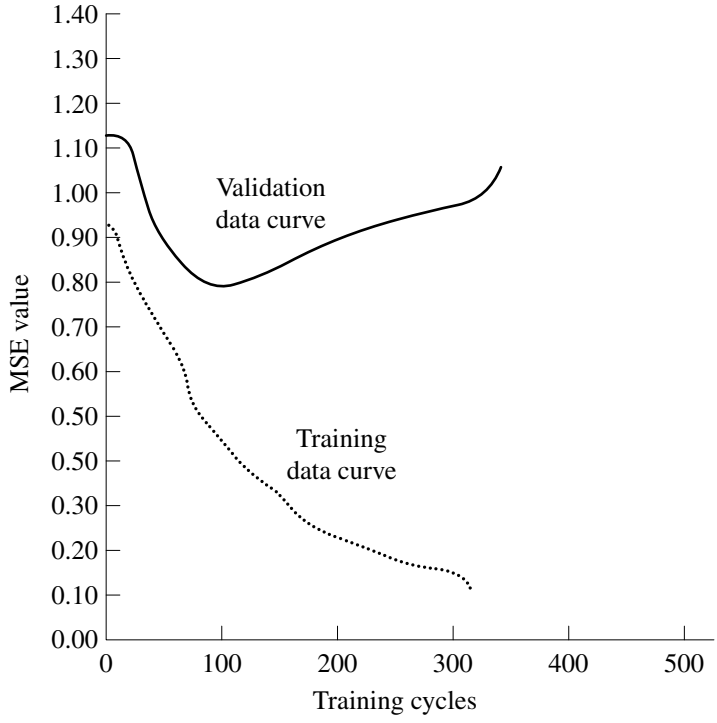


Figure 10.14 Backpropagation neural network training session

Counterpropagation neural networks require specification of three parameters α, β and θ. The first learning parameter, α, corresponds to the learning rate of the Kohonen layer. The second, β, corresponds to the learning rate of the Grossberg layer. The threshold value of the output units is represented by the final parameter, θ.

Like the backpropagation architecture the counterpropagation neural network utilized 12–15 hidden units. Training parameters were set to $\alpha = 0.2$, $\beta = 0.1$ and $\theta = 0.1$. The first training session was set to run for 1000 epochs (training cycles) with 34 aggregated delay reports. The validation test set, consisting of 49 patterns, was presented to the neural network every 100 cycles. A significant advantage observed here is the speed with which the counterpropagation neural network is trained. The neural network appears to reach the minimum error within the first 100 training cycles. This reduces the processing time needed to train a counterpropagation neural network to between 10–20 per cent of that required for a backpropagation neural network. The savings potential is particularly large when training with larger sets of data.

Like the backpropagation algorithm described above, the counterpropagation algorithm tends to produce a result that bisects the actual delay curve (Figure 10.15). A summary of results for some of the counterpropagation neural networks is presented in Table 10.4. A greater number of hidden units improved the training data recognition. However, generalization performance decreased and the validation data results were not as good compared to neural networks having a smaller number of hidden units.

5.4 Cross-street Neural Network Models

The cross-street neural network models utilize the additional information provided by the detectors located on cross-streets. These data increase the number of input nodes to 75, with 25 nodes representing each detector. Since the signals are semi-actuated, including additional information on cross-street traffic will enable the neural network to make a more informed decision.

The correlation values (Table 10.5) for these networks were similar and in some cases

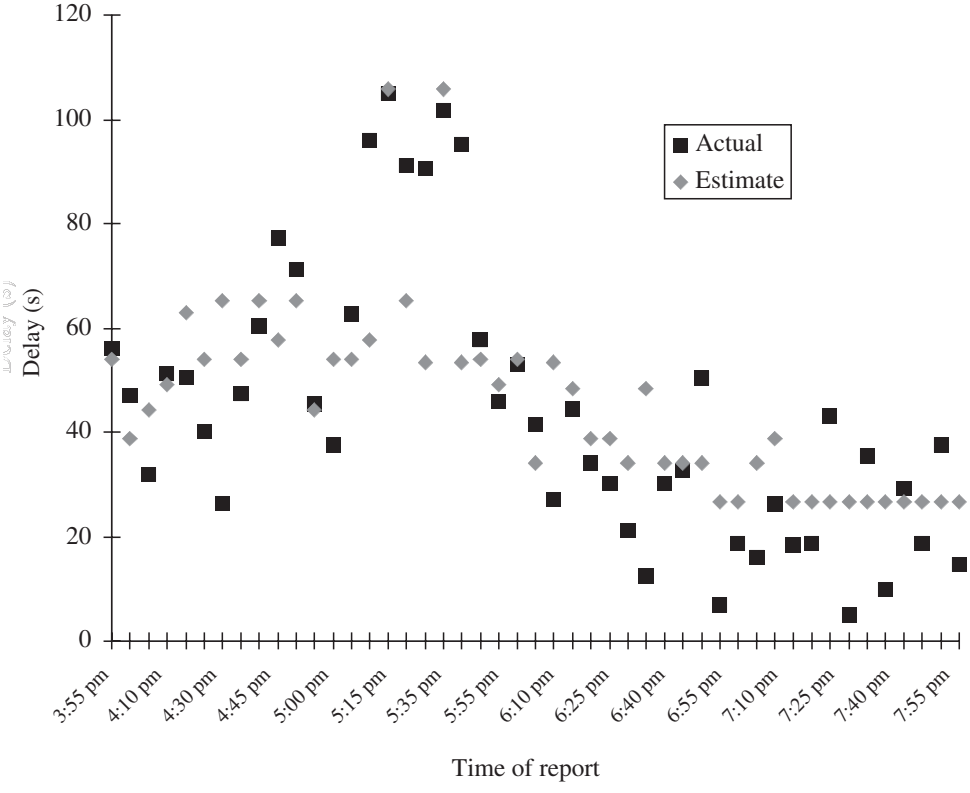


Figure 10.15 Validation data plot of delays for cp_out *neural network*

Table 10.4 Counterpropagation results summary

Neural network	Hidden units	α	β	θ	Correlation validation	Correlation training	RMSE validation	RMSE training
Cp_out	12	0.2	0.10	0.1	0.768	0.807	17.002	12.559
Cp_out2	12	0.4	0.20	0.1	0.736	0.724	17.954	14.685
Cp_out3	12	0.3	0.15	0.1	0.810	0.733	15.550	14.485
Cp_out4	15	0.4	0.20	0.1	0.652	0.753	20.125	14.001

Table 10.5 Comparison of backpropagation neural networks with 12 hidden units ($\eta = 0.2$, and $d_{max} = 0.1$)

Training cycle	With cross-street information		Without cross-street information	
	Correlation validation	Correlation training	Correlation validation	Correlation training
0	–	–	–	–
100	0.812	0.902	0.814	0.761
200	–	–	0.814	0.792
300	–	–	0.814	0.809
400	–	–	0.814	0.825
500	–	–	0.813	0.843

slightly lower when calibrating a neural network with the same data set and the same training parameters. However, it was found that a network architecture which included cross-street and cross-flow information consistently trained with approximately 5–10 times fewer training cycles than those networks which do not incorporate the additional information. Also, the correlation of the training data set for these neural networks was between 5 and 10 per cent higher (Table 10.6), without a significant decline in correlation values for validation data. This is significant because the networks' ability to classify known patterns without losing generalization is a key factor in their usefulness.

5.5 Backpropagation-through-time Neural Networks

Backpropagation through time (BPTT) is a fully recurrent neural network architecture that incorporates the following key elements of recurrent network architectures:

1. sequence recognition,
2. sequence reproduction, and
3. temporal association (Hertz et al. 1991).

The dynamic nature of arterial traffic conditions indicates a need to incorporate these characteristics into the architecture.

The results for BPTT were inferior to those of standard backpropagation. The network was able to classify portions of time periods, but was unable to generalize the overall estimates for the link. Lower correlation values were observed for both *bptt_out* and *bptt_out2* networks, particularly with the training data sets. These low values may be an indication that there is insufficient data to train a BPTT network or that there is too much noise in the input data. Considering that much higher values were obtained using other architectures, the former is more likely to be the case.

Table 10.6 Comparison of backpropagation neural networks with and without cross-street information

	Without cross-street information					With cross-street information				
Neural network	Training cycles	Correlation validation	Correlation training	RMSE validation	RMSE training	Training cycles	Correlation validation	Correlation training	RMSE validation	RMSE training
Bp_out	500	0.813	0.843	15.452	11.437	100	0.812	0.902	15.474	9.209
Bp_out2	400	0.807	0.870	15.682	10.493	100	0.742	0.948	17.795	6.748
Bp_out3	500	0.804	0.861	15.765	10.828	100	0.793	0.929	16.154	7.866
Bp_out4	900	0.782	0.895	16.542	9.513	100	0.756	0.957	17.367	6.199

5.6 Cascade Correlation Neural Networks

The cascade correlation learning architecture (Fahlman and Lebiere 1991) was developed to overcome the limitations of ordinary feed-forward neural networks. These limitations are referred to as the step-size problem and the moving target problem. The step-size problem has to do with the speed at which the traditional backpropagation algorithm learns. The gradient descent parameter controls the speed of learning during a neural network training session. As discussed previously, if too large a value is chosen there is the possibility that the network will not converge upon a good solution; if the value is too small the network will train slowly and may get stuck in a local minimum. The moving target problem has to do with the adjustment of weights in the network's hidden layer. As the network is trained, the weights of the network connections are adjusted based on the error propagated back through the network. While the network is adjusting the weights to fit one traffic pattern, it is presented with a second pattern. When adjustments are made to accommodate the second pattern the network may lose the information acquired about the first. Cascade correlation addresses both of these problems while decreasing both the network size and training time.

Cascade correlation works in the following manner. A specified number of hidden units are developed in parallel using a particular learning algorithm. After some criterion is met (for example, training cycles, error value) the best unit is selected to be placed in the network. This unit is frozen and does not change once it has been added. The learning algorithm that is used in cascade correlation is designed to maximize the correlation between the output of the unit that has just been added and the residual error signal that we are trying to eliminate. These two factors interact to eliminate the step-size problem and the moving target problem. Another feature of cascade correlation is that it eliminates much of the guesswork that is necessary to develop neural network architectures, since only the inputs and outputs need to be defined prior to training the network. The hidden units are added as needed and training is stopped when these units no longer contribute to minimizing the residual error.

Cascade correlation is also different from other neural architectures in that it is not purely a feed-forward learning algorithm. The architecture actually incorporates other feed-forward learning algorithms (that is, backpropagation). The dynamic addition of single-unit hidden layers creates the need for additional computation which is not present in ordinary feed-forward networks.

The cascade correlation algorithm is complex and requires over a dozen parameters. However, the neural network developed for the link being studied used the default settings for all but two parameters in order to adapt the network to a variety of data sets developed to train the link. These parameters are the number of hidden units (chosen to be five for this particular link) and the learning algorithm (quickpropagation or backpropagation) used for forward learning. Quickpropagation (Fahlman 1988) provided improved performance over backpropagation as expected, as quickpropagation is designed to converge faster than standard backpropagation. However, in the given data there was not a significant improvement in the training time or the correlation values when quickpropagation was used. While in some cases the network performed better on the training data sets, backpropagation maintained a slight advantage on its ability to generalize the validation data set.

The results (Table 10.7) of cascade correlation with cross-street information are

Table 10.7 Summary of cascade correlation neural network results

		Without cross-street information				With cross-street information			
Neural network	Learning algorithm	Correlation validation	Correlation training	RMSE validation	RMSE training	Correlation validation	Correlation training	RMSE validation	RMSE training
Cc_out	Backprop	0.830	0.749	14.802	14.100	0.832	0.864	14.710	10.698
Cc_out2	Quickprop	0.823	0.730	15.083	14.540	0.826	0.879	14.950	10.133

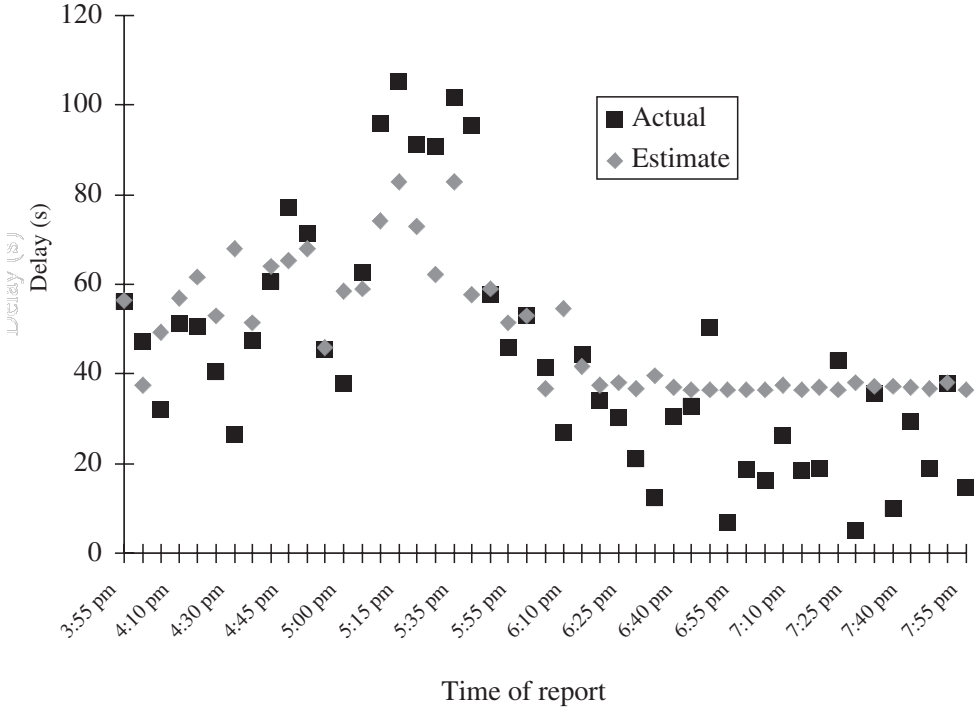


Figure 10.16 Validation data plot of delays for cc_out *neural network with cross-flow inputs*

comparable to any of the neural networks presented earlier. The graph in Figure 10.16 demonstrates the network's ability to follow the traffic pattern observed over the link. The ability of the network to perform well with a variety of data sets without modifying the training parameters makes it the preferred choice over any of the other architectures examined. The results of the remaining data sets are discussed later in the comparison with regression models.

5.7 Recurrent Cascade Correlation Neural Networks

Recurrent cascade correlation (RCC) is a recurrent version of the feed-forward cascade correlation architecture (Fahlman 1991). It provides the ability to develop a recurrent neural network while maintaining the advantages of the traditional cascade architecture. Like other recurrent neural networks, RCC can enable a network to learn time sequences. And like its feed-forward counterpart, RCC produces a near minimal network that is quick and simple to design and train. In fact, the training parameters utilized for the regular cascade neural networks proved to be the best choice also for the RCC neural network.

Like the other recurrent neural network architecture, BPTT, RCC did not perform better than the non-recurrent cascade neural networks. However, it did perform better

than backpropagation and counterpropagation. RCC appears to suffer from similar traits as BPTT, where very good correlation values are observed for the training data sets but generalization to the validation data is not greatly improved. A limitation of recurrent networks is the need to present data in chronological sequence; this is not the case with non-recurrent networks. Although recurrent cascade correlation outperforms cascade correlation with respect to training data, the latter has a higher correlation value for validation data.

6. COMPARING NEURAL NETWORKS WITH STATISTICAL MODELS

The regression model developed was unique in its representation of independent variables and was able to predict delays quite accurately; it represents a new standard in the use of regression for data fusion. The neural network chosen for direct comparison, the cascade correlation network, performed equally well.

The final choice on whether to use either a neural network or a regression model is left to the discretion of the ITS analyst. However, the project investigation did bring about the following observations which favor the use of neural networks for delay estimates:

1. The regression model's inability to handle all outliers which must be excluded.
2. The difficulty of calibrating regression models.
3. The relative ease of construction of the cascade correlation neural network.

Outlying data points are quite influential in regression and need to be carefully screened before the development of a regression model. In the regression model the data required an additional data screening step, beyond reasonability data screening, to exclude or modify data points used to develop a model. Outliers that are not removed with this additional step can give rise to skewed regression models. It was observed, however, that neural networks could appropriately handle most outliers, which were identified and mostly ignored. This observation was verified during the development of the neural network architectures by training networks with and without most outliers. The resulting correlations were within one hundredth of a point.

The graphs in Figures 10.17–23 provide comparisons between the cascade correlation neural network and the final regression model. Figures 10.17 and 10.18 depict graphs of the field, regression and neural network delays for training data. The neural network performed better than the regression model, estimating delays which are closer to the actual delay values (average correlations of 0.926 and 0.794, respectively). On closer inspection of the figure over the entire data collection the neural network more closely followed the peaks and valleys of the actual average field delay. However, the network and regression performed equally well with the validation data, estimating the delays quite similarly (Figures 10.19 and 10.20). Both neural networks and regression have a reasonably high degree of average validation correlations with the actual field delay (0.778 and 0.774, respectively). In Figure 10.21, the graph of neural network delay versus the actual field delay for validation data is shown. This plot shows a close tendency towards the 45-degree line. By comparison (Figure 10.22) regression exhibits a tendency to overestimate low field

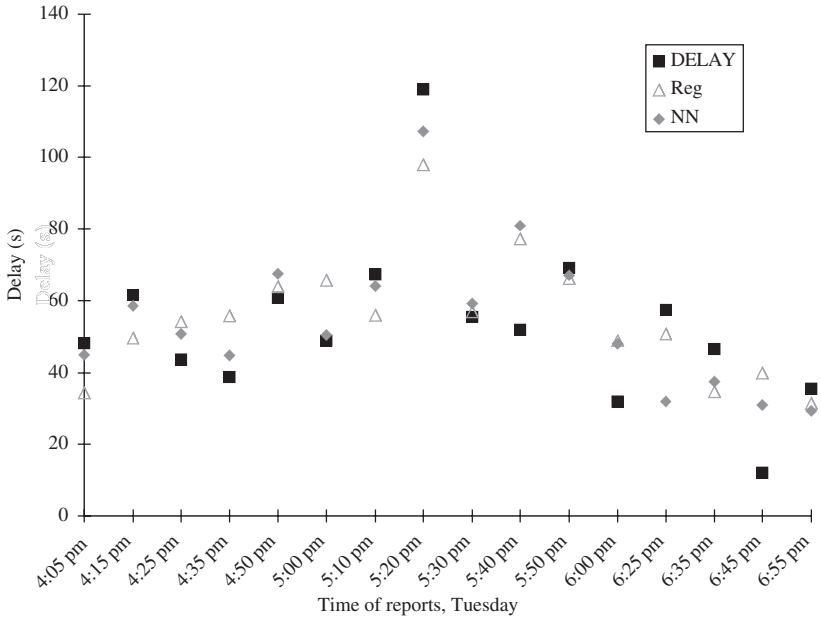


Figure 10.17 Plot of actual delay, regression (Reg) estimated delay, and neural network (NN) estimated delay versus time of report for training data on Tuesday, 13 February

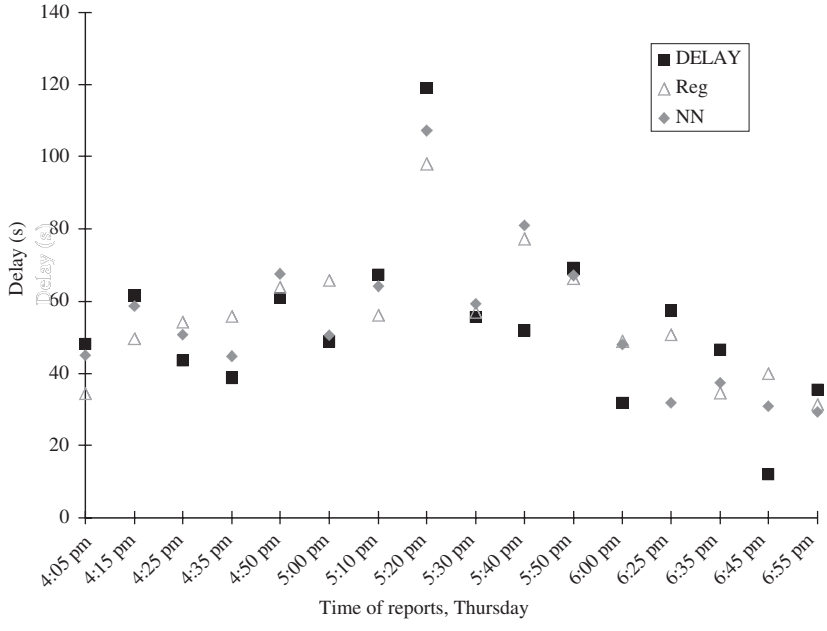


Figure 10.18 Plot of actual delay, Reg estimated delay, and NN estimated delay versus time of report for training data on Thursday, 15 February

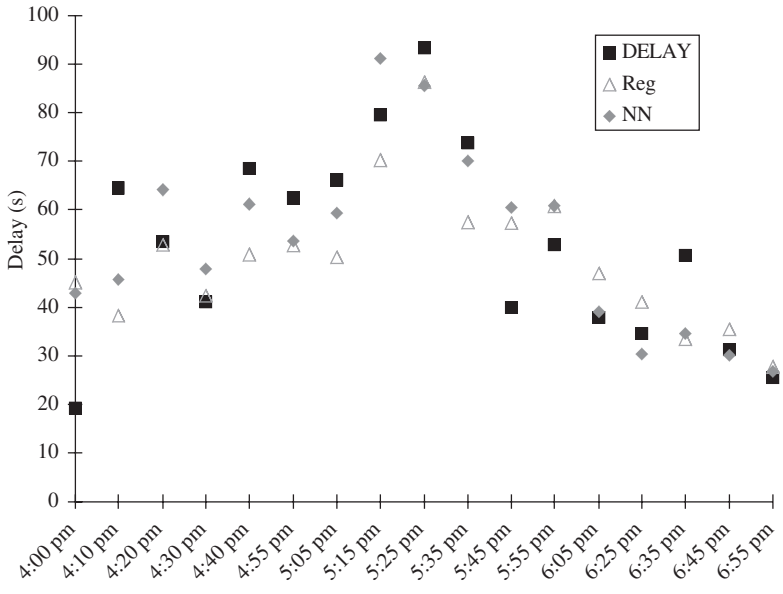


Figure 10.19 Plot of actual delay, Reg estimated delay, and NN estimated delay versus time of report for validation data on Tuesday, 13 February

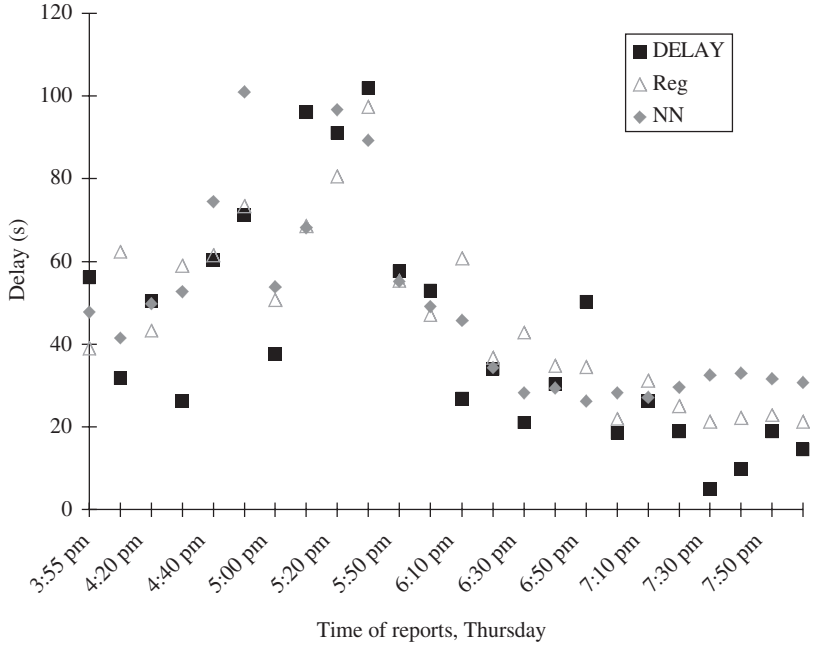


Figure 10.20 Plot of actual delay, Reg estimated delay, and NN estimated delay versus time of report for validation data on Thursday, 15 February

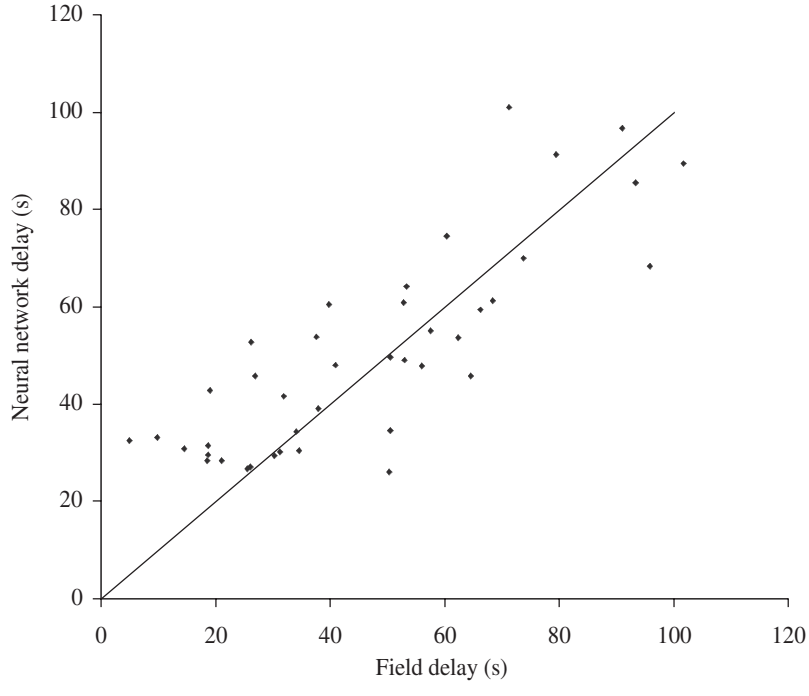


Figure 10.21 Plot of NN estimated delay versus actual field delay for validation data

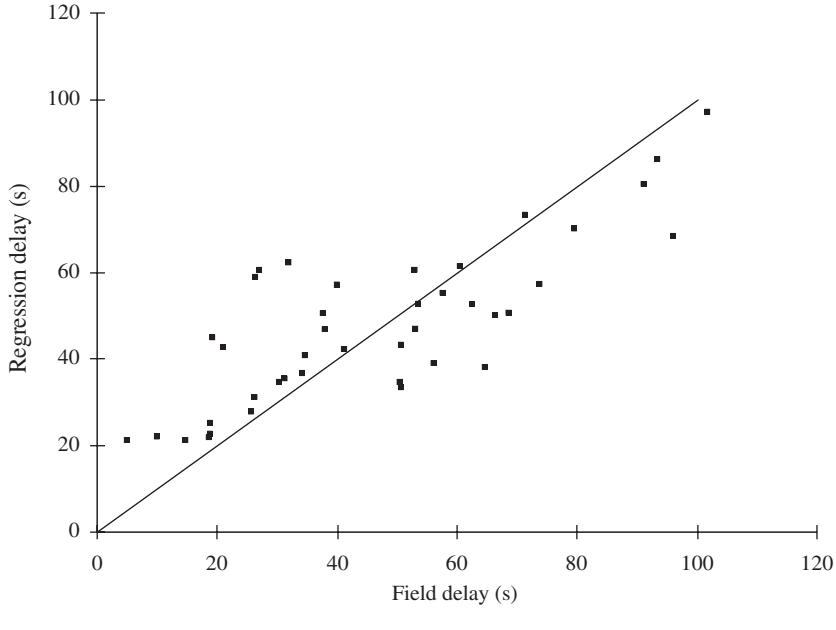


Figure 10.22 Regression estimated delay versus actual field delay

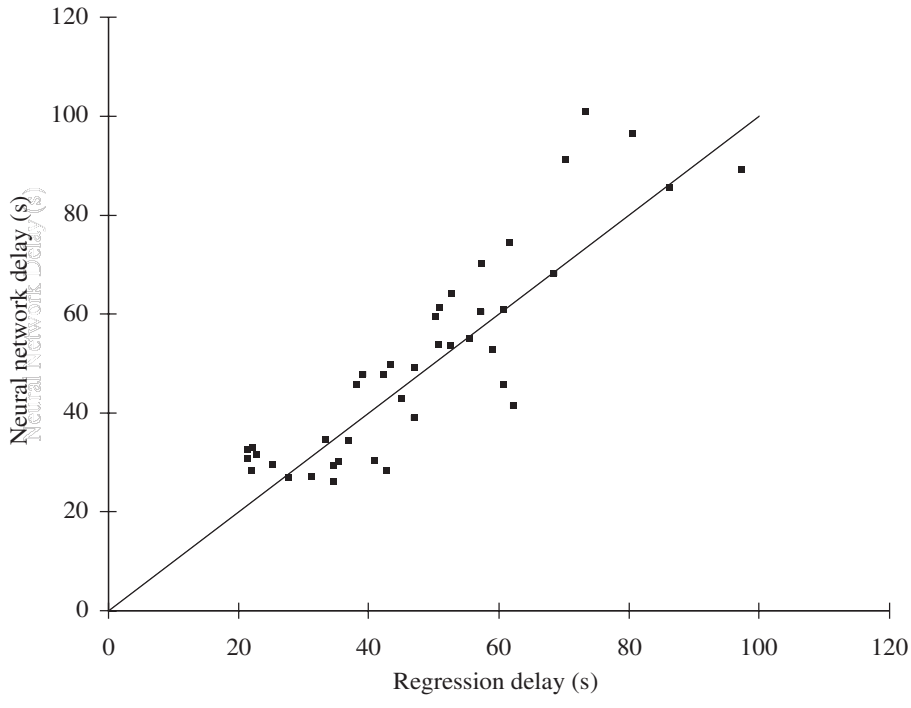


Figure 10.23 Plot of NN estimated delay versus Reg estimated delay

delays, and underestimate at high values of field delays. Finally, the regression delay values were plotted against the neural network delay values (Figure 10.23). In general, a tendency around the 45-degree line is observed, showing that both are quite highly correlated to each other (around 0.89).

Thus neural networks seem viable as a good alternate approach to regression model building and other traditional delay estimate models, especially in terms of their ease of construction and use.

7. CONCLUSIONS

This chapter has reported the results of an investigation into the challenging realm of arterial street delay estimation using artificial neural networks. Significant testing of the models/networks developed was performed by calibration and validation with actual field data. Traditional approaches to delay estimates such as regression model building, simple data selection and statistical weighted averaging have lacked the ability to classify syntactically diverse raw data which is often incomplete and inconsistent. Furthermore, some of these traditional methods have involved tedious and painstaking efforts in constructing models with properly selected variables. In a regression model, once we determine the model form, then it is simple to run the data through a statistical software package to get

the results. The difficulty arises in determining what form and combination of variables should be tested. This leads to an unlimited number of possibilities that should be tested. With neural nets, however, the strength of association and form of the explanatory variables are directly derived from the weights. This study has developed two methods, a neural network model and a statistical regression model. Our research activities included the following:

1. The development of several different neural network architectures for arterial street delay estimates.
2. The development of a statistical delay regression model based solely on field data.
3. Identifying cascade correlation and recurrent cascade correlation as the neural networks of choice for arterial street delay estimates.
4. A comparison of the best neural network architecture against statistical regression models.

Both the neural network architecture and the regression models were based on the knowledge of just the flow and occupancies collected on the main street and the cross-street alone. The variables selected for development in the models were the flows and occupancies on the main street and the cross-street. This produces models which are quite simple and easy to implement without the need to know the signal timings. Key variables which affected the travel time/delays were the occupancy on the main street and the cross-flow on the side street. For one of the data sets, occupancy alone yielded a 0.7 correlation for the prediction of the delays with the actual delays when applied to the validation data. By the inclusion of cross-flow as an additional independent variable, the correlation of the delays increased significantly by 10 per cent to 0.80. Previous regression models for delay/travel time have attempted and failed to represent flow and occupancy simultaneously due to their high correlation and have not focused on the cross-street variables. It was found that travel time/delay was best represented by the square term of occupancy on the main street (indicating that it increases non-linearly with increasing main street traffic).

The models developed are based on the current location of the detectors upstream of the intersection. A change in the placement location of the embedded inductive loop detectors could affect the computation of flows and occupancies over the time period studied and their relation to the delays (with no change in delays). A change in the geometry of the link (addition of left-turn bays and so on) or the type of signal plans will directly change delays, flows and occupancies. Newly trained neural network architectures and models need to be developed for either case in a similar fashion.

Regression model development proved to be tedious and painstaking. The appropriate selection of the independent variables based on the correlations, plots and so on and interpretation took time to develop. For example, cross-flow taken as the maximum of the cross-flow on either cross-leg proved to be more appropriate (and meaningful with respect to the way a semi-actuated intersection works) in comparison to the sum of the cross-leg flows. However, initial experimentation with the latter interpretation took much time and redoing the regression with the newly interpreted variable needed to be performed later. All important variables were presented directly to the neural networks for training and validation. It was found that restricting variables did not improve the training/validation and in some cases it had a negative impact.

Neural networks performed on an equal level with the regression models. However, neural networks proved easier to develop and adapt particularly well when new data were presented. The average correlation for the training data used to compare neural network and regression are 0.926 and 0.794, respectively, whereas for validation they are 0.778 and 0.774, respectively. In terms of regression, a different link with different data would require a complete statistical analysis of the field data in identifying the traffic flow relationships. A change in the link could present different types of traffic flow and one may have to redo the whole procedure followed above including data screening, selection of variables, checking for their multicollinearity and weighting of the residuals. Neural networks should only require minor adjustments to the training parameters and re-calibration.

The neural network architectures recommended are the cascade correlation learning architecture and recurrent cascade correlation learning architecture. These architectures were chosen over backpropagation, counterpropagation, and backpropagation through time. Backpropagation provided good results and a simple set of training parameters. However training time for backpropagation in cases where the data sets are greater than 500 reports may become an issue. Counterpropagation provided much faster training, as expected, but there was a reduction in the correlation values in most of the data sets presented. Backpropagation through time was the weakest of the architectures presented and its ability to handle time sequences did not appear to provide any improvement in our observations. As an alternative recurrent neural network, recurrent cascade correlation proved to be much better suited to delay estimates. Cascade correlation and recurrent cascade correlation provided the easiest setup and the fastest training times. The cascade architectures technique of adding hidden units to build the hidden layer during training proved a preferable method to making a best guess as to the size of the hidden layer. In the link investigated for this study the cascade neural network with five hidden units, and a maximum of 200 training cycles per unit, provided the best results when applied over a variety of data sets. Selecting the correct number of hidden units will be dependent on the link being examined and is easily determined by stopping the training session and validating the current neural network before adding additional units. If the validation values go up, training should be stopped.

Future work may include other factors such as:

- The study of multiple links to determine instantaneous route travel times/delays and to study platoon dispersion. Additional flow and occupancy variables may be easily represented with neural networks.
- The determination of traffic turning movements based on detector information.
- The effect of a change in location of detectors on the computed delays.
- The detection of incidents by comparison of computed delays to historical delays under non-incident conditions.
- The inclusion of additional sources of information, such as weather and anecdotal information may also be useful to improve the predictive ability of the models.

NOTE

* This research was supported by the National Research Council Transportation Research Board and the ADVANCE dynamic route guidance project. The authors also wish to thank the developers of the Stuttgart Neural Network Simulator (Zell et al. 1995) which, was used extensively for this project.

REFERENCES

Berka, S., X. Tian and A. Tarko (1995), 'Data fusion algorithm for ADVANCE release 2.0', ADVANCE Working Paper Series, Number 48, TRF-DF-202 and TRF-DF-204, May.

Boyce, D.E., A. Kirson and J. Schofer (1991), 'ADVANCE: The Illinois Dynamic Route Guidance Demonstration Program', paper presented at The Institute of Management Science/Operations Research Society of America TIMS/ORSA Joint National Meeting, Nashville, TN, 12–15 May.

Caudill, M. and C. Butler (1992), *Understanding Neural Networks: Computer Explorations*, Vols 1 and 2, Cambridge, MA: MIT Press.

Dillenburg, J.F., C. Lain, P.C. Nelson and D. Rorem (1995), 'The design of the ADVANCE Traffic Information Center', in *ITS America Proceedings*.

Fahlman, S.E. (1988), 'Faster learning variations on the backpropagation: an empirical study', in D. Taretzky, G. Hinton and T. Sejnawski (eds) *Proceedings of the 1988 Connectionist Models Summer School*, San Mateo: Morgan Kaufmann.

Fahlman, S.E. (1991), 'The recurrent Cascade–correlation learning architecture', Technical Report CMU-CS-91-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Fahlman, S.E. and C. Lebiere (1991), 'The Cascade-correlation learning architecture', Technical Report CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Freund, R.J. and R.C. Littell (1986), *SAS System for Regression*, SAS Institute Cary, NC.

Grossberg, S. (1982), *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control, Boston*, MA: Reidel.

Hecht-Nielsen, R. (1988), 'Application of counterpropagation networks', *Neural Networks*, 1, 131–9.

Hecht-Nielsen, R. (1990), *Neurocomputing*, Reading, MA: Addison-Wesley.

Hertz, J.A., A.S. Krogh and R.G. Palmer (1991), *Introduction to the Theory of Neural Computation*, Reading, MA: Addison-Wesley.

Kohonen, T. (1988), *Self-Organization and Associative Memory*, 2nd edn, New York: Springer-Verlag.

Kosko, B. (1992), *Neural Networks and Fuzzy Systems: A Dynamical Approach to Machine Intelligence*, Englewood Cliffs, NJ: Prentice-Hall.

Lippmann, R.P. (1987), 'An introduction to computing with neural nets', IEEE *ASSP Magazine*, April, 4–21.

Minsky, M.L. and S.A. Papert (1969), *Perceptrons*, Cambridge, MA: MIT Press.

Palacharla, P.V. (1995), 'A pattern recognition approach to data fusion in intelligent vehicle highway systems', PhD thesis, University of Illinois at Chicago.

Palacharla, P.V. and P.C. Nelson (1995), 'On-line travel time estimation using fuzzy neural networks', in *Proceedings of the Second ITS World Congress*, Yokohama, Japan, November.

Parker, D.B. (1985), 'Learning logic', Technical Report, TR-47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA.

Ritchie, S., R.L. Cheu and W.W. Recker (1992), 'Freeway incident detection using artificial neural networks', International Conference on Artificial Intelligence Applications in Transportation Engineering, San Buenaventura, CA, June.

Rumelhart, D.E., G.E. Hinton and R.J. Williams (1986), 'Learning internal representations by error propagation', in D.E. Rumelhart and J.L. McClelland (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Cambridge, MA: MIT Press, pp. 318–62.

Sen, A. and N. Srivastava (1990), *Regression Analysis, Theory, Methods and Analysis*, New York: Springer-Verlag.
Sisiopiku, V.P. and N.M. Rouphail (1993), 'Review of alternative procedures for using loop detector output to estimate arterial link travel times', ADVANCE Working Paper Series, No. 23, Task TRF-DF-02, May.
Sisiopiku, V.P. and N.M. Rouphail (1994a), 'Towards the use of detector output for arterial link travel time estimation: a literature review', Transportation Research Record Series, Washington, DC.
Sisiopiku, V.P. and N.M. Rouphail (1994b), 'Travel time estimation from loop detector data for advanced traveler information systems applications', Technical Report to IDOT, Urban Transportation Center, University of Illinois, Chicago, June.
Wasserman, P. (1987), *Neural Computing: Theory and Practice*, New York: Van Nostrand Reinhold.
Werbos, P.J. (1974), 'Beyond regression: new tools for prediction and analysis in the behavioral sciences', PhD thesis, Harvard University, Cambridge, MA.
Zadeh, L.A. (1965), 'Fuzzy sets', *Information and Control*, **8**, 338–53.
Zell, A. et al. (1995), *SNNS: Stuttgart Neural Network Simulator User Manual*, Version 4.1, Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, Germany, July.

11. Modeling travel times along signalized streets using expected cumulative counts

Andrzej P. Tarko and Gopalakrishnan Rajaraman*

1. INTRODUCTION

Highway traffic can be conveniently represented with cumulative counts (Newell 1993). Newell showed that this representation is equivalent to the Lighthill–Whitham–Richards theory (Lighthill and Whitham 1955; Richards 1956) if the traffic flow speed–density relationship $u = u(k)$ exists and vehicles do not pass each other (FIFO assumption: first in first out). After formal introduction by Newell, cumulative counts were used to measure freeway flows (Cassidy and Windover 1995). Cumulative counts provide appealing simplicity in measuring travel times and in formulating hypotheses about traffic flows between the points with known cumulative counts if FIFO is assumed.

Although the FIFO traffic may occur at a certain level of density even on multiple highways (Munoz and Daganzo 2000), it is also possible that traffic flows in different lanes move at different speeds. This would not be a problem if vehicles did not change lanes and cumulative counts were used for individual lanes. Indeed, vehicles do change lanes, particularly when traffic is sparse and drivers have the opportunity to pass slower vehicles to maintain their preferred speeds. So the question arises about the validity of travel time measurements if vehicles entering and exiting the segment are matched based on the assumption that they enter and exit the segment in the same order.

Another issue arises if cumulative counts are used to predict travel times. Because the entry and exit times of individual vehicles cannot be predicted with certainty, expected values must be used. Unlike cumulative counts, expected cumulative counts represent traffic in a macroscopic manner by cumulating expected flow rates. The macroscopic traffic representation is consistent with the original Newell concept of continuous flows.

2. THEORETICAL CONSIDERATIONS

This section proposes the use of expected cumulative counts for estimating and predicting vehicle travel times between two points with the application to signalized intersections. Although the idea is not new, this is the first time that expected cumulative counts have been examined formally, and tested to see whether they can be used to estimate the unbiased travel times of vehicles with known entry times if the traffic does not follow the FIFO principle (Section 2.1). A short discussion of estimating expected travel times of vehicles with known preferred speeds follows (Section 2.2).

Section 2.3 presents a practical procedure for estimating travel times along signalized streets, and a description of a field test and its results are given. The proposed method can also be used for short-term prediction of travel times along signalized streets if specific traffic and control information is known to the traffic center.

2.1 Travel Times along Roads with Non-FIFO Traffic

Let $f(t, t')$ be a joint density function of time when a vehicle enters and exits a segment, where t is entry time and t' is exit time. The joint density function is defined for entry times t such as $0<t<T$. The equivalent function is $f(t, t+\tau)=h(t,\tau)$, where τ is travel time along the segment. It is reasonable to assume that τ is bounded between τ_1 and τ_2. A vehicle entering the segment during the T period has a distribution of entry time equal to marginal distribution

$$g(t)=\int_{\tau_1}^{\tau_2} h(t,\tau)\,d\tau.$$

If we assume that $g(t)$ applies to any of M vehicles that enter the segment during the T period, the expected number of vehicles entering between t and $t+dt$ is $M\cdot g(t)dt=\lambda(t)dt$, where M is the number of vehicles entering the segment during period T and $\lambda(t)$ is entry flow rate at time t. Cumulating $\lambda(t)$ in period T yields cumulative rate $L(t)$ at time t which in fact is the expected value of cumulative count $N(t)$. In some cases, expected curve L is known or can be derived from available information.

Similarly, exit rate profile $\lambda'(t)$ can be obtained as a marginal distribution for t' derived from $f(t, t')$. It must be stressed that $\lambda'(t)$ is the rate of vehicles generated by vehicles that entered the section sufficiently earlier before time t but between 0 and T. The corresponding cumulative rate curve is denoted as L'.

Travel time in uncertain conditions is usually represented through its expected value. If traffic changes with time (non-constant entry rate), then we would like to know the expected travel time of a vehicle that enters the segment at time t. Following Newell's proposal for deterministic cases, it would be tempting to use L curves to estimate the travel times using the following relationship between cumulative curves:

$$L(t)=L'(t+\tau_0|t), \tag{11.1}$$

where L and L' are cumulative entry and exit rates, respectively, and τ_0 is travel time estimate for a vehicle that enters the segment at time t. The question arises, however, about τ_0. If the traffic is non-FIFO and vehicles pass each other, then the obtained value might not be the correct (unbiased) estimate of the expected travel time.

Let us first consider stochastic non-FIFO traffic at the microscopic level. A vehicle's entry label is equal to the number of vehicles that have entered earlier plus one. A vehicle's exit label is equal to the number of vehicles that have exited earlier plus one. If the vehicle has been passed by m other vehicles and the vehicle has passed n other vehicles before reaching the exit, its exit label is the entry label plus m minus n. If the vehicle passes and is passed by the same number of vehicles, then the vehicle preserves its entry label. If expected values are considered, then equation (11.1) is valid if τ_0 is a travel time such that

the expected numbers of vehicles passing and being passed are equal. Let us determine conditions for τ_0 to be an unbiased estimate of the expected travel time.

To estimate the expected number of vehicles passing a given vehicle, the travel time distribution of a vehicle starting at time t is needed:

$$y(\tau|t)=\frac{h(t,\tau)}{g(t)}.$$

Since the travel time is bounded between τ_1 and τ_2, any two vehicles selected from the flow can pass each other along the segment if the entry time gap between them is not too long. A vehicle entering at time t_0 and exiting at time $t_0+\tau_0$ can pass another vehicle if the latter enters the segment after $t_0+\tau_0-\tau_2$ but before t_0. To be passed, a vehicle entering at time t_0 and exiting at time $t_0+\tau_0$ must be followed by a vehicle that enters after t_0 but before $t_0-\tau_0-\tau_1$. If a vehicle enters at time t, its chance of being passed depends on whether its travel time is sufficiently long:

$$\int_{t_0+\tau_0-t}^{\tau_2} y(\tau|t)\,d\tau.$$

Since the expected number of vehicles entering between t and $t+dt$ is $\lambda(t)dt$, the expected number of vehicles being passed is

$$\lambda(t)dt \int_{t_0+\tau_0-t}^{\tau_2} y(\tau|t)\,d\tau.$$

Integrating along period $(t_0-\tau_0-\tau_2,\ t_0)$ yields the expected number of vehicles to be passed by the given vehicle:

$$\int_{t_0+\tau_0-\tau_2}^{t_0} \lambda(t)dt \int_{t_0+\tau_0-t}^{\tau_2} y(\tau|t)\,d\tau.$$

Similarly, the expected number of vehicles passing the given vehicle can be derived. The two expected values have to be equal to claim that τ_0 is the solution of equation (11.1):

$$\int_{t_0+\tau_0-\tau_2}^{t_0} \lambda(t)dt \int_{t_0+\tau_0-t}^{\tau_2} y(\tau|t)\,d\tau = \int_{t_0}^{t_0+\tau_0-\tau_1} \lambda(t)dt \int_{\tau_1}^{t_0+\tau_0-t} y(\tau|t)\,d\tau. \tag{11.2}$$

The stochastic steady state will be considered first and then traffic with a rapidly changing flow rate. Traffic is considered stochastically steady if the traffic rate is fixed and all vehicles have the same distribution of travel times. Note that independence of travel times is not assumed. Travel times may depend on those previously observed, but this dependence cannot change with time. If we assume steady-state traffic, then $y(\tau|t)=y(\tau)$ and $\lambda(t)=\lambda$.

$$\int_{t_0+\tau_0-\tau_2}^{t_0} \lambda dt \int_{t_0+\tau_0-t}^{\tau_2} y(\tau)\, d\tau = \int_{t_0}^{t_0+\tau_0-\tau_1} \lambda dt \int_{\tau_1}^{t_0+\tau_0-t} y(\tau)\, d\tau.$$

Dropping λ and solving the internal integrals gives:

$$\int_{t_0+\tau_0-\tau_2}^{t_0} [Y(\tau_2) - Y(t_0+\tau_0-t)\, dt = \int_{t_0}^{t_0+\tau_0-\tau_1} [Y(t_0+\tau_0-t) - Y(\tau_1)]dt,$$

where $Y(\tau)$ is a cumulative distribution of travel times. Substitutions $Y(\tau_1)=0$ and $Y(\tau_2)=1$ and further transformations yield:

$$(\tau_2-\tau_0) - \int_{t_0+\tau_0-\tau_2}^{t_0} Y(t_0+\tau_0-t)dt - \int_{t_0}^{t_0+\tau_0-\tau_1} Y(t_0+\tau_0-t)dt = 0 \tag{11.3}$$

and

$$(\tau_2-\tau_0) - \int_{t_0+\tau_0-\tau_2}^{t_0+\tau_0-\tau_1} Y(t_0+\tau_0-t)dt = 0.$$

Substituting τ for $(t_0+\tau_0-t)$ and $-d\tau$ for dt, the following is finally obtained:

$$\tau_0 = \tau_2 - \int_{\tau_1}^{\tau_2} Y(\tau)d\tau. \tag{11.4}$$

Equation (11.4) returns expected values of τ regardless of the distribution of travel times. This result indicates that if a steady flow rate persists at the segment entry for time $(\tau_2-\tau_1)$ then estimates of the travel times for vehicles entering in the middle of this period converge to the expected values.

In the next considered case, entry flow rate $\lambda(t)$ increases rapidly at time t_0 from λ_1 to λ_2. Transformations similar to the previous ones have yielded:

$$\lambda_1(\tau_2-\tau_0) - \lambda_1 \int_{\tau_0}^{\tau_2} Y(\tau)d\tau - \lambda_2 \int_{\tau_1}^{\tau_0} Y(\tau)d\tau = 0 \tag{11.5}$$

and

$$\tau_0 = E\tau - \frac{\lambda_2-\lambda_1}{\lambda_1} \int_{\tau_1}^{\tau_0} Y(\tau)d\tau. \tag{11.6}$$

Similarly for a rapid reduction in the flow rate ($\lambda_1 > \lambda_2$),

$$\tau_0 = E\tau + \frac{\lambda_1 - \lambda_2}{\lambda_2} \int_{\tau_0}^{\tau_2} Y(\tau) d\tau. \tag{11.7}$$

Although equations (11.6) and (11.7) do not give solutions for τ_0 because τ_0 appears on the right-hand side of the equations, they indicate the presence of bias in $E\tau$ estimates. Since the second components on the right-hand side of the equations are always positive, it can be said that an increase in flow rate causes underestimation of the expected travel time, while a decrease in flow rate causes overestimation.

At traffic signals, green periods with significant flow rates may neighbor red periods with near-zero flow rates. To check the effect of this traffic pattern, let us solve equation (11.5) for τ_0 assuming $\lambda_1 = 0$ and then again assuming $\lambda_2 = 0$. The solutions are $\tau_0 = \tau_1$ and $\tau_0 = \tau_2$ which means that for vehicles departing at the beginning and at the end of green signals, L curves tend to produce the shortest and the longest travel times, respectively, instead of the expected values. This bias is caused by a phenomenon described in the literature as 'platoon dispersion' (Hillier and Rothery 1967; Hunt et al. 1981).

Section 2.1 has introduced a concept of L curves that suit stochastic representation of traffic. It has been shown that L curves produce unbiased estimates of the expected travel times for non-FIFO traffic if it is in a stochastic steady state. Bias occurs when a vehicle enters a road segment during or near the time when a flow rate changes rapidly and considerably. In Section 2.2, an extreme case of traffic signals with rapidly changing flow rates at the beginning and end of green signals will be tested using field observations.

2.2 *L* Curves for Non-conserved Traffic – A Practical Approach

In the previous consideration, traffic represented by the two L curves was conserved. This requirement is easily met on freeways with full access control. For other roads, this assumption is often violated, with the L curve representing vehicles that enter the segment regardless of where they exit the road and the L' curve representing vehicles that exit the segment regardless of where they enter the road.

For signalized urban streets, where it is desirable to estimate travel times between two stop-lines and for particular traffic maneuvers, we must ask the following question: what is the expected travel time of a vehicle that enters a segment at some time t and performs a particular maneuver at the downstream intersection? Two L curves are useful and obtainable for this task: an L curve of vehicles that enter the segment from the upstream intersection and an L' curve of vehicles that perform a selected maneuver at the downstream intersection. Among the vehicles that enter the segment through the upstream intersection, many may perform other maneuvers at the downstream intersection or leave the segment between the intersections. Among the vehicles that perform the maneuver at the downstream intersection, some have entered the segment between the intersections.

The obvious result of the lack of traffic conservation, is that typically $L(t) > L'(t + E\tau|t)$ or, less frequently, $L(t) < L'(t + E\tau|t)$. In the former case, the L and L' curves can be reconciled by introducing an adjustment factor $r(t)$ such that $L(t) = r(t) \cdot L'(t + E\tau|t)$ for $t \in (0, T)$. Although the adjustment factor $r(t)$ produces two balanced curves, traffic conservation is regained if $L'(t)$ includes only vehicles that enter at the upstream intersection. If the $L'(t)$ curve includes vehicles that enter between intersections, then the reconciled curve

$r(t)\cdot L(t)$ includes an equal rate of vehicles that do not perform the particular maneuver at the downstream intersection. These vehicles are a source of additional variance in expected travel time estimates. To make it worse, only a crude estimation of r may be possible in many cases. One practical solution is to use a fixed value $r = L(T)/L'(T + E\tau \mid T)$ for the entire T period.

2.3 *L* Curves for Signalized Streets

Sections 2.1 and 2.2 discussed two possible sources of bias in travel time estimates: rapid changes in flow rates in non-FIFO conditions and lack of traffic conservation which calls for practical but crude reconciliation of cumulative counts. The bias will be evaluated using travel time measurements along selected signalized urban streets. In order to do so, a method of modeling travel times along urban streets based on expected cumulative counts is needed. This subsection presents the method developed by the authors for vehicles moving between two signalized intersections and performing a specific maneuver at the downstream intersection.

Traffic entering a street segment after passing the upstream intersection can be represented with a pulse-like flow rate profile $a(t)$ which includes high- and low-traffic periods (Figure 11.1). The high-traffic periods are effective green times for arterial traffic, while the low-traffic periods are effective red times. Flow rates are assumed fixed during individual periods and represent time-average values. Using a half-hour period as an example, the flow rate profile for that period can be determined if the traffic volume for this period is known from measurement or prediction. The high-traffic rate can be calculated as the total number of vehicles entering the segment during green divided by the total green time in the period. Similarly, the low-traffic rate can be calculated for a red signal.

Usually, only a portion of the vehicles entering a segment at the upstream intersection perform the specific maneuver at the downstream intersection. This portion, denoted as r, can be determined as the ratio of the total number of vehicles performing the maneuver at the downstream intersection and the total number of vehicles entering the segment at the upstream intersection. The reconciled flow rate $r \cdot a(t)$ of vehicles entering the segment is used in further calculations.

Figure 11.1 also presents a simplified profile of capacity rate $c(t)$ for a traffic maneuver at the downstream intersection. The high- and low-capacity periods are again effective green and effective red times for the particular traffic movement. In most cases, the capacity rate during red time is zero. The capacity profile limits the rate at which vehicles exit the street segment at the downstream intersection. The capacity rate will be used to build a traffic profile for the exiting traffic.

For the purpose of constructing expected cumulative counts L' from expected flow rates $r \cdot a(t)$ and capacity rates $c(t)$, time is divided into small intervals Δt. Rates $r \cdot a(t)$ and $c(t)$ are measured by the number of vehicles per interval Δt. Traffic is represented through the following quantities:

1. *entry rate* a(t) – the number of vehicles entering the segment during Δt, starting at time t (Figure 11.1);
2. *exit rate* d(t) – the number of vehicles exiting the segment during Δt, starting at time t;

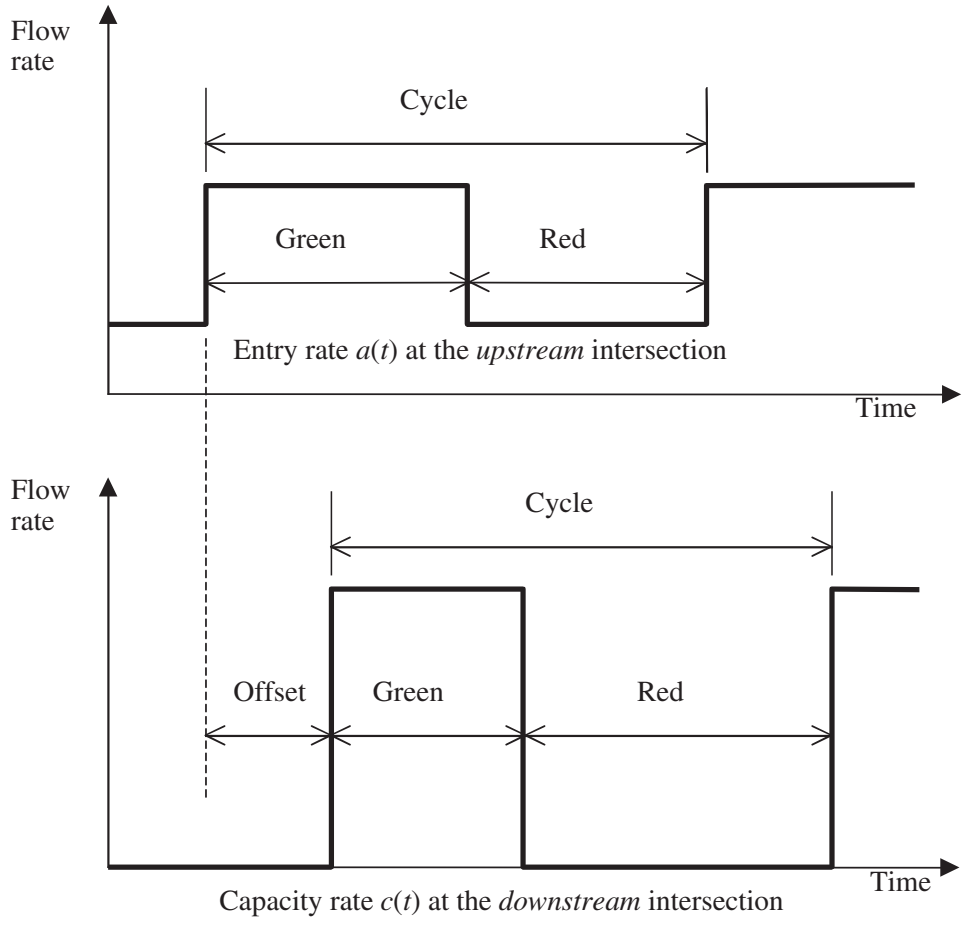


Figure 11.1 Simplified entry and capacity rates at signalized intersections

3. *capacity rate* c(t) – the maximum rate of exiting vehicles at time t as determined by the downstream bottleneck (Figure 11.1);
4. *movement occupancy* O(t) – the number of vehicles in the segment at time t that will perform the particular maneuver at the downstream intersection; occupancy takes a minimum value if there is no traffic obstruction along the segment;
5. *movement queue* Q(t) – the difference between the current movement occupancy and the minimum movement occupancy at time t.

The following procedure of constructing the cumulative curves is consistent with Newell's simplified flow theory. The calculations advance interval by interval. The entry rate $a(t)$ for $t \in [0, T]$ and capacity rate $c(t)$ for the corresponding period are known. The calculation process starts as soon as the movement occupancy is determined but not earlier than at $t \geq \tau_m$. Travel time τ_m is the expected travel time when traffic is not slowed

down by either queues or red signals. The travel time τ_m is a travel time of progressive traffic waves in non-congested traffic with the assumption of no or weak interactions between vehicles as postulated by Newell in his simplified theory of traffic flows (Newell 1993). This travel time will be called 'minimum travel time'.

Movement occupancy can be estimated if the entry and exit times of some typical vehicle are known. The movement occupancy is simply the number of movement vehicles (vehicles performing a particular maneuver at the downstream intersection) entering the segment when the typical vehicle is traveling along the segment. The estimated movement occupancy applies to the exit time of the typical vehicle. Let the time with known movement occupancy be t. The value of upstream cumulative $L(t)$ curve is set at the movement occupancy $O(t)$, while the value of downstream cumulative curve $L'(t)$ is set at zero.

After the initiation, an updating process for consecutive intervals starts, which yields values of L, L', and movement occupancy. First, the minimum movement occupancy $O_m(t)$ is estimated similarly to the movement occupancy, with the difference that the travel time is assumed to be minimum travel time τ_m:

$$O_m(t) = \sum_{i=t-\tau_m-1}^{t-1} r \cdot a(i).$$

Minimum movement occupancy is the number of movement vehicles on the segment when none of these vehicles experiences delays (zero queue). Next, the queue is calculated, which is actually the excess number of vehicles over the minimum movement capacity:

$$Q(t) = \mathrm{O}(t) - O_m(t).$$

At this point, all needed input is known to calculate the exit rate $d(t)$. If the signal is green and there is no queue of vehicles, then the rate of vehicles $d(t)$ performing the particular maneuver equals the rate of vehicles arriving at the downstream intersection. This rate is approximately equal to the rate of vehicles entering τ_m seconds earlier the segment at the upstream intersection $d(t) = r \cdot a(t-\tau_m)$. When a long queue is present on the downstream approach or the signal is red, the rate of exiting vehicles equals the capacity rate:

$$d(t) = \begin{cases} r \cdot a(t-\tau_m) + Q(t) \text{ if } r \cdot a(t-\tau_m) + Q(t) < c(t), \\ c(t) \quad \text{otherwise} \end{cases}$$

Expected cumulative entering flow $L(t)$ and expected cumulative exiting flow $L'(t)$ are calculated as:

$$L(t) = L(t-1) + r \cdot a(t),\ L'(t) = L'(t-1) + d(t).$$

The last step is to update the movement occupancy, which will be used to repeat the above set of calculations for the next interval $t+1$:

$$O(t+1) = O(t) + a(t) - d(t).$$

After the L and L' values have been calculated for all intervals $t \in [\tau_m, T]$, the expected travel times $E\tau|t$ are determined from the condition:

$$L(t) = L'(t + E\tau \mid t).$$

The cumulative curves L and L' may be subject to growing bias due to errors that sum up over time (drift error). The re-initialization based on known movement occupancy resets the curves and eliminates the bias.

3. FIELD EVALUATION

This section presents field evaluation data to test the proposed method. Travel times and other data were collected on three urban segments; the travel times were calculated using the methods described in the previous section, and the measured and calculated values were compared. More information about the data collection and the results can be found in Tarko et al. (2000).

3.1 Data Collection and Processing

Test sites were selected to cover various highway and traffic characteristics. The characteristics considered in the site selection are listed below:

1. *Congestion level* determines which component of the method is evaluated. If queues are negligible, the effect of progression is prevailing. On the other hand, the effect of long queues prevails in congested traffic. We have selected sites with various levels of congestion.
2. *The type of signal* controller is particularly important where there is a lack of congestion. At actuated signals, the progression effect can vary significantly from one cycle to another while at pre-timed signals it is kept fixed.
3. *Road segment length* influences the dispersion of initially dense vehicle columns. The method of constructing the L' curve assumes no dispersion since all vehicles move along a non-congested segment at the same speed. The effect of this simplification will be tested by incorporating short and long segments.
4. *Signal progression* may influence travel times significantly. The sites with various levels of quality of progression were included.
5. *Traffic entering and exiting between intersections* is often present (parking, driveways, less important unsignalized intersections). It was present at some of the selected sites.

Three selected sites are described in Table 11.1. Observations at the selected segments were performed in two-hour afternoon sessions to cover peak and off-peak traffic. Collected data include the inputs required in the tested method and the measured travel times for comparison.

Traffic at the two intersections was videotaped using color VHS camcorders. The observers recorded the clock times when traffic signals for the street segments turned green or red. The videotapes were then observed and the times when vehicles entered and exited the segments were recorded. The vehicles' maneuvers were noted too. This data allowed for counting vehicles during green-plus-yellow and during red signals.

Vehicles exiting the segment should be counted during a period that corresponds to the

Table 11.1 Selected sites

Site name	Location	Congestion level during observation	Type of signal	Segment length (km)	Signal progression	Mid-block traffic
Broadway	Broadway St. from 49th Av. to 53rd Av., Gary, IN	No permanent overflow	Semi-actuated	0.8	Very good	Moderate
North-western	Northwestern Av. from Grant St. to Stadium Av., West Lafayette, IN	Moderate queues	Pre-timed	0.5	Average	Heavy
Sagamore	Sagamore Parkway from Duncan Rd to SR 25, Lafayette, IN	Long queue with spillback	Actuated	1.2	None	None

one when the same vehicles are counted when entering the segment. If the count period for entering vehicles is $[0,T]$, then the corresponding period for departing vehicles should be $[t+E\tau|0,\ T+E\tau|T]$. Two vehicles believed to be typical were selected from those recorded on videotape: one vehicle entering the segment at the beginning of the observation period and one entering at the end. The two vehicles were identified at the exit intersections, and the exit times of the two vehicles determined the count period for exiting vehicles. A value of reconciliation factor r was determined for each street segment based on the number of vehicles entering and exiting segments during corresponding count periods. This has been calculated as the ratio of the smaller to the larger count. The reconciliation factor r was applied to the curve with the higher count.

Entering vehicles rates $a(t)$ were calculated by dividing vehicle counts during red or green signal periods by the measured lengths of these signals. The obtained values were multiplied by the reconciliation factor r if the number of entering vehicles was higher than the number of exiting vehicles.

The running speeds of the vehicles were measured by recording the travel times of the vehicles between two marked points on a segment and located beyond the influence of the queues. A manual technique (stopwatches) was used. The running times (minimum travel times) were estimated by dividing the distance between the stop-lines by the running speed measured at spots free of queue and red signal impacts.

The initial movement occupancy was estimated based on the entrance and exit times of the vehicle used to determine the starting time of the count periods. The calculations have been explained earlier.

Capacity rates $c(t)$ for signals were estimated as the total number of vehicles departing a segment from the queue divided by the total duration of the periods with a queue. A fixed capacity rate was applied to all green signals during the counting period, and zero capacity was assumed for red signals at all sites. The obtained values were multiplied by the reconciliation factor r if the number of exiting vehicles was higher than the number of those entering. The reduced capacity represents the capacity of vehicles that entered at the upstream intersection and crossed the downstream stop-line in a saturated flow including other vehicles. Measured signals were used when constructing the $c(t)$ curve.

The cumulative curve $L(t)$, the departure rates $c(t)$, and the corresponding cumulative curve $L'(t)$ were calculated using the procedure described earlier. These curves and the condition in equation (11.1) were used to estimate the expected travel times. The actual travel times along the segments for randomly selected vehicles were measured by watching videotapes, identifying the same vehicles at both intersections, and by subtracting the entry from the exit times. The evaluation of the method was based on comparison of the estimated and measured travel times.

3.2 Broadway Street

At this site, the straight-ahead movement was investigated. This movement was very light with no overflow queues. The length of the road segment was 0.8 km. The signal controllers were semi-actuated with very good coordination between the upstream and downstream signals. The variability of traffic signals is shown in Figure 11.2. Signal cycles varied on both intersections around a common background cycle enforced by the signal controller (Figure 11.3).

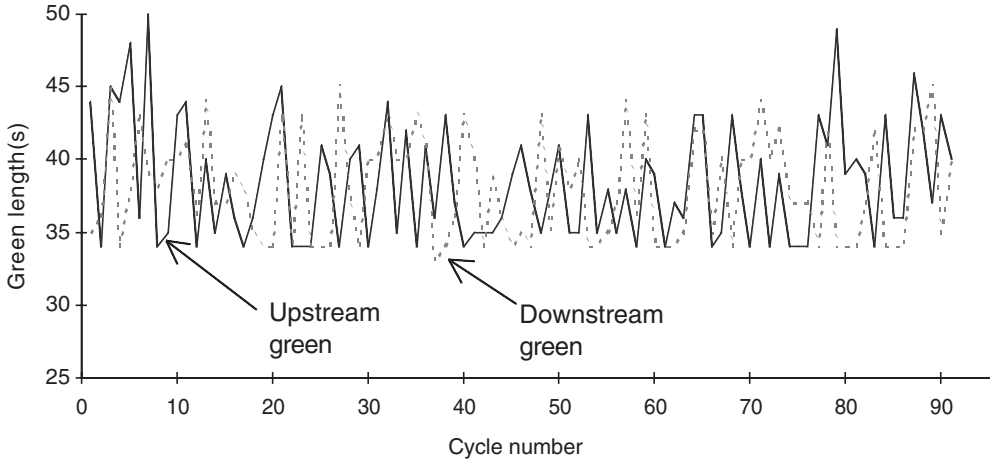


Figure 11.2 Green signals for the Broadway Street segment

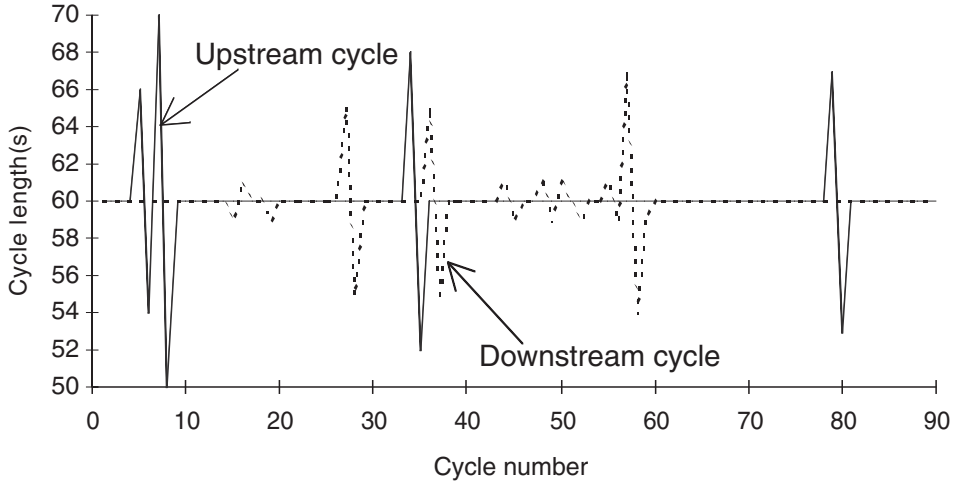


Figure 11.3 Signal cycles for the Broadway Street segment

The total count of vehicles entering the segment was lower than the total count of entering vehicles (Figure 11.4). The reconciled curves are shown in Figure 11.5. Due to the lack of congestion and very good progression, practically all vehicles hit a green signal at the downstream intersection. The estimated travel times that represent expected values are almost equal to the minimum travel time (running time) regardless of the vehicle's entry time (Figure 11.6). On the other hand, the measured travel times varied randomly due to an opportunity for passing along the segment. Figure 11.7 compares the estimated and measured travel times for randomly selected vehicles. The average estimation error is less than -1 second and the standard deviation is around 5 seconds (Table 11.2).

Table 11.2 Summary of the evaluation based on the simulation experiments

Site	Sample	No. of observations	Av. error (%)	Std error (%)	Av. travel time (s)
Broadway	All	210	−0.9	8.6	62.1
Northwestern	All	67	3.9	25.5	131.9
Northwestern	Reduced*	57	0.5	5.7	129.6
Sagamore	All	67	8.3	16.6	339.1
Sagamore	Reduced*	60	2.8	3.9	339.7

Note: *Excluded biased travel time estimations for vehicles arriving at the stop-line at the end of green.

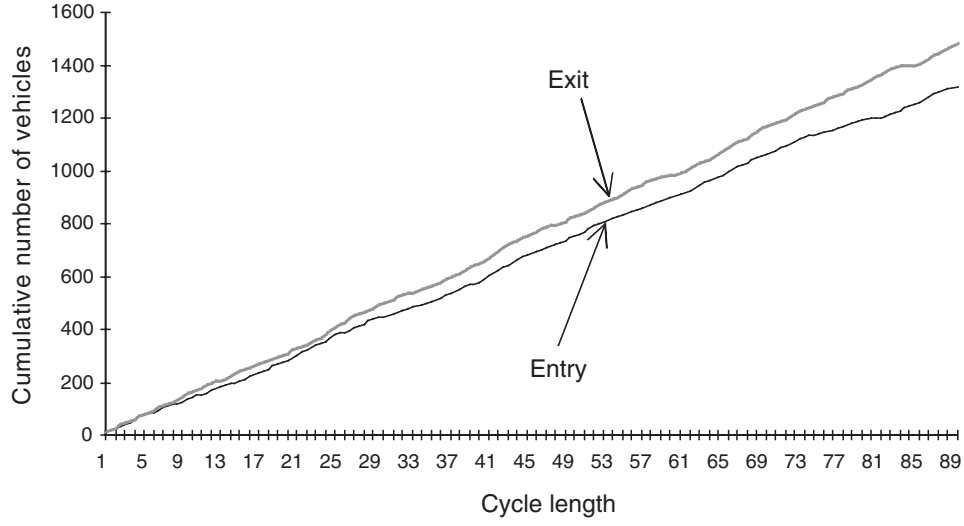


Figure 11.4 Cumulative count of vehicles entering and exiting the Broadway Street segment

3.3 Northwestern Avenue

At this site, the unopposed left-turn movement at the Northwestern Avenue–Stadium Street junction was selected. The link length was 0.5 km. Movement was moderately congested, but the queue did not reach the upstream intersection. The overflow queues for all cycles are presented in Figure 11.8.

At the time of observation, the traffic signals were to maintain fixed lengths. The cycle lengths, offsets, and green periods are plotted in Figures 11.9 and 11.10. The signal timings were fixed initially, but then started varying for a long time. It should be stressed that the signals were not dependent on traffic. The unstable signals were caused by technical difficulties in communication between the master and the local controllers.

The travel times tend to be longer during the unstable control than during stable control (Figure 11.11). The observed range of travel times during the entire count period was

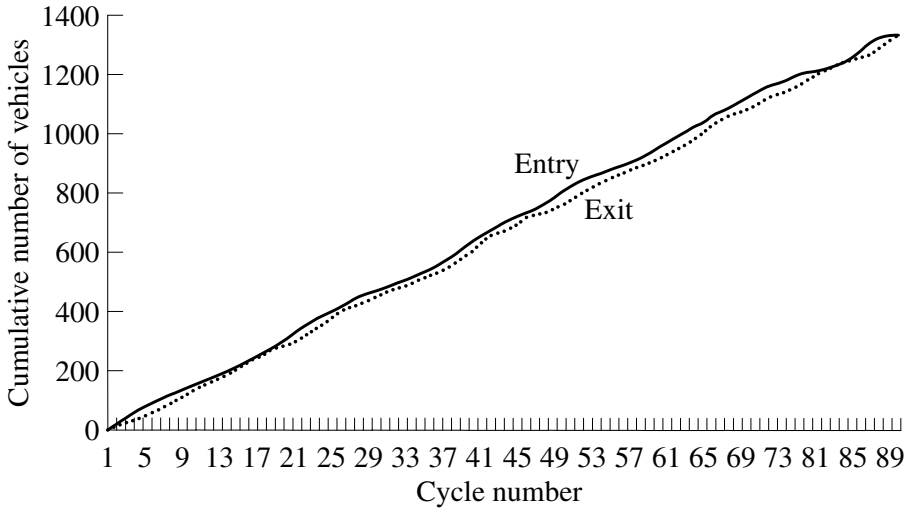


Figure 11.5 Reconciled cumulative count of vehicles entering and exiting the Broadway Street segment

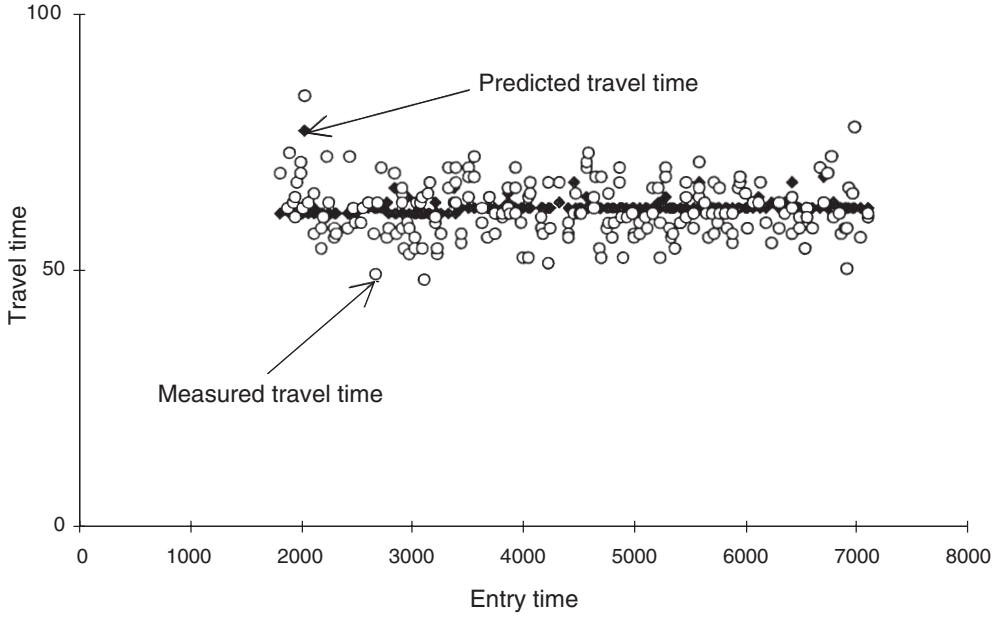


Figure 11.6 Predicted and measured travel times along the Broadway Street segment

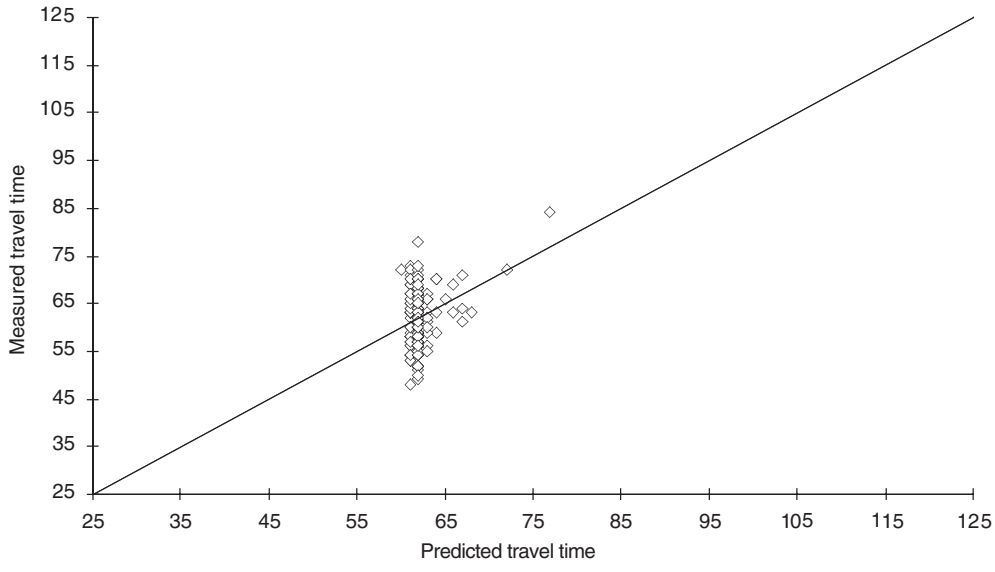


Figure 11.7 Comparison of predicted and measured travel times along the Broadway Street segment

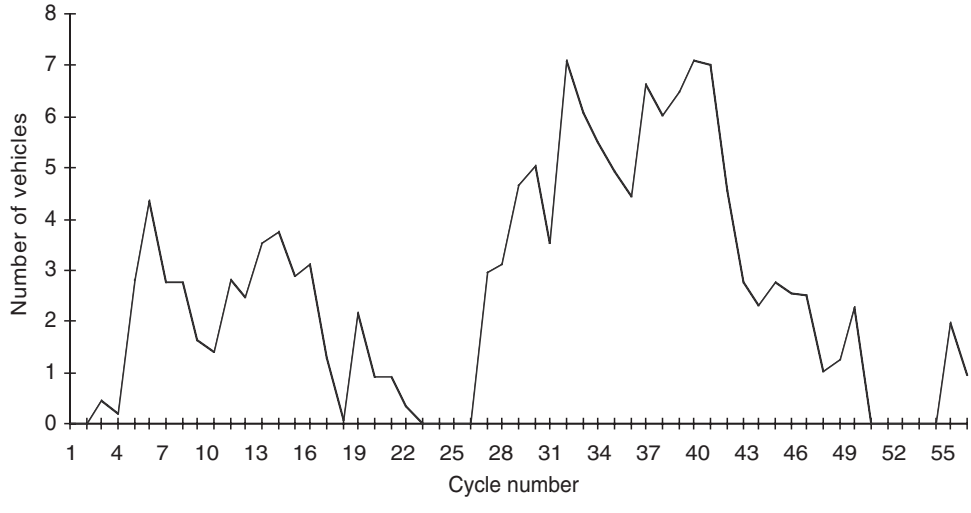


Figure 11.8 Overflow left-turn queues estimated for the Northwestern Avenue segment

between 80 and 250 seconds (Figure 11.12). Unlike the Broadway case, the predicted values varied considerably, but in most cases followed the measured values fairly closely. There were several exceptions, though, where predictions were considerably different from the measurements. These are cases where a driver was approaching a stop-line at the end of the green signal and had to decide whether to stop the car or to continue driving. This deci-

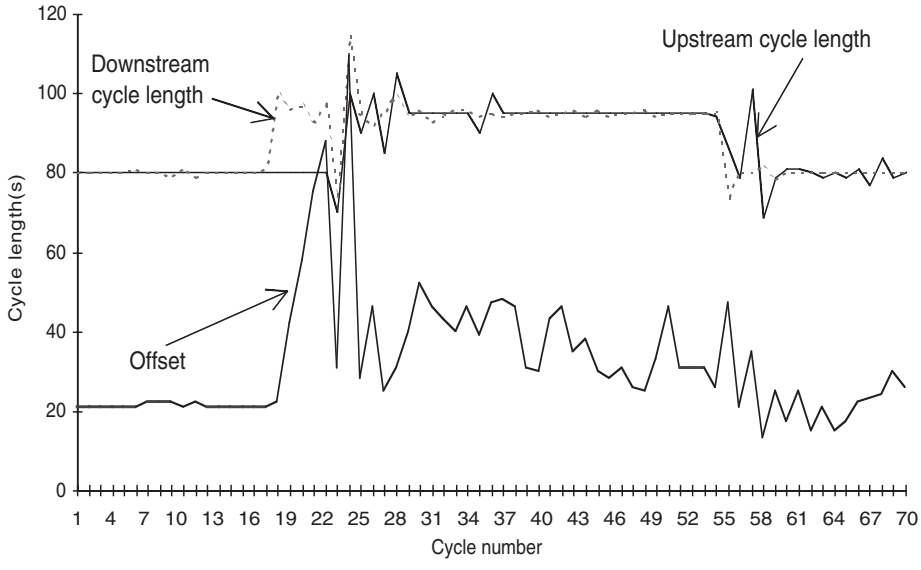


Figure 11.9 Cycles and offsets for the Northwestern Avenue segment

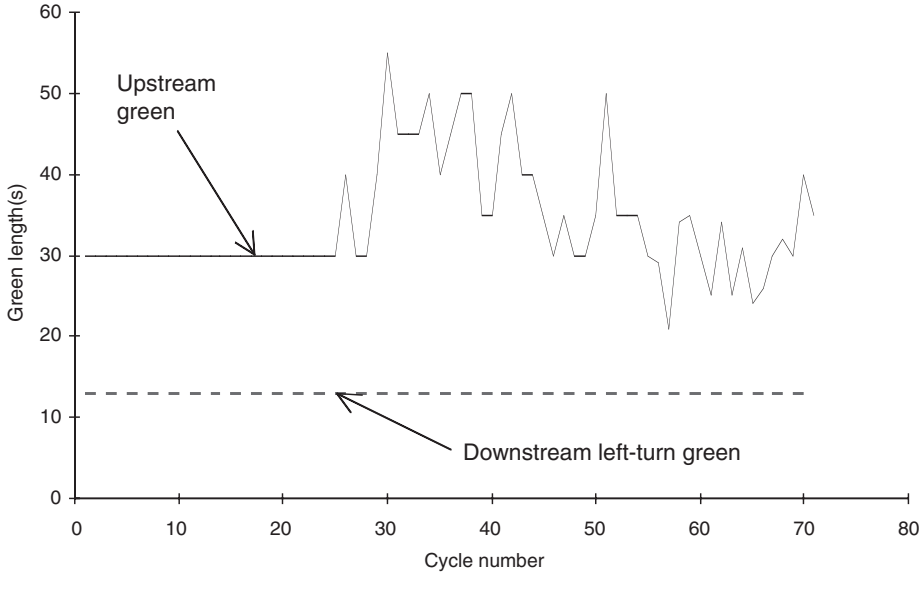


Figure 11.10 Green signals for the Northwestern Avenue segment

sion is very difficult to predict and probably any model of individual travel times may fail in this situation. The discrepancies between the predicted and actual decisions introduced errors approximately equal to the red signal lengths. The standard error of estimation is 26 per cent. This large value was caused by the noted end-of-green problem. The error reduced to 6 per cent if the end-of-green cases were ignored (reduced sample in Table 11.2, above).

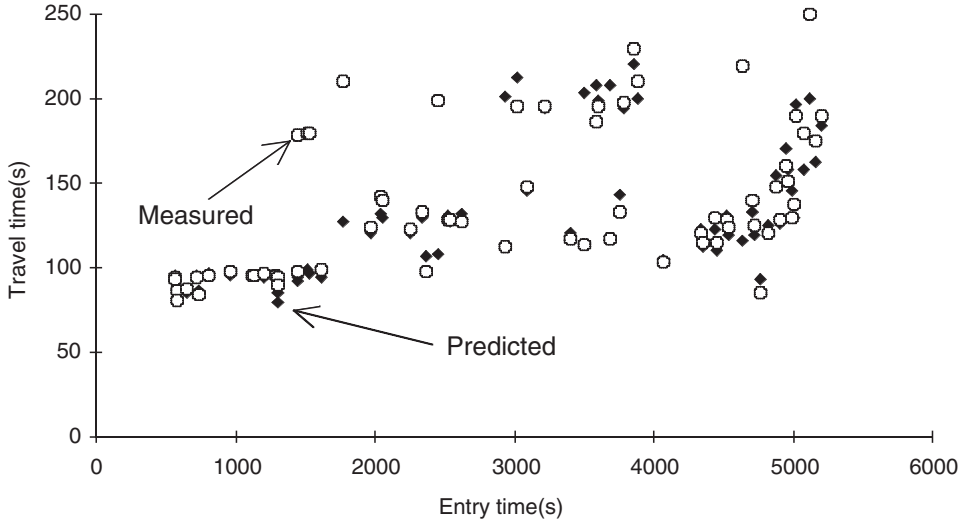


Figure 11.11 Predicted and measured travel times along the Northwestern Avenue segment

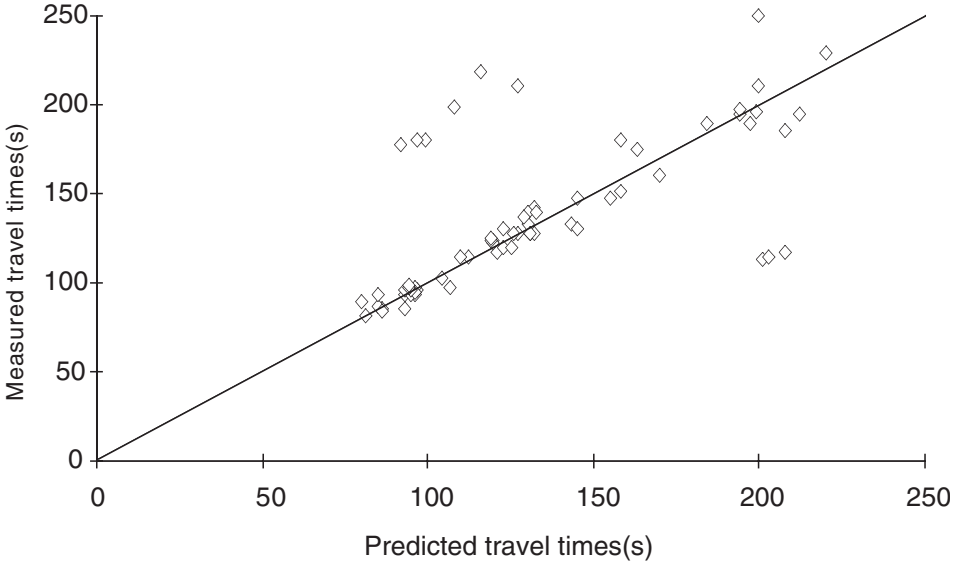


Figure 11.12 Comparison of the predicted and measured travel times along the Northwestern Avenue segment

3.4 Sagamore Parkway

This site was heavily congested during the observation due to a construction zone occupying one lane at some distance from the downstream intersection. The lack of capacity caused a queue spillback to the upstream intersection. The maximum number of left-turning vehicles present on the segment varied by around 25 vehicles during the spillback condition (Figure 11.13). The link length was 1.2 km and could accommodate a much larger number of left turns if they were queued in a dedicated traffic lane. Due to a short left-turn lane and a long queue of through vehicles, the left-turning vehicles were trapped in the joint queue and they could not proceed to the stop-line even when there was a green signal for left turns. The signals displayed to left-turning vehicles were different from the signals displayed to through and right-turning vehicles. There was no coordination between the signals, which is manifested through different cycle lengths, and different cycle numbers during the count period (Figure 11.14). The downstream signal was semi-actuated, though the left-turn green was almost constant (Figure 11.15).

The predicted and measured travel times are plotted in Figures 11.16 and 11.17. It is seen that the travel time follows a trend similar to the occupancy of the road segment. The calculated travel times are close to the observed values, except in a few cases. The value of the discrepancy in these cases is approximately equal to the length of the red signal. This indicates the end-of-green problem. The standard error of prediction is 17 per cent and drops to 4 per cent after excluding the end-of-green cases.

3.5 Additional Remarks

The estimation errors given in Table 11.2 apply to individual travel times and not to their expected values since only individual values can be observed. Since the method was intended to return the expected travel times, the actual estimation errors may be lower.

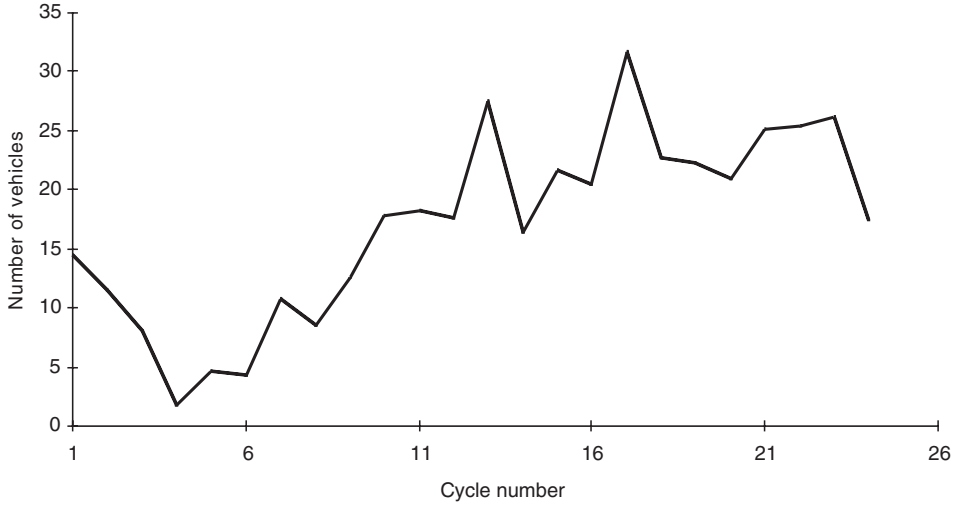


Figure 11.13 Overflow left-turn queues predicted for the Sagamore Parkway segment

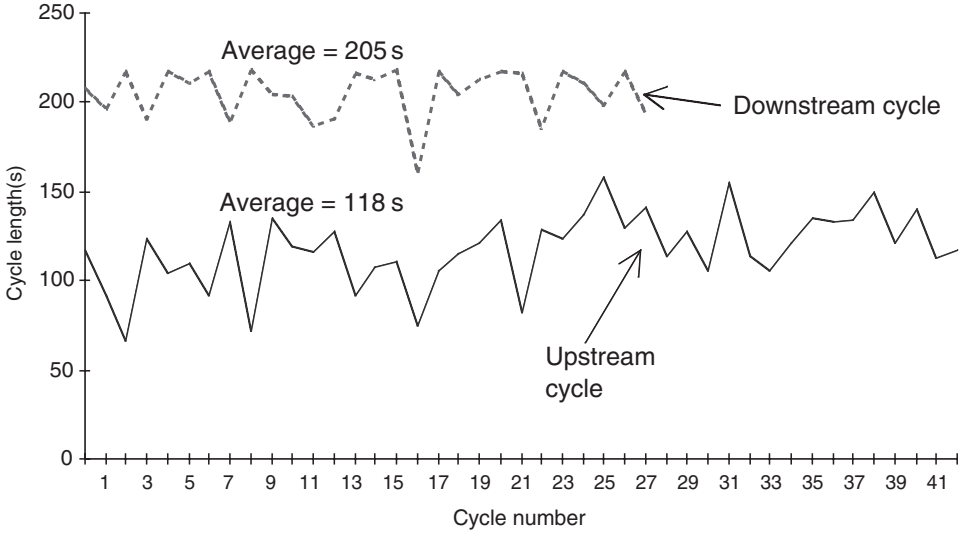


Figure 11.14 Signal cycles for the Sagamore Parkway segment

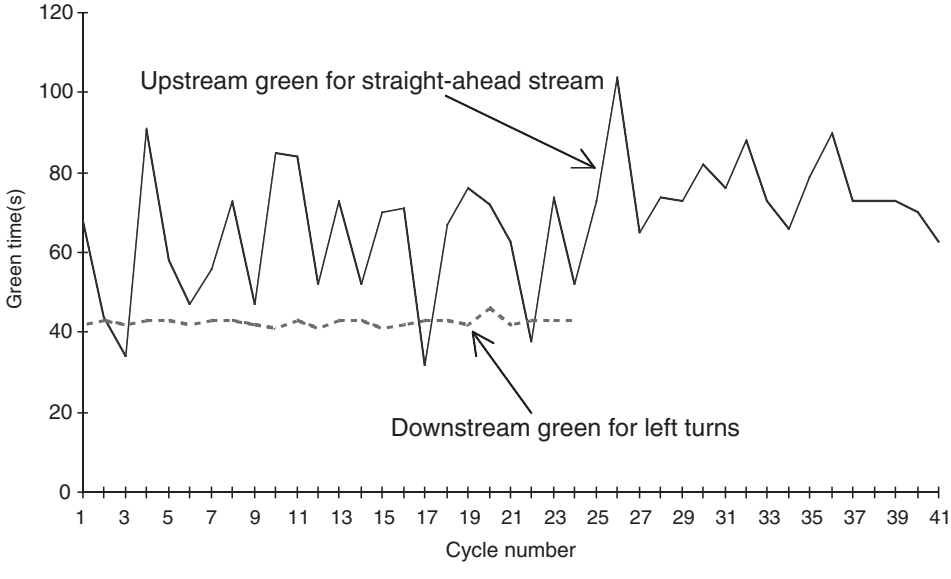


Figure 11.15 Green signals for the Sagamore Parkway segment

The estimation errors in Table 11.2 decrease with the growing congestion observed on the segments. This trend is easy to understand when one considers that the predictions are compared to measured individual travel times. High randomness of individual travel times as observed in non-FIFO conditions (light traffic) must cause high discrepancies between the results. On the other hand, heavy congestion with strongly imposed FIFO

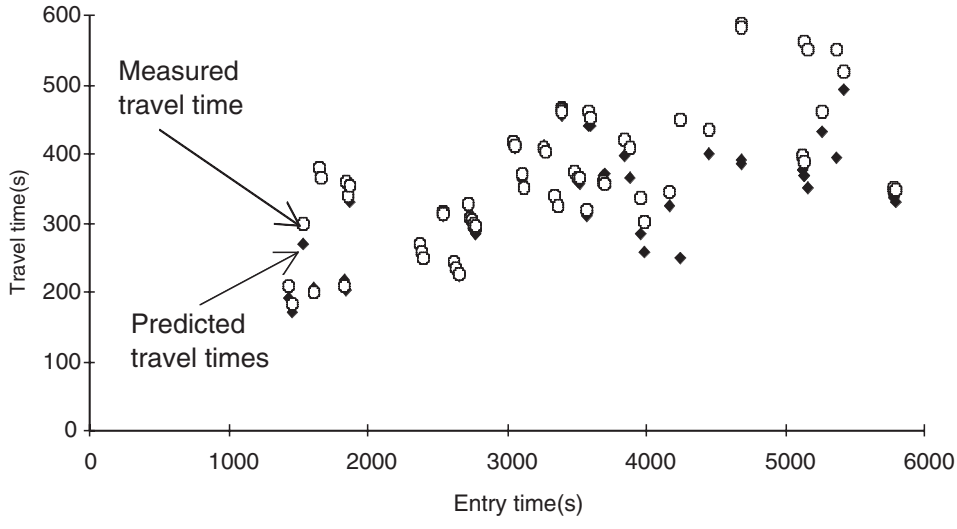


Figure 11.16 Predicted and measured travel times along the Sagamore Parkway segment

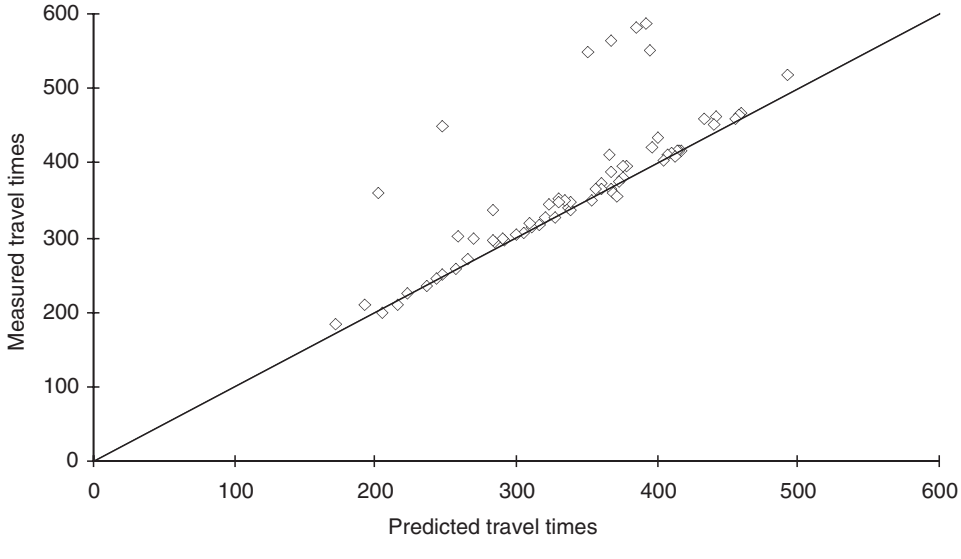


Figure 11.17 Comparison of the predicted and measured travel times along the Sagamore Parkway segment

reduces the randomness of individual travel times and at the same time reduces the estimation error.

The results of the field test indicate a 4 per cent overestimation bias. Courage et al. (1995), who evaluated the existing analytical method in the *Highway Capacity Manual* (Transportation Research Board 1994), reported an average overestimation error of 37

per cent. In the cited research, the average travel time over a longer period was estimated and the expected individual travel times were not.

4. CONCLUSION

Estimated or predicted *L* curves, which are cumulative flow rates, return expected travel times of vehicles entering the highway segment at known times, if the FIFO conditions persist along the segment. In the non-FIFO conditions, the sufficient condition for the estimates to converge to the expected value is that the flow rate does not change rapidly. This requirement is violated in traffic flows discharging from traffic signals.

The travel times of vehicles with known preferred speeds can be modeled using a modified downstream *L* curve if the net number of passing maneuvers along the segment is known. *L* curves do not provide sufficient traffic representation to estimate this value; nor is other useful theory known for non-congested traffic with considerable inter-vehicle interactions.

In our test, the predicted expected travel times were compared to the measured travel times of individual vehicles. The prediction standard errors in our test were quite high if the results included vehicles arriving at the stop-line at the end of the green signal. It is difficult to avoid these errors since they are dependent on a driver's decision to stop a car or continue driving, which is impossible to predict.

The prediction errors for vehicles that did not arrive at the end of the green signal were much lower and fell within the range of 4 per cent and 9 per cent. The highest standard error was observed on segments with light traffic and no congestion, while the lowest was at the location with heavily congested segments with FIFO traffic. The trend in the prediction errors follows the trend in the variability of individual travel times under various levels of congestion.

The field test focused on estimation of the travel times where entering volumes, signal timing and capacity at bottleneck were known. The error would be higher if these values had to be predicted. Nevertheless, the results are encouraging and indicate that cumulative *L* curves should be considered as a feasible approach to modeling travel times along signalized streets for a wide range of traffic conditions.

NOTE

* This research was partially funded by the ITS-IDEA Program,which is jointly supported by the Federal Highway Administration, the National Highway Traffic Safety Administration, and the Federal Railroad Administration. The authors would like to thank David Boyce and Bader Hafeez for their assistance in collecting and processing field data used in this chapter.

REFERENCES

Cassidy, M.J. and J. Windover (1995), 'Methodology for assessing dynamics of freeway flow', *Transportation Research Record*, no.1484, 73–9.

Courage, K.G., R.H. Showers and D.S. McLeod (1995), 'Reconciling estimated and measured travel

times on urban arterial streets', paper presented at the 74th Annual Meeting of the Transportation Research Board, Washington, DC.

Hillier, J.A. and R. Rothery (1967). 'The synchronization of traffic signals for minimum delays', *Transportation Science*, **1** (2), 81–94.

Hunt P.B., D.I. Robertson, R.D. Bretherton and R.I. Winton (1981), 'SCOOT – a traffic responsive method of coordinating signals', TRRL Report LR1014, Crowthorne.

Lighthill, M.J. and G.B. Whitham (1955), 'On kinematic waves. I: Flow movement in long rivers. II: A theory of traffic flow on long crowded roads', *Proceedings of the Royal Society A*, no. 229, pp. 281–345.

Munoz, J.C. and C.F. Daganzo (2000), 'Experimental characterization of multi-lane freeway traffic upstream of an off-ramp bottleneck', Partners for Advanced Transit and Highways (PATH) Working Paper 2000-13, Institute of Transportation Studies, University of California, Berkeley, CA.

Newell, G.F. (1993), 'Simplified theory of kinematic waves in highway traffic. I: General theory. II: Queuing at freeway bottlenecks. III: Multi-dimensional flows', *Transportation Research*, **27B**, 281–313.

Richards, P.I. (1956), 'Shockwaves on the highway', *Operations Research*, **4**, 42–51.

Tarko, A.P., G. Rajaraman and D.E. Boyce (2000), *Travel Time Prediction in Intelligent Transportation Systems*, ITS-IDEA Final Report, ITS-49, Transportation Research Board, National Research Council, Washington, DC.

Transportation Research Board (1994), *Highway Capacity Manual*, Special Report 209, Washington, DC.

12. System performance in network with parking and/or route information systems

William H.K. Lam, K.S. Chan and B.F. Si*

1. INTRODUCTION

In view of the traffic congestion problem in most metropolitan areas around the world and rapid developments in telecommunications and information processing technologies, the advanced traveler information systems (ATIS) are currently being developed in order to alleviate traffic congestion and enhance the performance of road networks. A fundamental requirement for ATIS applications is the development of drivers' choice behavior models in the presence or absence of parking/route information for assessment of the system impacts. Therefore, an understanding of drivers' behavior when they make travel choices including route and parking choices is important in modeling traffic on road networks. In the past decade, attention has been given to the development of models for evaluating ATIS (Hall 1996; Yang 1998; Lo et al. 1999; Chan and Lam 2002; Hong and Szeto 2002; Yin and Yang 2003). Most of this previous research is concerned mainly with the impacts of ATIS on route choices, but little has been done in modeling the effects of ATIS on parking choices. In practice, parking information can be collected more easily from the car park operators, and in general, is also much more accurate than the route information. So parking information is more likely to be available to drivers at lower cost. Recently, the parking information system (PIS) has been widely integrated with ATIS. Under the PIS environment, drivers can be provided with parking information (such as the available parking spaces, parking charges, access times and so on) so as to reduce the uncertainty of searching for parking spaces, particularly in densely-populated urban areas such as in Hong Kong.

In the literature, there are several papers investigating the route or parking choice models (Van Der Goot 1982; Lambe 1996; Asakura 1997; Yang 1998). In these studies, route choice and parking choice are considered separately. In fact, drivers have to choose both the route and the car park simultaneously, based on the perceptions of the route travel time and parking delay. As there always exists a random factor on drivers' perceptions in affecting their choice behaviors, in practice, drivers cannot receive perfect information of road traffic conditions, regardless of whether they have ATIS. Hence, drivers would select alternatives in a stochastic manner but with different perception variances. In principle, those with ATIS have a smaller perception variance for receiving the better traffic information, while those without would have a larger perception variance on both route travel time and parking delay.

An important factor that controls the benefits of ATIS is the level of market penetration, defined as the proportion of drivers equipped with the system. Most studies have

focused on the investigation of the benefits by assuming varied levels of market penetration. It was found that the benefits of ATIS to the road network, equipped and unequipped drivers all depend strongly on the level of market penetration (Emmerink et al. 1995a, 1995b). Emmerink et al. (1994) provided a framework for analysing the market potential of ATIS from an economic point of view. However, the demand relationship between equipped and unequipped drivers has not yet been incorporated into the ATIS modeling explicitly. Yang (1998) developed a combined equilibrium model that captures this demand relationship endogenously and equilibrates the market penetration to the net benefit of the road network, analogous to supply–demand equilibrium. He assumed that the drivers equipped with ATIS receive perfect information and hence would always be able to find the minimum travel time routes in a user equilibrium (UE) manner. However, in reality, ATIS cannot provide drivers with absolute accurate (or 'error-free') information and there are always stochastic perception errors on route travel times and parking delays. Usually, parking information is more accurate than route travel time information.

The previous related studies have found substantial benefits through provision of ATIS (10–20 per cent reduction in travel time), along with maximum benefits at relatively low market penetration (about 30 per cent). Counterintuitively, it was found in some simulation experiments that the average travel time in a road network increases when the level of ATIS market penetration is greater than 50 per cent (Hall 1996). These results show that a high level of market penetration could possibly lead to overreaction and a deterioration in the network performance (Ben-Akiva et al. 1991; Mahmassani and Jayakrishnan 1991). In view of this, some studies have been undertaken to investigate the effects of exogenous market penetrations of ATIS on the overall network travel time (Boyce 1988; Arnott et al. 1991; Bonsall et al. 1991; Halati and Boyce 1991; Watling 1994). Most of these studies, however, considered market penetration as a constant exogenous parameter, regardless of the potential travel time savings.

In practice, drivers are willing to pay for the service of ATIS when its benefit is perceived as greater than its cost. Provided that the drivers equipped with ATIS can perceive benefits in reducing the uncertainty of route and parking choices, the market penetration of ATIS will be increased. Contrarily, if ATIS cannot provide drivers with significant net benefits in terms of the difference between total travel time saving and the ATIS service cost, the market penetration will then be reduced. Thus, the demand for ATIS should be an elastic function of the net benefits. Subsequently, it leads to an optimal market penetration of ATIS that could achieve a 'win–win' situation for both the individual drivers and the overall system. In this chapter, the market penetration of ATIS is modeled in an elastic manner but the effects of parking and route information will be considered separately in the ATIS. Analogous to supply–demand equilibrium, market penetration is determined by the cost of ATIS, quality of service (or accuracy of the information system) and willingness of drivers to pay for the service. The willingness to pay is dependent on the sensitivity of drivers to the benefits of the ATIS service. So, in order to assess the effects of ATIS on network performance, all these three factors should be taken into account explicitly.

In addition, the decision of the traffic authority (that is, the leader) and the responses of the drivers (that is, the followers) should also be considered simultaneously so as to fully analyse the effects of the ATIS service. Note that from the viewpoint of the traffic

authority, the objective of implementing ATIS is to reduce the road network congestion; while the objective of the drivers is only to minimize their own travel time. Therefore, in a network that has ATIS, drivers can be divided into two classes: those with ATIS and those without. It is assumed that those equipped with ATIS have smaller perception errors on route travel times and parking delays but at different degrees as parking information is more accurate than route information.

In this chapter, a bi-level programming model is proposed to determine the optimal network performance in a network with parking and/or route information systems while the drivers' parking and route choice behaviors are explicitly taken into account. The lower-level problem is a multi-class probity-based stochastic user equilibrium (SUE) model that describes the combined parking and route choice problem. The advantage of using a probit- rather than a logit-based SUE model is that the independence of irrelevant alternatives (IIA) problem can be removed. There are objections to using a logit-based SUE model because of the axiom of the IIA, which is: 'where any two alternatives have a non-zero probability of being chosen, the ratio of one probability over the other is unaffected by the presence or absence of any additional alternative in the choice set' (Luce and Suppes 1965). This property causes the model to fail in the presence of the correlated alternatives. It is particularly important to the route choice problem. Unrealistic route choice probabilities can be derived from the IIA property of the logit model (Daganzo and Sheffi 1977; Sheffi 1985) although a C-logit model has been proposed by Cascetta et al. (1996) to alleviate the IIA problem in the logit-based traffic assignment model. Another advantage of the probit-based SUE model is that the use of the simulation method for solving the probit-based approach is flexible if the travel time error distribution is varied by route and road type.

The upper-level problem represents a network performance optimization problem, the objective of which is to minimize the total network travel time. The sensitivity analysis method is presented for solving the bi-level programming problem. A numerical example is used to illustrate the application of the proposed model and of the solution algorithm.

2. ASSUMPTIONS

Throughout this study, the following assumptions are adopted:

1. The study period is assumed to be a one-hour (unit time) period, such as the morning peak hour. It is known that the morning peak hour is usually the most critical period in a day and all the car trips are home-based work trips. It is also assumed that no round trips occur during the one-hour study period (Lam et al. 1999).
2. Although the type of car parks can be classified as private or public, in this chapter, only public car parks are considered to cater for the parking demand at the destination or at the attraction end (Lam et al. 1999).
3. Link travel times are continuous, strictly increasing functions of link flows, and the link travel time functions are assumed to be differentiable and separable; the same assumption is applied to the delays encountered in searching for a space in the car park (that is, the parking delay functions). The travel times on walk links are assumed to be given and fixed. Note that both the equipped and unequipped drivers share the

same links and car parks (with the corresponding link travel time functions and parking delay functions). These assumptions ensure the existence of equivalent mathematical programs for the combined SUE problem and the existence, uniqueness and stability of the solution (Daganzo 1983; Van Vuren and Watling 1991; Bennett 1993). In order to facilitate the presentation of the essential ideas in this chapter, the travel costs such as road tolls and operating costs are ignored. It is further assumed that there is one person per private car (or passenger car unit).

4. The total origin–destination (O–D) demands for all O–D pairs are given and fixed.
5. Neither equipped nor unequipped drivers have perfect ('error-free' knowledge of route travel times and parking delays in a network with ATIS. Drivers would make route and parking choices consistent with SUE, while the equipped drivers have smaller perception errors on route travel times and parking delays. The extent of these perception variations of the equipped drivers can be considered as a measure of the information quality of the ATIS services. It is assumed that the accuracy (or quality) or parking information is better than that of route information in ATIS.
6. The probit-based approach is adopted in the SUE assignment model in which the perceived link travel times and perceived parking delays are normally distributed.
7. The split function between equipped and unequipped drivers is presented in a logit-type choice model. The market penetration of ATIS is defined as the proportion of drivers equipped with ATIS, and is affected by three factors: the cost of (or charge for) the ATIS service; its quality; and the willingness of drivers to pay for it. The willingness to pay is dependent on the sensitivity of drivers to the benefits of ATIS.

3. NOTATION

In this chapter, we use superscript '–' for the variables of drivers equipped with ATIS and superscript '^' for those variables of drivers who are unequipped. The following notation will be used throughout:

A	set of links including road links, parking links and walk links in the road network;
N	set of nodes such as home, school or work locations and so on;
I	set of car parks in the road network;
O	set of origins in the road network;
D	set of destinations in the road network;
K_{ri}	set of routes between the origin $r \in O$ and the car park $i \in I$;
I_s	set of car parks that is associated with the destination of $s \in D$;
x_a	vehicular flow (in passenger car units per hour, pcu/hr) on road link $a \in A$;
t_a^0	free-flow travel time (hrs) on road link $a \in A$;
C_a	capacity (pcu/hr) of road link $a \in A$;
d_i^0	free-flow parking access time (hrs) to car park $i \in I$;
d_i^1	a pre-determined parameter of parking delay (hrs) at car park $i \in I$;

H_i available parking spaces (pcu) in car park $i \in \mathrm{I}$;

y_i parking flow or demand (pcu) in the car park $i \in \mathrm{I}$;

w_{is} walk time (hrs) from car park $i \in \mathrm{I}$ to the destination $s \in \mathrm{D}$;

q_{rs} total demand (pcu/hr) from $r \in \mathrm{O}$ to $s \in \mathrm{D}$;

$\bar{q}_{rs}$ ($\hat{q}_{rs}$) O–D demand of equipped (unequipped) drivers between $r \in \mathrm{O}$ and $s \in \mathrm{D}$;

b_{rs} market penetration of ATIS (that is, proportion of drivers equipped with ATIS, $\bar{q}_{rs}/q_{rs}$);

$\bar{\alpha}$ ($\hat{\alpha}$) parameter of perceived quality (error) of route information (route travel time) by equipped (unequipped) drivers;

$\bar{\beta}$ ($\hat{\beta}$) parameter of perceived quality (error) of parking information (parking delay) by equipped (unequipped) drivers;

$\bar{c}_k^{ri}$ ($\hat{c}_k^{ri}$) perceived travel time or cost (hrs) on route $k \in \mathrm{K}_{ri}$ by equipped (unequipped) drivers;

$\bar{d}_i$ ($\hat{d}_i$) perceived parking delay (hrs) at car park $i \in \mathrm{I}$; by equipped (unequipped) drivers;

$\bar{C}_{ki}^{rs}$ ($\hat{C}_{ki}^{rs}$) perceived total travel time or cost (hrs) from $r \in \mathrm{O}$ to $s \in \mathrm{D}$, via route $k \in \mathrm{K}_{ri}$ and car park $i \in \mathrm{I}_s$ by equipped (unequipped) drivers, which is equal to the sum of the route travel time $\bar{c}_k^{ri}$ ($\hat{c}_k^{ri}$), the corresponding parking delay $\bar{d}_i$ ($\hat{d}_i$) and the walk time w_{is};

$\bar{p}_{ki}^{rs}$ ($\hat{p}_{ki}^{rs}$) probability of choosing route $k \in \mathrm{K}_{ri}$ and car park $i \in \mathrm{I}_s$ for equipped (unequipped) drivers from $r \in \mathrm{O}$ to $s \in \mathrm{D}$;

$\bar{f}_k^{ri}$ ($\hat{f}_{ki}^{ri}$) vehicular flow (pcu/hr) of equipped (unequipped) drivers on route $k \in \mathrm{K}_{ri}$, traveling from $r \in \mathrm{O}$ to $s \in \mathrm{D}$ via car park $i \in \mathrm{I}_s$.

4. LOWER-LEVEL PROBLEM

Consider a road network G = [(N,I),A], we assume that the actual travel time on each link $a \in \mathrm{A}$ is an increasing and strictly convex function of the traffic flow x_a on the link and that the actual parking delay at each car park $i \in \mathrm{I}$ is also an increasing and strictly convex function of the parking flow (demand) y_i at the car park:

$$t_a = t_a(x_a) \tag{12.1a}$$

$$d_i = d_i(y_i). \tag{12.1b}$$

The travel by O–D pair is composed of the trip from the origin to the car park, searching for a parking lot and walking from the car park to the destination. So the actual travel time on a route $k \in \mathrm{K}_{ri}$ between O–D pair rs basically consists of three components: (i) route travel time from the origin to car park; (ii) parking delay; and (iii) walk time from car park to destination:

$$C_{ki}^{rs} = c_k^{ri} + d_i + w_{is}, \tag{12.2}$$

where c_k^{ri} is the travel time on the route $k \in \mathrm{K}_{ri}$, d_i is the parking delay (hrs) at car park $i \in \mathrm{I}$, and w_{is} is the walk time from car park $i \in \mathrm{I}$ to destination $s \in \mathrm{D}$. c_k^{ri} can be obtained by:

$$c_k^{ri} = \sum_{a \in \mathrm{A}} t_a(x_a) \delta_{a,k}^{ri}, \tag{12.3}$$

where $\delta_{a,k}^{ri} = 1$ if link a is a part of route k between ri, and 0 otherwise.

We now consider the route and parking choice behaviors of drivers with and without ATIS. In the literature, perception error of total travel time can be expressed by a random variable (Bifulco 1993; Lambe 1996). However, the perceived route travel time and perceived parking delay are influenced by different information sources of ATIS. Therefore, in this chapter, we consider the perception errors of the route travel time and parking delay as two independent random variables. The perceived route travel time (parking delay) for equipped and unequipped drivers is equal to the sum of the route travel time $c_k^{ri}, k \in \mathrm{K}_{ri}$ (parking delay, $d_i, i \in \mathrm{I}$) and a random variable (that is, different random errors for route travel time and parking delay). These errors are assumed normal distributed at the zero mean value, but with different variances perceived by equipped and unequipped drivers for route travel times and parking delays.

First, for equipped drivers, we have:

$$\bar{c}_k^{ri} = c_k^{ri} + \bar{\varepsilon}_k^{ri}, \ \bar{\varepsilon}_k^{ri} = \sum_{a \in \mathrm{A}} \bar{\varepsilon}_a \delta_{a,k}^{ri}, \ \bar{\varepsilon}_a \sim \mathrm{N}(0, \bar{\alpha} t_a^0) \tag{12.4}$$

$$\bar{d}_i = d_i + \bar{\varsigma}_i, \ \bar{\varsigma}_i \sim \mathrm{N}(0, \bar{\beta} d_i^1), \tag{12.5}$$

where $\bar{\varepsilon}_k^{ri}$, $\bar{\varepsilon}_a$, respectively, are referred to as the random variable of travel time on route k from the origin r to car park i and the random variable of travel time on road link a; similarly, $\bar{\varsigma}_i$ is the random variable of the parking delay for the car park i; t_a^0 is free-flow travel time on the link a; d_i^1 is a pre-determined parameter of parking delay at the car park i; $\bar{\alpha}$ and $\bar{\beta}$ are parameters of perceived quality of route and parking information, respectively. Note that a lower value of $\bar{\alpha}$ and $\bar{\beta}$ implies that the equipped drivers have a smaller variability in their route travel times and parking delays due to higher degree of accuracy of route and parking information.

The total travel time (or cost) for those equipped drivers can be formulated as below:

$$\bar{C}_{ki}^{rs} = \bar{c}_k^{ri} + \bar{d}_i + w_{is}, \tag{12.6}$$

and the choice probability of car park i and route k for equipped drivers is:

$$\bar{p}_{ki}^{rs} = \Pr(\bar{C}_{ki}^{rs} \le \bar{C}_{hl}^{rs}, \forall h \in \mathrm{K}_{ri}, \forall l \in \mathrm{I}_s, h \ne k \text{ or } l \ne i). \tag{12.7}$$

For unequipped drivers, similar to (12.4)–(12.7), we have:

$$\hat{c}_k^{ri} = c_k^{ri} + \hat{\varepsilon}_k^{ri}, \ \hat{\varepsilon}_k^{ri} = \sum_{a \in \mathrm{A}} \hat{\varepsilon}_a \delta_{a,k}^{ri}, \ \hat{\varepsilon}_a \sim \mathrm{N}(0, \hat{\alpha} t_a^0) \tag{12.8}$$

$$\hat{d}_i = d_i + \hat{\varsigma}_i, \ \hat{\varsigma}_i \sim \mathrm{N}(0, \hat{\beta} d_i^1) \tag{12.9}$$

$$\hat{C}_{ki}^{rs} = \hat{c}_{k}^{ri} + \hat{d}_{i} + w_{is} \tag{12.10}$$

$$\hat{p}_{ki}^{rs} = \Pr(\hat{C}_{ki}^{rs} \le \hat{C}_{hl}^{rs}, \forall h \in \mathrm{K}_{ri}, \forall l \in \mathrm{I}_{s}, h \ne k \text{ or } l \ne i). \tag{12.11}$$

Note that $\hat{\alpha}$, $\hat{\beta}$ are parameters related to perceived errors of unequipped drivers (based on their past experiences) on route travel times and parking delays. Therefore, the route flows and parking demands based on the SUE principle are given by:

$$\bar{f}_{ki}^{rs} = \bar{p}_{ki}^{rs} \bar{q}_{rs} \tag{12.12a}$$

$$\hat{f}_{ki}^{rs} = \hat{p}_{ki}^{rs} \hat{q}_{rs}, \tag{12.12b}$$

where $\bar{f}_{ki}^{rs}, \hat{f}_{ki}^{rs}$, respectively, are the traffic flows of equipped and unequipped drivers on the route $k \in \mathrm{K}_{ri}$ and choosing car park $i \in \mathrm{I}_{s}$; $\bar{q}_{rs}$, $\hat{q}_{rs}$, respectively, are the O–D demand of equipped and unequipped drivers between $r \in \mathrm{O}$ and $s \in \mathrm{D}$.

An important factor that affects the benefits of ATIS is the level of market penetration that is defined as the proportion of drivers equipped with ATIS (Yang 1998), represented by b_{rs} in this chapter. Suppose that the total travel demand between each O–D pair denoted by q_{rs} is given and fixed. Obviously,

$$\bar{q}_{rs} + \hat{q}_{rs} = q_{rs} \tag{12.13}$$

$$b_{rs} q_{rs} = \bar{q}_{rs} \tag{12.14a}$$

$$(1 - b_{rs}) q_{rs} = \hat{q}_{rs}. \tag{12.14b}$$

In principle, the drivers with ATIS can receive more accurate traffic information and can then make better travel choices as compared to the drivers without ATIS. So, drivers will decide whether to equip with ATIS based on the total travel time saving generated by the system and the cost of the ATIS service. Suppose that the market penetration of ATIS is an endogenous variable, and can be determined by the cost of the ATIS service, the quality of the service (or accuracy of information) and the willingness of drivers to pay for the service. The quality of the ATIS service can be measured by the benefit derived from the system and the benefit is defined by the travel time saving of the drivers with ATIS.

The average travel time of equipped drivers and unequipped drivers with SUE choice behavior, respectively, is formulated as follows:

$$\bar{C}_{rs} = \sum_{k \in \mathrm{K}_{ri}} \sum_{i \in \mathrm{I}_{s}} \bar{p}_{ki}^{rs} C_{ki}^{rs} \tag{12.15a}$$

$$\hat{C}_{rs} = \sum_{k \in \mathrm{K}_{ri}} \sum_{i \in \mathrm{I}_{s}} \hat{p}_{ki}^{rs} C_{ki}^{rs}. \tag{12.15b}$$

Therefore, the travel time saving generated by the ATIS can be derived by:

$$S_{rs} = \hat{C}_{rs} - \bar{C}_{rs}. \tag{12.16}$$

Generally, S_{rs} ($r \in \mathrm{O}$, $s \in \mathrm{D}$) is positive, because buying the ATIS service should be beneficial to the equipped drivers. Obviously, the more accurate the ATIS service is, under a certain level of ATIS market penetration, the larger is the travel time saving, S_{rs}.

The proportion of drivers buying ATIS can be determined by the following exponential function (Yang 1998):

$$b_{rs} = \frac{1}{1 + \exp(\lambda - \theta S_{rs})}, \tag{12.17}$$

where λ is the service cost of acquiring and using the ATIS service, and θ is a positive parameter which can reflect the sensitivity of the drivers to the benefits of the ATIS service. It implies that the larger the θ, the higher is the willingness of drivers to pay for ATIS.

5. UPPER-LEVEL PROBLEM

In the upper-level problem, the traffic authority aims to minimize the total network travel time by varying the ATIS service cost, thus driving the market penetration to the expected level. This can be formulated as the following nonlinear programming problem, in which the cost of the ATIS service is the decision variable of the upper-level problem:

$$\min Z(\lambda) = \sum_{a \in \mathrm{A}} x_a(\lambda) t_a[x_a(\lambda)] + \sum_{i \in \mathrm{I}} y_i(\lambda) d_i[y_i(\lambda)] + \sum_{s \in \mathrm{D}} \sum_{i \in \mathrm{I}_s} g_{is}(\lambda) w_{is}, \tag{12.18}$$

where g_{is} is the flow on the walk link between car park i, $i \in \mathrm{I}$ and destination s, $s \in \mathrm{D}$, so we have:

$$g_{is}(\lambda) = \sum_{r \in \mathrm{O}} \sum_{k \in \mathrm{K}_{ri}} [\bar{f}_{ki}^{rs}(\lambda) + \hat{f}_{ki}^{rs}(\lambda)]. \tag{12.19}$$

6. BI-LEVEL MODEL FOR PARKING AND ROUTE CHOICE PROBLEM WITH EQUILIBRIUM MARKET PENETRATION

The optimal network performance problem under an ATIS environment with parking and route information with elastic market penetration can be represented as a leader–follower game problem, where the traffic authority is the decision maker or the leader, and the drivers who can freely be equipped or unequipped with ATIS are the followers. It is assumed that the traffic authority can influence, but cannot control the drivers' travel choice behaviors. The drivers make their own parking and route choices in a SUE manner. This interaction game can be formulated as the following bi-level programming problem in which the upper-level problem is to:

$$\mathrm{P}_1 \quad \min Z(\lambda) = \sum_{a \in \mathrm{A}} x_a(\lambda) t_a[x_a(\lambda)] + \sum_{i \in \mathrm{I}} y_i(\lambda) d_i[y_i(\lambda)] + \sum_{s \in \mathrm{D}} \sum_{i \in \mathrm{I}_s} g_{is}(\lambda) w_{is},$$

where $x_a(\lambda)$, $y_i(\lambda)$, $g_{is}(\lambda)$ are obtained by solving the lower-level problem.

The upper-level problem minimizes the total network travel time, while the lower-level

problem characterizes the multi-class probit-based SUE model for parking and route choices in a network with elastic or endogenous market penetrations of ATIS. By solving the above bi-level programming problem P_1, the optimal ATIS cost and the market penetration can then be derived so as to minimize the total network travel time, in which the quality of the ATIS service (accuracy of parking and route information) and willingness of drivers to pay for the service are considered implicitly as shown in the previous sections. Note that the above problem P_1, similar to other forms of bi-level mathematical programming problem, is intrinsically non-convex, and hence it might be difficult to solve for a global optimum (Yang and Yagar 1994).

7. SOLUTION ALGORITHM

The bi-level programming approach has emerged as an important area for progress in handling complicated traffic problems. Typical examples include traffic signal setting (Allsop 1974), optimal road capacity improvement (Abdulaal and LeBlanc 1979), estimation of O–D matrices from traffic counts (Yang et al. 1992), ramp metering in freeway-arterial corridor (Yang and Yagar 1994), and optimization of road tolls (Yang and Bell 1997).

Due to the intrinsic complexity of model formulation, the bi-level programming problem has been recognized as one of the most difficult, yet challenging problems for global optimality in transportation systems. In the past, researchers have developed alternative solution algorithms for this problem. Abdulaal and LeBlanc (1979) applied the Hook–Jeeves heuristic algorithm for direct search of the solution of the network design problem. Fisk (1984) developed an alternative single-level optimization model for the optimal signal control problem using a gap function. Suwansirikul et al. (1987) developed an alternative heuristic method, referred to as the 'equilibrium decomposed optimization algorithm', by approximating the derivative of the objective function in the upper-level problem. Subsequently, a sensitivity analysis method was proposed by Tobin and Friesz (1988), which makes use of the derivatives of the equilibrium link flows with respect to perturbation parameters, is widely used for network equilibrium problems. Friesz et al. (1993) applied this sensitivity analysis method to solve the network design problems. Yang and Yagar (1994) extended this method for solving the inflow control problem on a freeway. Yang (1995) used it for solving the queuing equilibrium network assignment problem, while the derivatives of equilibrium link flows and equilibrium queuing times with respect to traffic control parameters are derived. In this chapter, a sensitivity-analysis-based (SAB) solution algorithm similar to the one proposed by Chan and Lam (2002) is presented for solving the proposed bi-level mathematical problem P_1. The mechanism of the solution algorithm is an iterative process between the upper- and lower-level problems. The lower-level problem is repeatedly solved by varying the value of the ATIS service cost λ until the optimal λ is obtained. During the iteration, it is necessary to calculate the derivatives of equilibrium link flows and parking flows with respect to the cost. However, the sensitivity analysis method does not guarantee to obtain the global optimal solutions. Therefore, the SAB algorithm could stop at a local minimum. The proposed SAB solution algorithm for solving the bi-level problem is outlined as follows:

Step 0 Determine an initial ATIS service cost $\lambda^{(0)}$. Set $n=0$.

Step 1 Solve the lower-level problem (multi-class probit-based SUE problem) and get the link-flow pattern $\mathbf{x}^{(n)}$, car park flow pattern $\mathbf{y}^{(n)}$ and market penetration $\mathbf{b}^{(n)}$.

Step 2 Calculate the derivatives of equilibrium link flows and parking flows with respect to the ATIS service cost based on the sensitivity analysis method.

Step 3 Solve the upper-level problem by using the derivative information and obtain the auxiliary value of the ATIS service cost $\omega^{(n)}$.

Step 4 Update the ATIS service cost:

$$\lambda^{(n+1)}=\lambda^{(n)}+\frac{1}{n+1}[\omega^{(n)}-\lambda^{(n)}]. \tag{12.20}$$

Step 5 If $\max\{|\lambda^{(n+1)}-\lambda^{(n)}|\}\leq\varepsilon$ or $n=m$ then stop. Otherwise, let $n=n+1$ and go to Step 1. m is the maximum number of iterations. ε is a stopping parameter with a small value of 0.0001.

In this chapter, we adopt an extension of the method of successive average (EMSA) algorithm for solving the lower-level problem (multi-class probit-based SUE problem) so as to determine the route and parking in a network with equilibrium market penetration of ATIS. The EMSA algorithm is described as below:

Step 0 Initialization. Determine an initial value of market penetration $b_{rs}^{(0)}$, $r\in\mathrm{O}$, $s\in\mathrm{D}$. Perform Monte Carlo stochastic network loadings based on the initial free-flow travel time t_a^0, $a\in\mathrm{A}$ and parking delays d_i^0, $i\in\mathrm{I}$. This generates a set of link flows $x_a^{(0)}$, $a\in\mathrm{A}$ and of parking flows $y_i^{(0)}$, $i\in\mathrm{I}$. Set $n=0$.

Step 1 Calculate the link travel times $t_a^{(n)}$ and parking delays $d_i^{(n)}$, based on the link flows $x_a^{(n)}$ and parking flows $y_i^{(n)}$.

Step 2 Perform Monte Carlo stochastic network loadings based on the link travel times $t_a^{(n)}$ and parking delays $d_i^{(n)}$. This yields an auxiliary link-flow pattern $u_a^{(n)}$, $a\in\mathrm{A}$ and an auxiliary parking-flow pattern $v_i^{(n)}$, $i\in\mathrm{I}$.

Step 3 Update the link and parking flows:

$$x_a^{(n+1)}=x_a^{(n)}+\left(\frac{1}{n+1}\right)\left[u_a^{(n)}-x_a^{(n)}\right] \tag{12.21}$$

$$y_i^{(n+1)}=y_i^{(n)}+\left(\frac{1}{n+1}\right)\left[v_i^{(n)}-y_i^{(n)}\right]. \tag{12.22}$$

Update the market penetration of ATIS using equation (12.17):

$$b_{rs}^{(n)} = \frac{1}{1 + \exp[\lambda - \theta S_{rs}^{(n)}]} \text{ where } S_{rs}^{(n)} = \hat{C}_{rs}^{(n)} - \bar{C}_{rs}^{(n)}.$$

Step 4 Convergence criterion. If $|x_a^{(n+1)} - x_a^{(n)}| < \varepsilon_x, |y_i^{(n+1)} - y_i^{(n)}| < \varepsilon_y$ then stop. Otherwise, let $n = n + 1$ and go to Step 1.

8. NUMERICAL EXAMPLE

A numerical example is used to illustrate the application of the proposed model and solution algorithm. In this simple example, the market penetration of ATIS services is determined by varying the service cost, quality of service (or accuracy of parking and route information) and willingness of users to pay for the service. We analyse the results of this numerical example and investigate under what circumstances (including the levels of demand, supply and the quality of ATIS) the combined route and parking information system is effective in improving the total network travel time. Similar analyses are extended to the scenarios with route information system only and with parking information only. The example network shown in Figure 12.1 is similar to the example adopted by Yang (1998). It consists of one O–D pair, 12 nodes and 18 links. In addition, two car parks and two walk links are incorporated.

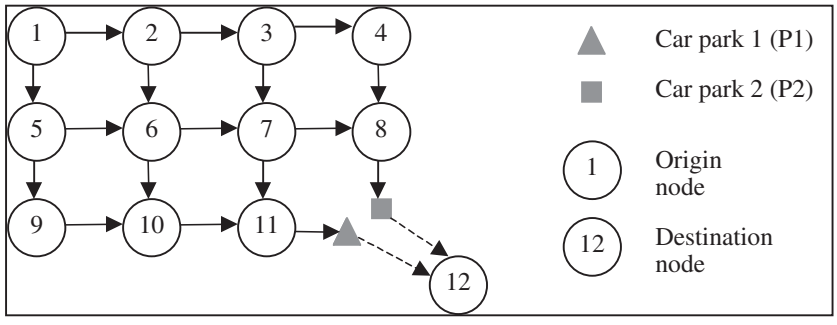


Figure 12.1 The test network

The following BPR (Bureau of Public Roads) link travel time function is used in this example:

$$t_a(x_a) = t_a^0 \left[1 + 0.15 \left(\frac{x_a}{C_a} \right)^4 \right]. \tag{12.23}$$

The free-flow travel time t_a^0 and link capacity C_a of the example network are presented in terms of hours (hrs) and passenger car units per hour (pcu/hr), respectively. They are shown in Table 12.1.

The following parking delay (in hrs) function is adopted (Lam et al. 1999):

$$d_i(y_i) = d_i^0 \left[1 + 0.31 \left(\frac{y_i}{H_i} \right)^{4.03} \right] \tag{12.24}$$

Table 12.1 Network input data

Link	t_a^0 (hrs)	C_a (pcu/hr)	Link	t_a^0 (hrs)	C_a (pcu/hr)
(1,2)	0.111	1000	(6,7)	0.094	1000
(2,3)	0.128	700	(7,8)	0.072	1000
(3,4)	0.094	700	(5,9)	0.133	900
(1,5)	0.100	1500	(6,10)	0.111	700
(2,6)	0.106	700	(7,11)	0.144	700
(3,7)	0.089	700	(9,10)	0.039	900
(4,8)	0.122	700	(10,11)	0.100	900
(5,6)	0.078	1000			

Table 12.2 Parking link input data

Parking link	Car park	d_i^0 (hrs)	H_i (pcu)
(11P1)	P1	0.69	800
(8,P2)	P2	0.85	1000

where H_i is the capacity (pcu) of car park $i \in$ I. Since illegal parking is not allowed in the model for the planning purpose, an artificial shadow link (Lam and Zhang 2000) can be introduced to each car park so as to store the excess parking-demand vehicles. The data for the example are given in Table 12.2.

The value of the parameter in equation (12.17), θ, is initially assumed as 30. It should be borne in mind that θ is a positive parameter which can reflect the sensitivity of the drivers to the benefits of the ATIS service. It implies that the larger the θ, the higher is the willingness of drivers to pay for the service. Since the perception errors on parking delays are comparatively smaller than those of route travel times in practice, the values of related parameters are $\hat{\alpha} = 1.0$, $\hat{\beta} = 0.5$ for unequipped drivers with reference to the perceived route travel time and parking delay errors, respectively. For equipped drivers, there are three scenarios: (I) $\bar{\alpha} < 1.0$ and $\bar{\beta} < 0.5$ implies that ATIS provides equipped drivers with route and parking information; (II) $\bar{\alpha} < 1.0$ and $\bar{\beta} = 0.5$ implies that ATIS provides equipped drivers with route information only; (III) $\bar{\alpha} = 1.0$ and $\bar{\beta} < 0.5$ implies that ATIS provides equipped drivers with parking information only. In this example, we examine market penetration of ATIS by varying the cost of the ATIS service, quality of information (or accuracy of information) and willingness of drivers to pay for it. Moreover, we carry out the sensitivity tests on both the supply and the demand sides, so as to assess their impacts on the network performance. In addition, the quality of PIS service (that is, accuracy of parking information) is also considered simultaneously in this numerical example. For brevity, we show two demand cases and two supply cases. 'D1' and 'D2' refer to the total travel demands of 1000 and 2000 units, respectively. 'S1' and 'S2' refer to the cases of $H_1 = 700$, $H_2 = 900$ and $H_1 = 1000$, $H_2 = 1200$, respectively.

Figure 12.2 shows the solution in terms of the objective function value of the upper-level problem against the iteration number for different O–D demands under the above

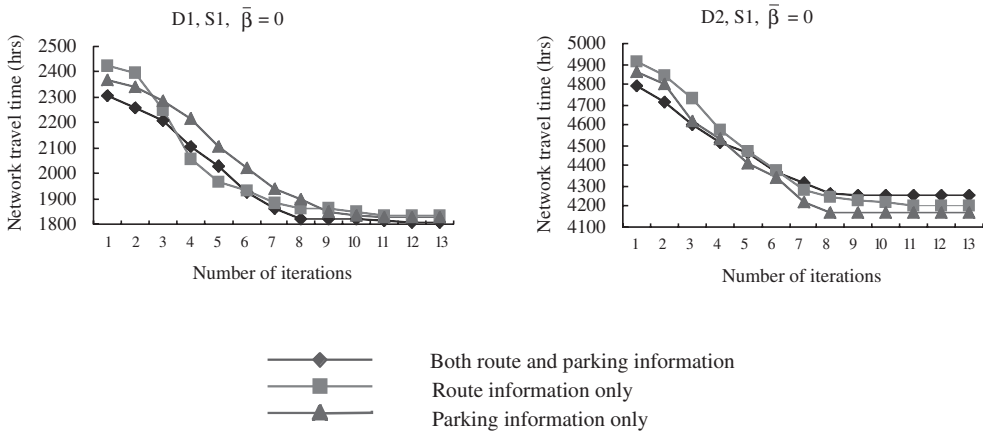


Figure 12.2 Convergence of the solution algorithm

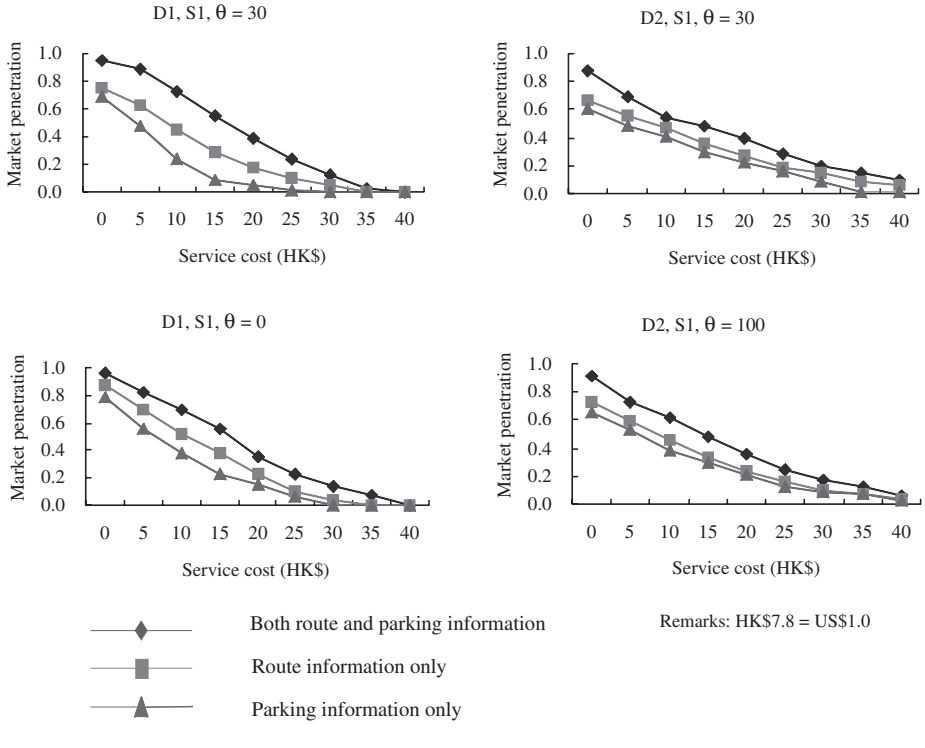


Figure 12.3 Market penetration of ATIS against the ATIS service cost

three scenarios. It can be observed that the solution algorithm has a fast convergence: the solutions for all the cases converge within 10 iterations.

Next we examine the market penetration by varying the service cost of ATIS with different levels of demand and supply. Figure 12.3 shows the solution of the market penetration against the ATIS service cost for the above three scenarios. It was found that the market penetration of ATIS is always decreasing with an increase in the service cost. It can also be seen that the market penetration for scenario III (that is, with parking information only) is lower than that of the other two scenarios at the same service cost level. This is because the difference of perception error on parking delays perceived by drivers with or without ATIS is comparatively smaller than that of the route travel times perceived by drivers in practice. Thus, the total travel time saving with parking information only is comparatively insignificant compared to the total travel time saving under the other two scenarios at the same service cost. And the benefit obtained by drivers equipped with parking information only is then comparatively insignificant compared to that of drivers with route information only or with both route and parking information at the same level of service cost. Therefore, in scenario III, the probability of drivers purchasing the ATIS service will be reduced and subsequently the market penetration of ATIS is lower.

In Figure 12.3, it is noted that the sensitivity of market penetration under the larger demand case (D2) is lower even with a larger value of θ for all three scenarios. The larger value of the parameter θ implies that the willingness of drivers to pay for ATIS is higher. The low sensitivity is due to the fact that under the larger demand level (D2), the congestion in the example network is relatively heavy, and the benefit for a driver equipped with ATIS is comparatively significant. Therefore the sensitivity of the market penetration against the cost of ATIS is comparatively lower when the demand level is larger. Similarly, the sensitivity of the market penetration against the cost is also lower when the value of the parameter θ is larger. For example, when $\theta = 30$, the reduction of market penetration for scenario I is 0.18 with an increase in the cost of the ATIS service from HK\$5.0 to HK\$10.0 (HK\$7.8 = US\$1.0). When $\theta = 100$, the reduction of market penetration for scenario I is only 0.13 with an increase in the cost of the service from HK\$5.0 to HK\$10.0. It is shown that the greater willingness of drivers to pay for ATIS can lead to lower sensitivity of the market penetration against its cost.

Now we examine the total network travel time by varying the cost of the ATIS service. We analyse the example results with different levels of total demand, the supply of car parks and the quality (or accuracy) of the parking information system. Figure 12.4 shows the effects of the ATIS service cost on the network performance for the above three scenarios given that ATIS provides the equipped drivers with complete parking information only (that is, $\bar{\beta} = 0$).

It was found in Figure 12.4 that the optimal network performance (that is, minimum network travel time) is achieved at a certain ATIS service cost. For example, under the condition of (D1, S1, $\bar{\beta} = 0$), the optimal cost of the ATIS service is HK\$15.0 with the optimal market penetration of 0.54, which implies that a high level of market penetration could possibly lead to an increase in the total network travel time. In Figure 12.4, it can also be seen that with parking information only, the system optimization can be reached with a lower ATIS service cost than those with route information only or with both. Under the condition of (D1, S1, $\bar{\beta} = 0$) and (D1, S2, $\bar{\beta} = 0$), the difference in minimal

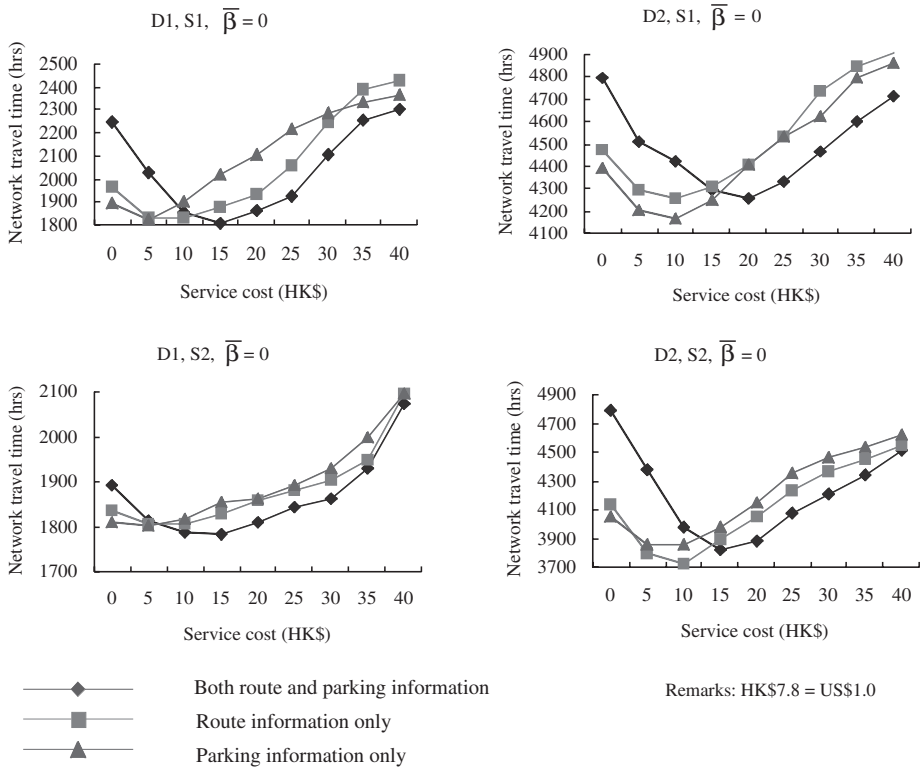


Figure 12.4 Effect of the ATIS service cost on the network travel time when $\bar{\beta} = 0$

network travel time between scenario III and the other two scenarios is very small, while the ratio of parking delay and total travel time is around 0.47. But this difference will become larger under the conditions of (D2, S1, $\bar{\beta} = 0$) and (D2, S2, $\bar{\beta} = 0$), while the ratios of parking delay and total travel time are 0.53 and 0.45, respectively. These results imply that the parking information effectively reduces the network travel time, particularly under the condition of larger total demand for and smaller supply of the number of parking spaces. In other words, ATIS with parking information only should be most effective when the roads are not congested but the parking demand is approaching or exceeding the parking capacity. Similar conclusions can be extended to ATIS with route information only and both route and parking information. The route information could be effective in improving the network performance under the condition of travel demand greater than capacity of road links. The ATIS service with both parking and route information would be effective in improving the network performance under the condition of transport supply slightly greater than travel demand (that is, excess capacities on car parks and road links).

Figure 12.5 shows the effects of the ATIS service cost on the network performance for scenarios I and III, given that ATIS provides the equipped drivers with imperfect parking information only (that is, $\bar{\beta} = 0.1$). It can be seen that the quality (or accuracy) of the ATIS

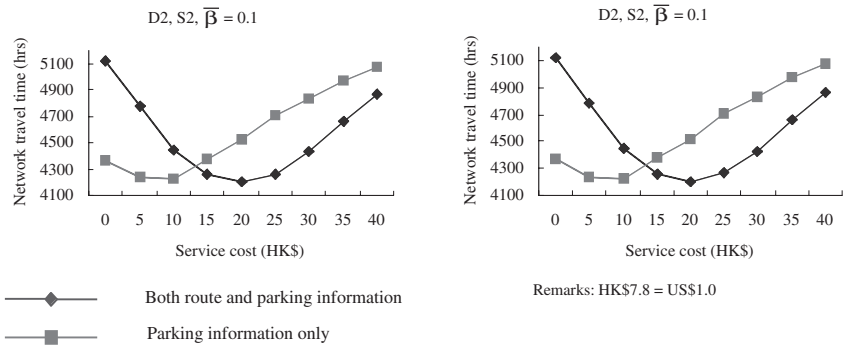


Figure 12.5 Effect of the ATIS service cost on the network travel time when $\bar{\beta}=0.1$

Table 12.3 Example results for scenarios I, II and III under different conditions

q_{rs} (pcu/hr)	(H_1,H_2)	$\bar{\beta}$	Scenario	λ^* (HK$)	b_{rs}^*	Network travel time (hrs)	Parking delay/total travel time
1000	(700, 900)	0	I	15	0.54	1804.26	0.4783
1000	(700, 900)	0	II	12	0.44	1831.47	0.4776
1000	(700, 900)	0	III	7	0.40	1901.56	0.4794
2000	(700, 900)	0	I	20	0.39	4255.51	0.5306
2000	(700, 900)	0	II	12	0.42	4253.74	0.5298
2000	(700, 900)	0	III	11	0.40	4168.50	0.5317
1000	(1000, 1200)	0	I	15	0.55	1784.28	0.4738
1000	(1000, 1200)	0	II	11	0.47	1805.40	0.4738
1000	(1000, 1200)	0	III	6	0.50	1801.28	0.4741
2000	(1000, 1200)	0	I	15	0.48	3821.21	0.4542
2000	(1000, 1200)	0	II	9	0.45	3729.27	0.4538
2000	(1000, 1200)	0	III	7	0.42	3858.15	0.4543
2000	(700, 900)	0.1	I	15	0.48	4573.52	0.5839
2000	(700, 900)	0.1	III	10	0.38	4492.31	0.5810
2000	(1000, 1200)	0.1	I	21	0.37	4203.78	0.5290
2000	(1000, 1200)	0.1	III	11	0.34	4226.56	0.5326

information would affect the performance of the road network. The more inaccurate the ATIS information, the less effective is the reduction in network travel time.

Table 12.3 shows the results of optimal ATIS service costs (λ^*), optimal market penetrations (b_{rs}^*), optimal network travel times, and the ratio of parking delay and total travel time for all three scenarios under various demand and supply conditions with a different quality of ATIS information. It should be noted that this set of results is preliminary, valid only for this simple example, and cannot be generalized to other networks. However, the model and solution algorithm proposed in this chapter surely provides a tool for investigating the effects of parking and route information (in terms of quality and service cost) on the network performance as well as the impacts on the market penetration of ATIS.

Table 12.4 Travel time savings for a different number of Monte-Carlo simulations

λ (HK$)	Number of Monte-Carlo simulations						Absolute difference				
	100	500	1000	2000	5000	10 000	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5
0	0.8492	0.8715	0.8724	0.8745	0.8743	0.8736	0.0223	0.0009	0.0021	0.0002	0.0007
5	0.9876	0.9822	0.9622	0.9619	0.9609	0.9606	0.0054	0.0202	0.0003	0.0010	0.0003
10	0.6838	0.6730	0.6896	0.6881	0.6871	0.6874	0.0108	0.0166	0.0015	0.0007	0.0003
15	0.5628	0.5644	0.5332	0.5352	0.5359	0.5348	0.0016	0.0312	0.0020	0.0007	0.0011
20	0.5235	0.4996	0.4981	0.4973	0.4960	0.4957	0.0239	0.0015	0.0008	0.0013	0.0003
25	0.2974	0.4648	0.4643	0.4618	0.4627	0.4620	0.1674	0.0005	0.0025	0.0009	0.0007
30	0.3481	0.4497	0.4233	0.4227	0.4233	0.4235	0.1016	0.0264	0.0010	0.0006	0.0002
35	0.2254	0.2213	0.2396	0.2378	0.2367	0.2362	0.0041	0.0183	0.0018	0.0011	0.0005

Furthermore, the traffic authority can determine an optimal ATIS service cost, and consequently an optimal market penetration, which could create a 'win–win' situation for both the individual drivers and the entire network.

Finally, the solution convergence of the Monte-Carlo simulation in the numerical example is examined. The travel time savings generated by ATIS at different costs of services for a different number of Monte-Carlo simulations are shown in Table 12.4, where $q_{rs} = 1000$, $H_1 = 700$, $H_2 = 900$, $\bar{\alpha} = 1.0$, and $\bar{\beta} = 0$. Δ_1, Δ_2, Δ_3, Δ_4, and Δ_5 are the absolute differences between the travel time saving for simulations 100 and 500, 500 and 1000, 1000 and 2000, 2000 and 5000, 5000 and 10 000, respectively. The maximum Δ_1, Δ_2, Δ_3, Δ_4, and Δ_5 are 0.1674, 0.0312, 0.0025, 0.0013 and 0.0011, respectively. The change of the travel time savings from 2000 to 5000 simulations is comparatively small and can be negligible. The maximum number of Monte-Carlo simulation is then set to 2000 for this numerical example.

9. CONCLUSIONS

In this chapter, the market penetration of ATIS is modeled in an elastic manner and is determined by the cost of ATIS, the quality of service (or accuracy of the information system) and the willingness of drivers to pay for the service. A bi-level programming model is proposed to determine the optimal network performance in a network with parking and/or route information systems while the drivers' parking and route choice behaviors are explicitly taken into account. The lower-level problem is a multi-class probit-based SUE model that describes the combined parking and route choice problem with endogenous market penetration of ATIS. The upper-level problem represents a network performance optimization problem, the objective of which is to minimize the total network travel time. In particular, we have investigated the traffic behaviors under three scenarios of the ATIS environment: (I) equipped drivers have both parking and route information; (II) equipped drivers have route information only; and (III) equipped drivers have parking information only.

A simple numerical example is used to illustrate the application of the model and solution algorithm. We examine the market penetration by varying the service cost of ATIS and the willingness of drivers to pay for the ATIS service with different levels of total travel demand and transport supply capacities (that is, road links and car parks). The willingness to pay is dependent on the sensitivity of drivers to the benefits of the service. Also we investigate the total network travel time by varying the ATIS service cost with different levels of demand and supply as well as the quality (or accuracy) of the parking information system. We can draw the following key conclusions by analysing the results of the numerical example.

1. The market penetration of ATIS always decreases with an increase in the service cost, while the quality of the ATIS service and the willingness of drivers to pay for it are given and fixed.
2. The market penetration for scenario III (that is, with parking information only) is lower than the other scenarios at the same service cost level.
3. Parking information can be a key contributor in reducing network travel time, particularly under the condition of larger total demand for but smaller supply of the number of parking spaces. In other words, ATIS with parking information only should be most effective when the roads are not congested but parking demand is approaching parking capacity.
4. ATIS with route information only is effective in improving the network performance, particularly under the condition of travel demand greater than road capacity.
5. ATIS with both parking and route information would be effective in improving the network performance under the condition of transport supply slightly greater than travel demand (that is, excess capacities on car parks and road links).
6. The more inaccurate the ATIS information, the less effective is the ATIS in reducing the network travel time.

The proposed model can be applied to a real network as a case study. Further work should be conducted to consider the effects of the parking and route information in a dynamic assignment framework.

NOTE

* The work described in this chapter was mainly supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region (Project No. PolyU 5040/01E).

REFERENCES

Abdulaal, M. and L.J. LeBlanc (1979), 'Continuous equilibrium network design models', *Transportation Research-B*, **13** (1), 19–32.

Allsop, R.E. (1974), 'Some possibilities for using traffic control to influence trip destinations and route choice', in D.J. Buckley (ed.), *Proceedings of the Sixth International Symposium on Transportation and Traffic Theory*, Amsterdam: Elsevier, pp. 345–74.

Arnott, R., A. De Palma and R. Lindsey (1991), 'Does providing information to drivers reduce traffic congestion', *Transportation Research-A*, **25** (5), 309–18.

Asakura, Y. (1997), 'Comparison of spatial location patterns of PGI message board: a microscopic network simulation model', in W.H.K. Lam (ed.), *Proceedings of the Second Conference of Hong Kong Society for Transportation Studies*, Hong Kong, Department of Civil and Structural Engineering, The Hong Kong Polytechnic University, pp. 9–14.

Ben-Akiva, M., A. De Palma and I. Kaysi (1991), 'Dynamic network models and driver information systems', *Transportation Research-A*, **25** (5), 251–66.

Bennett, L.D. (1993), 'The existence of equivalent mathematical programs for certain mixed equilibrium traffic assignment problems', *European Journal of Operational Research*, **71** (2), 177–87.

Bifulco, G.N. (1993), 'A stochastic user equilibrium assignment model for the evaluation of parking policies', *European Journal of Operational Research*, **71** (2), 269–87.

Bonsall, P., L. Pickup and A. Stathopoulos (1991), 'Measuring behavioral responses to road transport informatics', in Commission of the European Communities (ed.), *Advanced Telematics in Road Transport*, Amsterdam: Elsevier, pp. 1457–87.

Boyce, D.E. (1988), 'Route guidance systems for improving urban travel and location choices', *Transportation Research-A*, **22** (4), 275–81.

Cascetta, E., A. Nuzzolo, F. Russo and A. Vitetta (1996), 'A new route choice logit model overcoming IIA problems: specification and some calibration results for interurban networks', in J.B. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, Oxford, New York: Pergamon, pp. 691–711.

Chan, K.S. and W.H.K. Lam (2002), 'Optimal speed detector density for the network with travel time information', *Transportation Research-A*, **36** (3), 203–23.

Daganzo, C.F. (1983), 'Stochastic network equilibrium with multiple vehicle types and asymmetric, indefinite link cost Jacobians', *Transportation Science*, **17** (3), 282–300.

Daganzo, C.F. and Y. Sheffi (1977) 'On stochastic models of traffic assignment', *Transportation Science*, **11**, 253–74

Emmerink, R.H.M., K.W. Axhausen, P. Nijkamp and P. Rietveld (1995a), 'Effects of information in road transport networks with recurrent congestion', *Transportation Research-A*, **22** (1), 21–53.

Emmerink, R.H.M., K.W. Axhausen, P. Nijkamp and P. Rietveld (1995b), 'The potential of information provision in a simulated road transport network with non-recurrent congestion', *Transportation Research-C*, **3** (5), 293–309.

Emmerink, R.H.M., P. Nijkamp, P. Rietveld and K.W. Axhausen (1994), 'The economics of motorist information systems revisited', *Transport Review*, **14** (2), 363–88.

Fisk, C.S. (1984), 'Game theory and transportation systems modeling', *Transportation Research-B*, **18** (4), 301–13.

Friesz, T.L., A. Anandalingham, N.J. Mehta, K. Nam, S.J. Shah and R.L Tobin (1993), 'The multiobjective equilibrium network design problem revisited: a simulated annealing approach', *European Journal of Operational Research*, **65** (1), 44–57.

Halati, A. and D.E. Boyce (1991), 'Effectiveness of in-vehicle navigation systems in alleviating nonrecurring congestion'. *Proceedings of Vehicle Navigation and Information Systems Conference*, Vol. 2, *Dearborn, Michigan*, Warrendale, PA.: Society of Automotive Engineers, pp. 871–89.

Hall, R.W. (1996), 'Route choice and advanced traveler information systems on a capacitated and dynamic network', *Transportation Research-C*, **4** (5), 289–306.

Hong, K. and W.Y. Szeto (2002), 'A methodology for sustainable traveler information services', *Transportation Research-B*, **36** (2), 113–30.

Lam, W.H.K., M.L. Tam, H. Yang and S.C. Wong (1999), 'Balance of demand and supply of parking spaces', *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*, Jerusalem, Israel, 20–23 July, Amsterdam, New York: Pergamon pp. 707–31.

Lam, W.H.K. and Y. Zhang (2000), 'Capacity-constrained traffic assignment in networks with residual queues', *Journal of Transportation Engineering*, **126** (2), 121–8.

Lambe, T.A. (1996), 'Driver choice of parking in the city', *Socio-Economics Planning Sciences*, **30** (3), 207–19.

Lo, H.K., A. Chen and H. Yang (1999), 'System time minimization in route guidance with elastic market penetration', *Transportation Research Record*, **1667**, 25–32.

Luce, R.D. and P. Suppes (1965) 'Preference, utility and subjective probability', in Luce R.D., R.R.

Bush and E.H. Galanter (eds), *Handbook of Mathematical Psychology*, New York: John Wiley & Sons, pp. 250–410.
Mahmassani, H. and R. Jayakrishnan (1991), 'System performance and user response under real-time information in a congested traffic corridor', *Transportation Research-A*, **25** (2), 293–307.
Sheffi, Y. (1985), *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Englewood Cliffs, NJ: Prentice-Hall.
Suwansirikul, C., T.L. Friesz and R.L. Tobin (1987), 'Equilibrium decomposed optimization: a heuristic for the continuous equilibrium network design problem', *Transportation Science*, **21** (4), 254–63.
Tobin, R.L. and T.L. Friesz (1988), 'Sensitivity analysis for equilibrium network flows', *Transportation Science*, **22** (4), 242–50.
Van Der Goot, D. (1982), 'A model to describe the choice of parking places', *Transportation Research-A*, **16** (2), 109–15.
Van Vuren, T. and D. Watling (1991), 'A multiple user class assignment model for route guidance', *Transportation Research Record*, **1306**, 22–31.
Watling, D. (1994), 'Urban traffic network models and dynamic driver information systems', *Transport Review*, **14** (1), 219–46.
Yang, H. (1995), 'Sensitivity analysis for queuing equilibrium network flow and its application to traffic control', *Mathematical and Computer Modeling*, **22** (4), 247–58.
Yang, H. (1998), 'Multiple equilibrium behaviors and advanced traveler information systems with endogenous market penetration', *Transportation Research-B*, **32** (3), 205–18.
Yang, H. and M.G.H. Bell (1997), 'Traffic restraint, road pricing and network equilibrium', *Transportation Research-B*, **31** (4), 303–14.
Yang, H., T. Sasaki, Y. Iida and Y. Asakura (1992), 'Estimation of origin–destination matrices from link traffic counts on congested networks', *Transportation Research-B*, **26** (6), 417–34.
Yang, H. and S. Yagar (1994), 'Traffic assignment and traffic control in general freeway-arterial corridor systems', *Transportation Research-B*, **28** (6), 463–86.
Yin, Y. and H. Yang (2003), 'Simultaneous determination of the equilibrium market penetration and compliance rate of advanced traveler information systems', *Transportation Research-B*, **37** (2), 165–81.

13. Real-time spatiotemporal data mining for short-term traffic forecasting

Hongyu Sun, Heng Xiao and Bin Ran

1. INTRODUCTION

As defined by Cabena and IBM (1998), data mining involves extracting previously unknown but valid information from large databases and then using it to arrive at important business decisions. The term 'predictive data mining' means a search for clear-cut patterns present in data that allows a generalization to be accurately made about future decisions (Weiss and Indurkhya 1998). Among the goals of predictive data mining is identification of time-ordered patterns that are predictive of events of special interest. Such data patterns appear in various formats such as spatiotemporal patterns that are found in traffic data.

There is a large literature about short-term forecasting algorithms (Van Arem et al. 1997). Typical methods used by transportation researchers include the following: time series models (Ahmed and Cook 1979), simulation models (Rathi and Santiago, 1989), Kalman filtering theory (Okutani and Stephanedes 1984), nonparametric methods (Davis and Nihan 1991; Smith et al. 1999), dynamic traffic assignment models (Ran 2000) and neural network models (Park and Rilett 1998).

Intuitively, traffic data of the neighboring links (upstream and downstream) are correlated with the traffic on the link of interest. Incorporation of such data into the prediction model inputs could lead to better prediction. Rilett et al. (1999) included such inputs in their spectral basis neural network models for freeway path travel time prediction. Ishak (2003) studied the methodology for the identification of a viable set of measures that is capable of extracting important features from spatiotemporal traffic contour maps in a manner analogous to extracting features from digital images in the field of image analysis and pattern recognition. However, the feature extraction process in his study is for revealing texture characteristics of spatiotemporal contour maps used in performance monitoring systems and assessment of level of service, rather than for pattern recognition application or prediction.

From the traffic flow theory, the inversed triangle is often found in the time space domain traffic contour map (May 1990). These can happen very often at fixed bottlenecks (Zhang 2001). Figure 13.2 (below) illustrates the inversed triangle shape displayed from the case study data. When demand exceeds the bottleneck capacity, queues form and spread upstream. This accounts for the left leg of the triangle. When demand subsides, queuing starts to dissipate. This accounts for the right leg of the triangle.

There is a need for the formulation of both accurate and real-time models that approximate the nonlinear and time-variant prediction functions online. There is a fundamental

difference between the parametric and the nonparametric approaches. This is that the nonparametric approach allows the data to speak for itself and compute a whole function. The parametric approach, however, assumes that the function takes a form represented by a finite number of parameters. Rather than predefining the prediction model subjectively, nonparametric methods are data-driven. Nonparametric models based upon pattern recognition also allow easy incorporation of exogenous variables such as data on neighbor links to identify the spatiotemporal pattern.

Although neural network models belong to nonparametric methods, they are not real-time models. Studies of smoothing and time series have resulted in the nonparametric regression approach to the problem of prediction (Györfi 1989). The nonparametric regression models include local models. A local model such as the k nearest neighbor model is the opposite of a global model such as a neural network. In other words, simple local models can be used as building blocks to approximate the complex global function. This is done by means of weighted regression analysis in which the weights represent the kernel functions. Local models allow for both real-time modeling and incorporation of exogenous variables into covariates (inputs). Thus the spatiotemporal prediction is made by slightly modifying the prediction using only traffic data on the link of interest. By real-time modeling, data pre-classification is avoided; this is also the case in offline training. Such training is required by neural networks and time-series models.

Very little work has been done on incorporating neighboring links data into nonparametric local regression traffic prediction models. Among the nonparametric local regression traffic prediction models, is our recent study (Sun et al. 2003) showing that local linear regression is a better model than local constant models. This chapter adds spatiotemporal data mining features to the local linear prediction model.

The chapter is organized as follows. In Section 2, the problem of definition of the short-term traffic forecasting and its local linear model formulation with spatiotemporal data is given. Section 3 gives the results of a case study of the spatiotemporal traffic prediction using a local linear model, a k-NN and a kernel regression model for comparison with the ones without neighboring link data inputs. Only one-step prediction is used in that comparison. In addition, the results comparing the local linear model with the benchmark models under multiple-step prediction are also given. Section 4 includes the conclusion and comments about forthcoming work.

2. METHODOLOGY

2.1 Problem Statement

The traffic prediction problem can be described in the following way: given the observed traffic data, $\mathrm{TT}(i)$, $i=1, \ldots t$, the prediction is to generate an estimate of $\mathrm{TT}(t+s)$, where s is the prediction horizon. In the corresponding regression model, the given inputs are also called covariates and the output variable is called the response.

The covariate vector $\mathbf{x}$ in this chapter is a set of $[\mathrm{TT}^{l}(t-d+1), \ldots, \mathrm{TT}^{l}(t)]^{T}$. Here ^{T} denotes the matrix transpose, l is the link number, $l=l_1, l_2, \ldots, l_k$. The response variable is $y=\mathrm{TT}^{l_i}(t+s)$, where l_i is the link of interest.

Next we give a description of the local linear model and two simple models: random

walk and historical profile. The *k*-NN and the kernel regression models that are used in the case study are not shown. They can be found in the earlier work (Sun et al. 2003).

2.2 Local Linear Model

Here we give a brief review of the local linear regression model (Fan et al. 1996) used to approximate the relationship of the future traffic with the past and current traffic data. Given multivariate covariate $\mathbf{X}$ and a univariate response Y, it is of interest to estimate the mean regression function, that is, the prediction of traffic variable, $m(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x})$, where $\mathbf{x}^T = (x_1, ..., x_d)$ is a point in R^d. Given the observations $\{(\mathbf{X}_i^T, Y_i): i = 1, \ldots, n\}$ and $\mathbf{X}_i = (\mathrm{X}_{i1}, \ldots, \mathrm{X}_{id})^T$, the estimator of $\beta = (\beta_0, \ldots, \beta_d)^T$ to minimize

$$\sum_{i=1}^{n} \left[Y_i - \beta_0 - \sum_{j=1}^{d} \beta_j (X_{ij} - x_j) \right]^2 K_B(\mathbf{X}_i - \mathbf{x})$$

is:

$$\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_d)^T = (\mathbf{X}_d^T \mathbf{W} \mathbf{X}_d)^{-1} \mathbf{X}_d^T \mathbf{W} \mathbf{y}, \tag{13.1}$$

where:

$$\mathbf{X}_d = \begin{pmatrix} 1 & X_{11} - x_1 & \ldots & X_{1d} - x_d \\ 1 & X_{21} - x_1 & \ldots & X_{2d} - x_d \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} - x_1 & \ldots & X_{nd} - x_d \end{pmatrix},$$

and

$$\mathbf{W} = \mathrm{diag}\,[K_B(\mathbf{X}_i - \mathbf{x})], \; \mathbf{y} = (Y_1, \ldots, Y_n)^T.$$

Here $K_B(\mathbf{u}) = [1/|B|]K(B^{-1}\mathbf{u})$, where K is a multivariate probability density function with mean zero and the covariance matrix of $\mu_2(K)\mathbf{I}_d$, with $\mathbf{I}_d$ the $d \times d$ identity matrix. B is called the bandwidth matrix and $|B|$ denotes its determinant. The weighting kernel K is chosen as the Gaussian function and $B = h\mathbf{I}_d$.

Thus,

$$\hat{m}(\mathbf{x}) = \hat{\beta}_0,$$

$$\left(\frac{\partial \hat{m}}{\partial x_j} \right)(\mathbf{x}) = \hat{\beta}_j, j = 1, \ldots, d.$$

and

$$\hat{j}_y = \hat{\beta}_0$$

is the prediction value (Atkeson et al. 1997).

We shall not repeat the procedures of selecting dimension d of the covariate vector and the bandwidth h. The details of parameter choice can be found in the work by Sun et al. (2003).

2.3 Two Simple Prediction Models

There are two simple prediction models that are often used as benchmarks. Neither of the models can incorporate exogenous data into inputs. But comparative studies among the local linear model and these two models will be shown to obtain a clear-cut view. Below is a description of the two models.

Random walk
The random walk model uses the current data as the prediction for the next time interval. For the multiple step prediction ($s>1$), we can treat the time series as a random walk model with intervals of s multiples of the original interval duration. Thus the estimation $\hat{y}$ for $\mathrm{TT}^{li}(t+s)$ for all s ($s\geqslant 1$), is given by $\hat{y}=\mathrm{TT}^{li}(t)$.

Note that the model cannot work when there is missing data. Depending on how missing data is processed, the results for random walk may be given partial favor. That is, if the missing data at time t is replaced by its preceeding data at time $(t-s)$ which is exactly the same policy as the s-step random walk algorithm, the algorithm's results will gain unfair advantage.

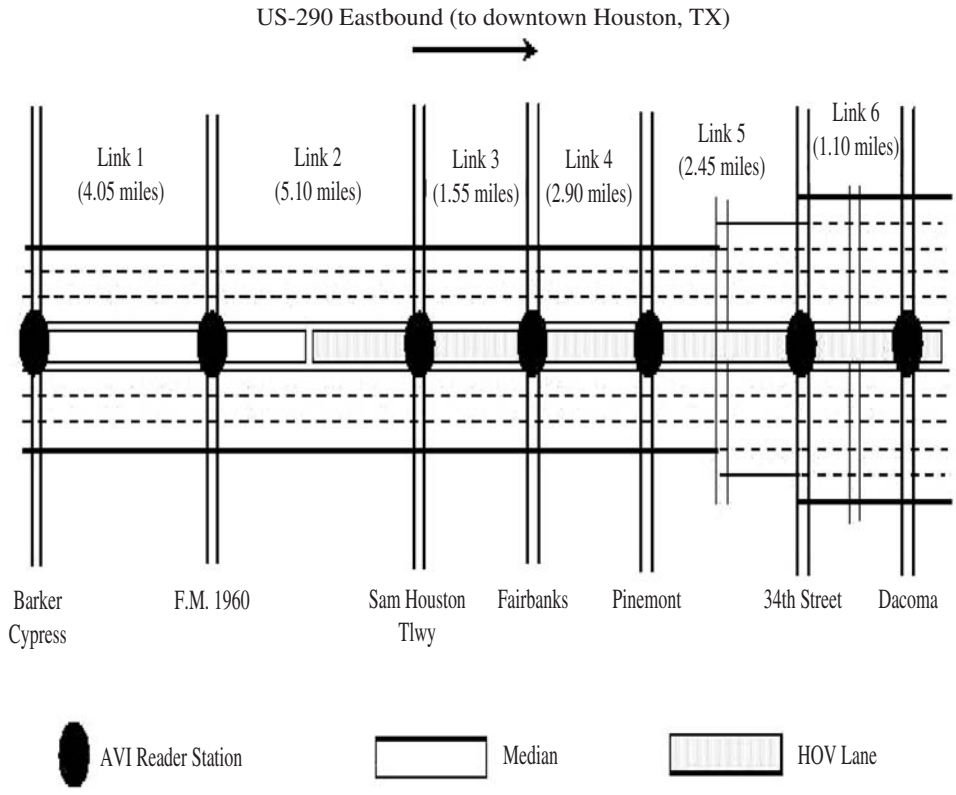
Historical profile
The historical profile model predicts the future traffic based on the time of day. It uses the calculated mean at that time as the prediction. Thus the estimation $\hat{y}$ for $\mathrm{TT}^{li}(t+s)$ for all s ($s\geqslant 1$), is given by $\hat{y}=\overline{TT^{li}(t+s)}$, where the $\bar{}$ denotes the arithmetic average operation.

3. CASE STUDY

This study is based on Houston US-290 Northwest freeway eastbound traffic speed data collected from February 2002 to July 2002 every five minutes. Figure 13.1 shows the map for the study road. The same freeway as in Park's research (Park and Rilett 1998) is chosen for results comparison in the future. The data are retrieved by software from the online real time information provided to the public by the Houston TranStar Automatic Vehicle Identification (AVI) traffic monitoring system (http://traffic.tamu.edu).

The selected road segment to predict is US-290 from the cross-street Sam Houston toll way to the cross-street Fairbanks, since this segment is the middle of the six segments and has most usable data. The length of this segment is 1.55 miles, the usual travel speed is about 68 miles per hour and the estimated free-flow travel time is about 1 minute 22 seconds. To adopt spatiotemporal prediction, one upstream and one downstream road segment are chosen. The upstream road segment is US-290 from the cross-street of F.M. 1960 to the cross-street Sam Houston toll way, with a length of 5.10 miles and an estimated free-flow travel time of about 4 minutes 30 seconds. The downstream road segment is US-290 from the cross-street Pinemont to the cross-street 34th street, with a length of 2.45 miles and an estimated free-flow travel time of about 2 minutes 6 seconds. Note that the selected downstream road segment is not next to the link to predict, the distance between their starting points is 3.45 miles or nearly 4 minutes free-flow travel time.

Since the estimated travel times between upstream and downstream links are based on free-flow status, the data aggregation interval of 5 minutes is adequate for our study of



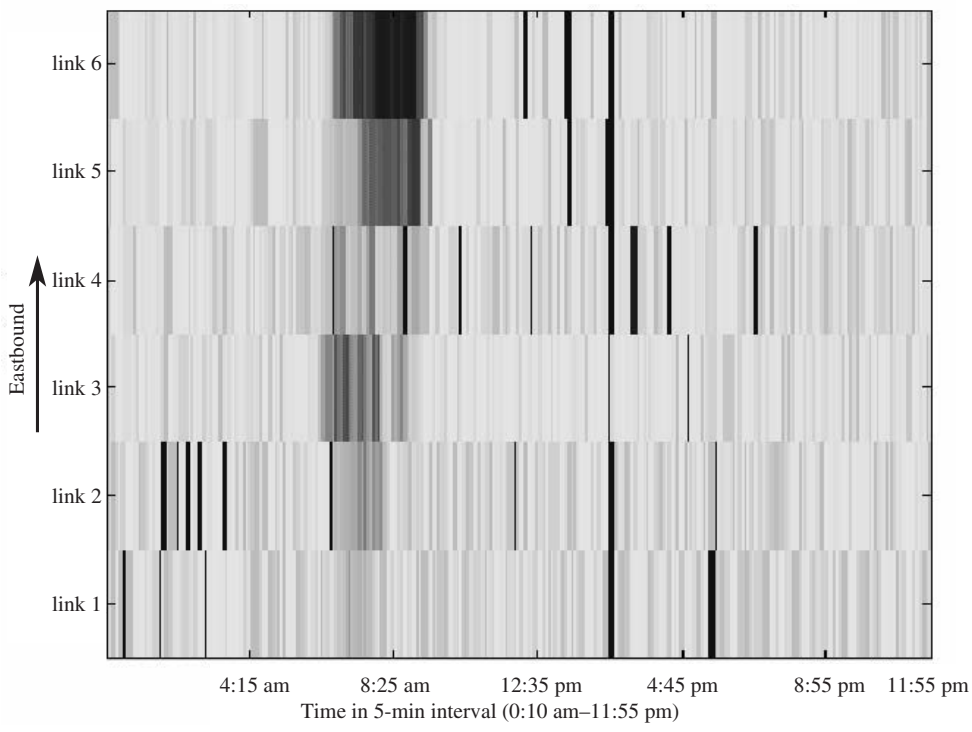
Source: Houston TranStar Automatic Vehicle Identification, http://traffic.tamu.edu.

Figure 13.1 Map of US-290 Eastbound, Houston, TX to illustrate the link to predict traffic (link 3) and the neighboring links used (links 2 and 5)

spatial effect. This is because the performance for free-flow traffic prediction is usually good and not of interest. We are concerned with prediction within peak hours or the transition phases, when the traffic is usually found with reduced speed. The traffic propagation time is greater than the free-flow traffic time (4 minutes 30 seconds) during those periods of interest. Therefore a 5-minute interval is chosen.

The space time diagram for raw speed data (without data cleaning) on the six links of US-290 Eastbound on 24 April 2002 for Houston, TX is shown in Figure 13.2. The dark areas in this figure denote missing data which are coded as –1. We can clearly find the well-known inverted triangular shape (May 1990) which corresponds to the onset, propagation and dissipation of traffic congestion.

One-step prediction has been studied to compare the model for the predictor using only single link data and the spatiotemporal model. The relative mean error (RME) is used as the performance index to compare the two models: the local linear model with and without neighboring link inputs.



Note: Raw data: darker gray scale corresponds to lower speed.

Figure 13.2 Space time speed field diagram for 6 links of US-290 Eastbound on 24 April 2002, Houston, TX

$$\mathrm{RME} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|,$$

where Y_i is the observation value, and $\hat{Y}_i$ is the predicted value.

After deleting data for holidays and weekends and screening out data with too many missing values for each of the three links, 14-day data were left. This is less than the data used in Sun et al. (2003) which only use a single-link data. For each day, the first two points are deleted because some days missed those two points. The other missing data are replaced by the nearest recent time data in that day. Thus each day has 286 data points. Data for 14 days are used as the training set and the other two days as the testing set. Since a 16-day database is not large, we leave out different combinations of two days to form the training set and the two days left to form the testing set. Thus we ran the 240 combinations. The error performance over all results from all runs is averaged to evaluate each method. The ridging parameter is chosen as 0.1 and d equals two. The data on one upstream and one downstream link are incorporated by simply adding two elements in the covariate vector without modifying other parts. The covariates dimension tripled to six.

Table 13.1 Comparison of the overall average relative error for predictors using only single-link data and the spatiotemporal predictors (prediction step = 1)

RME (%) \ Models	*k*-NN ($k=3$)	Kernel regression	Local linear model
		With ridging	
Using only single-link data	10.27	8.91	8.46
Spatiotemporal predictor	10.3	8.61	8.95

Table 13.2 Comparison of the overall average relative error for the local linear predictor, the historical profile predictor and the random walk predictor

Prediction horizon \ Models	RME (%)		
	Local linear	Historical profile	Random walk
1	8.46	13.8036	7.99
2	9.05	13.8283	9.92
3	9.82	13.8569	11.03
4	10.77	13.8881	12.07
5	11.38	13.9158	13.85

Table 13. 1 summarizes the results of the average RME over 240 runs using 14 training days for models with and without incorporating neighboring links data. The models in the table are the *k*-NN, kernel regression and local linear models. It is found that the spatiotemporal prediction has a comparable performance with the prediction model using only single-link traffic, but does not show much advantage over the latter.

Table 13.2 shows the results of the average RME over 240 runs using 14 training days for the local linear model and the two simple models with prediction horizons from one to five time intervals. The historical profile model shows about 14 per cent RME almost constantly for all prediction horizons since it does not consider any current data dynamics. The random walk model gives worse results than the local linear model except when the prediction horizon equals one. However, we should recall that the missing data preprocessing policy adopted in this case study replaces missing data with the recent data. This is exactly according to the random walk assumption. Therefore, such artificial data may make the advantage of the random walk model under this case unsound.

4. CONCLUSIONS AND FUTURE WORK

This study has tried to explore the performance of incorporating neighboring links traffic data into the local linear prediction model. However, the results for the case study data did not show much advantage. There are possible reasons: first, the chosen neighboring links may not be quite correlated with the link of interest in view of the flow propagation; second, the database is not large enough to store enough spatiotemporal patterns.

Further improvements could be achieved by using more data. Other neighboring links will be tested. We shall study the performance under other aggregation time intervals. Missing data were preprocessed in the present chapter, thus future work includes coping with the missing data issue online. The methods such as adaptively adding the upstream and downstream links data into the covariates when necessary (in the triangle area) and removing them when not needed (out of the triangle area) will be investigated. Future work will also include studying the peak- and non-peak-hour data separately.

REFERENCES

Ahmed, M. and A. Cook (1979), 'Analysis of freeway traffic time series data by using Box–Jenkins techniques', *Transportation Research Board*, No. 722, 1–9.

Atkeson, C., A. Moore and S. Schaal (1997), 'Locally weighted learning', *Artificial Intelligence Review*, **11** (1–5), 11–73.

Cabena, P. and International Business Machines Corporation (IBM) (1998), *Discovering Data Mining: From Concept to Implementation*, Upper Saddle River, NJ: Prentice-Hall.

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and its Applications*, London: Chapman & Hall.

Györfi, L. (1989), *Nonparametric Curve Estimation from Time Series*, New York: Springer-Verlag.

Ishak, S. (2003), 'Deriving traffic performance measures and levels of service from second-order statistical features of spatiotemporal traffic contour maps', Transportation Research Board Annual Meeting, Washington, DC, Preprints CDROM 2003.

May, A. (1990), Traffic Flow Fundamentals, Englewood Cliffs, N.J. Prentice Hall.

Okutani, I. and J. Stephanedes (1984), 'Dynamic prediction of traffic volume through Kalman filter theory', *Transportation Research*, **18B**, 1–11.

Park, D. and L. Rilett (1998), 'Forecasting multiple-period freeway link travel times using modular neural networks', Transportation Research Board Annual Meeting, Washington DC.

Ran, B. (2000), 'Using traffic prediction models for providing predictive traveller information', *International Journal of Technology Management*, 20 (3/4), pp. 326–39.

Rathi, A. and A. Santiago (1989), 'The new NETSIM: TRAF-NETSIM 2.00 Simulation Program', Transportation Research Board 68th Annual Meeting, Washington, DC.

Rilett, L., D. Park and G. Han (1999), 'Spectral basis neural networks for real-time travel time forecasting', *Journal of Transportation Engineering*, **125** (6), 515–23.

Smith, B., B. Williams and K. Oswald (1999), 'Parametric and nonparametric traffic volume forecasting', Transportation Research Board Annual Meeting, Washington DC., Preprints CDROM.

Sun, H., X. Liu, H. Xiao, R. He and B. Ran (2003), 'Short-term traffic forecasting using the local linear regression model', Transportation Research Board Annual Meeting,Washington DC., Preprints CDROM.

Van Arem, B., H. Kirby, M. Van Der Vlist and J. Whittaker (1997), 'Recent advances and applications in the field of short-term traffic forecasting', *International Journal of Forecasting*, **13**, 1–12.

Weiss, S. and N. Indurkhya (1998), *Predictive Data Mining: A Practical Guide*, San Francisco: Morgan Kaufman.

Zhang, H.M. (2001), Guest Editor, 'Special Issue on Traffic Flow Theory', Vol. 1&2, *Journal of Networks and Spatial Economics*, Kluwer Academic.

14. On-line traffic assignment and network loading

Pitu Mirchandani, Rohit Syal, David Lucas and Yang He*

1. INTRODUCTION

Consider the following scenario, which is very applicable, for example, to cities like Tucson, Arizona, that have several sections within a 'work zone' to upgrade network infrastructure. Work zones invariably lead to changes in traffic patterns providing a challenge to traffic management and planning departments to minimize the disruption caused to motorists. This is especially so in the immediate neighborhood of the work zone, although it is conceivable that a work zone at a critical location can impact a large segment of the network. In particular, one needs to estimate the re-routing and resulting traffic loading for this purpose. Similar concerns are also raised by major traffic incidents whose impacts last for several days, local flooding and, in general, when a portion of network infrastructure is damaged to the point where some motorists avoid it. In order to forecast the network load, several issues arise: (i) Which routes will carry the displaced loads? (ii) Will there be some sort of *traffic equilibrium* among the routes? (iii) If so, how long will it take to reach this equilibrium? (iv) What is the role of traffic measurements, such as detector data and travel times from probe vehicles? These are some of the issues addressed in this chapter.

2. EQUILIBRIUM MODELING

To estimate current and future use of traffic networks, planners use what is referred to as *static traffic assignment (STA)*, where, for given origin–destination predicted flow volumes, an *equilibrium* concept is utilized to assign routes and load flows on these routes. The most accepted equilibrium condition is the so-called Wardrop's First Principle (1952), or the *user equilibrium* where it is assumed that minimizing travel time is the only attribute of concern and any traveler cannot unilaterally decrease his/her travel time at equilibrium by choosing an alternative route.

There are two categories of issues that arise in the use of an STA model: behavioral modeling and analytical modeling issues. With respect to the former category, we note that it is difficult to develop a mathematical model that replicates all the human decision processes in route choice and encompasses various factors about the trip and possible choices. The major behavioral approximations made in STA are that (B1) each traveler knows the state of the network at all times, (B2) the state does not change during the trip,

(B3) the traveler computes the travel times on all possible routes and (B4) the traveler chooses one of the routes that gives the shortest travel times. Some of these assumptions have been relaxed in models that have appeared in the literature, such as each traveler has some measurement errors in his/her knowledge of the state of the system and chooses a route that is perceived as being the shortest (Daganzo and Sheffi 1977), or each traveler perceives the network as having random travel times with known distributions and chooses the route that minimizes the expected disutility of travel time where the traveler's disutility function is given (Mirchandani and Soroush 1987).

Major analytical modeling approximations include (A1) the traffic network can be modeled as a directed network, where (A2) the travel time on each directed link is a known (or calibrated from data) function of only the traffic volume on the link, (A3) the travel time on a route is simply the summation of the travel times of the links on the route and (A4) the origin–destination demand is distributed uniformly over time. Approximation A4 is a physical rationalization for approximation B2, which in turn translates to the network being 'static' in the sense that instantaneous link travel times are always the same and therefore the travel time on the route is the sum of these link travel times, leading to approximation A3.

A major weakness in the use of STA models is approximation A2. In such models, *volume–delay* or *BPR (Bureau of Public Roads)* functions, also referred to as *impedance functions*, have been developed to model congestion dynamics. They represent link travel times as nonlinear, monotonically increasing functions of traffic volume on the link. These functions provide nice convex optimization formulations and convergence properties to the equilibrium assignment models and have found favor universally with modelers.

Notwithstanding their popularity, volume–delay functions have certain modeling limitations. These functions take a simplistic view of congestion; they do not capture queues or incorporate network or traffic characteristics such as lane changing and gap acceptance behavior, intersection control strategies, start-up loss times and vehicle headways, pedestrian traffic, different driver types and modes of transport and so on. Further, such a model makes a simplifying assumption that the volume–delay function of a link is independent of other links, which is not necessarily true, especially for urban networks.

The other major weakness in the use of STA is the implication of behavioral approximations B1, B3 and B4 that result in the network always being in user equilibrium. First, there are always some capacity changes in the network, so that travelers would find difficulty in adjusting daily to an equilibrium. In other words, the state of the traffic at any point in time is a transient condition that is tending to an equilibrium. Moreover, with the availability of traffic information systems and advisories, based on, for example, work-zone activities and temporary loss of traffic capacity due to a major incident, travelers could decide to switch routes during the prevailing traffic situation, or even not travel. Consequently, the traffic forecasts obtained from the assignment models may not be as realistic as anticipated and could struggle to match the observed data.

This chapter, in fact, attempts to present an approach to respond to the above two weaknesses – the assumption that a calibrated volume-delay function is available and that the network is always in an equilibrium – so that a STA model may be used for short-term traffic planning of a locally impacted area due to, for example, work-zone activities. Although not precisely defined here, one may define the subnetwork that is impacted by

such activities based on traffic observations and engineering judgment. In the following we shall refer to the locally impacted area as an *impacted neighborhood network* (INN).

The approach presented also responds to the limitation of approximation A4, which has already been addressed by researchers in the context of wide-area traffic prediction and management using advanced traveler information systems (ATIS) like *Dynamic Variable Message Signs* and *In-Vehicle Route Guidance*. Planners consider 'peak demands' in order to design the network. Hence, to use STA, approximation A4 translates to a uniform peak period during which there is a constant demand per unit time, and the static equilibrium, or more appropriately *steady-state* conditions, are based on this uniform demand. Since the 1970s, researchers (for example, Merchant and Nemhauser 1978) have realized that during a traveler's trip the state of the network changes and the travel time is not simply the sum of instantaneous link travel times, but rather the sum of link travel times that are dynamically changing. To model the route travel times and the speed profiles during the trip actually experienced by the traveler, researchers have developed the so-called *dynamic traffic assignment* (DTA) approaches; based on both analytical optimization (for example, Janson 1991; Ran and Boyce 1996) and simulation (for example, Mahmassani et al. 1993; Jayakrishna et al. 1994; and Ben-Akiva et al. 1997a,b). In this scenario, if all drivers optimize their own travel times, one comes up with the *dynamic user equilibrium,* the counterpart to the static user equilibrium, where, as before, a traveler cannot unilaterally decrease his/her travel time by switching to another route. The richness of DTA models allows one to include, besides route choice, (i) departure time choice, (ii) en-route decision making in cases where en-route traffic advisory and/or route guidance is available, and, in simulation-based models, (iii) a large variety of traveler behavior assumptions.

Most of the available DTA models, especially the analytical optimization ones, still need calibrated volume–delay functions and assume that the network operates at a deterministic equilibrium. Versions of some prototype simulation DTA models claim that they do not need explicit volume–delay functions, but these are not commercially available.

3. RESEARCH OBJECTIVES AND APPROACH

The primary goals of this chapter are (i) to propose a model that captures the dynamics of route-choice equilibration, (ii) to provide an analytical traffic assignment approach to manage an INN when count detectors and traffic probes are available, which considers implicitly the actual dynamics of traffic and congestion rather than the use of explicit BPR functions, and (iii) to demonstrate the feasibility and applicability of the approach. The idea is to develop a traffic assignment process to manage an INN which is more responsive to behavioral aspects and inherent randomness of the transportation system in the presence of new interventions, such as work zones. Moreover, the rapid deployment of ATIS technologies has resulted in a shift in the way individuals choose their routes in such situations.

Ideally, the process should be demonstrated in the field with actual detector data and travel times to calibrate the model and compare its results; but that would be expensive and probably not appropriate at this stage of model development. Thus, we demonstrate the proposed approach using a microsimulation model. Although any microsimulator

that assigns and loads vehicles on specified routes can be used, we chose to use the widely-used CORSIM simulation package that was appropriately modified to load vehicles on user-specified routes (we refer to this as Route-based CORSIM or PF-CORSIM). Of the currently available simulators in the market, AIMSUN2, PARAMICS and VISSIM claim to also allow route-based loading (see, for example, Oh et al. 1999).

To capture the dynamics of route-choice equilibration we propose a simple feedback model that can be depicted as shown in Figure 14.1. We shall assume that the INN is con-

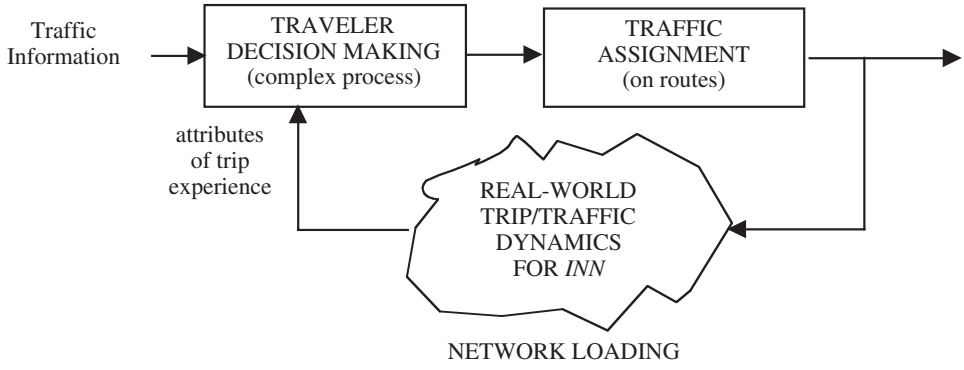


Figure 14.1 Schematic for on-line traffic assignment and loading

stantly monitored with count detectors and vehicle probes that provide trip times. This information is available to travelers who use the INN. After each period, for example a day, each traveler experiences a trip time and chooses another route if he/she has current knowledge of some trip times. We refer to the modeling of the route choices of traveler population as *traffic assignment.* The travelers' trips are then *loaded* on the network where, because of transportation supply–demand congestion effects, each traveler experiences a trip time. This is repeated in the next period and the process continues. Note that rather than assume a volume–delay function, the traveler makes his/her decision based on the actual experience on the previous trip and, hence, can account for whatever factors that influence his/her decision such as travel times (as in most traffic assignment models), travel distances, aversion to left turns, avoidance of traffic signals, lane blockages due to buses, pedestrian traffic, and so on.

This dynamic system model actually represents several nested dynamic processes, predominately the vehicle movements on network trips that capture traffic and queuing delays, and day-to-day discrete changes in trips that capture dynamics in route-choice behavior.

To test this approach, we replaced the 'real-world' block with a route-based simulator, in our case the *Route-based CORSIM*. We replaced the 'traveler decision making' with a 'route-choice model that seeks to minimize experienced travel times', which is more in line with the equilibrium models that are commonly used. This was done only to compare this approach with such equilibrium models, but we emphasize again that 'traveler decision making' could be based on other attributes besides travel times. In the evaluation of our scheme, we examined how close the travel times were with one other over time, to reflect

the process of equilibration and level of equilibrium in the network. If, in fact, travelers did choose routes to minimize their travel times, then it is expected that most of the experienced travel times for any given origin–destination pair would tend to a steady state where they are approximately equal.

4. BRIEF REVIEW OF VOLUME–DELAY FUNCTIONS

Although the proposed approach does not utilize volume–delay functions, we briefly review some functions used by traffic planners, only to make the reader aware of their basis and limitations in modeling traffic flow. Many types of volume–delay functions have been proposed and are used in practice (Branston 1976). By far the most widely used are the BPR functions, where the travel time has the form:

$$t^{BPR}(v) = t_0 \cdot [1 + (v/c)^{\alpha}]$$

where t_0 is the free-flow travel time, v is traffic volume, c is the link capacity and α is a calibration parameter. With higher values of α, the onset of congestion effects becomes more and more sudden. The simplicity of these BPR functions is certainly one reason for their widespread use. Unfortunately, these functions also have some inherent drawbacks, especially when used with high values of α. These drawbacks were pointed out by Spiess (1990). According to Spiess (pp. 153–8):

> While for any realistic set of travel volumes, we can assume that $v/c \leq 1$ (or at least not much larger than 1) this is usually not the case during the first few iterations of an equilibrium assignment. Values of v/c may well reach values of 3, 5 or even more. A BPR function with $\alpha = 12$ and a v/c ratio of 3 has congested link travel-time equal to $1 + (3)^{12}$ or 531 442 times the free-flow time, which means that every minute of free flow time becomes roughly equal to one year of congested time! These aberrations slow down convergence by giving undue weight (cost) to links with high α value.

For highly underutilized link conditions, the BPR functions, especially with high α values, yield free-flow times independent of actual traffic volume. Therefore, the equilibrium model will locally degenerate to an all-or-nothing assignment, where the slightest change (or error) in free-flow time may result in a complete shift of volume from one path to another.

A different volume–delay function was presented by Davidson (1966) as a general purpose travel time formula for transport planning purposes:

$$t = t_0 \, [1 + J_D \, x/(1 - x)],$$

where:

t = average travel time per unit distance (for example, in seconds per km),
t_0 = minimum (zero-flow) travel time per unit distance (for example, in seconds per km),
J_D = delay parameter,

x = q/Q = degree of saturation,
q = demand (arrival) flow rate in vehicle/hour,
Q = capacity (in vehicle/hour).

A critical analysis of this volume delay function was presented by Akçelik (1991). Akçelik claims that:

> Davidson derived this function from concepts of queuing theory. He modified the well-known steady-state delay equation, which, for a single channel queuing system with Poisson arrivals and exponentially distributed service rates, is
>
> $$d = (1/Q) + x/[Q\,(1-x)]$$
>
> where the first term is the service time (reciprocal of mean service rate) and the second term is the queuing delay. Davidson used saturation flow rather than capacity (Q) as mean service rate in his equation. These two parameters have the same value for uninterrupted traffic facilities (e.g., freeways), but capacity rather than saturation flow needs to be used for interrupted facilities (e.g., traffic signals where capacity equals saturation flow multiplied by the ratio of green time to cycle time).

For a detailed discussion on this issue, see Tisato (1991).

5. THE METHOD OF SUCCESSIVE AVERAGES AND ITERATIVE NETWORK LOADING

To capture day-to-day route changing and equilibration process for the dynamic model proposed, we need to develop or use a procedure where traffic loading may change based on last period (iteration) observations and travelers' route-choice objectives. Further, the convergence of the equilibration process needs to be controllable by the modeler since it is conceivable that in some locations or scenarios this process converges quickly (for example, availability of ATIS and high percentage of aggressive commuters) and in other locations/scenarios this converges slowly (for example, in retirement communities with little congestion). For this purpose, we chose to use the *method of successive averages* (MSA) both because this can be used without resorting to volume–delay functions and can be modified so that convergence can be controlled to range from 'very fast' to 'very slow'.

MSA has been applied to a wide variety of traffic assignment problems that arise in transportation analysis and planning. Basically, in MSA, at each iteration (that is, each day in our model) new flows are generated which are averaged with flows from earlier iterations (earlier days) to come up with a new solution at that iteration. In some applications, it can be proved to converge to a solution, whereas in others it is used as a heuristic that usually gives good results in practice. There are many cases where the MSA displays poor convergence properties. Although it begins promisingly, the initial iterations are followed by a pronounced 'tail' effect, resulting in slow overall convergence.

Bottom and Chabini (2001) extensively researched various attempts at improving convergence properties of fixed-point methods like the MSA. They list several interesting

methods employed over the years to develop fixed-point models with better convergence rates. Reporting one such effort by Cascetta and Postorino (2001) they state that:

> Cascetta and Postorino observed that in the MSA, an iteration's estimate is affected by the results from all the previous iterations, including those from early iterations that are presumably far from the solution. Furthermore, later iterations, which are presumably closer to the solution, receive smaller weights when computing a new estimate. Accordingly, Cascetta and Postorino propose a heuristic method that from time to time restarts the MSA. (i.e., resets the iteration counter to 1) using the last computed value as the new initial point. By restarting, the direct influence of earlier iterations is eliminated, and larger step sizes are applied to subsequent iterations. The frequency with which these restarts are carried out decrease as the number of iterations increase, via a user specified 'refreshing modulus' η. The first restart is done after η iterations, the second after 2 η iterations following the first restart, and so on. (Bottom and Chabini 2001)

In assignment algorithms like MSA, each successive iteration provides a new flow, or a *design point*, while some sort of averaging of the design points provides a new solution at that iteration. Frees and Ruppert (1990) point out the advantages of using one method to select design points, and a different method to estimate the solution. Use of the methods adapted to each purpose allows a more aggressive exploration of the feasible space and a more effective exploitation of the results generated during that exploration. These types of method have been classified as *iterate averaging*.

Polyak (1990) introduced a method to implement iterate averaging. In this method one also computes, 'in parallel' with, and independent of, the MSA's weighted average, a running average of the design points. The effect of this additional step is that the MSA's large step size tends to prevent the algorithm from getting stuck in an early stage, while the 'off-line' parallel averaging takes care of the increased oscillations that the larger step sizes produce. The averaging is 'off-line' in the sense that the iterations of the recursive averaging process (of the MSA) make use of the weighted average and not the running average.

Bottom and Chabini applied the Polyak and MSA averaging to a variety of problem instances and analysed the resulting convergence behavior. They report that 'In most cases considered, *accelerated averaging* can significantly outperform the MSA in terms of both the noise at convergence as well as the number of iterations needed to converge. This performance can be obtained at a modest incremental cost to the MSA'. They observed that the application of the Polyak method to dynamic traffic assignment problems resulted in noise at convergence considerably lower than for the MSA with convergence rates four or more times faster than the MSA.

The above properties of MSA, especially the fact that the modeler has a reasonable ability to control the convergence of the equilibration process, makes MSA an appropriate candidate to model the day-to-day route assignment in network loading process (see Figure 14.1, above), to capture the dynamics of congestion formation and dissipation associated with physical vehicular traffic flow and the variety of route-choice behaviors.

6. MODEL EVALUATION APPROACH

As discussed in Section 3, the ideal way to evaluate the proposed traffic assignment approach would be in the field, where one sets up a data collection system that includes

detector counts and travel time probes. This test would be quite expensive and extensive given the scope of the research. Hence, the traffic assignment approach was tested 'in the laboratory' using a computer representation for network traffic loading.

One may study the above dynamic model for route assignment and network loading using an analytical assignment procedure (for example, Ran and Boyce 1996). This requires volume–delay functions for each link. As discussed previously, the volume–delay or impedance functions used in analytical traffic assignment approaches have their limitations in modeling congestion and dissipation. In these models the impedance on a link is simply a function of volume of traffic, capacity of link, mean travel time on the path-link, saturation rate and some parameters obtained empirically. Although the parameters of these impedance functions may be obtained after many empirical observations and extensive calibration, the ability of such functions to comprehensively capture all characteristics and dynamics of a specific link of a particular route is limited, especially when one needs to include a new intervention such as a work zone.

In order to overcome the shortcomings in capturing the traffic dynamics details for modeling route choices, an alternate simulation-based traffic-loading model was used to evaluate our dynamic model. This approach does not use volume–delay functions but, rather, uses travelers' trip experiences from a simulation model. That is, we use iterative route-based simulation as the 'field' in conjunction with an iterative averaging algorithm for traffic assignment. By avoiding the use of impedance functions and instead simulating the traffic on the network, we can expect more realistic traffic loading where network and traffic characteristics, which have a significant impact on network congestion, can be included in the model. Further, the efficiency of the assignment procedure is improved by avoiding the limitations of the volume–delay functions in highly congested or underutilized link conditions and by introducing accelerated averaging for route assignment. Note that the Frank–Wolfe algorithm (FWA) cannot be used in this model because calculation of the step size in FWA requires a one-dimensional search to determine the step size that minimizes the objective function. This line search requires evaluation of volume–delay functions, which are not available for on-line traffic assignment and loading.

To use any evaluation model, whether analytical or simulation, one still needs to assume some sort of route-choice model. For example, do travelers on a rectilinear grid choose minimum travel time routes or do they choose routes with the smallest number of left turns? Although our model can consider any route-choice criterion, to compare our approach with existing static traffic equilibrium models we used the conventional route-choice criterion of selecting routes with minimum travel times. Specifically, our approach was tested on a small network (INN) to examine whether the on-line traffic assignment process eventually results in nearly equal travel times when travelers are made to choose routes that minimize their experienced travel times.

For the assignment procedure, at each iteration a minimum path tree is constructed for each specified origin node to all its destination nodes, using a shortest path algorithm. However, unlike conventional assignment models, this model does not calculate link impedances using BPR or Davidson functions. Instead, at each iteration, the route costs are found by simulating the traffic network using the flows and paths generated by the assignment algorithm. The network is simulated such that traffic loads between origin–destination pairs follow exactly the paths generated in that iteration. This is accomplished by using a route-based simulation (PF-CORSIM). The iterative procedure continues until a defined

convergence criterion is attained or when the number of iterations reaches a pre-specified upper bound. Figure 14.2 presents a flow chart for *traffic assignment and loading on a route-based simulation.*

6.1 Route-based CORSIM Model (PF-CORSIM)

The evaluation of the proposed traffic assignment model requires a mechanism to simulate traffic loads on user-defined paths. We used a route-based implementation of the CORSIM microscopic simulation package (FHWA 1998) called PF-CORSIM which has been developed by ITT Industries (ITT maintains CORSIM) with support from the ATLAS Research Center at the University of Arizona.

For readers familiar with the CORSIM package, the Path-Following implementation (PF-CORSIM) works with CORSIM releases 4.32 and 5.0 and retains all the features of the original simulation package while providing additional features like path assignment and vehicle injection at user-defined times.

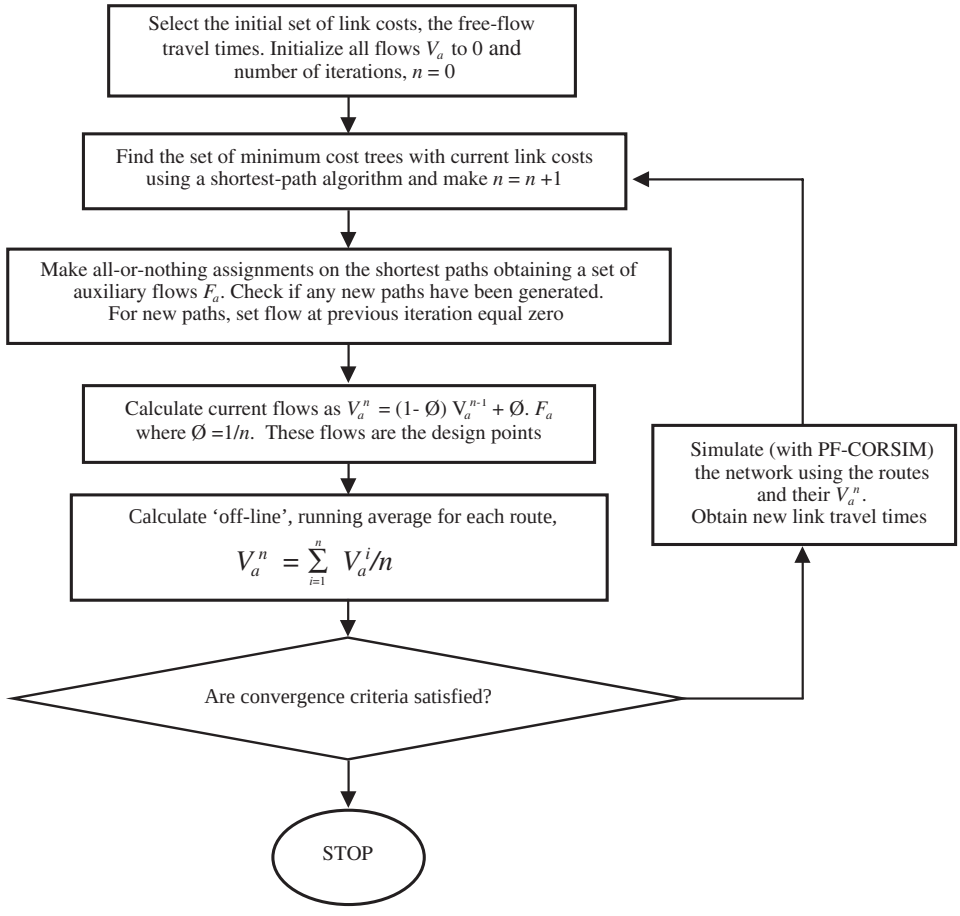


Figure 14.2 Flow chart for traffic assignment using iterative route-based simulation

The following is a brief description of the test network model and traffic flow parameters used in our experiments.

- A 40-node, 104-link, 4×5 grid network (including entry/exit nodes and links), modeled using CORSIM Version 4.32.
- Ten origin–destination (O–D) pairs with a cumulative flow of 720 vehicles.
- The origins and destinations are on the borders of the test network, the grid structure allowing several alternative routes for each O–D pair.
- Specified O–D traffic volumes are uniformly loaded, with 2-second headways, only through the PF-CORSIM tool, that is, there is no background traffic.
- The free-flow speed is 45 miles per hour for entry/exit links and 35 miles per hour for the rest of the links.
- The mean values of start-up loss time and queue discharge headway are 2.5 and 1.7 seconds, respectively, and are the same throughout the network.
- A pre-timed, four-phase signal control strategy for all nodes (intersections). The signal duration interval for all through movements is 30 seconds and for all left-turning movements is 15 seconds.
- Right turns on red are allowed.

We note that this test network model was used only to evaluate the assignment procedure and does not represent an actual site.

7. ANALYSIS AND RESULTS

Several traffic assignments and loadings were performed on the simulation model, using a different random seed for each simulation run. The different random seed for each experiment results in stochastic variations in lane-changing and car-following behavior, as well as queue discharge and start-up loss-time characteristics, to simulate variability that may occur in the field. For standard CORSIM models, a random seed is also used to stochastically generate the vehicle entry headways at origins; routes are not explicitly specified for each vehicle – they result only from the intersection turning ratios as prespecified by the user. In PF-CORSIM an external agent (for example, our traffic assignment procedure) specifies the traffic routes and loads. The fact that stochastic attributes are incorporated in the traffic assignment/simulation model makes the network loading model more realistic than the case where these characteristics are assumed to have been modeled by a simple volume–delay relationship.

In each experiment, the traffic assignment/simulation iterative process continued until some convergence criteria were satisfied to indicate that the dynamic system was close to 'steady state'. The convergence criteria for this particular network and scenarios included: (i) travel times of all routes of an O–D pair should be similar (since we are assuming minimum travel-time route choices), and (ii) the flows assigned to each route should not vary by more than 0.5 vehicles for two successive iterations. We specify 0.5 vehicles as the acceptable upper limit on the absolute difference in flows for consecutive iterations because most assignments generally have a few routes with fractional flows assigned to them. Also, if the difference in flow is fractional, it will not impact the link travel times

obtained at the end of simulation, and consequently not impact the flow assignments in subsequent iterations. Continuing, the other convergence criteria include: (iii) no new routes are generated, and (iv) more than 90 per cent of the trips should have *excess travel time* less than δ per cent of the *average trip travel time* or τ seconds. Excess travel time of a route is defined as the difference between its route travel time and the travel time of the shortest route for that O–D pair. In this test case, the average trip time ranges from 326 to 360 seconds for different random seeds. Therefore, $\tau = 10$ seconds of excess travel time is approximately $\delta = 3$ per cent of the average trip time. We chose the statistic 'percentage of trips with excess travel time less than 10 seconds' because it appeals as a reasonably good indicator of the state of convergence. This statistic can have different excess travel time (cut-off) values depending on the nature of the network under consideration.

To gain insight into the effect of simulated randomness in each experimental run, travel-time distributions for a given O–D pair, the routes and their travel times, were plotted for the different experiments to observe the variations induced. Figure 14.3 presents route travel-time distributions for O–D Pair 8 for different experimental runs, where each bar

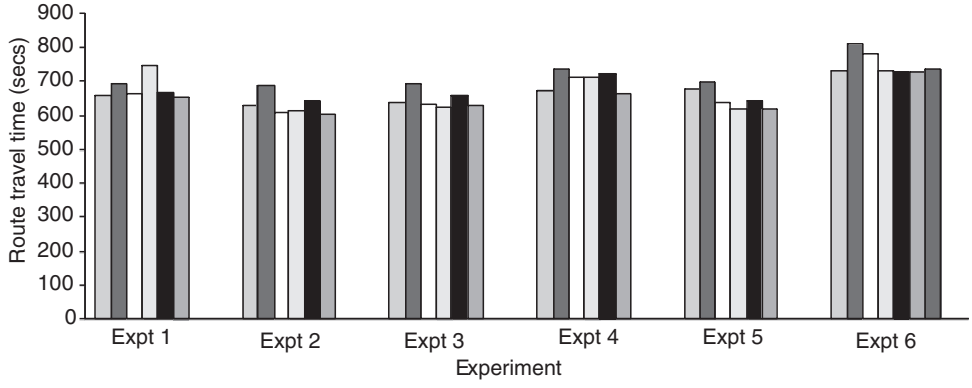


Figure 14.3 Routes and their travel times for O–D Pair 8 in six experiments

represents the travel time of a route. We can see from the figure that the number of routes generated and the distributions of route-travel times in the different experiments have some variations, but not very much, as would be expected in the field. Observe that in Experiment 6 some travelers in fact chose a rather long route of over 800 seconds; but, as will be shown later, very few (three out of 160) used that route.

7.1 Validation of Assignment/Simulation Process

To further validate the assignment/simulation process, we examined whether the results provide a reasonable representation of the system, given the particular objectives of the travelers. Therefore, for this model, we felt that the output data should satisfy the following condition: As the model converges to steady state, the travel times of all routes for a given O–D pair must approach a consensual value. In case the values differ widely, the routes with the lowest travel times should be assigned almost all the flow. Figure 14.4 shows distributions of route travel times for O–D Pairs 1 to 10 in Experiment 6. The dis-

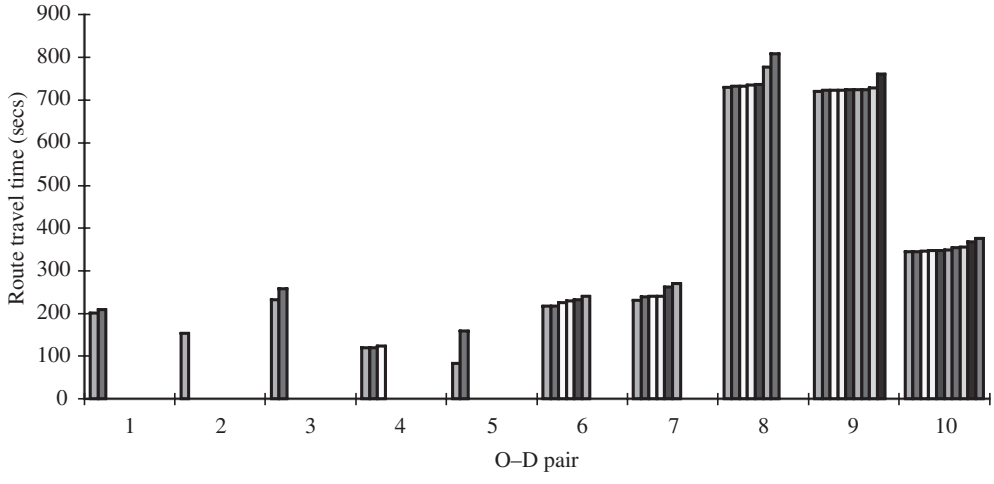


Figure 14.4 Route travel times (in ascending order) for all O–D pairs in Experiment 6

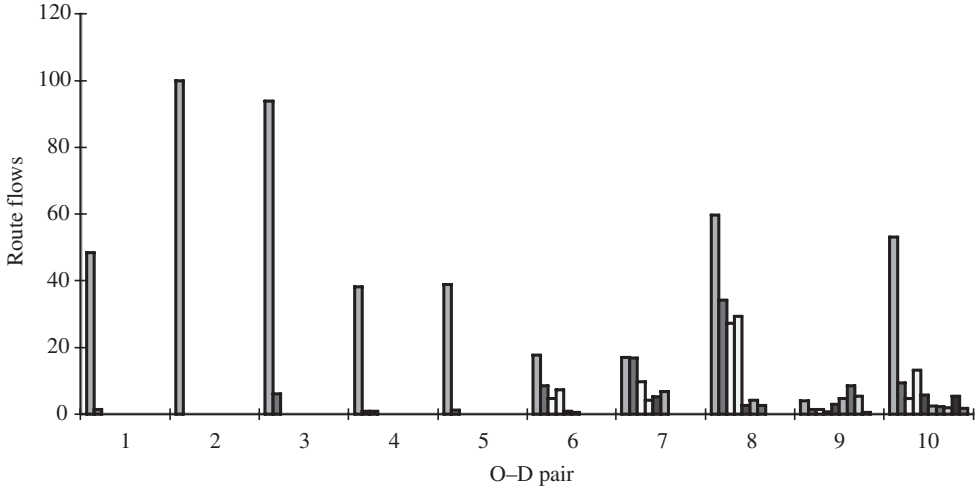


Figure 14.5 Corresponding flows for the routes shown in Figure 14.4 for each O–D pair

tributions of route flows for this traffic assignment are presented in Figure 14.5, where the route flows are given as the total number of vehicles for that period, on each route for each O–D pair. We can make the following observations from these figures:

- The number of routes for each O–D pair ranged between 1 (Pair 2) and 10 (Pair 10).
- The two routes for O–D Pair 3 showed perceptible difference in route travel times; the shorter route had 94 vehicles (out of 100) assigned to it, while the larger had six.
- The two routes for O–D Pair 5 showed significant variation in travel times. As expected almost 97 per cent of the trips were assigned to the shorter path.

- Of the six routes for O–D Pair 8, two had somewhat longer travel times. The total flow on these two long routes was only seven vehicles (out of 160); three on the longest and four on the second longest.

7.2 Excess Travel-time Analysis

To further study convergence dynamics, excess travel-time analysis was carried out. Figures 14.6 and 14.7 give the excess travel-time distribution of 720 O–D trips for Experiment 6. In these figures, the horizontal axis is the excess travel time in seconds, and, for each excess travel time, the number of trips that have excess travel time greater than this is represented by the y coordinates. Figure 14.6 shows the excess travel-time distribution after the 15th iteration (midway to convergence) and Figure 14.7 gives the distribution at convergence (31st iteration) of the assignment process.

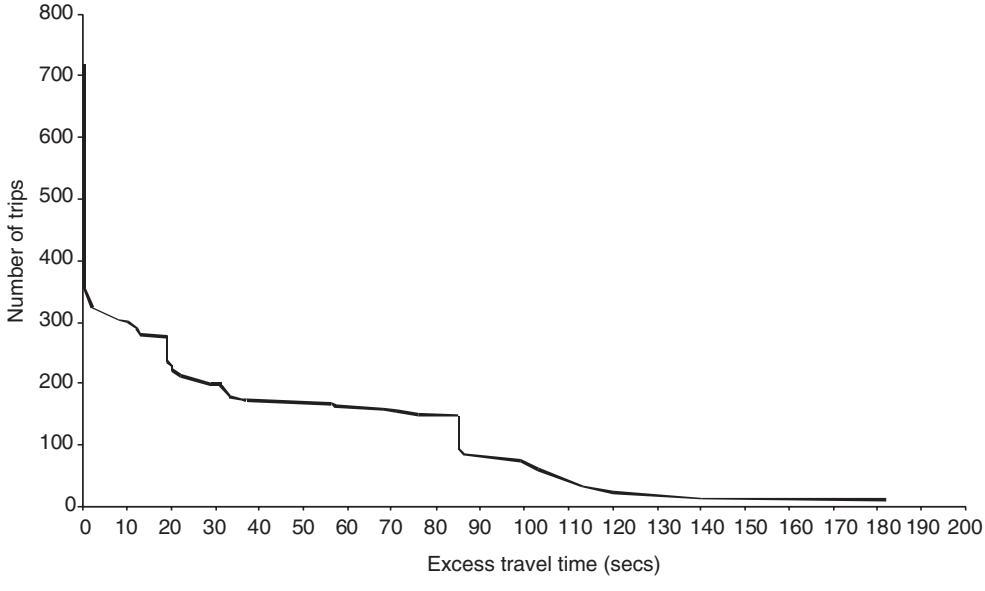


Figure 14.6 Distribution of excess travel time after 15th iteration for Experiment 6

The following observations can be made from Figure 14.6 at the end of 15 iterations:

- Some 58 per cent of the trips have excess travel time of less than 10 seconds.
- About 50 per cent of the trips have zero excess travel time.
- Approximately 22 per cent of trips have excess travel time of more than a minute and just above 3 per cent of the trips have excess travel time of more than 2 minutes.
- The average excess travel time per trip is 2.29 seconds.

According to Figure 14.7, at the end of 31 iterations,

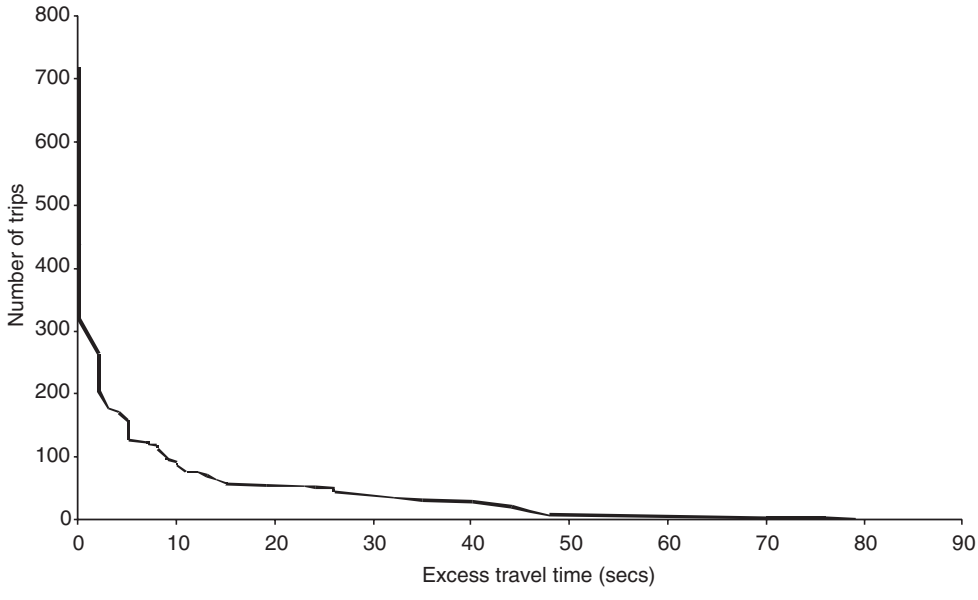


Figure 14.7 Distribution of excess travel time at convergence (after 31 iterations) for Experiment 6

Table 14.1 Excess travel-time statistics for six random seeds

	Experiment					
	1	2	3	4	5	6
Trips with excess travel time less than 10 secs (%)	92.1	94.6	96.3	94.9	90.5	90.5
Average excess travel time per trip (secs)	0.96	0.59	0.79	0.71	1.72	0.9

- Some 90.5 per cent of the trips have an excess travel time of less than 10 seconds.
- About 63.2 per cent of the trips have zero excess travel time.
- Less than seven, or 0.96 per cent, of the trips have excess travel time of a minute or more. No trip has excess travel time greater than 79 seconds.
- The average excess travel time per trip is 0.9 seconds.

Table 14.1 shows the excess travel-time statistics for six random seeds. The results are as anticipated, considering the fact that the model is based on a simulated network, which can induce some randomness. In essence, the excess travel-time analysis clearly shows that the assignment/simulation approach converges to steady state as the number of iterations (representing time periods) increases.

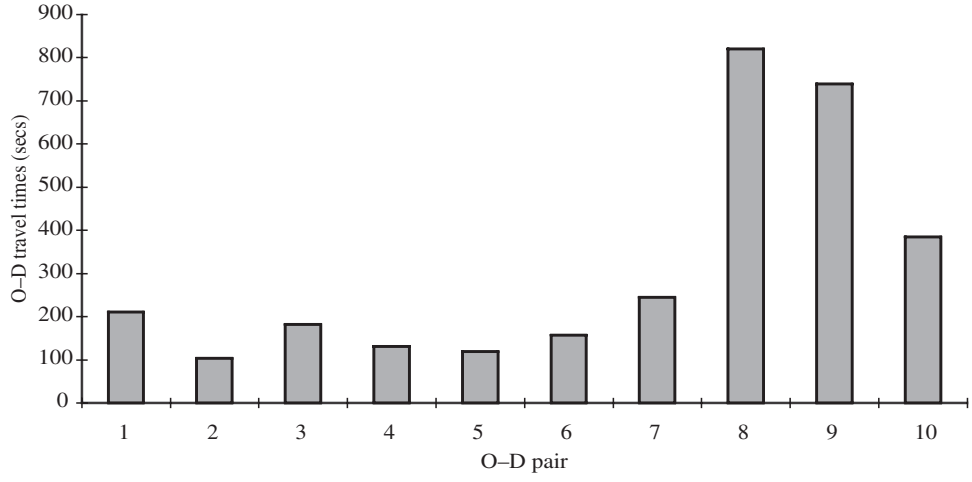


Figure 14.8 Average travel time for all O–D pairs from STA

7.3 Comparison with the Static Traffic Assignment Approach

Finally, we compared the results at steady state with those obtained with the STA approach. Since our experimental scenario with the dynamic model assumed minimum travel-time route choice, we expect the results at steady state to be similar to those obtained by STA.

We used the final link volumes as 'observations' from the real world and tuned STA parameters so that link volumes matched – as done by those who use STA for network planning. Normally, these tuning parameters relate to the link impedance functions, or the BPR functions, used by STA. In our case, we added left-turn penalties to account for the longer delays drivers experience when making left-turn movements at intersections, compared to the delays traveling straight through or making right turns.

Figure 14.8 shows the average travel time for each O–D pair and Figure 14.9 gives the distributions of route flows for the corresponding O–D pairs as obtained by STA. In general the travel times and route flow distributions are similar to those obtained by our dynamic model. However, there are slight differences, as would be anticipated since there are some differences in the underlying behavioral and analytical assumptions. As expected, for each O–D pair there is only one travel time in the STA approach due to explicit use of Wardrop's First Principle for user equilibrium. For O–D Pairs 2 and 5, the numbers of routes used were the same for both approaches, while for the other O–D pairs the numbers differed slightly; however, the total number of routes used by all vehicles was approximately 40 in both approaches.

8. CONCLUSIONS

The primary objectives of this study were (i) to propose an analytical traffic assignment approach to dynamically monitor an impacted neighborhood network (INN) when count

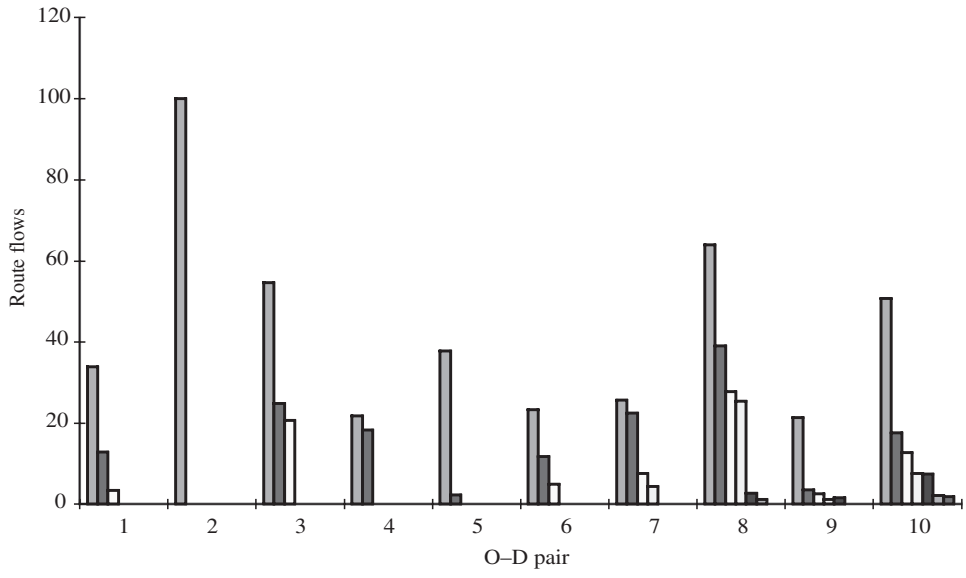


Figure 14.9 Route flow distributions for O–D pairs from STA

detectors and traffic probes are available and (ii) to demonstrate the feasibility and applicability of the model. The model considers the dynamics of travel through the network, as well as day-to-day congestion-forming processes, rather than the use of explicit BPR functions. The basic goal was to re-engineer the traffic assignment process for an INN to make it more responsive to behavioral aspects and inherent randomness of the transportation system in the presence of new interventions such as work zones, considering that the rapid deployment of ATIS technologies has resulted in a shift in the way individuals choose their routes in such situations.

Our dynamic model assumes that the INN is constantly monitored with count detectors and vehicle probes that provide trip attributes such as travel times. This information is available to travelers who use the INN. After each period, for example a day, each traveler experiences attributes of his/her route and chooses another route if he/she has current knowledge of attributes of other routes. This is repeated in the next period and the process continues. Note that, rather than assume a volume–delay function, the traveler makes his/her decision based on the actual experience on the previous trip and, hence, can account for whatever factors influence his/her decision, such as aversion to left turns, avoidance of traffic signals, lane blockages due to buses, and so on.

To test our approach, as well as to compare with conventional traffic assignment models, we assumed that travelers choose routes that minimize travel times experienced by them. In the evaluation of our scheme, we examined how close the travel times were with one another at steady state, to reflect the level of equilibrium reached in the network. If, in fact, travelers did choose routes to minimize their travel times, then it is expected that most of the experienced travel times for any given O–D pair would be approximately equal. We also compared the steady-state results with those that may be obtained using a

conventional static traffic assignment. In addition, we studied the convergence process of the dynamic model to see how 'excess travel time' decreases at each iteration.

Analysis of the results from the assignment/simulation model indicate that our dynamic equilibration process does behave as we would expect in the field: (i) the process exhibits dynamic travel times and day-to-day convergence to a steady state; (ii) routes most used for an O–D pair, at steady state, have nearly equal travel times; and (iii) if a long route is chosen for an O–D pair then only a very small fraction of travelers choose it. The statistics observed from excess travel-time distribution graphs further support the claim that the assignment process provides reasonable network loads over time and that the process tends to converge to a traffic equilibrium as the number of iterations increases.

A natural extension to this research effort would be the application of this model to a real-life traffic network. This is essential to validate the qualitative improvements this model seeks to bring to short-term traffic planning of a locally impacted area due to situations such as work-zone activities. Further, the travel demand forecasts obtained from this model could be compared to those from a conventional assignment process and their performance against field data could be tested.

Perhaps the most important contribution of this chapter is the introduction of an approach that could track the transient behavior in the process of reaching a traffic equilibrium when an intervention such as a work zone, or short-term disruption of capacity, is introduced in the network. Here the 'iterations' in the assignment process represent the time periods in the *equilibration dynamics.* However, as is the case of developing a validated model of any complex dynamic system, especially when human decision processes significantly influence the dynamic characteristics of the system, considerable effort is necessary to develop an operational process to fit the dynamic model's parameters from field observations (that is, data from detectors and traffic probes) over time and validate the model's results. By some careful, and somewhat extensive, data collection during the intervention, we may be able to calibrate a model that predicts the traffic loads for the next day as a function of the previous few days' (i) trip times, (ii) traffic volumes, and (iii) equilibration dynamics' parameters that represent the inertia among decision makers in re-routing.

NOTE

* The authors acknowledge funding from the Federal Highway Administration and the Arizona Department of Transportation (ADOT Contract KR98-2083TRN/FHWA Agreement ITS-9804-001 and ADOT Contract KR00-0972TRN) and the National Science Foundation (Grant CMS-0231458) for the partial support of the research reported here. The contents of this chapter reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein, and do not necessarily reflect the official views of the FHWA, Arizona DOT, or the NSF. The authors would like to thank and acknowledge Larry Owens and Dave Holmgren of ITT Systems for their assistance on the development and use of the route-based simulation model PF-CORSIM.

REFERENCES

Akçelik, R. (1991), 'Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function', *Australian Road Research*, **21** (3), 49–59.

Ben-Akiva, M., M. Bierlaire, J. Bottom, H.N. Koutsopoulos and R.G. Mishalani (1997a), 'Development of a route guidance generation system for real-time applications', *Proceedings of the 8th International Federation of Automatic Control (IFAC) Symposium on Transportation Systems*, Chania, Greece.

Ben-Akiva, M., M. Bierlaire, H.N. Koutsopoulos, R.G. Mishalani and Q. Yang (1997b), 'Simulation laboratory for evaluating dynamic traffic management systems', *Journal of Transportation Engineering*, **123**, 283–9.

Bottom, J.A. and I. Chabini (2001), 'Accelerated averaging methods for fixed point problems in transportation analysis and planning', Triennial Symposium on Transportation Analysis (TRISTAN IV), São Miguel, Azores, Preprints 1/3, 69–74

Branston, D. (1976), 'Link capacity functions: a review', *Transportation Research*, **10**, 223–36.

Cascetta, E. and M. Postorino (2001), 'Fixed-point models for estimating or updating origin/destination matrices from traffic counts', *Transportation Science*, **35**, 134–47.

Daganzo, C.F. and Y. Sheffi (1977), 'On stochastic models of traffic assignment', *Transportation Science*, **11**, 253–74.

Davidson, K.B. (1966), 'A flow travel time relationship for use in transportation planning', *Proceedings of 3rd Australian Road Research Board Conference*, **3** (1), 183–94.

Federal Highway Administration (FHWA) (1998), *CORSIM User's Manual*, Version 4.2, ITS Research Division, FHWA, Washington, DC.

Frees, E.W. and D. Ruppert (1990), 'Estimation following a Robbins–Monro designed experiment', *Journal of the American Statistical Association*, **85**, 1123–9.

Janson, B.N. (1991), 'Dynamic traffic assignment for urban road networks', *Transportation Research*, **25B**, 143–61.

Jayakrishna, R., H.S. Mahmassani and T. Hu (1994), 'An evaluation tool for advanced traffic information and management systems in urban networks', *Transportation Research*, **3C**, 127–47.

Mahmassani, H.S., T. Hu, S. Peeta and A. Ziliaskopoulos (1993), 'Development and testing of dynamic traffic assignment and simulation procedures for ATIS/ATMS applications', Technical Report DTFH61-90-R-00074-FG, Center for Transportation Research, University of Texas at Austin, TX.

Merchant, D.K. and G.L. Nemhauser (1978), 'A model and an algorithm for the dynamic traffic assignment problems', *Transportation Science*, **12**, 183–99.

Mirchandani, P.B. and H. Soroush (1987), 'Generalized traffic equilibrium with probabilistic travel times and perceptions', *Transportation Science*, **21**, 133–52.

Oh, Jun-Seok, C.E. Cortes, R. Jayakrishnan and D.H. Lee (1999), 'Microscopic simulation with large network path dynamics for advanced traffic management and information systems', Working Paper UCI-ITS-WP-99-16, Institute of Transportation Studies, University of California at Irvine, CA.

Polyak, B.T. (1990), 'New method of stochastic approximation type', *Automation and Remote Control*, **51** (7), 937–46.

Ran, B. and D.E. Boyce (1996), *Modeling Dynamic Transportation Networks*, Berlin: Springer-Verlag.

Spiess, H. (1990), 'Conical volume-delay functions', *Transportation Science*, **24**, 153–8.

Tisato, P. (1991), 'Suggestions for an improved Davidson travel time function', *Australian Road Research*, **21** (2), 85–100.

Wardrop, J.G. (1952), 'Some theoretical aspects of road traffic research', *Proceedings of the Institute of Civil Engineering*, Part II, 325–78.

15. Multi-modal routing and navigation cost functions for location-based services (LBS)

Tschangho John Kim*

1. INTRODUCTION

It is widely believed that about 80 per cent of public and private decisions are related to some sort of spatial and locational consideration, leaving only a few areas that are not affected by locational considerations. The Internet puts an unprecedented amount of locational information of all kinds at a user's fingertips, information that can be used for personal production activities in a mind-boggling variety of ways (Ostensen 2001).

The location-based services (LBS) are the new face of the wireless Internet (see Kim et al. 2002 for a detailed introduction to LBS). LBS – sometimes called location-based mobile services (LBMS) – are an emerging technology combining information technology, geographic information systems (GIS), positioning technology, intelligent transportation systems (ITS) technology and the Internet. LBS combine hardware devices, wireless communication networks, geographic information and software applications that provide location-related guidance for customers. It differs from mobile position determination systems, such as global positioning systems (GPS) in that LBS provide much broader, application-oriented location services, such as the following:

> 'You are about to join a 10-kilometer traffic queue, turn right on Washington Street, 1 km ahead.'
> 'Help, I'm having a heart attack!' or 'Help, my car has broken down!'
> 'I need to buy a dozen roses and a birthday cake. Where can I buy the least expensive ones while spending the minimum amount of time on my way home from the office?'

A typical example of LBS for personal navigation would include the following:

- Entering address to desired destination (Geocode).
- Subscriber wishes to start from his/her current position and add one stop along the way (Gateway).
- Determining the route (Route Determination).
- Presenting route summary to subscriber.
- Presenting turn-by-turn directions to subscriber.
- Subscriber wants to see a map overview with the route shown (Presentation).
- Subscriber is now in transit and wants to see maneuvers (Gateway and Presentation).

An advertising and e-commerce consulting firm predicts that by 2005, the LBS market will reach $11–$15 billion in revenue and as many as one billion Internet-enabled hand-

sets will be in use.[1] With this growth potential, LBS present a substantial emerging market opportunity for wireless providers.[2] While the market for LBS seems to be rapidly emerging and some initial, but primitive services have been introduced, there remain many basic issues that have yet to be researched and developed in order to provide efficient services for users and providers (see http://pulver.com/lbsreport/bissues.html). Among these many issues, this chapter focuses on developing functional forms for costs for providing multi-modal routing and navigation services and on searching for feasible directions to solve the functions heuristically.

2. A USE CASE: REQUEST AND RESPONSE FOR ROUTING AND NAVIGATION SERVICES

2.1 Cost Model

An example in LBS is a request for a routing and navigation service: 'find the least-cost routes using all available modes of transportation from my current position, stopping at a gas station for 10 gallons of gas, a pharmacy to pick up a prescription medicine, and a flower shop for a dozen roses before arriving home' (Figure 15.1).

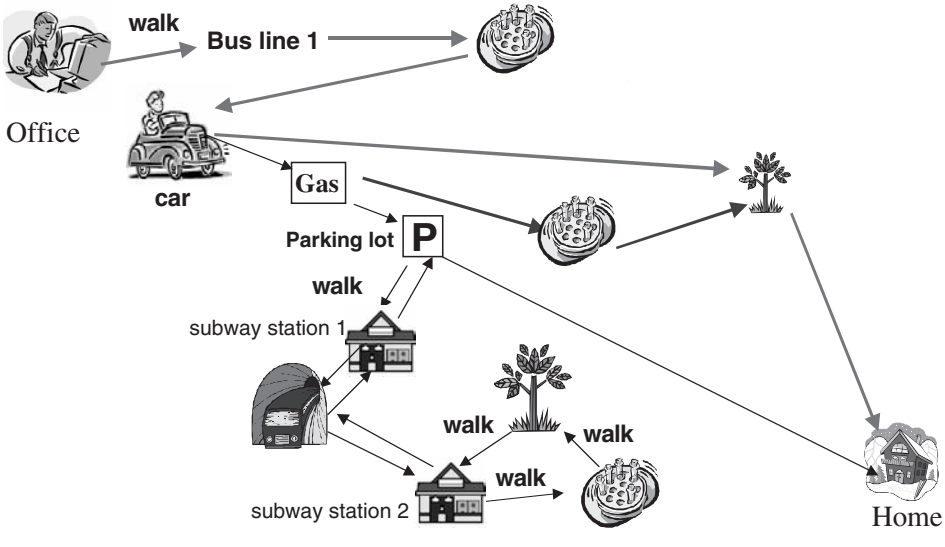


Figure 15.1 Possible choices for completing planned activities

In this example, there are three types of costs involved: (i) the purchasing costs for needed items and stopping costs including parking, (ii) costs related to the time spent on the road or on transit, and (iii) distance-related costs such as gasoline used, wear and tear from the use of a car, and transit fares.

Purchasing and stopping costs

There are three items to shop for, which we shall call activities, denoted by (B^m), meaning that activity one ($B^1 = 10$) is to buy 10 gallons of gas, activity two ($B^2 = 1$) is to pick up the prescribed medicine, and activity three ($B^3 = 12$) is to buy a dozen roses. Suppose that there are three gas stations, two flower shops and one pharmacy to choose from. Assume that the pharmacy and flower shops can also be reached by subway. The unit cost for a gallon of gas at the three different locations is denoted by (b_j^m), meaning that cost per gallon at location 1 ($j = 1$) is denoted by (b_1^1), at location two is denoted by (b_2^1), and the third as (b_3^1). Likewise, the unit cost for a rose at the flower shop at location one is denoted by (b_1^2) and the other as (b_2^2). The unit cost of the medicine at the pharmacy is (b_1^3). If an item 'm' is not available at stop 'j', then $b_j^m = \infty$ (unbounded). The matrix B_j^m represents the decision to purchase B amount of item m at stop j. Let s_j represent initial stopping costs that include parking costs once a decision has been made to stop at j. The marginal costs for stopping at location j for purchasing m are represented by s_j^m which include walking, waiting, queuing and other added costs to purchase m at location j. The decision to stop at location j is given by $d_j = 1$ if any B_j^m is non-zero. The decision to purchase item m at stop j is $d_j^m = 1$ for all B_j^m that are non-zeros.

Thus, the total cost of purchasing needed items at all locations j (C_j) including stopping costs in this example can be written as:

$$C_j = \sum_j d_j s_j + \sum_m (b_j^m B_j^m + d_j^m s_j^m), \tag{15.1}$$

where $d_j s_j$ is the cost of making an initial stop at j which includes parking cost, $b_j^m B_j^m$ is the total cost of purchasing m at location j, and $d_j^m s_j^m$ is the marginal cost incurred for purchasing item m including queuing at j.

Once the shopping is done, the total items purchased should be at least the same as the original intention to buy (B^m), that is, 10 gallons of gas, a dozen roses and the medicine.

This is expressed as:

$$\sum_j B_j^m = B^m. \tag{15.2}$$

Time costs

Time spent on the road to go to one of three gasoline stations, on road or on transit to go to one of two flower shops and to the pharmacy will depend on which shop is visited and in which order.

The road and transit network is represented by two types of element: a set of points called nodes and a set of line segments connecting these points called links. Figure 15.2 depicts a combined road and transit network including five nodes connected by 11 links. It is not important at this point which links are transit and which are roads since what matters is the travel time in this example. Nodes are numbered by ordinary Arabic numer-

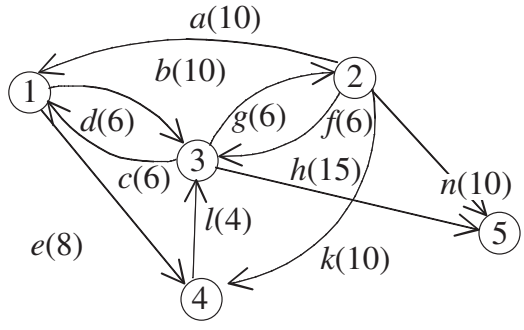


Figure 15.2 Sample combined network with fixed-link travel time

als, 1 to 5, and links are numbered by the alphabet, from *a* to *n*. The link travel time in minutes is given in parentheses, right after the link number.

It is possible to reach node 5 from node 1 by several routes (or paths) through the network. In fact, there are $(n-1)!$ possible routes if no one way exists. A route is a sequence of directed links leading from one node to another. For example, in the above network, to get from node 1 to node 5, the following routes are available excluding those routes that require stopping at the same node more than once:

- Route 1: $1 \rightarrow b \rightarrow 2 \rightarrow n \rightarrow 5$ (total time ($\sum_a t_a$) to be on roads: 20 minutes)
- Route 2: $1 \rightarrow b \rightarrow 2 \rightarrow f \rightarrow 3 \rightarrow h \rightarrow 5$ ($\sum_a t_a$ is 31 minutes)
- Route 3: $1 \rightarrow b \rightarrow 2 \rightarrow k \rightarrow 4 \rightarrow l \rightarrow 3 \rightarrow h \rightarrow 5$ ($\sum_a t_a$ is 39 minutes)
- Route 4: $1 \rightarrow c \rightarrow 3 \rightarrow h \rightarrow 5$ ($\sum_a t_a$ is 21 minutes)
- Route 5: $1 \rightarrow c \rightarrow 3 \rightarrow g \rightarrow 2 \rightarrow n \rightarrow 5$ ($\sum_a t_a$ is 22 minutes)
- Route 6: $1 \rightarrow e \rightarrow 4 \rightarrow l \rightarrow 3 \rightarrow h \rightarrow 5$ ($\sum_a t_a$ is 27 minutes)
- Route 7: $1 \rightarrow e \rightarrow 4 \rightarrow l \rightarrow 3 \rightarrow g \rightarrow 2 \rightarrow n \rightarrow 5$ ($\sum_a t_a$ is 28 minutes).

This is not all of the possible routes with stops. Several possible routes pass through two gas stations and would therefore give alternative routes. The ones not listed are not optimal, but that is not obvious a priori.

Let us assume that we know each traveler's unit cost of time per hour and denote it as (α^w) for a traveler of occupation type *w*, then the total travel cost for traveler type *w* (C_w) is the unit cost of time (α^w) multiplied by the total time spent on all links *a* on mode *k* ($\sum_a t_a^k$), if link *a* is in route *r* of mode *k* between the origin (O) and destination (D) pair denoted by *i* and *j*. That is:

$$C_w = \alpha^w \sum_k \sum_{a \in r_{ij}^k} t_a^k, \tag{15.3}$$

where $a \in r_{ij}^k$: link *a* is in route *r* of mode *k* between the O–D pair denoted by *i* and *j*. Mode *k* includes walking, and link travel time t_a^k includes transfer time connecting the two modes on foot if link *a* is a connecting link.

Assuming that a traveler's unit time cost (α^w) is \$20/hour or 33 cents per minute, the total cost for taking route 1 is 33 cents times 20 minutes, that is, \$6.60. If route 2 is the minimum route and thus is chosen, then the total costs for taking route 2 is 33 cents times 31 minutes, that is, \$10.23.

Distance-related costs

Unit cost per mile for distance-related costs such as gasoline used, and the wear and tear from using a car and/or transit fare is denoted by d^k, and the total distance traveled by mode k is denoted by (d_a^k), then the total distance-related costs (C_d) are:

$$C_d = \sum_k d^k \sum_{a \in r_{ij}^k} d_a^k, \tag{15.4}$$

where $a \in r_{ij}^k$: link a is in route r of mode k between the O–D pair denoted by i and j.

Total costs for shopping, routing and navigation

The total cost for traveler type w (W) for stopping at a gas station for 10 gallons of gas, a pharmacy to pick up a prescribed medicine, and a flower shop for a dozen roses before reaching home can now be calculated by summing up the minimum purchasing costs (C^j) over all location j, the minimum routing and navigation costs (C^w) and the minimum distance costs (C^d), on condition that all items are successfully purchased.

The solution of the problem for type w traveler can be found by minimizing the total costs, that is:

$$W = C^j + C^w + C^d$$

or

$$W = \sum_j d_j s_j + \sum_m (b_j^m B_j^m + d_j^m s_j^m) + \alpha^w \sum_k \sum_{a \in r_{ij}^k} t_a^k + \sum_k d^k \sum_{a \in r_{ij}^k} d_a^k, \tag{15.5}$$

subject to equation (15.2).

This is a typical linear programming (LP) problem, however, solving an LP with tens of thousands of links and nodes is not easy. Many scholars have found efficient algorithms for solving it differently from solving it as an LP problem (see Boyce et al. 1998; Lee et al. 2002 for a detailed discussion of formulating and solving dynamic route-choice problems).

The next question is, how can we obtain the link travel time (t_a)?

2.2 Estimating Link Travel Time, t_a

Estimating link travel time will depend on whether or not real-time traffic data including volume and speed can be obtained and be made available for service brokers/users.

When real-time link speed is available

When real-time speed data are available for each link, then we can easily estimate the link travel time by the following equation:

$$t_a = (60 \text{ mins} \times \text{link distance in miles})/[\text{Speed (miles per hour)}]. \tag{15.6}$$

For example, if the current speed on link a is 30mph and the link length is 2 miles, the current link travel time is (60 mins × 2)/30mph = 4 mins.

When real-time link traffic volume is available

Many cities have now installed devices such as loop detectors to obtain real-time traffic volume on certain links. In such cases, real-time link speed may not usually be obtainable from loop detectors, but real-time link traffic volume can be. We can convert real-time link volume to link travel time by using a function such as the BPR (Bureau of Public Roads) function as shown below:

$$t_a = t_a^0 [1 + \eta(v_a/(c_a)^{\lambda}], \tag{15.7}$$

where:

- t_a = current link travel time;
- t_a^0 = uncongested free-flow travel time on link a;
- v_a = real-time traffic volume on link a;
- c_a = capacity of link a in number of vehicles per lane (refer to the Highway Capacity Manual (HCM) by the Federal Highway Administration (FHWA) (Bureau of Transportation Statistics, 1998a);
- η = a coefficient to be calibrated. The usual value used for US city roads is 0.88 (Bureau of Transportation Statistics, 1998b) and also see Horowitz (1991);
- λ = a coefficient to be calibrated. The usual value used for US city roads is 5.5 (Bureau of Transportation Statistics, 1998b).

For example, assume that there is a link on which uncongested link travel time (t_a^0) is 40 miles per hour for a 2-mile link (or 3-minute link travel time), and has two lanes with the capacity of handling 1600 passenger car-equivalent units (PCUs) per lane. Further, assume that loop detectors indicate that there are 4000 PCUs passing by in that link now, then the estimated link travel time is:

$$t_a = 3 \text{ mins } [1 + 0.88\ (4000/3200)^{5.5}] = 12 \text{ mins.}$$

If there are only 1000 PCUs traveling, then link travel time is:

$$t_a = 3 \text{ mins } [1 + 0.88\ (1000/3200)^{5.5}] = 3 \text{ mins.}$$

For detailed descriptions on the other type of link travel-time functions, see Suh et al. (1990).

3. TOWARD DEVELOPING HEURISTIC SOLUTION ALGORITHMS

3.1 Developing a Node–Node Adjacency Matrix

Let us define $h^{m,w}_{a\in r^k_{ij}}$ as costs for taking link a on mode k to purchase B amounts of item m at location j including purchasing costs ($b^m_j B^m_j$), initial stopping costs ($d_j s_j$), marginal stopping costs ($d^m_j s^m_j$), time costs ($t^k_{a\in r^k_{ij}}$), and distance-related costs ($d^k_{a\in r^k_{ij}}$) for a traveler type w where link a belongs to the shortest route of mode k chosen between origin i and destination j. That is:

$$h^{m,w}_{a\in r^k_{ij}} = (d_j s_j + b^m_j B^m_j + d^m_j s^m_j) + \alpha^w(t^k_{a\in r^k_{ij}}) + d(d^k_{a\in r^k_{ij}}). \tag{15.8}$$

In other words, $h^{1,2}_{a\in r_{1,1}}$ indicates the cost for purchasing item 1 (10 gallons of gasoline) by taking link a, which is within the shortest route connecting origin 1 to destination 2. Let us look at our problem again. Suppose that a gas station exists in each of nodes 2, 3 and 4, and a flower shop is available in each of nodes 2 and 4. One pharmacy is located at each of nodes 3 and 4. The unit costs are assumed as:

Gasoline costs:	Gas station at node 2, that is,	$b^1_2 = \$1.5$/gallon
	Gas station at node 3, that is,	$b^1_3 = \$2.0$/gallon
	Gas station at node 4, that is,	$b^1_4 = \$1.3$/gallon
Flower costs:	Flower shop at node 2, that is,	$b^2_2 = \$2.0$/rose
	Flower shop at node 4, that is,	$b^2_4 = \$2.5$/rose
Pharmacy costs:	Pharmacy at node 3, that is,	$b^3_3 = \$1.0$/bottle
	Pharmacy at node 4, that is,	$b^3_3 = \$10.0$/bottle

Ignoring stopping and distance-related costs for the moment (it is easy to include them in the actual estimation, however), and assuming that the traveler's unit time cost (α^w) is \$0.33 per minute, then the travel-time and purchasing costs, by taking various links, are as shown in Figure 15.3. In the figure, link travel-time costs are shown next to each link. Costs for purchasing item m at j are shown in the boxes. For instance, [3^1: \$20] indicates that it costs \$20 to purchase 10 gallons of gasoline ($m = 1$) at location 3. As we can see in the boxes, pseudo node numbers were assigned to each eligible shop with the same node number, but with different superscripts indicating different shops available in node j. For instance, 2^1 is gasoline station located in node 2 and 2^2 indicates a flower shop located in node 2. 2^0. 2^0, 3^0 and 4^0 are nodes with no shops, indicating that they are passing nodes with no stopping for shopping.

For a typical traveler w, travel time and purchasing cost, $h^m_{a\in r^k_{ij}}$ can be rearranged as the node–node adjacency matrix representation of the network given in Figure 15.3, as shown in Table 15.1. In the table, the rows and columns in the matrix correspond to the nodes on the network. A non-zero element in the ith row and jth column in the matrix represents the cost for link travel time and purchasing costs at the end node. A zero element in the matrix indicates that there is no link from node i to node j.

Once the travel time and purchasing costs are arranged as shown in Table 15.1, many efficient solution algorithms exist for this type of problem. Zhan (1997) and Zhan and

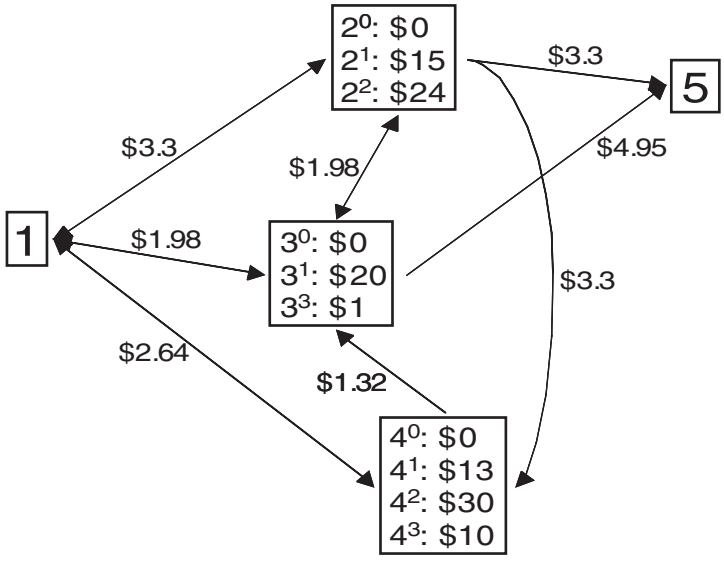


Figure 15.3 Travel-time and purchasing costs at different locations

Table 15.1 Node-node adjacency matrix indicating costs for travel time and purchasing B *amounts of items* m *between pairs of nodes*

	1	2^0	2^1	2^2	3^0	3^1	3^3	4^0	4^1	4^2	4^3	5
1	0	3.3	18.3	27.3	1.98	21.98	2.98	2.64	15.64	32.64	12.64	0
2^0	3.3	0	15.0	24.0	1.98	21.98	2.98	3.3	16.3	33.3	13.3	3.3
2^1	3.3	0	0	24.0	1.98	21.98	2.98	3.3	16.3	33.3	13.3	3.3
2^2	3.3	0	15.0	0	1.98	21.98	2.98	3.3	16.3	33.3	13.3	3.3
3^0	1.98	1.98	16.98	25.98	0	20.0	1.0	0	0	0	0	4.95
3^1	1.98	1.98	16.98	25.98	0	0	1.0	0	0	0	0	4.95
3^3	1.98	1.98	16.98	25.98	0	20.0	0	0	0	0	0	4.95
4^0	2.64	0	0	0	1.32	21.32	2.32	0	13.0	30.0	10.0	0
4^1	2.64	0	0	0	1.32	21.32	2.32	0	0	30.0	10.0	0
4^2	2.64	0	0	0	1.32	21.32	2.32	0	13.0	0	10.0	0
4^3	2.64	0	0	0	1.32	21.32	2.32	0	13.0	30.0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0

Noon (2000) rearranged this type of data structure by means of the Forward Star and the Reverse Star representations. The Forward Star representation of data can be used to efficiently determine the set of arcs going out from any node. On the flip side, the Reverse Star representation is a data structure that provides an efficient means to determine the set of incoming arcs for any node. The Reverse Star representation of a network can be constructed in a manner similar to the Forward Star representation. The only difference is that *incoming* arcs at each node are numbered sequentially. Past research has demonstrated that

the Forward and Reverse Star representations are the most efficient among all existing network data structures for representing a network (Ahuja et al. 1993; Cherkassky et al. 1993; Zhan 1997; Zhan and Noon 2000). Zhan and Noon (1996) evaluated 15 different algorithms for solving this type of problem and recommended the following:

1. The fastest algorithms for computing shortest paths on real road networks are:
   a. the Pallottino graph growth algorithm implemented with two queues (TWO-Q), and
   b. the Dijkstra algorithm implemented with approximate buckets (DIKBA).
2. Avoid the Bellman–Ford–Moore implementations (BF and BFP) and the naive implementation of the Dijkstra algorithm (DIKQ).

3.2 Estimating Spatiotemporal Link Travel Time

In either situation described above, the more accurate link travel time for a given O–D pair can be estimated by using a spatiotemporal function. If the current traffic volume and speed indicate that it would take 10 minutes to travel on link *a* connected from node 1 to node 2 in Figure 15.1, travel time on link *n* connecting from node 2 to node 5 has to be estimated since what we have now is the current link travel time, not the link travel time 10 minutes later. Figure 15.4 illustrates the situation.

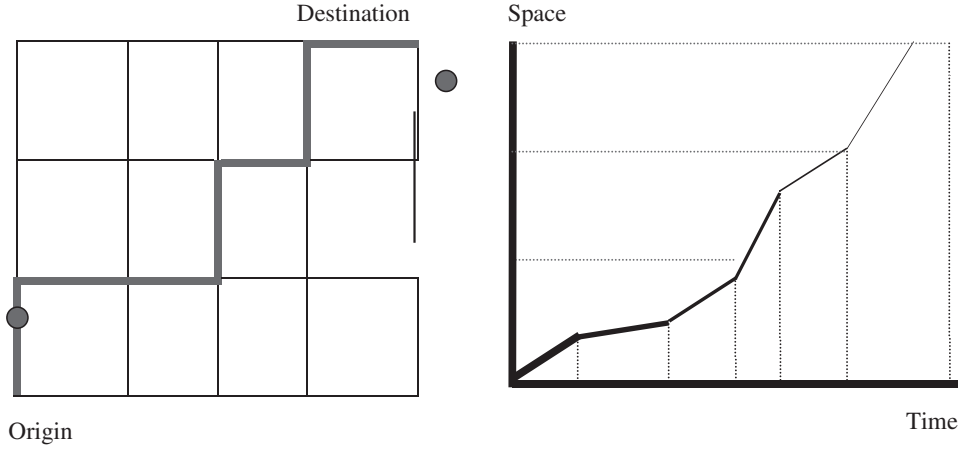


Figure 15.4 Forecasting spatiotemporal link travel time with real-time data

A general spatiotemporal function can be written as follows:

$$t_{a+1} = f(t_a). \tag{15.9}$$

In the absence of real-time link speed and volume, we could construct link travel time based on the past link time data. I would suggest that at least three link travel timetables be made:

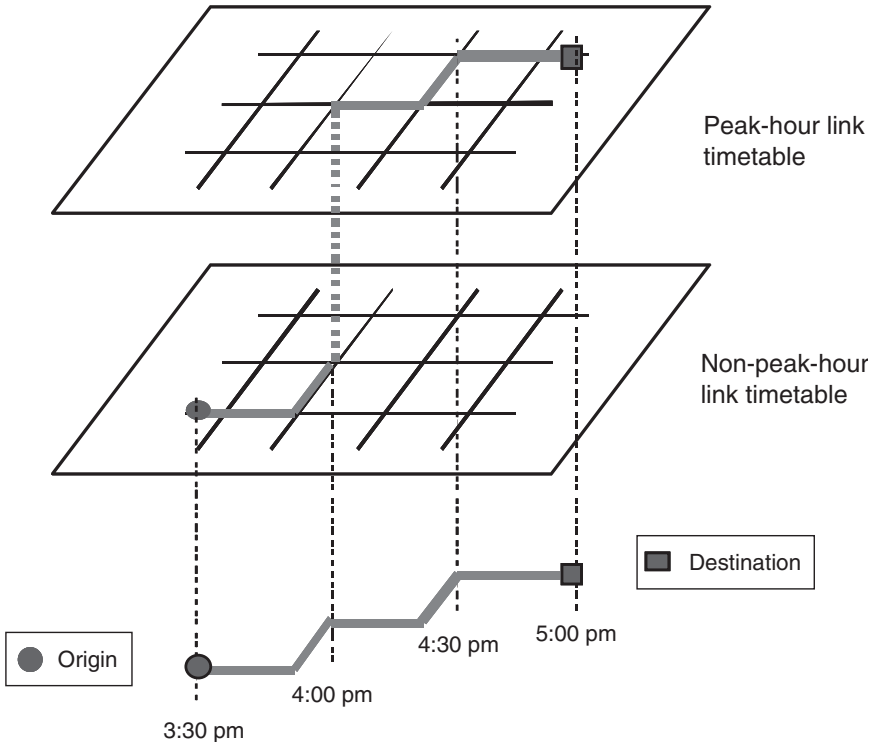


Figure 15.5 Estimating spatiotemporal link travel time with past data

1. peak-hour link travel timetable for weekdays,
2. non-peak-hour link travel timetable for weekdays, and
3. link travel timetable for weekends.

Each link time can then be used in any functions described above, depending on the time of the day and time of the week. When travel between origin and destination is involved in both peak and non-peak hours, both peak and non-peak tables from the past data can be used as shown in Figure 15.5.

4. SUMMARY

In the past, GIS have mostly been used for institutional purposes. Now, rapidly evolving wireless technology makes it possible to utilize GIS for personal productivity such as finding the best routes to follow to a destination, points of interest, friends, current traffic conditions, and a wide variety of other convenience (concierge) services for entertainment, leisure, sports, shopping, travel, local information, community interest, health, education, banking, hobbies, services and so on. An enormous market can be foreseen in this field, including the market for tracking, route-finding and guiding, notification and alert services which would reach $15 billion per year by 2005 as was reported earlier.

While the LBS market seems to be rapidly emerging, there are many basic research issues to be addressed. This chapter has focused on developing functional forms for providing services for multi-modal routing and navigation services. A feasible set of functional forms has been presented. The chapter also addressed issues related to solving the cost functions in a serviceable time, say within 15 seconds, before responding to users, in the hope that this would shed light on the development of heuristic, but efficient solution algorithms in the near future.

NOTES

* This chapter is dedicated to Professor David Boyce, who has made numerous and important contributions to the field of transportation planning and engineering and from whom I have personally benefited immensely. Funding for this research is from the Ministry of Commerce, Industry, and Energy, Republic of Korea through Korea Standards Association, and is gratefully acknowledged.

1. See Jim VanderMeer, 2001, 'Location content drives wireless telecommunications', www.geoplace.com/bg/2001/0201/0201pay.asp.
2. See www.isotc211.org.

REFERENCES

Ahuja, R.K., T.L. Magnanti and J.B. Orlin (1993), *Network Flows: Theory, Algorithms and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

Boyce, D.E., D.H. Lee and B.N. Janson (1998), 'Variational inequality model of ideal dynamic user-optimal route choice', in M.G.H. Bell (ed.), *Transportation Networks: Recent Methodological Advances*', Oxford: Elsevier, pp. 289–302.

Bureau of Transportation Statistics (1998a), 'Delay–volume relations for travel forecasting: based on the 1985 Highway Capacity Manual. Delay functions for uncontrolled road segments', www.bts.gov/tmip/papers/general/dvrt/ch4.htm (accessed 5 March 2003).

Bureau of Transportation Statistics (1998b), 'Delay–volume relations for travel forecasting: based on the 1985 Highway Capacity Manual. Recommendations', www.bts.gov/tmip/papers/general/dvrt/ch10.htm (accessed 5 March 2003).

Cherkassky, B.V., A.V. Goldberg and T. Radzik (1993), 'Shortest paths algorithms: theory and experimental evaluation', Technical Report 93–1480, Computer Science Department, Stanford University, Stanford, CA.

LBS-related issues are taken from http://pulver.com/lbsreport/bissues.html.

Lee, D.H., D.E. Boyce and B.N. Janson (2002), 'Analysis of lane-blocking events with an analytical dynamic traffic assignment model', *Intelligent Transportation Systems*, **6**, 351–74.

Ostensen, O. (2001), 'The expanding agenda of geographic information standards', www.iso.ch/iso/en/commcentre/pdf/geographic0107.pdf, (accessed 5 March 2003).

Suh, S., C.-H. Park and T.J. Kim (1990), 'A highway capacity function in Korea: measurement and calibration', *Transportation Research*, **24A** (3), 176–86.

Zhan, F.B. and C.E. Noon (1996), 'Shortest path algorithms: an evaluation using real road networks', *Transportation Science*, **32** (1), 65–73.

Zhan, F.B. (1997), 'Three fastest shortest path algorithms on real road networks: data structures and procedures', *Journal of Geographic Information and Decision Analysis*, **1** (1), 70–82. The article can be seen in http://publish.uwo.ca/njmalczew/gida 1/Zhan/Zhan-htm, (accessed 5 March 2003).

Zhan, F.B. and C.E. Noon (2000), 'A comparison between label-setting and label-correcting algorithms for computing one-to-one shortest paths', *Journal of Geographic Information and Decision Analysis*, **4** (2), 1–13. The article can be also found in http://publish.uwo.ca/njmalczew/gida_8/Zhan/Zhan_Noon.html, (accessed 5 March 2003).

16. Supply chain supernetworks with random demands

June Dong, Ding Zhang and Anna Nagurney*

1. INTRODUCTION

The study of supply chain network problems through modeling, analysis, and computation has been an active area of research due to the complexity of the relationships among the various decision makers, such as suppliers, manufacturers, distributors, and retailers as well as the practical importance of the topic for the efficient movement of products. The topic is multidisciplinary by nature since it involves particulars of manufacturing, transportation and logistics, retailing/marketing, as well as economics.

Recently, the introduction of electronic commerce has unveiled new opportunities in terms of research and practice in supply chain analysis and management (see, for example, Kuglin and Rosenbaum 2001). Notably, electronic commerce (e-commerce) has had a huge effect on the manner in which businesses order goods and have them transported with the major portion of e-commerce transactions being in the form of business-to-business (B2B). Estimates of B2B e-commerce range from approximately $0.1 trillion to $1 trillion in 1998 and with forecasts reaching as high as $4.8 trillion in 2003 in the United States (see Federal Highway Administration 2000; Southworth 2000). Moreover, it has been emphasized by Handfield and Nichols (1999) and by the National Research Council (2000) that the principal effect of B2B commerce, estimated to be 90 per cent of all e-commerce by value and volume, is in the creation of new and more profitable supply chain networks.

In this chapter, we introduce the first supernetwork supply chain model with random demands. The term *supernetwork* here refers to a network in which decision making regarding transportation and telecommunications tradeoffs (such as those that arise in e-commerce) are modeled in a unified fashion. This concept has been explored to date in supply chains and in financial networks with intermediation, as well as in other applications relevant to the Information Age. For an introduction to the subject, as well as numerous citations, see Nagurney and Dong (2002).

In particular, in this chapter, we build upon the recent work of Nagurney et al. (2002c) who modeled supply chain networks with e-commerce but who assumed that all the underlying functions were known with certainty. Here, in contrast, we consider the more realistic situation in which the demands associated with the product at the retail outlets are now random. This chapter also generalizes the results of Dong et al. (2002), who considered two-tiered supply chains, consisting of manufacturers and retailers, in which the demand at the retail outlets are random. In this chapter, however, we extend that earlier framework to include another tier of decision makers consisting of distributors of the

product and, significantly, we introduce e-commerce, in the form of B2B transactions, between manufacturers and retailers. Nagurney et al. (2002c), in turn, considered dynamic supply chains viewed as multilevel networks, but did not consider e-commerce or random demands.

We emphasize that the interplay of transportation networks with telecommunication networks is a subject that has been studied in the context of other applications, notably, intelligent transportation systems (see Boyce 1988a,b; Boyce et al. 1994; Ran and Boyce 1996, 1999; and the references therein). In addition, the role of such networks as the foundation of our modern economies and societies has also received attention from the regional science communities (see Batten et al. 1995 and Beckmann et al. 1998). In this chapter, we take the perspective of network economics (see Nagurney 1999) in order to formalize the interactions among distinct decision makers on multitiered networks in the form of supply chains who can compete within a tier of nodes but cooperate between tiers of nodes.

The chapter is organized as follows. In Section 2, we develop the supernetwork supply chain model with random demands, derive the optimality conditions of the various decision makers, and establish that the governing equilibrium conditions can be formulated as a finite-dimensional variational inequality problem. We emphasize here that the concept of equilibrium, first explored in a general setting for supply chains by Nagurney et al. (2002a), provides a valuable benchmark against which prices of the product at the various tiers of the network as well as product flows between tiers can be compared.

In Section 3, we study qualitative properties of the equilibrium pattern, and, under reasonable conditions, establish existence and uniqueness results. We also provide properties of the function that enters the variational inequality that allows us to establish convergence of the proposed algorithmic scheme in Section 4. In Section 5, we apply the algorithm to several supply chain examples for the computation of the equilibrium prices and shipments. We conclude the chapter with Section 6, in which we summarize our results and present suggestions for future research.

2. SUPERNETWORK SUPPLY CHAIN MODEL

In this section, we present a supply chain supernetwork model with random demands. Specifically, we consider m manufacturers involved in the production of a homogeneous product which is then shipped to n distributors, who, in turn, ship the product to o retailers. The retailers can transact either physically (in the standard manner) with the distributors or directly, in an electronic manner, with the manufacturers.

We denote a typical manufacturer by i, a typical distributor by j, and a typical retailer by k. Note (Figure 16.1) that the manufacturers are located at the top tier of nodes of the network, the distributors at the middle tier, and the retailers at the third or bottom tier. The links in the supply chain supernetwork in the figure include classical physical links as well as Internet links to allow for e-commerce. The introduction of e-commerce allows for 'connections' that were, heretofore, not possible, such as, for example, those enabling retailers to purchase a product directly from the manufacturers.

The behavior of the various network decision makers represented by the three tiers of nodes in Figure 16.1 is now described. We first focus on the manufacturers. We then turn to the distributors, and, subsequently, to the retailers.

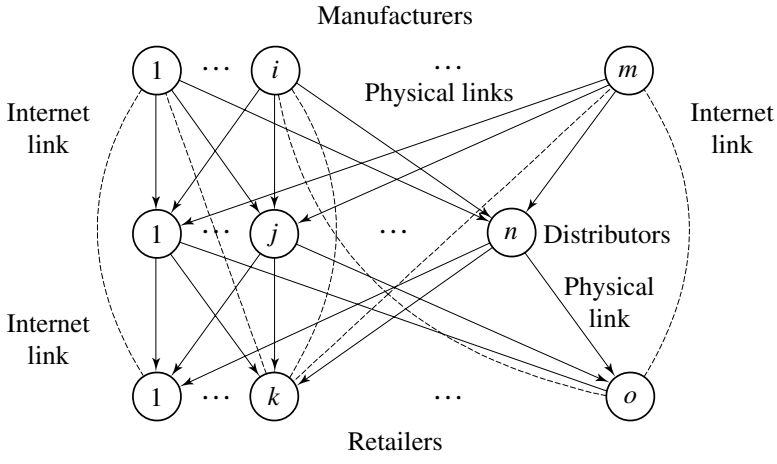


Figure 16.1 Supernetwork structure of the supply chain

2.1 Behavior of the Manufacturers and Their Optimality Conditions

Let q_i denote the non-negative production output of manufacturer i. Group the production outputs of all manufacturers into the column vector $q \in R_+^m$. Here it is assumed that each manufacturer i is faced with a production cost function f_i, which can depend, in general, on the entire vector of production outputs, that is,

$$f_i = f_i(q), \quad \forall i. \tag{16.1}$$

Hence, the production cost of a particular manufacturer can depend not only on his/her production output but also on those of the other manufacturers. This allows us to model competition.

The transaction cost associated with manufacturer i transacting with distributor j is denoted by c_{ij}. The product shipment between manufacturer i and distributor j is denoted by q_{ij}. The product shipments between all pairs of manufacturers and distributors are grouped into the column vector $Q^1 \in R_+^{mn}$. In addition, a manufacturer i may transact directly with the retailer k with this transaction cost associated with the Internet transaction being denoted by c_{ik} and the associated product shipment from manufacturer i to retailer k by q_{ik}. We group these product shipments into the column vector $Q^2 \in R_+^{mo}$.

The transaction cost between a manufacturer and distributor pair and the transaction cost between a manufacturer and retailer may depend upon the volume of transactions between each such pair, and are given, respectively, by:

$$c_{ij} = c_{ij}(q_{ij}), \quad \forall i, j, \tag{16.2a}$$

and

$$c_{ik} = c_{ik}(q_{ik}), \quad \forall i, k. \tag{16.2b}$$

The quantity produced by manufacturer i must satisfy the following conservation of flow equation:

$$q_i = \sum_{j=1}^{n} q_{ij} + \sum_{k=1}^{o} q_{ik}, \tag{16.3}$$

which states that the quantity produced by manufacturer i is equal to the sum of the quantities shipped from the manufacturer to all distributors and to all retailers.

The total costs incurred by a manufacturer i, thus, are equal to the sum of the manufacturer's production cost plus the total transaction costs. His/her revenue, in turn, is equal to the price that the manufacturer charges for the product times the total quantity obtained/purchased of the product from the manufacturer by all the distributors and all the retailers. Let ρ_{1ij}^* denote the price charged for the product by manufacturer i to distributor j who has transacted, and let ρ_{1ik}^* denote the price charged by manufacturer i for the product to the retailer k. Hence, manufacturers can price according to their locations, as to whether the product is sold to the distributor or to the retailers directly, and according to whether the transaction was conducted via the Internet. How these prices are arrived at is discussed later in this section.

Noting the conservation of flow equations (16.3) and the production cost functions (16.1), we can express the criterion of profit maximization for manufacturer i as:

$$\text{Maximize} \sum_{j=1}^{n} \rho_{1ij}^* q_{ij} + \sum_{k=1}^{o} \rho_{1ik}^* q_{ik} - f_i(Q^1, Q^2) - \sum_{j=1}^{n} c_{ij}(q_{ij}) - \sum_{k=1}^{o} c_{ik}(q_{ik}), \tag{16.4}$$

subject to $q_{ij} \geq 0$, for all j, and $q_{ik} \geq 0$, for all k.

The manufacturers are assumed to compete in a non-cooperative fashion. Also, it is assumed that the production cost functions and the transaction cost functions for each manufacturer are continuous and convex. The governing optimization/equilibrium concept underlying non-cooperative behavior is that of Nash (1950, 1951), which states, in this context, that each manufacturer will determine his/her optimal production quantity and shipments, given the optimal ones of the competitors. Hence, the optimality conditions for all manufacturers *simultaneously* can be expressed as the following inequality (see also Gabay and Moulin 1980; Dafermos and Nagurney 1987; Bazarra et al. 1993; and Nagurney 1999): determine the solution $(Q^{1*}, Q^{2*}) \in R_+^{mn+mo}$, which satisfies:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{\partial f_i(Q^{1*}, Q^{2*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q_{ij}^*)}{\partial q_{ij}} - \rho_{1ij}^* \right] \times (q_{ij} - q_{ij}^*)$$

$$+ \sum_{i=1}^{m} \sum_{k=1}^{o} \left[\frac{\partial f_i(Q^{1*}, Q^{2*})}{\partial q_{ik}} + \frac{\partial c_{ik}(q_{ik}^*)}{\partial q_{ik}} - \rho_{1ik}^* \right] \times (q_{ik} - q_{ik}^*) \geq 0, \quad \forall (Q^1, Q^2) \in R_+^{mn+mo}. \tag{16.5}$$

The inequality (16.5), which is a *variational inequality*, has a nice economic interpretation. In particular, from the first term we can infer that, if there is a positive shipment of the product transacted between manufacturer and a distributor, then the marginal cost of production plus the marginal cost of transacting must be equal to the price that the distributor is willing to pay for the product. If the marginal cost of production plus the marginal cost of transacting exceeds that price, then there will be zero volume of flow of the product between the two. The second term in (16.5) has a similar interpretation; in par-

ticular, there will be a positive volume of flow of the product from a manufacturer to a retailer if the marginal cost of production of the manufacturer plus the marginal cost of transacting with the retailer via the Internet is equal to the price the retailers are willing to pay for the product.

2.2 The Behavior of the Distributors and Their Optimality Conditions

The distributors, in turn, are involved in transactions both with the manufacturers since they wish to obtain the product for their inventory, as well as with the retailers. Thus, a distributor conducts transactions both with the manufacturers as well as with the retailers.

Let q_{jk} denote the amount of the product purchased by retailer k from distributor j. Group these shipment quantities into the column vector $Q^3 \in R_+^{no}$.

A distributor j is faced with what is termed a *handling* cost, which may include, for example, the loading/unloading and storage costs associated with the product. Denote this cost by c_j and then, in the simplest case, c_j is a function of $\Sigma_{i=1}^{m} q_{ij}$ and $\Sigma_{k=1}^{o} q_{jk}$ that is, the inventory cost of a distributor is a function of how much of the product he/she has obtained from the various manufacturers and how much of the product he/she has shipped out to the various retailers. However, for the sake of generality, and to enhance the modeling of competition, allow the function to, in general, depend also on the amounts of the product held by other distributors. Therefore, we may write:

$$c_j = c_j(Q^1, Q^3), \quad \forall j. \tag{16.6}$$

Distributor j associates a price with the product, which is denoted by γ_j^*. This price, as will be shown, will also be endogenously determined in the model and will be, given a positive volume of flow between a distributor and any retailer, equal to a clearing-type price. Assuming, that the distributors are also profit-maximizers, the optimization problem of a distributor j is given by:

$$\text{Maximize } \gamma_j^* \sum_{k=1}^{o} q_{jk} - c_j(Q^1, Q^3) - \sum_{i=1}^{m} \rho_{1ij}^* q_{ij} \tag{16.7}$$

subject to:

$$\sum_{k=1}^{o} q_{jk} \leq \sum_{i=1}^{m} q_{ij,} \tag{16.8}$$

and the non-negativity constraints: $q_{ij} \geq 0$, and $q_{jk} \geq 0$, for all i, and k. Objective function (16.7) expresses that the difference between the revenues and the handling cost and the payout to the manufacturers should be maximized. Constraint (16.8) expresses that the retailers cannot purchase more from a distributor than is held in stock.

The optimality conditions of the distributors are now obtained, assuming that each distributor is faced with the optimization problem (16.7), subject to (16.8), and the non-negativity assumption on the variables. Here it is also assumed that the distributors compete in a non-cooperative manner so that each maximizes his/her profits, given the actions of the other distributors. Note that, at this point, we consider that distributors seek to determine not only the optimal amounts purchased by the retailers, but, also, the

amount that they wish to obtain from the manufacturers. In equilibrium, all the shipments between the tiers of network decision makers will have to coincide.

Assuming that the handling cost for each distributor is continuous and convex as are the transaction costs, the optimal $(Q^{1*}, Q^{3*}, \rho_2^*) \in R_+^{mn+no+n}$ satisfy the optimality conditions for all the distributors or, equivalently, the variational inequality:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial c_j(Q^{1*},Q^{3*})}{\partial q_{ij}} + \rho_{1ij}^* - \rho_{2j}^*\right] \times (q_{ij} - q_{ij}^*)$$

$$+ \sum_{j=1}^{n}\sum_{k=1}^{o}\left[-\gamma_j^* + \frac{\partial c_j(Q^{1*},Q^{3*})}{\partial q_{jk}} + \rho_{2j}^*\right] \times (q_{jk} - q_{jk}^*) + \sum_{j=1}^{n}\left(\sum_{i=1}^{m} q_{ij}^* - \sum_{k=1}^{o} q_{jk}^*\right)$$

$$\times (\rho_{2j} - \rho_{2j}^*) \geq 0, \quad \forall Q^1 \in R_+^{mn}, \forall Q^3 \in R_+^{no}, \forall \rho_2 \in R_+^n, \tag{16.9}$$

where ρ_{2j} is the Lagrange multiplier associated with constraint (16.8) for distributor j and ρ_2 is the column vector of all the distributors' multipliers. In this derivation, as in the derivation of inequality (16.5), the prices charged were not variables. The $\gamma^* = (\gamma_1^*, \ldots, \gamma_n^*)$ is the vector of endogenous equilibrium prices in the complete model.

The economic interpretation of the distributors' optimality conditions is now highlighted. From the second term in inequality (16.9), if retailer k purchases the product from a distributor j, that is, if the q_{jk}^* is positive, then the price charged by retailer j, γ_j^*, plus the marginal handling cost, is precisely equal to ρ_{2j}^*, which, from the third term in the inequality, serves as the price to clear the market from distributor j. Also, note that, from the second term, we see that if no product is sold by a particular distributor, then the price associated with holding the product can exceed the price charged to the retailers. Furthermore, from the first term in inequality (16.9), we can infer that, if a manufacturer transacts with a distributor resulting in a positive flow of the product between the two, then the price ρ_{2j}^* is precisely equal to distributor j's payment to the manufacturer, ρ_{1ij}^*, plus his/her marginal cost of handling the product associated with transacting with the particular manufacturer.

2.3 Retailers and Their Optimality Conditions

The retailers, in turn, must decide how much to order from the distributors and from the manufacturers in order to cope with the random demand while still seeking to maximize their profits. A retailer k is also faced with what we term a *handling* cost, which may include, for example, the display and storage cost associated with the product. We denote this cost by c_k and, in the simplest case, we would have that c_k is a function of $s_k = \Sigma_{i=1}^m q_{ik} + \Sigma_{j=1}^n q_{jk}$, that is, the holding cost of a retailer is a function of how much of the product he/she has obtained from the various manufacturers directly and from the various distributors. However, for the sake of generality, and to enhance the modeling of competition, we allow the function to, in general, depend also on the amounts of the product held by other retailers and, therefore, we may write:

$$c_k = c_k(Q^2, Q^3), \quad \forall k. \tag{16.10}$$

Let ρ_{3k} denote the demand price of the product associated with retailer k. We assume that $\hat{d_k}(\rho_{3k})$ is the demand for the product at the demand price of ρ_{3k} at retail outlet k, where $\hat{d_k}(\rho_{3k})$ is a random variable with a density function of $\mathcal{F}_k(x, \rho_{3k})$, with ρ_{3k} serving as a parameter. Hence, we assume that the density function may vary with the demand price. Let P_k be the probability distribution function of $\hat{d_k}(\rho_{3k})$, that is, $P_k(x, \rho_{3k}) = P_k(\hat{d_k} \leq x) = \int_0^x \mathcal{F}_k(x, \rho_{3k})dx$.

Retailer k can sell to the consumers no more than the minimum of his/her supply or demand, that is, the actual sale of k cannot exceed $\min\{s_k, \hat{d_k}\}$. Let

$$\Delta_k^+ \equiv \max\{0, s_k - \hat{d_k}\} \tag{16.11}$$

and

$$\Delta_k^- \equiv \max\{0, \hat{d_k} - s_k\}, \tag{16.12}$$

where Δ_k^+ is a random variable representing the excess supply (inventory), whereas Δ_k^- is a random variable representing the excess demand (shortage).

Note that the expected values of excess supply and excess demand of retailer k are scalar functions of s_k and ρ_{3k}. In particular, let e_k^+ and e_k^- denote, respectively, the expected values: $E(\Delta_k^+)$ and $E(\Delta_k^-)$, that is,

$$e_k^+(s_k, \rho_{3k}) \equiv E(\Delta_k^+) = \int_0^{s_k} (s_k - x)\mathcal{F}_k(x, \rho_{3k})dx, \tag{16.13}$$

$$e_k^-(s_k, \rho_{3k}) \equiv E(\Delta_k^-) = \int_{s_k}^{\infty} (x - s_k)\mathcal{F}_k(x, \rho_{3k})dx. \tag{16.14}$$

Assume that the unit penalty of having excess supply at retail outlet k is λ_k^+ and that the unit penalty of having excess demand is λ_k^-, where the λ_k^+ and the λ_k^- are assumed to be non-negative. Then, the expected total penalty of retailer k is given by:

$$E(\lambda_k^+\Delta_k^+ + \lambda_k^-\Delta_k^-) = \lambda_k^+ e_k^+(s_k, \rho_{3k}) + \lambda_k^- e_k^-(s_k, \rho_{3k}).$$

Assuming, as already mentioned, that the retailers are also profit-maximizers, the expected revenue of retailer k is $E(\rho_{3k} \min\{s_k, \hat{d_k}\})$. Hence, the optimization problem of a retailer k can be expressed as:

$$\text{Maximize } E(\rho_{3k} \min\{s_k, \hat{d_k}\}) - E(\lambda_k^+\Delta_k^+ + \lambda_k^-\Delta_k^-) - c_k(Q^2, Q^3) - \sum_{i=1}^{m} \rho_{1ik}^* q_{ik} - \sum_{j=1}^{n} \gamma_j^* q_{jk}, \tag{16.15}$$

subject to:

$$q_{ik} \geq 0, \quad q_{jk} \geq 0, \text{ for all } i \text{ and } j. \tag{16.16}$$

Objective function (16.15) expresses that the expected profit of retailer k, which is the difference between the expected revenues and the sum of the expected penalty, the handling cost, and the payouts to the manufacturers and to the distributors, should be maximized.

Applying now the definitions of Δ_k^+, and Δ_k^-, we know that min $\{s_k, \hat{d}_k\} = \hat{d}_k - \Delta_k^-$. Therefore, the objective function (16.15) can be expressed as:

$$\begin{aligned}\text{Maximize} \quad & \rho_{3k} d_k(\rho_{3k}) - (\rho_{3k} + \lambda_k^-) e_k^-(s_k, \rho_{3k}) - \lambda_k^+ e_k^+(s_k, \rho_{3k}) \\ & - c_k(Q^2, Q^3) - \sum_{i=1}^{m} \rho_{1ik}^* q_{ik} - \sum_{j=1}^{n} \gamma_j^* q_{jk}, \end{aligned} \tag{16.17}$$

where $d_j(\rho_{3k}) \equiv E(\hat{d}_k)$ is a scalar function of ρ_{3k}.

We now consider the optimality conditions of the retailers assuming that each retailer is faced with the optimization problem (16.15), subject to (16.16), which represents the non-negativity assumption on the variables. Here, we also assume that the retailers compete in a non-cooperative manner so that each maximizes his/her profits, given the actions of the other retailers. Note that, at this point, we consider that retailers seek to determine the amount that they wish to obtain from the manufacturers and from the distributors. First, however, we make the following derivation and introduce the necessary notation:

$$\frac{\partial e_k^+(s_k, \rho_{3k})}{\partial q_{ik}} = \frac{\partial e_k^+(s_k, \rho_{3k})}{\partial q_{jk}} = P_k(s_k, \rho_{3k}) = P_k\left(\sum_{i=1}^{m} q_{ik} + \sum_{j=1}^{n} q_{jk,} \rho_{3k} \right) \tag{16.18}$$

$$\frac{\partial e_k^-(s_k, \rho_{3k})}{\partial q_{ik}} = \frac{\partial e_k^-(s_k, \rho_{3k})}{\partial q_{jk}} = P_k(s_k, \rho_{3k}) - 1 = P_k\left(\sum_{i=1}^{m} q_{ij} + \sum_{j=1}^{n} q_{jk,} \rho_{3k} \right) - 1. \tag{16.19}$$

Assuming that the handling cost for each retailer is continuous and convex, then the optimality conditions for all the retailers satisfy the variational inequality: determine $(Q^{2*}, Q^{3*}) \in R_+^{mo+no}$, satisfying:

$$\begin{aligned} & \sum_{i=1}^{m} \sum_{k=1}^{o} \left\{ \lambda_k^+ P_k(s_k^*, \rho_{3k}) - (\lambda_k^- + \rho_{3k})[1 - P_k(s_k^*, \rho_{3k})] + \frac{\partial c_k(Q^{2*}, Q^{3*})}{\partial q_{ik}} + \rho_{1ik}^* \right\} \\ & \times (q_{ik} - q_{ik}^*) + \sum_{j=1}^{n} \sum_{k=1}^{o} \left\{ \lambda_k^+ P_k(s_k^*, \rho_{3k}) - (\lambda_k^- + \rho_{3k})[1 - P_k(s_k^*, \rho_{3k})] + \frac{\partial c_k(Q^{2*}, Q^{3*})}{\partial q_{jk}} + \gamma_j^* \right\} \\ & \times (q_{jk} - q_{jk}^*) \geq 0, \quad \forall (Q^2, Q^3) \in R_+^{mo+no}. \end{aligned} \tag{16.20}$$

In this derivation, as in the derivation of inequality (16.5), we have not had the prices charged be variables. They become endogenous variables in the complete supply chain supernetwork model.

We now highlight the economic interpretation of the retailers' optimality conditions. In inequality (16.20), we can infer that, if a manufacturer i transacts with a retailer k resulting in a positive flow of the product between the two, then the selling price at retail outlet k, ρ_{3k}, with the probability of $1 - P_k(\Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^n q_{jk}^*, \rho_{3k})$, that is, when the demand is not less than the total order quantity, is precisely equal to the retailer k's payment to the manufacturer, ρ_{1ik}^*, plus his/her marginal cost of handling the product and the penalty of having excess demand with probability of $P_k(\Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^o q_{jk}^*, \rho_{3k})$, (which is the probability when actual demand is less than the order quantity), subtracted from

the penalty of having shortage with probability of $1 - P_k(\Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^o q_{jk}^*, \rho_{3k})$ (when the actual demand is greater than the order quantity).

Similarly, a distributor j transacts with a retailer k resulting in a positive flow of the product between the two, then the selling price at retail outlet k, ρ_{3k}, with the probability of $1 - P_k(\Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^n q_{jk}^*, \rho_{3k})$, that is, when the demand is not less than the total order quantity, is precisely equal to the retailer k's payment to the manufacturer, γ_j^*, plus his/her marginal cost of handling the product and the penalty of having excess demand with probability of $P_k(\Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^o q_{jk}^*, \rho_{3k})$, (which is the probability when actual demand is less than the order quantity), subtracted from the penalty of having shortage with probability of $1 - P_k(\Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^o q_{jk}^*, \rho_{3k})$ (when the actual demand is greater than the order quantity) .

2.4 Equilibrium Conditions

We now turn to a discussion of the market equilibrium conditions. Subsequently, we construct the equilibrium conditions for the entire supply chain.

The equilibrium conditions associated with the transactions that take place between the retailers and the consumers are the stochastic economic equilibrium conditions, which, mathematically, take on the following form: For any retailer k; $k = 1, \ldots, o$:

$$\hat{d}_k(\rho_{3k}^*) \begin{cases} \leq \Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^o q_{jk}^* & \textbf{a.e.}, \quad \text{if} \quad \rho_{3k}^* = 0 \\ = \Sigma_{i=1}^m q_{ik}^* + \Sigma_{j=1}^o q_{jk}^* & \textbf{a.e.}, \quad \text{if} \quad \rho_{3k}^* = 0, \end{cases} \tag{16.21}$$

where **a.e.** means that the corresponding equality or inequality holds almost everywhere.

Conditions (16.21) state that, if the demand price at outlet k is positive, then the quantities purchased by the retailer from the manufacturers and from the distributors in the aggregate is equal to the demand, with exceptions of zero probability. These conditions correspond to the well-known economic equilibrium conditions (see Nagurney 1999 and the references therein). Related equilibrium conditions, but without electronic transactions allowed, were proposed in Dong et al. (2002).

Equilibrium conditions (16.21) are equivalent to the following variational inequality problem, after taking the expected value and summing over all retailers k: determine $\rho_3^* \in R_+^o$ satisfying

$$\sum_{k=1}^{o} \left[\sum_{i=1}^{m} q_{ik}^* + \sum_{j=1}^{n} q_{jk}^* - d_j(\rho_{3k}^*) \right] \times (\rho_{3k} - \rho_{3k}^*) \geq 0, \quad \forall \rho_3 \in R_+^o, \tag{16.22}$$

where ρ_3 is the o-dimensional column vector with components: $\{\rho_{31}, \ldots, \rho_{3o}\}$.

2.5 The Equilibrium Conditions of the Supply Chain

In equilibrium, we must have that the sum of the optimality conditions for all manufacturers, as expressed by inequality (16.5), the optimality conditions of the distributors, as expressed by condition (16.9), the optimality conditions for all retailers, as expressed by inequality (16.20), and the market equilibrium conditions, as expressed by inequality (16.22) be satisfied. Hence, the shipments that the manufacturers ship to the retailers must be equal to the shipments that the retailers accept from the manufacturers. In addition,

the shipments shipped from the manufacturers to the distributors must be equal to those accepted by the distributors, and, finally, the shipments from the distributors to the retailers must coincide with those accepted by the retailers. We state this explicitly in the following definition:

Definition 1: Supply chain network equilibrium with random demands The equilibrium state of the supply chain with random demands is one where the product flows between the tiers of the decision makers coincide and the product shipments and prices satisfy the sum of the optimality conditions (16.5), (16.9), and (16.20), and the conditions (16.22). The summation of inequalities (16.5), (16.9), (16.20), and (16.22) (with the prices at the manufacturers, the distributors, and at the retailers denoted, respectively, by their values at the equilibrium, and denoted by ρ_1^*, ρ_2^*, and ρ_3^*), after algebraic simplification, yields the following result:

Theorem 1: Variational inequality formulation A product shipment and price pattern $(Q^{1*}, Q^{2*}, Q^{3*}, \rho_2^*, \rho_3^*) \in \mathcal{K}$ is an equilibrium pattern of the supply chain model according to Definition 1 if and only if it satisfies the variational inequality problem:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1*}, Q^{2*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q_{ij}^*)}{\partial q_{ij}} + \frac{\partial c_j(Q^{1*}, Q^{3*})}{\partial q_{ij}} - \rho_{2j}^*\right] \times (q_{ij} - q_{ij}^*)$$

$$\sum_{i=1}^{m}\sum_{k=1}^{o}\left\{\frac{\partial f_i(Q^{1*}, Q^{2*})}{\partial q_{ik}} + \frac{\partial c_{ik}(q_{ik}^*)}{\partial q_{ik}} + \frac{\partial c_k(Q^{2*}, Q^{3*})}{\partial q_{ik}}\right.$$

$$\left. + \lambda_k^+ P_k(s_k^*, \rho_{3k}^*) - (\lambda_k^- + \rho_{3k}^*)[1 - P_k(s_k^*, \rho_{3k}^*)]\right\} \times (q_{ik} - q_{ik}^*)$$

$$+ \sum_{j=1}^{n}\sum_{k=1}^{o}\left\{\lambda_k^+ P_k(s_k^*, \rho_{3k}^*) - (\lambda_k^- + \rho_{3k}^*)[1 - P_k(s_k^*, \rho_{3k}^*)] + \frac{\partial c_j(Q^{1*}, Q^{3*})}{\partial q_{jk}}\right.$$

$$\left. + \frac{\partial c_k(Q^{2*}, Q^{3*})}{\partial q_{jk}} + \rho_{2j}^*\right\} \times (q_{jk} - q_{jk}^*) + \sum_{j=1}^{n}\left(\sum_{i=1}^{m} q_{ij}^* - \sum_{k=1}^{o} q_{jk}^*\right) \times (\rho_{2j} - \rho_{2j}^*)$$

$$+ \sum_{k=1}^{o}\left[\sum_{j=1}^{n} q_{jk}^* + \sum_{i=1}^{m} q_{ik}^* - d_k(\rho_3^*)\right] \times (\rho_{3k} - \rho_{3k}^*) \geq 0, \quad \forall (Q^1, Q^2, Q^3, \rho_2, \rho_3) \in \mathcal{K}, \tag{16.23}$$

where $\mathcal{K} = [(Q^1, Q^2, Q^3, \rho_2, \rho_3) | (Q^1, Q^2, Q^3, \rho_2, \rho_3) \in R_+^{mn+mo+no+n+o}]$.

For easy reference in the subsequent sections, variational inequality problem (16.23) can be rewritten in standard variational inequality form (see Nagurney 1999) as follows: determine $X^* \in \mathcal{K}$ satisfying:

$$\langle F(X^*), X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K} \equiv R_+^{mn+mo+no+n+o}, \tag{16.24}$$

where $X \equiv (Q^1, Q^2, Q^3, \rho_2, \rho_3)$, $F(X) \equiv (F_{ij}, F_{ik}, F_{jk}, F_j, F_k)_{i=1,\ldots,m; j=1,\ldots,n; k=1,\ldots,o,}$ and the specific components of F are given by the functional terms preceding the multiplication

signs in (16.24). The term $\langle \cdot, \cdot \rangle$ denotes the inner product in N-dimensional Euclidean space.

Note that the variables in the model (and which can be determined from the solution of either variational inequality (16.23) or (16.24)) are: the equilibrium product shipments between manufacturers and the distributors given by Q^{1*}, the equilibrium product shipments transacted electronically between the manufacturers and the retailers denoted by Q^{2*}, and the equilibrium product shipments between the distributors and the retailers given by Q^{3*}, as well as the equilibrium demand prices ρ_3^* and the equilibrium distributor prices γ^*. We now discuss how to recover the prices ρ_1^* associated with the top tier of nodes of the supply chain supernetwork and the prices ρ_2^* associated with the middle tier.

First note that from (16.5), we have that (as already discussed briefly) if $q_{ij}^* > 0$, then the price $\rho_{1ij}^* = [\partial f_i(Q^{1*}, Q^{2*})/\partial q_{ij}] + [\partial c_{ij}(q_{ij}^*)/\partial q_{ij}]$. Also, from (16.5) it follows that if $q_{ik}^* > 0$, then the $\rho_{1ik}^* = [\partial f_i(Q^{1*}, Q^{2*})/\partial q_{ij}] + [\partial c_{ik}(q_{ik}^*)/\partial q_{ik}]$. Hence, the product is priced at the manufacturer's level according to whether it has been transacted physically or electronically; and also according to the distributor or retailer with which the transaction has taken place. On the other hand, from (16.9) it follows that if $q_{jk}^* > 0$, then $\gamma_j^* = \rho_{2j}^* + [\partial c_j(Q^{1*}, Q^{2*})/\partial q_{jk}]$.

3. QUALITATIVE PROPERTIES

In this section, we provide some qualitative properties of the solution to variational inequality (16.23) (equivalently, variational inequality (16.24)). In particular, we derive existence and uniqueness results. We also investigate properties of the function F (cf. (16.24)) that enters the variational inequality of interest here.

Since the feasible set is not compact we cannot derive existence simply from the assumption of continuity of the functions. Nevertheless, we can impose a rather weak condition to guarantee existence of a solution pattern.

Let

$$\mathcal{K}_b = [(Q^1, Q^2, Q^3, \rho_2, \rho_3) | 0 \leq Q^l \leq b_l,\ l = 1, 2, 3;\ 0 \leq \rho_2 \leq b_4;\ 0 \leq \rho_3 \leq b_5], \quad (16.25)$$

where $b = (b_1, \ldots, b_5) \geq 0$ and $Q^l \leq b_l$; $\rho_2 \leq b_4$; $\rho_3 \leq b_5$ means that $q_{ij} \leq b_1$, $q_{ik} \leq b_2$, $q_{jk} \leq b_3$, and $\rho_{2j} \leq b_4$, $\rho_{3k} \leq b_5$, for all i, j, k. Then $\mathcal{K}_b$ is a bounded closed convex subset of $R^{mn+mo+no+n+o}$. Thus, the following variational inequality:

$$\langle F(X^b), X - X^b \rangle \geq 0, \quad \forall X^b \in \mathcal{K}_b, \quad (16.26)$$

admits at least one solution $X^b \in \mathcal{K}_b$, from the standard theory of variational inequalities, since $\mathcal{K}_b$ is compact and F is continuous. Following Kinderlehrer and Stampacchia (1980) (see also Theorem 1.5 in Nagurney 1999), we then have:

Theorem 2 Variational inequality (16.23) admits a solution if and only if there exists a $b > 0$, such that variational inequality (16.26) admits a solution in $\mathcal{K}_b$ with

$$Q^{1b} < b_1, \quad Q^{2b} < b_2, \quad Q^{3b} < b_3, \quad \rho_2^b < b_4, \quad \rho_3 < b_5. \quad (16.27)$$

Theorem 3: Existence Suppose that there exist positive constants M, N, R with $R>0$, such that:

$$\frac{\partial f_i(Q^1, Q^2)}{\partial q_{ij}} + \frac{\partial c_{ij}(q_{ij})}{\partial q_{ij}} + \frac{\partial c_j(Q^1, Q^3)}{\partial q_{ij}} \geq M, \quad \forall Q^1 \text{ with } q_{ij} \geq N, \quad \forall i, j \tag{16.28a}$$

$$\frac{\partial f_i(Q^1, Q^2)}{\partial q_{ik}} + \frac{\partial c_{ik}(q_{ik})}{\partial q_{ik}} + \frac{\partial c_k(Q^2, Q^3)}{\partial q_{jk}} + \lambda_k^+ P_k(s_k, \rho_{3k}) - (\lambda_k^- + \rho_{3k})[1 - P_k(s_k, \rho_{3k})] \geq M,$$
$$\forall Q^2 \text{ with } q_{ik} \geq N, \quad \forall i, k, \tag{16.28b}$$

$$\lambda_k^+ P_k(s_k, \rho_{3k}) - (\lambda_k^- + \rho_{3k})[1 - P_k(s_k, \rho_{3k})] + \frac{\partial c_j(Q^1, Q^3)}{\partial q_{jk}} + \frac{\partial c_k(Q^2, Q^3)}{\partial q_{jk}} \geq M,$$
$$\forall Q^3 \text{ with } q_{jk} \geq N, \quad \forall j, k, \tag{16.28c}$$

and

$$d_k(\rho_{3k}) \leq N, \quad \forall \rho_3 \text{ with } \rho_{3k} \geq R, \quad \forall k. \tag{16.29}$$

Then, variational inequality (16.24) admits at least one solution.

Proof Follows using analogous arguments as the proof of existence for Proposition 1 in Nagurney and Zhao (1993) (see also existence proof in Nagurney et al. 2002a).

Assumptions (16.28a), (16.28b), (16.28c) and (16.29) can be economically justified as follows. In particular, when the product shipment, q_{ij}, between manufacturer i and distributor j, and the shipment, q_{ik}, between manufacturer i and retailer k, are large, one can expect the corresponding sum of the marginal costs associated with the production, transaction, and holding to exceed a positive lower bound, say M. At the same time, the large q_{ij} and q_{ik} causes a greater s_k, which in turn causes the probability distribution $P_k(s_k, \rho_{3k})$ to be close to 1. Consequently, the sum of the last two terms on the left-hand side of (16.28b), $\lambda_k^+ P_k(s_k, \rho_{3k}) - (\lambda_k^- + \rho_{3k})[1 - P_k(s_k, \rho_{3k})]$ is seen to be positive. Therefore, the left-hand sides of (16.28b) and (16.28c), respectively, are greater than or equal to the lower bound M. On the other hand, a high price ρ_{3k} at retailer k will drive the demand at that retailer down, in line with the decreasing nature of any demand function, which ensures (16.29).

We now recall the concept of additive production cost, which was introduced by Zhang and Nagurney (1996) in the stability analysis of dynamic spatial oligopolies, and has also been employed in the qualitative analysis by Nagurney et al. (2000) for the study of spatial economic networks with multicriteria producers and consumers and in the case of supply chains by Nagurney et al. (2002a).

Definition 2: Additive production cost Suppose that for each manufacturer i, the production cost f_i is additive, that is,

$$f_i(q) = f_i^1(q_i) + f_i^2(\bar{q}_i), \tag{16.30}$$

where $f_i^1(q_i)$ is the internal production cost that depends solely on the manufacturer's own output level q_i, which may include the production operation and the facility maintenance, and so on, and $f_i^2(\bar{q}_i)$ is the interdependent part of the production cost that is a function of all the other manufacturers' output levels $\bar{q}_i = (q_1, \cdots, q_{i-1}, q_{i+1}, \cdots, q_m)$ and reflects the impact of the other manufacturers' production patterns on manufacturer i's cost. This interdependent part of the production cost may describe the competition for the resources, consumption of the homogeneous raw materials and so on.

We now explore additional qualitative properties of the vector function F that enters the variational inequality problem. Specifically, we show that F is monotone as well as Lipschitz continuous. These properties are fundamental in establishing the convergence of the algorithmic scheme in the subsequent section.

Lemma 1 Let $g_k(s_k, \rho_{3k})^T = \{P_k(s_k, \rho_{3k}) - \rho_{3k}[1 - P_k(s_k, \rho_{3k})], s_k - \rho_{3k})\}$, where P_k is a probability distribution with the density function of $\mathcal{F}_k(x, \rho_{3k})$. Then $g_k(s_k, \rho_{3k})$ is monotone, that is,

$$\{-\rho'_{3k}[1 - P_k(s'_k, \rho'_{3k})] + \rho''_{3k}[1 - P_k(s''_k, \rho''_{3k})]\} \times (q'_{jk} - q''_{jk})$$
$$+ [s'_k - d_k(\rho'_{3k}) - s''_k + d_k(\rho''_{3k})] \times (\rho'_{3k} - \rho''_{3k}) \geq 0, \quad \forall (s'_k, \rho'_{3k}), (s''_k, \rho''_{3k}) \in R^2_+ \quad (16.31)$$

if and only if $d'_k(\rho_{3k}) \leq -(4\rho_{3k}\mathcal{F}_k)^{-1}[P_k + \rho_{3k}(\partial P_k/\partial \rho_{3k})]^2$.

Proof In order to prove that $g_k(s_k, \rho_{3k})$ is monotone with respect to s_k and ρ_{3k}, we only need to show that its Jacobian matrix is positive semidefinite, which will be the case if all eigenvalues of the symmetric part of the Jacobian matrix are non-negative real numbers.

The Jacobian matrix of g_k is:

$$\nabla g_k(s, \rho_{3k}) = \begin{bmatrix} \rho_{3k}\mathcal{F}_k(s_k, \rho_{3k}) & -1 + P_k(s_k, \rho_{3k}) + \rho_{3k}[\partial P_k(s_k, \rho_{3k})/\partial \rho_{3k}] \\ 1 & -d'_k(\rho_{3k}) \end{bmatrix}, \quad (16.32)$$

and its symmetric part is:

$$\tfrac{1}{2}[\nabla g_k(s_k, \rho_{3k}) + \nabla^T g_k(s_k, \rho_{3k})] =$$
$$\begin{Bmatrix} \rho_{3k}\mathcal{F}_k(s_k, \rho_{3k}), & \tfrac{1}{2}[\rho_{3k}[\partial P_k/\partial \rho_{3k}] + P_k(s_k, \rho_{3k})] \\ \tfrac{1}{2}[\rho_{3k}[\partial P_k/\partial \rho_{3k}] + P_k(s_k, \rho_{3k})], & -d'_k(\rho_{3k}) \end{Bmatrix}. \quad (16.33)$$

The two eigenvalues of (16.33) are:

$$\gamma\min(s_k, \rho_{3k}) = \tfrac{1}{2}\left\{(\rho_{3k}\mathcal{F}_k - d'_k) - \sqrt{\left[(\alpha_k\mathcal{F}_k - d'_k)^2 + \left(\rho_{3k}\frac{\partial P_k}{\partial \rho_{3k}} + P_k\right)^2 + 4\rho_{3k}\mathcal{F}_k d'_k\right]}\right\}, \quad (16.34)$$

$$\gamma\max(s_k, \rho_{3k}) = \tfrac{1}{2}\left\{(\rho_{3k}\mathcal{F}_k - d'_k) + \sqrt{\left[(\rho_{3k}\mathcal{F}_k - d'_k)^2 + \left(\rho_{3k}\frac{\partial P_k}{\partial \rho_{3k}} + P_k\right)^2 + 4\rho_{3k}\mathcal{F}_k d'_k\right]}\right\}. \quad (16.35)$$

Moreover, since what is inside the square root in both (16.34) and (16.35) can be re-written as:

$$(\rho_{3k}\mathcal{F}_k + d'_k)^2 + \left(\rho_{3k}\frac{\partial P_k}{\partial \rho_{3k}} + P_k\right)^2$$

and can be seen as being non-negative, both eigenvalues are real. Furthermore, under the condition of the lemma, d'_k *is* non-positive, so the first item in (16.34) and in (16.35) is non-negative. The condition further implies that the second item in (16.34) and in (16.35), the square root part, is not greater than the first item, which guarantees that both eigenvalues are non-negative real numbers.

The condition of Lemma 1 states that the expected demand function of a retailer is a non-increasing function with respect to the demand price and its first-order derivative has an upper bound.

Theorem 4: Monotonicity The function that enters the variational inequality problem (16.24) is monotone, if the condition assumed in Lemma 1 is satisfied for each k; $k = 1,\ldots,o$, and if the following conditions are also satisfied.

Suppose that the production cost functions f_i; $i=1,\ldots,m$, are additive, as defined in Definition 2, and that the f_i^1; $i=1,\ldots,m$, are convex functions. If the c_{ij}, c_{ik}, c_k and c_j functions are convex, for all i, j, k, then the vector function F that enters the variational inequality (16.24) is monotone, that is,

$$\langle F(X') - F(X''), X' - X''\rangle \geq 0, \quad \forall X', X'' \in \mathcal{K}. \tag{16.36}$$

Proof Let $X' = (Q^{1\prime}, Q^{2\prime}, Q^{3\prime}, \rho'_2, \rho'_3)$, $X'' = (Q^{1\prime\prime}, Q^{2\prime\prime}, Q^{3\prime\prime}, \rho''_2, \rho''_3)$. Then, inequality (16.36) can been seen in the following deduction:

$$\begin{aligned}
&\langle F(X') - F(X''), X' - X''\rangle \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1\prime}, Q^{2\prime})}{\partial q_{ij}} - \frac{\partial f_i(Q^{1\prime\prime}, Q^{2\prime\prime})}{\partial q_{ij}}\right] \times (q'_{ij} - q''_{ij}) \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial c_j(Q^{1\prime}, Q^{3\prime})}{\partial q_{ij}} - \frac{\partial c_j(Q^{1\prime\prime}, Q^{3\prime\prime})}{\partial q_{ij}}\right] \times (q'_{ij} - q''_{ij}) \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial c_{ij}(q'_{ij})}{\partial q_{ij}} - \frac{\partial c_{ij}(q''_{ij})}{\partial q_{ij}}\right] \times (q'_{ij} - q''_{ij}) \\
&+ \sum_{i=1}^{m}\sum_{k=1}^{o}\left[\frac{\partial f_i(Q^{1\prime}, Q^{2\prime})}{\partial q_{ik}} - \frac{\partial f_i(Q^{1\prime\prime}, Q^{2\prime\prime})}{\partial q_{ik}}\right] \times (q'_{ik} - q''_{ik}) \\
&+ \sum_{i=1}^{m}\sum_{k=1}^{o}\left[\frac{\partial c_k(Q^{2\prime}, Q^{3\prime})}{\partial q_{ik}} - \frac{\partial c_k(Q^{2\prime\prime}, Q^{3\prime\prime})}{\partial q_{ik}}\right] \times (q'_{ik} - q''_{ik}) \\
&+ \sum_{i=1}^{m}\sum_{k=1}^{o}\left[\frac{\partial c_{ik}(q'_{ik})}{\partial q_{ik}} - \frac{\partial c_{ik}(q''_{ik})}{\partial q_{ik}}\right] \times (q'_{ik} - q''_{ik})
\end{aligned}$$

$$+\sum_{i=1}^{m}\sum_{k=1}^{o}[\lambda_k^+ P_k(s'_k,\rho'_{3k}) - \lambda_k^+ P_k(s''_k,\rho''_{3k})]\times(q'_{ik}-q''_{ik})$$
$$+\sum_{i=1}^{m}\sum_{k=1}^{o}\{-\lambda_k^-[1-P_k(s'_k,\rho'_{3k})]+\lambda_k^-[1-P_k(s''_k,\rho''_{3k})]\}\times(q'_{ik}-q''_{ik})$$
$$+\sum_{i=1}^{m}\sum_{k=1}^{o}\{-\rho'_{3k}[1-P_k(s'_k,\rho'_{3k})]+\rho''_{3k}[1-P_k(s''_k,\rho''_{3k})]\}\times(q'_{ik}-q''_{ik})$$
$$+\sum_{j=1}^{n}\sum_{k=1}^{o}\left[\frac{\partial c_j(Q^{1\prime},Q^{3\prime})}{\partial q_{jk}}-\frac{\partial c_j(Q^{1\prime\prime},Q^{3\prime\prime})}{\partial q_{jk}}\right]\times(q'_{jk}-q''_{jk})$$
$$+\sum_{j=1}^{n}\sum_{k=1}^{o}\left[\frac{\partial c_k(Q^{2\prime},Q^{3\prime})}{\partial q_{jk}}-\frac{\partial c_k(Q^{2\prime\prime},Q^{3\prime\prime})}{\partial q_{jk}}\right]\times(q'_{jk}-q''_{jk})$$
$$+\sum_{j=1}^{n}\sum_{k=1}^{o}[\lambda_k^+ P_k(s'_k,\rho'_{3k})] - da_k^+ P_k(s''_k,\rho''_{3k})]\times(q'_{jk}-q''_{jk})$$
$$+\sum_{j=1}^{m}\sum_{k=1}^{o}\{-\lambda_k^-[1-P_k(s'_k,\rho'_{3k})+\lambda_k^-[1-P_k(s''_k,\rho''_{3k})]\}\times(q'_{jk}-q''_{jk})$$
$$+\sum_{j=1}^{n}\sum_{k=1}^{o}\{-\rho'_{3k}[1-P_k(s'_k,\rho'_{3k})]+\rho''_{3k}[1-P_k(s''_k,\rho''_{3k})]\}\times(q'_{jk}-q''_{jk})$$
$$+\sum_{k=1}^{o}[s'_k-d_k(\rho'_{3k})-s''_k+d_k(\rho''_{3k})]\times(\rho'_{3k}-\rho''_{3k})$$
$$=(I)+(II)+(III)+\cdots+(XV). \tag{16.37}$$

Since the f_i; $i=1,\ldots,m$, are additive, and the f_i^1; $i=1,\ldots,m$, are convex functions, we have

$$(I)+(IV)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i^1(Q^{1\prime},Q^{2\prime})}{\partial q_{ij}}-\frac{\partial f_i^1(Q^{1\prime\prime},Q^{2\prime\prime})}{\partial q_{ij}}\right]\times(q'_{ij}-q''_{ij})$$
$$+\sum_{k=1}^{o}\left[\frac{\partial f_i(Q^{1\prime},Q^{2\prime})}{\partial q_{ik}}-\frac{\partial f_i(Q^{1\prime\prime},Q^{2\prime\prime})}{\partial q_{ik}}\right]\times(q'_{ik}-q''_{ik})\geq 0. \tag{16.38}$$

The convexity of c_j, $\forall j$; c_{ij}, $\forall i,j$; c_k, $\forall k$, and c_{ik}, $\forall i,k$ gives, respectively,

$$(II)+(X)=\sum_{j=1}^{n}\sum_{i=1}^{m}\left[\frac{\partial c_j(Q^{1\prime},Q^{3\prime})}{\partial q_{ij}}-\frac{\partial c_j(Q^{1\prime\prime},Q^{3\prime\prime})}{\partial q_{ij}}\right]\times(q'_{ij}-q''_{ij})$$
$$\sum_{k=1}^{o}\left[\frac{\partial c_j(Q^{1\prime},Q^{3\prime})}{\partial q_{jk}}-\frac{\partial c_j(Q^{1\prime\prime},Q^{3\prime\prime})}{\partial q_{jk}}\right]\times(q'_{jk}-q''_{jk})\geq 0 \tag{16.39}$$

$$(III)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial c_{ij}(q'_{ij})}{\partial q_{ij}}-\frac{\partial c_{ij}(q''_{ij})}{\partial q_{ij}}\right]\times(q'_{ij}-q''_{ij})\geq 0 \tag{16.40}$$

$$(V)+(XI)=\sum_{k=1}^{o}\left\{\sum_{i=1}^{m}\left[\frac{\partial c_k(Q^{2\prime},Q^{3\prime})}{\partial q_{ik}}-\frac{\partial c_k(Q^{2\prime\prime},Q^{3\prime\prime})}{\partial q_{ik}}\right]\times(q'_{ik}-q''_{ik})\right.$$
$$\left.+\sum_{j=1}^{n}\left[\frac{\partial c_k(Q^{2\prime},Q^{3\prime})}{\partial q_{jk}}-\frac{\partial c_k(Q^{2\prime\prime},Q^{3\prime\prime})}{\partial q_{jk}}\right]\times(q'_{jk}-q''_{jk})\right\}\geq 0 \tag{16.41}$$

$$(VI)=\sum_{i=1}^{m}\sum_{k=1}^{o}\left[\frac{\partial c_{ik}(q'_{ik})}{\partial q_{ik}}-\frac{\partial c_{ik}(q''_{ik})}{\partial q_{ik}}\right]\times(q'_{ik}-q''_{ik})\geq 0. \tag{16.42}$$

Since the probability function P_k is an increasing function with respect to s_k, for all k, and $s_k=\Sigma_{i=1}^{m}\,q_{ik}+\Sigma_{j=1}^{n}\,q_{jk}$, hence, we have the following:

$$(VII)+(XII)=\sum_{k=1}^{o}[\lambda_k^{+}P_k(s'_k,\rho'_{3k})-a_k^{+}P_k(s''_k,\rho''_{3k})]\times(s'_k-s''_k)\geq 0 \tag{16.43}$$

$$(VIII)+(XIII)+\sum_{k=1}^{o}\{-\lambda_k^{-}[1-P_k(s'_k,\rho'_{3k})+\lambda_k^{-}[1-P_k(s''_k,\rho''_{3k})]\}\times(s'_k-s''_k)\geq 0. \tag{16.44}$$

Since for each k, applying Lemma 1, we can see that $g_k(s_k, \rho_{3k})$ is monotone, hence, we have:

$$(IX)+(XIV)+(XV)=\sum_{k=1}^{o}\{-\rho'_{3k}[1-P_k(s'_k,\rho'_{3k})]+\rho''_{3k}[1-P_k(s''_k,\rho''_{3k})]\}\times(s'_k-s''_k)$$
$$+\sum_{k=1}^{o}[s'_k-d_k(\rho'_{3k})-s''_k+d_k(\rho''_{3k})]\times(\rho'_{3k}-\rho''_{3k})\geq 0. \tag{16.45}$$

Therefore, we conclude that (16.37) is non-negative in $\mathcal{K}$. The proof is complete.

Theorem 5: Strict monotonicity The function that enters the variational inequality problem (16.24) is strictly monotone, if the conditions mentioned in Lemma 1 for $g_k(s_k, \rho_{3k})$ are satisfied strictly for all k and if the following conditions are also satisfied.

Suppose that the production cost functions f_i; $i=1,\ldots,m$, are additive, as defined in Definition 2, and that the f_i^1; $i=1,\ldots,m$, are strictly convex functions. If the c_{ij}, c_{ik}, c_k and c_j functions are strictly convex, for all i, j, k, then the vector function F that enters the variational inequality (16.24) is strictly monotone, that is,

$$\langle F(X')-F(X''),\, X'-X''\rangle>0, \quad \forall X',\, X''\in\mathcal{K}. \tag{16.46}$$

Theorem 6: Uniqueness Under the conditions indicated in Theorem 5, the function that enters the variational inequality (16.24) has a unique solution in $\mathcal{K}$.

From Theorem 6 it follows that, under the above conditions, the equilibrium product shipment pattern between the manufacturers and the retailers, as well as the equilibrium price pattern at the retailers, is unique.

Theorem 7: Lipschitz continuity The function that enters the variational inequality problem (16.24) is Lipschitz continuous, that is,

$$\|F(X') - F(X'')\| \leq L\|X' - X''\|, \quad \forall X', X'' \in K, \text{ with } L > 0, \tag{16.47}$$

under the following conditions:

(i) each f_i; $i = 1,\ldots,m$, is additive and has a bounded second-order derivative;
(ii) the c_{ij}, c_{ik}, c_k, and c_j have bounded second-order derivatives, for all i, j, k.

Proof Since the probability function P_k is always less than or equal to 1, for each retailer k the result is direct by applying a mid-value theorem from calculus to the vector function F that enters the variational inequality problem (16.24).

4. THE ALGORITHM

In this section, an algorithm is presented which can be applied to solve any variational inequality problem in standard form (see (16.24)), that is:

Determine $X^* \in \mathcal{K}$, satisfying:

$$\langle F(X^*), X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K}. \tag{16.48}$$

The algorithm is guaranteed to converge provided that the function F that enters the variational inequality is monotone and Lipschitz continuous (and that a solution exists). The algorithm is the modified projection method of Korpelevich (1977).

The statement of the modified projection method is as follows, where τ denotes an iteration counter:

Step 0: Initialization Set $X^0 \in \mathcal{K}$. Let $\tau = 1$ and let α be a scalar such that $0 < a \leq 1/L$, where L is the Lipschitz continuity constant (see Korpelevich 1977) (see (16.47)).

Step 1: Computation Compute $\bar{X}^\tau$ by solving the variational inequality subproblem:

$$\langle \bar{X}^\tau + aF(X^{\tau-1}) - X^{\tau-1}, X - \bar{X}^\tau \rangle \geq 0, \quad \forall X \in \mathcal{K}. \tag{16.49}$$

Step 2: Adaptation Compute X^τ by solving the variational inequality subproblem:

$$\langle X^\tau + aF(\bar{X}^\tau) - X^{\tau-1}, X - X^\tau \rangle \geq 0, \quad \forall X \in \mathcal{K}. \tag{16.50}$$

Step 3: Convergence verification If $\max |X_l^\tau - X_l^{\tau-1}| \leq \varepsilon$, for all l, with $\varepsilon > 0$, a prespecified tolerance, then stop; else, set $\tau =: \tau + 1$, and go to Step 1.

We now state the convergence result for the modified projection method for this model.

Theorem 8: Convergence Assume that the function that enters the variational inequality (16.23) (or (16.24)) has at least one solution and satisfies the conditions in Theorem 4 and in Theorem 7. Then the modified projection method described above converges to the solution of the variational inequality (16.23) or (16.24).

Proof According to Korpelevich (1977), the modified projection method converges to the solution of the variational inequality problem of the form (16.24), provided that the function F that enters the variational inequality is monotone and Lipschitz continuous and that a solution exists. Existence of a solution follows from Theorem 3. Monotonicity follows Theorem 5. Lipschitz continuity, in turn, follows from Theorem 7.

We emphasize that, in view of the fact that the feasible set $\mathcal{K}$ underlying the supply chain supernetwork model with random demands is the non-negative orthant, the projection operation encountered in (16.49) and (16.50) takes on a very simple form for computational purposes. Indeed, the product shipments as well as the product prices at a given iteration in both (16.49) and in (16.50) can be computed exactly and in closed form.

5. NUMERICAL EXAMPLES

In this section, we apply the modified projection method to six numerical examples. The algorithm was implemented in FORTRAN and the computer system used was a DEC Alpha system located at the University of Massachusetts at Amherst. The convergence criterion used was that the absolute value of the product shipments and prices between two successive iterations differed by no more than 10^{-4}. The parameter a in the modified projection method (see (16.49) and (16.50)) was set to 0.01 for all the examples.

In all the examples, we assumed that the demands associated with the retail outlets followed a uniform distribution. Hence, we assumed that the random demand, $\hat{d_k}(\rho_{3k})$, of retailer k, is uniformly distributed in $[0, b_k/\rho_{3k}]$, $b_k > 0$; $k = 1, \ldots, o$. Therefore,

$$P_k(x, \rho_{3k}) = \frac{x\rho_{3k}}{b_k}, \tag{16.51}$$

$$\mathcal{F}_j(x, \rho_{3k}) = \frac{\rho_{3k}}{b_k}, \tag{16.52}$$

$$d_k(\rho_{3k}) = E(\hat{d_k}) = 0.5\,\frac{b_k}{\rho_{3k}}; \quad k = 1, \ldots, o. \tag{16.53}$$

It is easy to verify that the expected demand function $d_k(\rho_{3k})$ associated with retailer k is a decreasing function of the price at the demand market.

The modified projection method was initialized as follows: all variables were set to zero, except for the initial retail prices ρ_{3k} which were set to 1 for all retailers k.

Example 1 The first and subsequent two numerical supply chain examples with e-commerce consisted of two manufacturers, two distributors, and two retailers, as depicted in Figure 16.2.

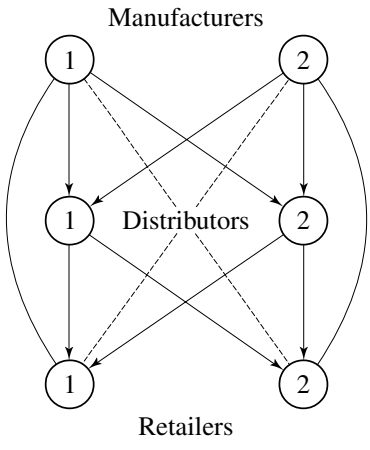


Figure 16.2 Supply chain supernetwork structure for Examples 1 to 3

The data for this example were constructed for easy interpretation purposes. The production cost functions for the manufacturers were given by:

$$f_1(q) = 2.5q_1^2 + q_1q_2 + 2q_1, \quad f_2(q) = 2.5q_2^2 + q_1q_2 + 2q_2.$$

The transaction cost functions faced by the manufacturers and associated with transacting with the distributors were given by:

$$c_{ij}(q_{ij}) = 0.5q_{ij}^2 + 3.5q_{ij}, \quad \text{for } i = 1, 2; j = 1, 2.$$

The transaction cost functions faced by the manufacturers but associated with transacting electronically with the retailers were given by:

$$c_{ik}(q_{ik}) = 0.5q_{ik}^2 + 5q_{ik}, \quad \text{for } i = 1, 2; k = 1, 2.$$

The handling costs of the distributors, in turn, were given by:

$$c_j(Q^1, Q^3) = 0.5\left(\sum_{i=1}^{2} q_{ij}\right)^2, \quad \text{for } j = 1, 2,$$

whereas the handling costs of the retailers were given by:

$$c_k(Q^2, Q^3) = 0.5\left(\sum_{j=1}^{2} q_{jk}\right)^2, \quad \text{for } k = 1, 2.$$

The b_ks were set to 100 for both retailers yielding probability distribution functions as in (16.51) and the expected demand functions as in (16.53). The weights associated with the excess supply and excess demand at the retailers were: $\lambda_k^+ = \lambda_k^- = 1$ for $k = 1, 2$. Hence, we assigned equal weights for each retailer for excess supply and for excess demand.

The modified projection method converged and yielded the following equilibrium

pattern: the product shipments between the two manufacturers and the two distributors were: $q_{ij}^* = 0.3697$ for $i = 1, 2; j = 1, 2$, whereas the product shipments transacted electronically between the manufacturers and the retailers were: $q_{ik}^* = 0.3487$ for $i = 1, 2; k = 1, 2$, and, finally, the product shipments between the distributors and the retailers were: $q_{jk}^* = 0.3697$ for $j = 1, 2; k = 1, 2$. The computed equilibrium prices, in turn, were: $\rho_{2j}^* = 15.2301$ for $j = 1, 2$ and $\rho_{3k}^* = 34.5573$ for $k = 1, 2$. The expected demands (see (16.53)) were: $d_1(\rho_{31}^*) = d_2(\rho_{32}^*) = 1.4469$.

Example 2 Example 2 was constructed from Example 1 as follows. We retained all the data as in Example 1, except that we increased b_1 and b_2 from 100 to 1000. This has the interpretation that the expected demand at both retailers increased.

The modified projection method converged and yielded the following equilibrium pattern. The product shipments between the two manufacturers and the two distributors were now: $q_{ij}^* = 0.6974$ for $i = 1, 2; j = 1, 2$, whereas the product shipments transacted electronically between the manufacturers and the retailers were: $q_{ik}^* = 1.9870$ for $i = 1, 2; k = 1, 2$, and, finally, the product shipments between the distributors and the retailers were now: $q_{jk}^* = 0.6973$ for $j = 1, 2; k = 1, 2$. The computed equilibrium prices, in turn, were: $\rho_{2j}^* = 39.8051$ for $j = 1, 2$ and $\rho_{3k}^* = 92.9553$ for $k = 1, 2$. The expected demands increased (as expected) relative to those obtained in Example 1 with $d_1(\rho_{31}^*) = d_2(\rho_{32}^*) = 5.3789$.

Example 3 Example 3 was constructed from Example 2 as follows. We retained all the data as in Example 2, except that now we decreased the transaction costs associated with transacting electronically, where now $c_{ik}(q_{ik}) = q_{ik} + 1$, $i = 1, 2; k = 1, 2$.

The modified projection method converged and yielded the following equilibrium pattern: the product shipments between the two manufacturers and the two distributors were now: $q_{ij}^* = 0.0484$ for $i = 1, 2; j = 1, 2$, whereas the product shipments transacted electronically between the manufacturers and the retailers were: $q_{ik}^* = 2.7418$ for $i = 1, 2; k = 1, 2$, and, finally, the product shipments between the distributors and the retailers were now: $q_{jk}^* = 0.0483$ for $j = 1, 2; k = 1, 2$. The computed equilibrium prices, in turn, were: $\rho_{2j}^* = 39.1269$ for $j = 1, 2$ and $\rho_{3k}^* = 89.4390$ for $k = 1, 2$. Hence, the product shipments between the manufacturers and the retailers increased and the prices at the retailers decreased (relative to those obtained in Example 2).

Example 4 Examples 4 to 6 had the supernetwork structure depicted in Figure 16.3 and consisted of 3 manufacturers, 2 distributors, and 3 retailers.

The production cost functions for the manufacturers were given by:

$$f_1(q) = 2.5q_1^2 + q_1q_2 + 2q_1, \quad f_2(q) = 2.5q_2^2 + q_1q_2 + 2q_2, \quad f_3(q) = 0.5q_3^2 + 0.5q_1q_3 + 2q_3.$$

The transaction cost functions faced by the manufacturers and associated with transacting with the distributors were given by:

$$c_{11}(q_{11}) = 0.5q_{11}^2 + 1.5, \quad c_{12}(q_{12}) = 0.5q_{12}^2 + 3.5,$$

$$c_{21}(q_{21}) = 0.5q_{21}^2 + 5.5, \quad c_{22}(q_{22}) = 0.5q_{22}^2 + 3.5,$$

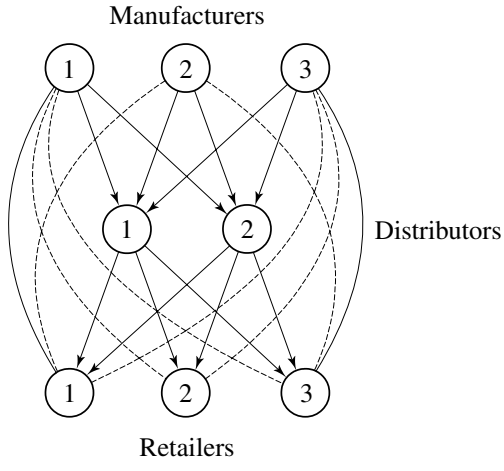


Figure 16.3 Supply chain supernetwork structure for Examples 4 to 6

$$c_{31}(q_{31}) = 0.5q_{31}^2 + 2, \quad c_{32}(q_{32}) = 0.5q_{32}^2 + 2.$$

The transaction cost functions faced by the manufacturers but associated with transacting electronically with the retailers were given by:

$$c_{ik}(q_{ik}) = 0.5q_{ik}^2 + 5q_{ik}, \quad \text{for } i = 1, 2; k = 1, 2, 3.$$

The handling costs of the distributors, in turn, were given by:

$$c_j(Q^1, Q^3) = 0.5\left(\sum_{i=1}^{2} q_{ij}\right)^2, \quad \text{for } j = 1, 2,$$

whereas the handling costs of the retailers were given by:

$$c_k(Q^2, Q^3) = 0.5\left(\sum_{j=1}^{2} q_{jk}\right)^2, \quad \text{for } k = 1, 2, 3.$$

The b_ks were set to 1000 for all three retailers yielding probability distribution functions as in (16.51) and the expected demand functions as in (16.53). The weights associated with the excess supply and excess demand at the retailers were: $\lambda_k^+ = \lambda_k^- = 1$ for $k = 1, 2, 3$.

The modified projection method converged and yielded the following equilibrium pattern: the product shipments between the two manufacturers and the two distributors were:

$$Q^{1*} := q_{11}^* = 1.2656, \; q_{12}^* = 0.0000, \; q_{21}^* = 0.0000, \; q_{22}^* = 0.2543, \; q_{31}^* = 0.0790, \; q_{32}^* = 0.7564,$$

whereas the product shipments transacted electronically between the manufacturers and the retailers were:

$$Q^{2*} := q_{11}^* = q_{12}^* = q_{13}^* = 0.4596; \; q_{21}^* = q_{22}^* = q_{23}^* = 0.7709; \; q_{31}^* = q_{32}^* = q_{33}^* = 4.7730,$$

and, finally, the product shipments between the distributors and the retailers were:

$$Q^{3*} := q_{11}^* = q_{12}^* = q_{13}^* = 0.4590;\ q_{21}^* = q_{22}^* = q_{23}^* = 0.3401.$$

The computed equilibrium prices, in turn, were: $\rho_{21}^* = 21.8951$ and $\rho_{22}^* = 22.2489$ and $\rho_{3k}^* = 73.6386$ for $k = 1, 2, 3$. The expected demands were: $d_1(\rho_{31}^*) = d_2(\rho_{32}^*) = 3(\rho_{33}^*) = 6.7899$.

Example 5 The data for Example 5 were identical to the data in Example 4 except now we modified the weights associated with excess supply and excess demand as follows. We set $\lambda_k^+ = 10$ for $k = 1, 2, 3$ and $\lambda_k^- = 0$ for $k = 1, 2, 3$. Hence, only excess supply (or inventory was penalized).

The modified projection method converged and yielded the following equilibrium pattern: the product shipments between the two manufacturers and the two distributors were:

$$Q^{1*} := q_{11}^* = 1.2941,\ q_{12}^* = 0.0000,\ q_{21}^* = 0.0000,\ q_{22}^* = 0.3040,\ q_{31}^* = 0.0000,\ q_{32}^* = 0.6330,$$

whereas the product shipments transacted electronically between the manufacturers and the retailers were:

$$Q^{2*} := q_{11}^* = q_{12}^* = q_{13}^* = 0.3798;\ q_{21}^* = q_{22}^* = q_{23}^* = 0.6843;\ q_{31}^* = q_{32}^* = q_{33}^* = 4.5133,$$

and, finally, the product shipments between the distributors and the retailers were:

$$Q^{3*} := q_{11}^* = q_{12}^* = q_{13}^* = 0.4347;\ q_{21}^* = q_{22}^* = q_{23}^* = 0.3097.$$

The computed equilibrium prices, in turn, were: $\rho_{21}^* = 20.6190$ and $\rho_{22}^* = 20.9572$ and $\rho_{3k}^* = 79.0529$ for $k = 1, 2, 3$. The expected demands were now: $d_1(\rho_{31}^*) = d_2(\rho_{32}^*) = d_3(\rho_{33}^*) = 6.3249$. Hence, the expected demand at the retailers decreased relative to the values obtained in the preceding example.

Since the penalty of having excess supply, λ_k^+, increased for all k (as compared to those values in Example 4) and the penalty of having excess demand, λ_k^-, decreased, as expected, the shipments between distributors and retailers, Q^{3*}, as well as the shipments between manufacturers and the retailers, Q^{2*}, decreased.

Example 6 The data for Example 6 were identical to the data in Example 5 except now we modified the weights associated with excess supply and excess demand as follows. We set $\lambda_k^+ = 0$ for $k = 1, 2, 3$ and $\lambda_k^- = 10$ for $k = 1, 2, 3$. Hence, only excess demand was penalized.

The modified projection method converged and yielded the following equilibrium pattern: the product shipments between the two manufacturers and the two distributors were:

$$Q^{1*} := q_{11}^* = 1.2130,\ q_{12}^* = 0.0000,\ q_{21}^* = 0.0000,\ q_{22}^* = 0.2017,\ q_{31}^* = 0.2043,\ q_{32}^* = 0.8816,$$

whereas the product shipments transacted electronically between the manufacturers and the retailers were:

$$Q^{2*} := q_{11}^* = q_{12}^* = q_{13}^* = 0.5522;\ q_{21}^* = q_{22}^* = q_{23}^* = 0.8636;\ q_{31}^* = q_{32}^* = q_{33}^* = 5.0435,$$

and, finally, the product shipments between the distributors and the retailers were:

$$Q^{3*} := q_{11}^* = q_{12}^* = q_{13}^* = 0.4691;\ q_{21}^* = q_{22}^* = q_{23}^* = 0.3644.$$

The computed equilibrium prices, in turn, were: $\rho_{21}^* = 23.2677$ and $\rho_{22}^* = 23.6216$ and $\rho_{3k}^* = 68.5502$ for $k = 1, 2, 3$. The expected demands were now: $d_1(\rho_{31}^*) = d_2(\rho_{32}^*) = d_3(\rho_2^*) = 7.2939$.

In this example, because of the decrease in the penalties of having excess supply (inventory) and the increase in the penalties of having excess demand (shortage), the shipments between manufacturers and retailers and the shipments between distributors and retailers increased, as opposed to Example 5.

6. SUMMARY AND CONCLUSIONS

This chapter has developed a three-tiered supply chain network equilibrium model consisting of manufacturers, distributors, and retailers. The model allows for physical transactions between the different tiers of decision makers as well as electronic transactions in the form of B2B commerce between manufacturers and the retailers. In addition, the demands for the product associated with the retailers are no longer assumed to be known with certainty but rather, are random. The model generalized previous supply chain network equilibrium models to include e-commerce, multiple tiers of decision makers as well as random demands within the same framework.

Finite-dimensional variational inequality theory was used to formulate the derived equilibrium conditions, to study the model qualitatively, and also to obtain convergence results for the proposed algorithmic scheme. Finally, numerical examples were presented to illustrate the model and computational procedure.

Future research will include empirical work as well as allowing for other components of the model to also be random.

NOTE

* The research of the first and third authors was supported, in part, by NSF Grant No. IIS-0002647 and that of the third author also, in part, by NSF Grant No. CMS-0085720 and by a 2001 AT&T Industrial Ecology Faculty Fellowship. This support is gratefully acknowledged.

The authors would like to dedicate this chapter to Professor David Boyce whose knowledge, enthusiasm for research, energy, guidance, and friendship have been a true inspiration.

The authors thank the editor and the anonymous reviewer for helpful comments and suggestions.

REFERENCES

Batten, D., J. Casti and R. Thord (eds) (1995), *Networks in Action*, Berlin: Springer-Verlag.

Bazarra, M.S., H.D. Sherali and C.M. Shetty (1993), *Nonlinear Programming: Theory and Algorithms*, New York: John Wiley & Sons.

Beckmann, M.J., B. Johansson, F. Snickars and R. Thord (eds) (1998), *Knowledge and Networks in a Dynamic Economy*, Berlin: Springer-Verlag.

Boyce, D.E. (1988a), 'Route guidance for improving urban travel and location choices', *Transportation Research*, **22A**, 275–81.

Boyce, D.E. (1988b) 'Combining communication and transport technology to improve urban travel choices', in *Information Technology: Social and Spatial Perspective*, I. Orishimo, G.J.D. Hewings and P. Nijkamp (eds), Berlin: Springer-Verlag, pp. 141–52.

Boyce, D.E., A.M. Kirson and J.L. Schofer (1994), 'ADVANCE: The Illinois Dynamic and Route Guidance Demonstration Program', in I. Catling (ed.), *Advanced Technology for Road Transport: IVHS and ATT*, London: Artech House, pp. 247–70.

Dafermos, S. and A. Nagurney (1987), 'Oligopolistic and competitive behavior of spatially separated markets', *Regional Science and Urban Economics*, **17**, 245–54.

Dong, J., D. Zhang and A. Nagurney (2002), 'A supply chain network equilibrium model with random demands', to appear in *European Journal of Operational Research.*

Federal Highway Administration 2000, 'E-commerce trends in the market for freight, Task 3 Freight Trends Scans', Draft, Multimodal Freight Analysis Framework, Office of Freight Management and Operations, Washington, DC.

Gabay, D. and H. Moulin (1980), 'On the uniqueness and stability of Nash equilibria in noncooperative games', A. Bensoussan, P. Kleindorfer and C.S. Tapiero (eds) in *Applied Stochastic Control of Econometrics and Management Science*, Amsterdam: North-Holland.

Handfield, R.B. and E.L. Nichols, Jr. (1999), *Introduction to Supply Chain Management*, Englewood Cliffs, NJ: Prentice-Hall.

Kinderlehrer D. and G. Stampacchia (1980), *An Introduction to Variational Inequalities and Their Application*, New York: Academic Press.

Korpelevich, G.M. (1977), 'The extragradient method for finding saddle points and other problems', *Matekon*, **13**, 35–49.

Kuglin, F.A. and B.A. Rosenbaum (2001), *The Supply Chain Network @ Internet Speed*, New York: American Management Association.

Nagurney, A. (1999), *Network Economics: A Variational Inequality Approach*, 2nd and rev. edn, Dordrecht: Kluwer Academic.

Nagurney, A. and J. Dong (2002), *Supernetworks: Decision-Making for the Information Age*, Cheltenham: Edward Elgar.

Nagurney, A., J. Dong and D. Zhang (2000), 'Multicriteria spatial price networks: statics and dynamics', in *Equilibrium Problems and Variational Models*, P. Daniele, A. Maugeri and F. Giannessi (eds), Dordrecht: Kluwer Academic.

Nagurney, A., J. Dong and D. Zhang (2002a), 'A supply chain network equilibrium model', *Transportation Research E*, **38**, 281–303.

Nagurney, A., K. Ke, J. Cruz, J. Hancock and F. Southworth (2002b), 'Dynamics of supply chains: a multilevel/logistical/informational/financial network perspective', *Environment & Planning B*, **29**, 795–818.

Nagurney, A., J. Loo, J. Dong and D. Zhang (2002c), 'Supply chain networks and electronic commerce: a theoretical perspective', *Netnomics*, **4**, 187–220.

Nagurney, A. and L. Zhao (1993), 'Networks and variational inequalities in the formulation and computation of market disequilibria: the case of direct demand functions', *Transportation Science*, **27**, 4–15.

Nash, J.F. (1950), 'Equilibrium points in *n*-person games', in *Proceedings of the National Academy of Sciences*, **36**, 48–9.

Nash, J.F. (1951), 'Noncooperative games', *Annals of Mathematics,* **54**, 286–98.

National Research Council (2000), *Surviving Supply Chain Integration: Strategies for Small*

Manufacturers, Committee on Supply Chain Integration, Board on Manufacturing and Engineering Design, Commission on Engineering and Technical Systems, Washington, DC.

Ran, B. and D.E. Boyce (1996), *Modeling Dynamic Transportation Networks*, 2nd and rev. edn, Heidelberg: Springer-Verlag.

Ran, B. and D.E. Boyce (1999), 'Modelling dynamic transportation networks with variational inequalities', in *Behavioural and Network Impacts of Driver Information Systems*, R. Emmerink and P. Nijkamp (eds), Aldershot: Ashgate, pp. 53–67.

Southworth, F. (2000), 'E-commerce: implications for freight', Oak Ridge National Laboratory, Oak Ridge, TN.

Zhang, D. and A. Nagurney (1996), 'Stability analysis of an adjustment process for oligopolistic market equilibrium modeled as a projected dynamical system', *Optimization*, **36**, 263–85.

17. An efficient path-based algorithm for a dynamic user equilibrium problem

Huey-Kuo Chen, Hsiao-Chi Peng and Cheng-Yi Chou*

1. INTRODUCTION

The trip distribution and traffic assignment (TDTA) problem characterizes travelers' choice of route with the lowest travel impedance from trip origin to destination given (fixed and known) trip productions and trip attractions. This problem *combines* two submodels into a unified framework, namely, trip distribution and traffic assignment, which appear in the traditional four-stage transportation planning process. The combined model accrues no internal inconsistency between two modules. At equilibrium, the combined model must meet the total number of trips generated from origins and the total number of trips attracted to destinations and, in the meantime, comply with the travelers' behavior of searching for the shortest path from trip origin to destination. To conserve the trips at both ends, a share model that preserves both the trip productions and trip attractions is needed. One of the most commonly used share formulas is based on the entropy maximization principle, which results in a joint entropy distribution/assignment model (JEDA).

Two algorithms, the Frank–Wolfe (FW) method and the Evans method (Evans 1976; Sheffi 1985), have been employed since the JEDA model became available. However, the issue of the efficiency of algorithms has seldom been taken up in the literature. In one of the first publications to state that the FW method is better suited than the Evans algorithm for the JEDA model, Sheffi (1985 p. 187) laid out the following argument:

> Unlike the case with singly constrained models, the double-stage algorithm (also known as the Evans algorithm) has no apparent advantage over the convex combinations (also known as the FW method), when both are applied to the solution of doubly constrained models. The reasons are twofold. Firstly in both algorithms the O–D flows from all origins to all destinations are determined simultaneously. In other words, even in the applications of the convex combinations method, more than one destination can be loaded from each origin at each iteration. Consequently, the convex combinations method is not as inefficient as it is in the case of the singly constrained model. Secondly, the auxiliary problem is a nonlinear problem representing the gravity model. It is considerably more difficult to solve than Hitchcock's transportation problem, which has to be solved as part of the convex combinations method. These considerations imply, then, that the singly constrained model should be solved by using the double-stage algorithm, whereas the doubly constrained model should be solved by using the convex combinations method.

While Sheffi's claim may be correct under certain circumstances, two questions about the FW method for solving the JEDA problem may arise: (i) no theoretical proofs or numeri-

cal examples are provided, and, therefore, the statements may not be conclusive; (ii) the Hitchcock transportation subproblem resulting from linearizing the JEDA problem may not be able to generate an initial (non-degenerate) solution without incurring intra-trips for each traffic centroid that represents both an origin and a destination. As a consequence, when the FW algorithm fails to solve the JEDA problem, no comparison of it can be made with other algorithms in terms of performance efficiency. A different statement regarding the performance of the FW and Evans methods for the JEDA problem was made by Chabini et al. (1994) and Chabini and Florian (1995) suggesting that the gravity model (that is, JEDA problem) can be efficiently solved by the RAS algorithm. However, in the case of the RAS algorithm in its original form (Bachem and Korte 1979; Schneider and Zenios 1990), the problem of incurring intra-trips remains. Chen and Chen (1999) proposed a modified procedure to eliminate the undesired intra-trips for each centroid that represents both an origin and a destination.

In this chapter, a path-based algorithm is proposed and compared with the Evans algorithm (one of the two aforementioned link-based algorithms) for the dynamic user equilibrium problem with doubly constrained origin–destination/departure time/route choice (DUE–DC–OD–D–R), which is a dynamic extension of the JEDA problem. The remaining sections are organized as follows. In Section 2, the DUE–DC–OD–D–R problem is described, including the equilibrium conditions and model formulation. In Section 3, an associated path-based solution algorithm is proposed, with the general scheme of the nested diagonalization (ND) described in Section 3.1, and the path-based algorithm herein named the nested diagonalization–augmented Lagrangian–gradient projection (ND–AL–GP) described in Section 3.2. In Section 4, three numerical examples are provided of the proposed path-based solution algorithm and the Evans algorithm in order to compare the computational efficiency of each method for the DUE–DC–OD–D–R problem. Remarks are given in Section 5.

2. EQUILIBRIUM CONDITIONS AND MODEL FORMULATION FOR THE DUE–DC–OD–D–R PROBLEM

The DUE–DC–OD–D–R problem is a dynamic extension of the JEDA problem, which characterizes travelers' choice of the time-dependent shortest path(s) for each O–D pair and the departure time subject to known trip origins and destinations. In the following, we shall first illustrate the equilibrium conditions and then the corresponding variational inequality (VI) model formulation.

Since the trip decision simultaneously determines the origin/destination and the route, the corresponding dynamic user equilibrium conditions may be characterized through the respective route choice behavior and time-independent O–D demands. For route choice behavior, the conditions state for each O–D pair that the actual route travel times experienced by travelers, regardless the departure times and routes, are equal and minimal. These equilibrium conditions can be mathematically expressed as follows (for summary of notation, see Appendix 17A1):

$$c_p^{rs*}(k)\begin{cases} = \bar{c}^{rs} & \text{if } h_p^{rs*}(k) > 0 \\ \geq \bar{c}^{rs} & \text{if } h_p^{rs*}(k) = 0 \end{cases} \quad \forall r,s,p,k, \tag{17.1}$$

where:

$$\bar{c}^{rs} = \min_{p,k}[c_p^{rs*}(k)] \quad \forall r,s. \tag{17.2}$$

For the time-independent O–D demands, the O–D demand function is assumed to be a strictly decreasing function of the O–D route travel time, $\bar{c}^{rs}$, subtracting origin-based travel time, $\{\pi^r\}$, and destination-based travel time, $\{\pi^s\}$, that is,

$$D^{rs} = f(\bar{c}^{rs} - \pi^r - \pi^s). \tag{17.3}$$

Since the inverse O–D demand function has a strictly monotone mapping, the above formula can be alternatively expressed as follows:

$$\bar{c}^{rs} = (D^{rs})^{-1} + \pi^r + \pi^s \quad \forall r,s. \tag{17.4}$$

The DUE–DC–OD–D–R problem is equivalent to finding a vector $(\mathbf{u}^*,\mathbf{q}^*) \in \Omega_{DUE-DC-OD-D-R}$ such that the following variational inequality problem (VIP) holds:

$$\sum_a \sum_t c_a^*(t)[u_a(t) - u_a^*(t)] - \sum_{rs} (D^{rs*})^{-1}(q^{rs} - q^{rs*}) \geq 0 \quad \forall [u,q] \in \Omega_{DUE-DC-OD-D-R}^* \tag{17.5}$$

where $\Omega_{DUE-DC-OD-D-R}^*$ is a subset of $\Omega_{DUE-DC-OD-D-R}$ with indicator variables being realized at equilibrium, that is, $\{\delta_{apk}^{rs}(t)\} = \{\delta_{apk}^{rs*}(t)\}$. The symbol $\Omega_{DUE-DC-OD-D-R}$ denotes the feasible region, delineated by the following constraints:

Trip production constraint:

$$\sum_s \sum_p \sum_k h_p^{rs}(k) = \bar{q}^r \quad \forall r \tag{17.6}$$

Trip attraction constraint:

$$\sum_r \sum_p \sum_k h_p^{rs}(k) = \bar{q}^s \quad \forall s \tag{17.7}$$

Flow propagation constraints:

$$u_{apk}^{rs}(t) = h_p^{rs}(k)\delta_{apk}^{rs}(t) \quad \forall r,s,a,p,k,t \tag{17.8}$$

Non-negativity constraint:

$$h_p^{rs}(k) \geq 0 \quad \forall r,s,p,k \tag{17.9}$$

Definitional constraints:

$$\sum_p \sum_k h_p^{rs}(k) = q^{rs} \quad \forall r,s \tag{17.10}$$

$$u_a(t) = \sum_{rs} \sum_{p} \sum_{k} h_p^{rs}(k)\,\delta_{apk}^{rs}(t) \quad \forall a,t \tag{17.11}$$

$$c_p^{rs}(k) = \sum_{a} \sum_{t} c_a(t)\,\delta_{apk}^{rs}(t) \quad \forall r,s,p,k \tag{17.12}$$

$$\delta_{apk}^{rs}(t) = \{0,1\} \quad \forall r,s,a,p,k,t \tag{17.13}$$

Equation (17.6) expresses the time-independent trip departures in terms of route flows and states that summing the route flows over all possible destinations s, routes p and intervals k must be equal to the trips departing from origin r. Equation (17.7) expresses the time-independent trip arrivals in terms of route flows and states that summing the route flows over all possible origins r, routes p and intervals k must be equal to the trips arriving at destination s. Equation (17.8) expresses the link flows in terms of the route flows through the use of the indicator variable with which the path flow propagates to each of the links along that path. The path flow into link a must continue to move onto its succeeding link b in route p after the actual link travel time $\tau_a(t)$ has passed. Equation (17.9) ensures that all route inflows are non-negative. Equation (17.10) conserves the O–D demand. Equation (17.11) expresses the link inflow in terms of route departure flows through the incidence relationship. Equation (17.12) expresses the route travel time in terms of the link travel times. Equation (17.13) designates that indicator variables are integer valued, implying that flow deformation is not possible in our model.

Under a certain flow propagation relationship $\{\delta_{apk}^{rs}(t)\} = \{\delta_{apk}^{rs\,*}(t)\}$, equilibrium conditions (17.1) and (17.4) imply VIP (17.5) and vice versa; a proof is provided in Appendix 17A2.

3. NESTED DIAGONALIZATION METHOD

3.1 Nested Diagonalization Method

For the DUE–DC–OD–D–R problem, the general scheme of the solution algorithm essentially consists of three loops. The first (outermost) loop estimates the actual link travel times. The second loop temporarily fixes the other time-dependent link inflows at the current level, yielding an optimization subproblem, that is, the time-dependent JEDA (TD–JEDA) model. In the third (innermost) loop, the resulting TD–JEDA subproblem is solved by any available optimization algorithm. This nested diagonalization (ND) method can be described as follows.

Step 0: Initialization

Step 0.1: Let $m = 0$. Set $\{\tau_a^0(t)\} = \{NINT[c_{a0}(t)]\}$.

Step 0.2: Let $n = 1$. Find the initial feasible solution $\{q^{rs^n}\}, \{u_a^n(t)\}$ and compute the associated *link* travel times $\{c_a^n(t)\}$.

Step 1: First loop operation Let $m = m + 1$. Update the estimated actual link travel times

$$\tau_a^m(t) = NINT[(1-\gamma)\tau_a^{m-1}(t) + \gamma c_a^n(t)], \ \forall a,t \tag{17.14}$$

where $0<\gamma\leq 1$. Construct the corresponding *feasible* time–space network based on the estimated actual link travel times.

Step 2: Second loop operation

Step 2.1: Let $n=1$. Find the *initial* feasible solution $\{q^{rs^n}\}$, $\{u_a^n(t)\}$ and the associated link travel times $\{c_a^n(t)\}$, based on the time–space network constructed by the estimated actual link travel times $\{\tau_a^m(t)\}$.

Step 2.2: Fix the other time–space link inflows at the current level, yielding the following time-*dependent* joint entropy distribution/assignment (TD–JEDA) problem.

$$\min_{(\mathbf{u},\mathbf{q})\in\bar{\Omega}_{DUE-DC-OD-D-R}} \sum_a \sum_t \int_0^{u_a^{n+1}(t)} c_a[\bar{\mathbf{u}}^n \backslash \bar{u}_a^n(t), \omega]d\omega \tag{17.15}$$

$$-\sum_r \sum_s \int_0^{q^{rs^{n+1}}} [D_{rs}(\omega)]^{-1} d\omega$$

where $\bar{\Omega}_{DUE-DC-OD-D-R}$ is a subset of $\Omega_{DUE-DC-OD-D-R}$ which is *defined* by constraints (17.5)–(17.13) with $\{\delta_{apk}^{rs}(t)\}$ realized at $\{\bar{\delta}_{apk}^{rs}(t)\}$.

Step 3: Third loop operation Solve the TD–JEDA *problem* (17.15) for the solution $\{q^{n+1}\}$, $\{u_a^{n+1}(t)\}$ and the associated link travel times $\{c_a^{n+1}(t)\}$ by any available optimization algorithm.

Step 4: Convergence check for the second loop operation If

$$\max_{a,t}\left|\frac{u_a^{n+1}(t)-u_a^n(t)}{u_a^{n+1}(t)}\right|\leq\varepsilon \text{ and } \max_{r,s}\left|\frac{q^{rs^{n+1}}-q^{rs^n}}{q^{rs^{n+1}}}\right|\leq\varepsilon,$$

go to Step 5; otherwise, set $n=n+1$, go to Step 2.2.

Step 5: Convergence check for the first loop operation If

$$\{\tau_a^m(t)\}=\{NINT[c_a^{n+1}(t)]\},$$

stop; the *current* solution is optimal. Otherwise, set $n=n+1$, go to Step 1.

Note that the ND method is not guaranteed to be convergent in solving the DUE–DC–OD–D–R problem due to discretization in time. However, as stated in Appendix 17A2, once the algorithm converges, it will converge to the equilibrium conditions. In Step 3, the TD–JEDA model (17.15) is to be solved. The second term of the objective involves the integral of the inverse demand function, which may appear in a variety of forms. For the sake of demonstration, the principle of entropy maximization is adopted to characterize O–D trip demands. Assuming $(D^{rs})^{-1}=-(1/\zeta)\ln q^{rs}$, where entropy parameter $\zeta>0$, the TD–JEDA problem can be rewritten as follows:

$$\min_{(\mathbf{u},\mathbf{q})\in\bar{\Omega}_{DUE-DC-D-R}} \sum_a \sum_t \int_0^{u_a^{n+1}(t)} c_a[\bar{\mathbf{u}}^n \backslash \bar{u}_a^n(t), \omega]d\omega$$

$$+\frac{1}{\zeta}\sum_{rs}\left(q^{rs^{n+1}}\ln q^{rs^{n+1}}-q^{rs^{n+1}}\right). \tag{17.16}$$

For solving the TD–JEDA model, two categories of solution algorithms can be considered, namely, link- and path-based methods. The link-based algorithms, such as the FW and Evans methods, have been analysed (Chen and Chen 1999) and their detailed steps are shown in Appendix 17A3. In the following section, we focus on the path-based method.

3.2 Path-based Algorithm for the TD–JEDA Problem

The JEDA problem can be treated as a standard user equilibrium problem by means of a supernetwork representation, and the TD–JEDA problem likewise can be treated as the standard time-dependent user equilibrium problem. A supernetwork is usually constructed by adding some dummy nodes and dummy links with artificial link costs so that the additional term other than that for the standard user equilibrium in the objective function can be automatically satisfied when traffic assignment is being carried out. Consider a two-link three-node 'supernetwork' in which O–D pair (r,s) is connected by a dummy link $r\rightarrow s$ with cost $c^{rs}=\left(D^{rs}\right)^{-1}$, as shown in Figure 17.1a. The corresponding time-

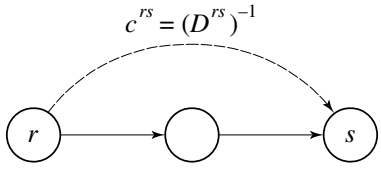


Figure 17.1a Static network

dependent supernetwork can be constructed by adding a dummy time-independent superorigin r connected to an original time-dependent origin $r(.)$ by a dummy time–space link $r\rightarrow r(.)$ with zero link cost, and a time-dependent destination $s(.)$ connected to a superdestination s by a dummy link $s(.)\rightarrow s$ with zero link cost, as shown in Figure 17.1b. The cost associated with the direct dummy link $r\rightarrow s$ for each O–D pair (r,s) is equal to the inverse demand function $(D^{rs})^{-1}$.

The TD–JEDA problem can be solved by two link-based algorithms, namely, the FW and Evans algorithms and a few path-based algorithms such as the gradient projection (GP) method. Here, we address only the path-based approach since it has already been dealt with more efficiently elsewhere (Tatineni et al. 1998; Chen et al. 2001). Note that the TD–JEDA problem in its current form cannot be solved by any path algorithm because trip loadings cannot conserve both trip productions and trip attractions while the assignment procedure is being carried out. However, by regarding either trip productions or trip attractions as a side constraint and by dualizing it into the objective function, the aforementioned difficulty of conserving both trip productions and trip attractions can be overcome. For any side-constrained problem like TD–JEDA, two dual-based algorithms can be employed, namely, the augmented Lagrangian method or the penalty method. For

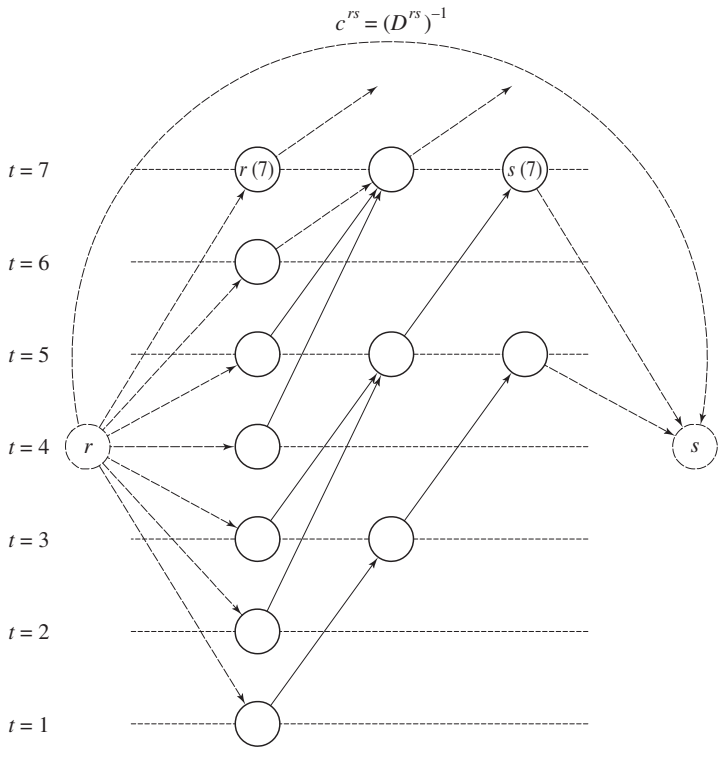


Figure 17.1b Time–space network

the purpose of demonstration, here we employ the augmented Lagrangian method (Luenberger 1984) coupled with the path-based GP method (Jayakrishnan et al. 1994; Chen et al. 2001). The corresponding nested diagonalization-augmented Lagrangian-GP (ND–AGP) method can be formally described as follows.

Step 0: Initialization.

Step 0.1: Network establishment Let $m = 0$. Set the estimated link travel times

$\{\tau_{rr(k)}, \tau_a^m(t), \tau_{s(k)s}\}$, where $\tau_a^m(t) = NINT[c_0(t)]$, $\forall a,t$, $\tau_{rr(k)} \equiv 0$, $\forall r,k$

and $\tau_{s(k)s} \equiv 0$, $\forall s,k$.

Step 0.2: Link interaction treatment.

Step 0.2.1: Let $n = 1$. Compute and reset the initial feasible flow $\{u_a^n(t)\}$.

Step 0.2.2: Fix other time–space link inflows temporarily at the current level, that is, $[\bar{\mathbf{u}}^n \backslash \bar{u}_a^n(t), u_a(t)]$.

Step 0.3: Initialize Lagrange multipliers and positive penalty parameter Let $l = 0$. Find inflow pattern $\{h_p^{rs}(k)^l\}$, $\{u_a(t)^l\} \in \Omega_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}R}$ based on free-flow travel times $\{c_{a_0}(t)\}$, where the constraint set $\bar{\Omega}_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}R}$ is the same as $\bar{\Omega}_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}R}$ except that the

trip attraction constraint is not included. In other words, the singly constrained feasible region $\bar{\Omega}_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}R}$ is delineated by constraints (17.6), (17.8)–(17.13). Find the initial Lagrange multipliers using the following formula:

$$\beta^{sl+1} = \begin{cases} \xi\left[\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^l\right]^2, \text{ if } \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^l\right| > \varepsilon \ \forall s \\ 0 \qquad \text{otherwise} \end{cases} \tag{17.17}$$

$$\xi^l = 1. \tag{17.18.}$$

Step 0.4: Solve the following augmented optimization subproblem by the GP method, yielding flows $\{h_p^{rs}(k)^{l+1}\}$, $\{u_a(t)^{l+1}\}$.

$$\max_{\beta \geq 0} \min_{(\mathbf{u},\mathbf{q}) \in \bar{\Omega}_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}R}} \sum_a \sum_t \int_0^{u_a^{n+1}(t)} c_a[\mathbf{u}^n \backslash u_a^n(t), \omega] d\omega$$

$$- \sum_r \sum_s \int_0^{q^{rs^{n+1}}} [D^{rs}(\omega)]^{-1} d\omega + \sum_s \bar{\beta}^s \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)\right| \tag{17.19}$$

$$+ \sum_s \frac{\xi}{2}\left[\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)\right]^2$$

Step 0.5: Feasibility check for the trip attractions constraint If the link capacity constraint is satisfied, that is, $\max_s \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^{l+1}\right| \leq 0.0001$, let $n = 1$, $\{c_a(t)^n\} = \{c_a(t)^{l+1}\}$, go to Step 1. Otherwise, update Lagrange multipliers $\{\beta^{sl+1}\}$ and penalty parameter ξ^{l+1} as follows:

$$\beta^{sl+1} = \beta^{sl} + \begin{cases} \xi\left[\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^l\right]^2, \text{ if } \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^l\right| > \varepsilon \ \forall s \\ 0 \qquad , \qquad \text{otherwise} \end{cases} \tag{17.20}$$

$$\xi^{l+1} = \begin{cases} \kappa\xi^l, \text{ if } \max_s \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^{l+1}\right| \geq \lambda \max_s \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^l\right| \\ \xi^l, \qquad \text{otherwise} \end{cases} \tag{17.21}$$

$$\kappa = 2 \tag{17.22}$$

$$\lambda = 0.25. \tag{17.23}$$

Set $l = l + 1$ and go to Step 0.4.

Step 1: First loop operation Let $m = m + 1$. Update the estimated actual link travel times as follows:

$$\tau_a^m(t) = NINT[(1-\gamma)\tau_a(t)m^{-1} + \gamma c_a(t)^n] \ \forall a,t \tag{17.24}$$

where γ, $0 < \gamma \leq 1$, is a predetermined constant. Construct the corresponding feasible time–space network based on the estimated actual link travel times $\{\tau_{rr(k)}, \tau_a^m(t), \tau_{s(k)s}\}$.

Step 2: Second loop operation

Step 2.1: Let $n=1$. Compute O–D demands $\{q^{rs^n}\}$ and reset the initial feasible flows $\{h_p^{rs}(k)^n\}$, $\{u_a^n(t)\}$.

Step 2.2: Fix other time–space link inflows temporarily at the current level, that is, $[(\bar{\mathbf{u}}^n \setminus \bar{u}_a^n(t), u_a(t)]$.

Step 3: Third loop operation Let $l=0$. Find inflow pattern $\{h_p^{rs}(k)^l\}$, $\{u_a(t)^l\} \in \Omega_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}R}$ based on free-flow travel times $\{c_{a_0}(t)\}$, where the constraint set $\bar{\Omega}_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}R}$ does not take into account the trip attraction constraint. Find initial Lagrange multipliers using (17.17) and (17.18).

Step 4: Fourth loop operation Solve the augmented Lagrangian problem (17.19) using the GP algorithm, yielding link flows $\{u_a(t)^{l+1}\}$.

Step 5: Feasibility check for the trip attraction constraint If the trip attraction constraint is satisfied, that is, $\max_s \left|\bar{q}^s - \sum_r \sum_p \sum_k h_p^{rs}(k)^{l+1}\right| \leq 0.0001$, let $n=1$, $\{c_a(t)^n\} = \{c_a(t)^{l+1}\}$, go to Step 6. Otherwise, use (17.20) and (17.21) to update Lagrange multipliers $\{\beta^{s^{l+1}}\}$ and penalty parameter ξ^{l+1}. Set $l=l+1$ and go to Step 4.

Step 6: Convergence check for the second loop operation If

$$\max_{a,t} \left| \frac{u_a^{n+1}(t) - u_a^n(t)}{u_a^{n+1}(t)} \right| \leq \varepsilon,$$

continue; otherwise, set $n=n+1$, go to Step 2.2.

Step 7: Convergence check for the first loop operation If

$$\{\tau_a^m(t)\} = \{NINT[c_a^{n+1}(t)]\},$$

stop; the current solution $\{u_a(t)^{(l+1)}\}$ is optimal. Otherwise, set $n=n+1$, go to Step 1.

An in-depth discussion for the augmented Lagrangian method can be seen in Luenberger (1984). In Step 4, the GP method is employed to solve the augmented Lagrangian problem (17.19). Since the tackled problem is convex and the associated cost function is continuous and differentiable, the GP method is guaranteed to converge. Unlike some other path-based algorithms, the GP method does not enumerate all possible paths in its algorithmic procedure and is hence very efficient in computation (Jayakrishnan et al. 1994).

4. NUMERICAL EXAMPLES

4.1 Input Data

Three networks are utilized for testing, as shown in Figures 17.2–4. The shaded nodes represent either origin, destination or both.

The link cost function for network 1 is assumed in equation (17.25), and the link cost functions for networks 2 and 3 are assumed in equation (17.26):

$$c_a(t) = \begin{cases} 1 + 0.01[u_a(t)]^2 + 0.01[x_a(t)]^2 & \forall a \in \{1,3,4,6\} \\ 2.5 + 0.01[u_a(t)]^2 + 0.01[x_a(t)]^2 & \forall a \in \{2.5\} \end{cases} \tag{17.25}$$

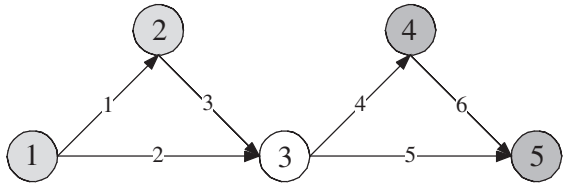


Figure 17.2 Test network 1

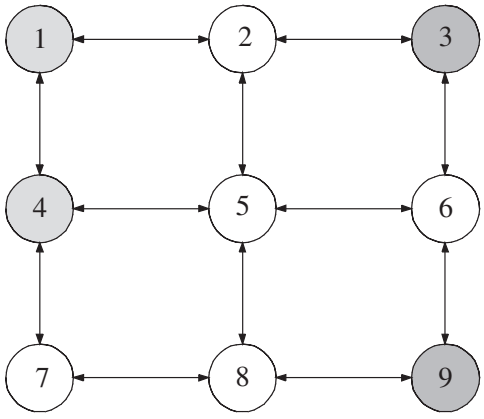


Figure 17.3 Test network 2

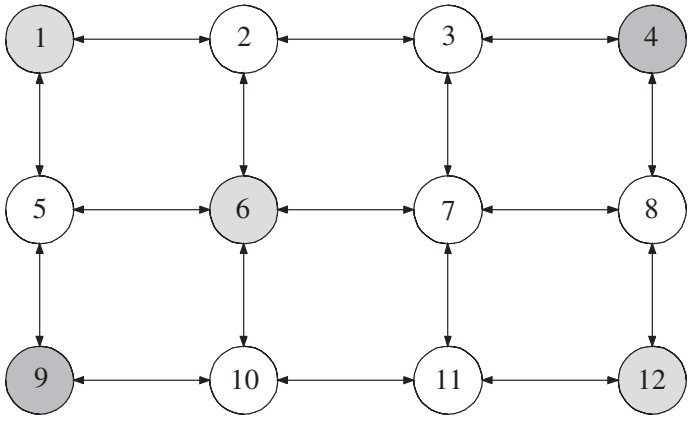


Figure 17.4 Test network 3

$$c_a(t) = 1 + 0.01[u_a(t)]^2 + 0.01[x_a(t)]^2 \quad \forall a. \tag{17.26}$$

Trip productions and attractions are shown in Table 17.1. The entropy parameter is set as $\xi = 1$ and inverse demand function $(D^{rs})^{-1} = -\ln q^{rs}$, $\forall r, s$.

4.2 Results

The equilibrium results for the three test networks were obtained by both ND–mRAS and ND–AL–GP methods. For the sake of demonstration, only the detailed results of network 1 are taken for comparison. Table 17.2 shows that the O–D demands for network 1 obtained by both methods are almost identical and, indeed, in compliance with trip production and trip attraction constraints.

The route travel times for test network 1 are summarized in Table 17.3. For each O–D pair the route travel times are equal and minimal, which satisfies the equilibrium conditions (17.1)–(17.4).

Table 17.4 compares time-independent O–D travel times and inverse demand functions that were obtained for network 1 by both ND–mRAS and ND–AL–GP methods. The deviation of the two results is within 3 per cent, that is, negligible. That means the level of the precision of the two methods is almost identical.

We now proceed to compare the computational efficiency of the two methods for all three test networks (Table 17.5). The computational times used by the ND–AL–GP method are 35.17, 2.57 and 3.35 times faster than those by the ND–mRAS method for networks 1, 2 and 3, respectively. The results imply that, to a great extent, the former method outperforms the latter, mainly due to the fact that the ND–AL–GP method

Table 17.1 Trip productions and trip attractions

Test network	Origin					Destination			
	1	2	4	6	12	3	4	5	9
1	15	20	–	–	–	–	25	10	–
2	30	–	25	–	25	34	–	–	21
3	40	–	–	50	35	–	50	–	75

Table 17.2 O–D demands (network 1)

Origin	Destination					
	4		5		Attractions	
	Evans–mRAS	AL–GP	Evans–mRAS	AL–GP	Evans–mRAS	AL–GP
1	10.71	10.72	4.29	4.28	15.00	15.00
2	14.29	14.28	5.71	5.72	20.00	20.00
Productions	25.00	25.00	10.00	10.00	–	–

Note: Evans–mRAS is the ND–Evans–mRAS method; AL–GP is the ND–AL–GP method.

Table 17.3 Route travel times (network 1)

Path		Departure time							
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
1→2→3→4	Evans–mRAS	3.98	3.98	3.97	3.98	3.98	3.98	3.98	–
	AL–GP	4.06	4.06	4.06	–	4.06	4.06	4.06	–
		(1.52)	(2.16)	(1.52)	–	(0.68)	(2.54)	(0.68)	–
1→3→4	Evans–mRAS	3.98	3.97	3.98	3.98	3.97	–	–	–
	AL–GP	4.06	–	4.06	–	–	4.06	–	–
		(0.34)		(0.34)	–	–	(0.94)	–	–
1→2→3→5	Evans–mRAS	5.01	5.01	5.01	5.01	5.01	–	–	–
	AL–GP	–	–	–	–	–	–	–	–
1→2→3→4→5	Evans–mRAS	5.01	5.01	5.01	5.01	5.01	5.01	5.01	–
	AL–GP	–	–	–	4.98	–	–	–	–
					(2.54)				
1→3→5	Evans–mRAS	5.01	5.01	5.01	5.01	–	–	–	–
	AL–GP	–	4.98	–	–	–	–	–	–
			(1.74)						
1→3→4→5	Evans–mRAS	5.01	5.01	5.01	5.01	5.01	–	–	–
	AL–GP	–	–	–	–	–	–	–	–
2→3→4	Evans–mRAS	2.92	NA	2.91	2.91	2.91	2.91	2.91	2.92
	AL–GP	2.99	–	2.99	–	2.99	2.99	2.99	2.99
		(3.88)			(1.39)	(3.86)	(0.83)	(4.03)	(0.29)
2–3–4–5	Evans–mRAS	3.95	NA	3.94	3.94	3.94	3.94	3.94	3.95
	AL–GP	3.91	–	3.91	–	–	–	–	–
		(2.86)		(2.86)					
2–3–5	Evans–mRAS	NA	NA	3.94	3.95	3.94	3.94	3.95	3.94
	AL–GP	–	–	–	–	–	–	–	–
Total travel cost					155.56				

Note: Numbers in parentheses denote path flows; Evans–mRAS is the ND–Evans–mRAS method; AL–GP is the ND–AL–GP method.

Table 17.4 O–D travel times and inverse demand functions (network 1)

O–D pair	O–D minimal travel time $\bar{c}^{rs} = \min_{p,k}\{c_p^{rs*}(k)\}$ (1)			Inverse demand function $(D^{rs*})^{-1}(2)$			(1)–(2)*		
	Evans–MRAS (3)	AL–GP (4)	((3)–(4)) ×100 /(4)	Evans–mRAS (5)	AL–GP (6)	((5)–(6)) ×100 /(6)	Evans–mRAS (7)	AL–GP (8)	((7)–(8)) ×100 /(8)
1–4	3.98	4.06	−2.0	−2.37	−2.37	0.0	6.35	6.43	1.2
1–5	5.01	4.98	0.6	−1.46	−1.45	0.7	6.47	6.43	0.6
2–4	2.91	2.99	−2.6	−2.66	−2.66	0.0	5.57	5.65	−1.4
2–5	3.94	3.91	0.8	−1.74	−1.74	0.0	5.68	5.65	0.5

Note: * By equation (17.4), $\bar{c}^{rs} = (D^{rs})^{-1} + \pi^r + \pi^s \; \forall r,s$. Evans–mRAS is the ND–Evans–mRAS method; AL–GP is the ND–AL–GP method.

Table 17.5 Computational efficiency of ND–mRAS and ND–AL–GP methods

Performance measures	Test networks								
	Network 1			Network 2			Network 3		
	Evans–mRAS (1)	AL–GP (2)	(1)/(2)	Evans–mRAS (3)	AL–GP (4)	(3)/(4)	Evans–mRAS (5)	AL–GP (6)	(5)/(6)
Number of iterations	21	6	3.5	7	9	0.78	18	11	1.64
Computation time per iteration (sec)	415	41	10.12	372	113	3.29	1475	721	2.05
Computation time (sec)	8717	248	35.17	2606	1016	2.57	26558	7929	3.35

Note: Evans–mRAS is the ND–Evans–mRAS method; AL–GP is the ND–AL–GP method.

requires much less computation time in each iteration for solving the time-dependent JEDA problem.

5. CONCLUSION AND SUGGESTIONS

In this chapter, we developed a thorough analysis of the DUE–DC–OD–D–R problem. (In Appendix 17A2, the equivalency of the VI model and the equilibrium conditions are proved under a certain flow propagation relationship.) We discussed both link- and path-based algorithms. We pointed out that, in the category of link-based algorithms, the FW–S method is apt to fail to provide an initial solution for the JEDA problem, whereas the Evans–mRAS method can cope with the difficulty of producing intra-trips for each 'Centroid'. Moreover, according to Chen and Chen (1999), the Evans–mRAS method outperforms the FW–S method with respect to computation times, which implies that the ND–FW–S may not be worth further study. Consequently, we compared only the ND–Evans–mRAS method with a certain path-based method – called the ND–AL–GP method – for the DUE–DC–OD–D–R problem in this study. Not surprisingly, in our experiments the ND–AL–GP to a large extent outperformed the ND–Evans–mRAS method, in a range of 2.57–35.17 times faster, mainly due to the zigzagging phenomenon, which appears to be less significant for the path-based algorithms. The superiority of the path- over the link-based algorithm may be equally applicable to the user equilibrium problem with singly constrained origin–destination/departure time/route choice (DUE–SC–OD–D–R), and other dynamic travel choice models (Chen 1999) as well. Note also that the computational performance of the ND–AL–GP algorithms may be further improved by adopting more efficient formulas to update Lagrangian multipliers and penalty parameters, which, needless to say, is a subject for future research.

APPENDIX 1 SUMMARY OF NOTATION

Notation in Text

a	link designation
$c_a(t)$	travel time for link a during time interval t
$c_{a_o}(t)$	free-flow travel time for link a during interval t
$c_p^{rs}(k)$	travel time for route p between O–D pair rs during time interval k
$\bar{c}^{rs}$	minimal time-independent travel time for O–D pair rs (subproblem)
π^r	dual variable associated with trip production constraint at origin r
π^s	dual variable associated with trip attraction constraint at destination s
$(D^{rs})^{-1}$	inverse demand function for O–D pair rs
$h_p^{rs}(k)$	flow on route p between O–D pair rs during time interval k
h	vector of route flow
k	time interval designation which usually denotes the departure time interval for a route
p	route designation
$\bar{q}^r$	trip production at origin r
$\bar{q}^s$	trip attraction at destination s
q^{rs}	traffic demand between O–D pair rs
q	vector of q^{rs}
r	origin designation
s	destination designation
t	time interval designation which usually denotes the link entering time interval
$u_a(t)$	inflow on link a during time interval t
$u_{apk}^{rs}(t)$	part of link inflow rate into link a during time interval t associated with path p between O–D pair rs during time interval k
$[\bar{\mathbf{u}}\backslash\bar{u}_a(t), u_a(t)]$	vector of inflows being fixed at current level except inflow on link a during time interval t
$\mathbf{u}$	vector of link inflows
$x_a(t)$	number of vehicles on link a at the beginning of time interval t
β^s	multiplier associated with trip production constraint at destination s
γ	weight associated with the current actual link travel time
$\delta_{apk}^{rs}(t)$	1, if inflow rate on link a during time interval t departs from origin r over route p toward destination s during time interval k; otherwise, 0
$\bar{\delta}_{apk}^{rs}(t)$	incidence variable in the subproblem where actual link travel times are temporarily fixed at the current level
ε	convergence criterion
λ	predetermined constant, $\lambda = 0.25$
ξ	coefficient in the formula for updating Lagrangian multiplier
ζ	coefficient defined for updating Lagrangian multiplier in the entropy formula $\zeta = 1$
$\tau_a(t)$	actual travel time for link a during time interval t
$\Omega_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}RI}$	feasible region associated with the dynamic user equilibrium with doubly constrained origin–destination/departure time/route choice
$\bar{\Omega}_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}RI}$	feasible region associated with the TD–JEDA problem; a subset of $\Omega_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}RI}$ with $\{\delta_{apk}^{rs}(t)\}$ realized at $\{\bar{\delta}_{apk}^{rs}(t)\}$.
$\bar{\Omega}_{DUE\text{–}SC\text{–}OD\text{–}D\text{–}RI}$	feasible region associated with the dynamic user equilibrium problem with singly constrained origin–destination/departure time/route choice
$*$	equilibrium condition.

Notation in Appendices

A^r	balancing factor associated with origin r
B^s	balancing factor associated with destination s
e(.)	vector of equality constraints
f(.)	vector of inequality constraints
$g_p^{rs}(k)$	flow on route p between O–D pair rs during time interval k (subproblem)
$P_a(t)$	inflow on link a during time interval t (subproblem)
r^i	super destination associated with origin r
s^r	super destination associated with origin r
v^{rs}	traffic demand between O–D pair rs (subproblem)
α	step size in the (current solution) updating formula
Ω_v	feasible region associated with the dynamic user equilibrium problem with doubly constrained origin–destination/departure time/route choice (subproblem).

APPENDIX 2 EQUIVALENCE ANALYSIS

The equivalence analysis is similar to those shown in Chen (1999), by stating the following theorem.

Theorem A1 Under a certain flow propagation relationship $\{\delta_{apk}^{rs}(t)\} = \{\delta_{apk}^{rs*}(t)\}$, DUE–DC–OD–D–R equilibrium conditions (17.1)–(17.4) imply VIP (17.5) and vice versa.

Proof of necessity We need to prove that under a certain propagation relationship $\{\delta_{apk}^{rs}(t)\} = \{\delta_{apk}^{rs*}(t)\}$, dynamic user equilibrium conditions (17.1)–(17.4) can be reformulated as VIP (17.5). We first rewrite equilibrium condition (17.1) as follows:

$$\left[c_p^{rs*}(k) - \bar{c}^{rs}\right] h_p^{rs*}(k) = 0 \quad \forall r,s,p,k. \tag{17A2.1}$$

Since $c_p^{rs*}(k) - \bar{c}^{rs}, \geq 0$, $\forall r,s,p,k$, and $h_p^{rs*}(k) \geq 0$, $\forall r,s,p,k$, it implies:

$$\left[c_p^{rs*}(k) - \bar{c}^{rs}\right] h_p^{rs*}(k) \geq 0 \quad \forall r,s,p,k. \tag{17A2.2}$$

Subtracting equation (17A2.1) from equation (17A2.2) results:

$$\left[c_p^{rs*}(k) - \bar{c}^{rs}\right]\left[h_p^{rs}(k) - h_p^{rs*}(k)\right] \geq 0 \quad \forall r,s,p,k. \tag{17A2.3}$$

Summing over r, s, p, k and considering equation (17.4) yields:

$$\sum_{rs}\sum_{p}\sum_{k} c_p^{rs*}(k)\left[h_p^{rs}(k) - h_p^{rs*}(k)\right]$$

$$-\sum_{r}\sum_{s}\left[(D^{rs*})^{-1} + \pi^r + \pi^s\right]\sum_{p}\sum_{k}\left[h_p^{rs}(k) - h_p^{rs*}(k)\right] \geq 0. \tag{17A2.4}$$

By making the substitution of $\sum_{p}\sum_{k} h_p^{rs}(k) = q^{rs}$, it yields:

$$\sum_{rs}\sum_{p}\sum_{k} c_p^{rs*}(k)\left[h_p^{rs}(k) - h_p^{rs*}(k)\right] - \sum_{rs}(D^{rs*})^{-1}(q^{rs} - q^{rs*})$$

$$-\sum_{r}\sum_{s}\pi^r\left[q^{rs} - q^{rs*}\right] - \sum_{r}\sum_{s}\pi^s\left(q^{rs} - q^{rs*}\right) \geq 0. \tag{17A2.5}$$

Since $\sum_{s} q^{rs} = \bar{q}^r$, $\forall r$ and $\sum_{r} q^{rs} = \bar{q}^s$, $\forall s$, we have:

$$\sum_{rs}\sum_{p}\sum_{k}\left[\sum_{a}\sum_{t}c_a^*(t)\delta_{apk}^{rs*}(t)\right]\left[h_p^{rs}(k)-h_p^{rs*}(k)\right]$$

$$-\sum_{rs}(D^{rs*})^{-1}(q^{rs}-q^{rs*})\geq 0. \tag{17A2.6}$$

According to equation (17.11), one obtains:

$$\sum_{a}\sum_{t}c_a^*(t)\left[u_a(t)-u_a^*(t)\right]-\sum_{rs}(D^{rs*})^{-1}(q^{rs}-q^{rs*})\geq 0. \tag{17A2.7}$$

The above inequality is identical to VI (17.5).

Proof of sufficiency We next prove that VIP (17.5) can induce dynamic equilibrium conditions (17.1)–(17.4). If the vector $\mathbf{u}^*\in\Omega^*_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}R}=\{\mathbf{u}\in R^n|\mathbf{f}(\mathbf{u})\geq\mathbf{0},\ \mathbf{e}(\mathbf{u})=\mathbf{0}\}$ is a solution to the VI (17.5) and the gradient $\nabla\mathbf{f}_i(\mathbf{u}^*)$, $\forall i$ such that $\mathbf{f}_i(\mathbf{u}^*)=\mathbf{0}$, $\forall i$ and $\nabla\mathbf{e}_i(\mathbf{u}^*)$, $\forall i$ are linear independent, then according to Tobin and Friesz (1988), $\mathbf{u}^*$ also solves the following nonlinear programming problem (linear objective function):

$$\min_{(\mathbf{u},\mathbf{q})\in\Omega_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}R}}\sum_{a}\sum_{t}c_a(t)\left[u_a(t)-u_a^*(t)\right]-\sum_{rs}(D^{rs*})^{-1}(q^{rs}-q^{rs*}), \tag{17A2.8}$$

where the feasible region $\bar{\Omega}_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}R}$ is a subset of $\Omega_{DUE\text{–}DC\text{–}OD\text{–}D\text{–}R}$ which is delineated by the constraints (17.6)–(17.13) with $\{\delta_{apk}^{rs}(t)\}$ being realized as indicator variables $\{\bar{\delta}_{apk}^{rs}(t)\}$.

The Lagrangian can be constructed by relaxing trip production and trip attraction constraints. Introducing the corresponding dual variables $\{\pi^r\}$ and $\{\pi^s\}$, respectively, we have:

$$L(\mathbf{h},\boldsymbol{\pi})=\sum_{a}\sum_{t}c_a(t)\left[u_a(t)-u_a^*(t)\right]-\sum_{rs}(D^{rs*})^{-1}(q^{rs}-q^{rs*}) \tag{17A2.9}$$

$$+\sum_{r}\pi^r\left[\bar{q}^r-\sum_{s}\sum_{p}\sum_{k}h_p^{rs}(k)\right]+\sum_{s}\pi^s\left[\bar{q}^s-\sum_{r}\sum_{p}\sum_{k}h_p^{rs}(k)\right].$$

The optimality conditions can thus be obtained by taking partial derivatives of the Lagrangian with respect to both primal and dual decision variables. Taking partial derivatives of the Lagrangian with respect to path flows yields:

$$\frac{\partial L(\mathbf{h},\boldsymbol{\pi})}{\partial h_p^{rs}(k)}=c_p^{rs}(k)-(D^{rs})^{-1}-\pi^r-\pi^s\geq 0\ \ \forall r,s,p,k. \tag{17A2.10}$$

By complementarity slackness, we have:

$$h_p^{rs}(k)\left[c_p^{rs}(k)-(D^{rs})^{-1}-\pi^r-\pi^s\right]=0\ \ \forall r,s,p,k. \tag{17A2.11}$$

By using equation (17.4), $\bar{c}^{rs}=(D^{rs})^{-1}+\pi^r+\pi^s$, $\forall r,s$, the above two expressions can be simplified as:

$$c_p^{rs}(k)-\bar{c}^{rs}\geq 0\ \ \forall r,s,p,k \tag{17A2.12}$$

$$h_p^{rs}(k)\left[c_p^{rs}(k)-\bar{c}^{rs}\right]=0\ \ \forall r,s,p,k. \tag{17A2.13}$$

The above two expressions are identical to expression (17.1). In addition, taking partial derivatives of the Lagrangian with respect to dual variables results in equations (17.6) and (17.7). The assumption of independence of the gradients of the binding constraints is a sufficient condition for the Karush–Kuhn–Tucker constraint qualification to be satisfied at $\mathbf{u}^*$. Therefore, by the Karush–Kuhn–Tucker necessity theorem, there exist $\{\pi^r\}$ and $\{\pi^s\}$ such that the equilibrium conditions corresponding to the DUE–DC–OD–D–R problem result. This completes the proof.

APPENDIX 3 LINK-BASED ALGORITHMS

Link-based algorithms can be further differentiated into the Frank–Wolfe (FW) or Evans (double stage) method depending on which linearization or partial linearization technique is adopted to construct the corresponding subproblem.

FW Method

The FW method decomposes the original TD–JEDA problem into a series of linearized subproblems – known as the Hitchcock transportation problem in the operations research literature – and proceeds repetitively through the designated steps of direction search, move size determination and update. The Hitchcock transportation problem in turn can be described as a minimum-cost flow problem over the network. The associated 'Simplex' algorithm begins with initial 'basis' and then improves the quality of the immediate solutions through a pivoting process that introduces a non-basic link with the smallest negative 'penalty' into the basis and moves the link with zero flow out of the basis. The time-dependent Hitchcock transportation subproblem can be formulated as follows:

$$\min \sum_{rs}\sum_{p}\sum_{k}\frac{\partial z(\mathbf{h})}{\partial h_p^{rs}(k)}g_p^{rs}(k)=\sum_{rs}\sum_{p}\sum_{k}\left[c_p^{rs}(k)+\frac{1}{\zeta}\ln q^{rs}\right]g_p^{rs}(k), \tag{17A3.1}$$

subject to:

$$\sum_{s}\sum_{p}\sum_{k}g_p^{rs}(k)=\bar{q}^{r}\ \ \forall r \tag{17A3.2}$$

$$\sum_{r}\sum_{p}\sum_{k}g_p^{rs}(k)=\bar{q}^{s}\ \ \forall s \tag{17A3.3}$$

$$g_p^{rs}(k)\geq 0\ \ \forall r,s,p,k \tag{17A3.4}$$

$$\sum_{p}\sum_{k}g_p^{rs}(k)=v^{rs}\ \ \forall r,s \tag{17A3.5}$$

where $\{v^{rs}\}$, $\{g_p^{rs}(k)\}$ are subproblem decision variables denoting O–D demands and path flows, respectively, for each O–D pair. This auxiliary problem cannot be decomposed by O–D pair unless the auxiliary O–D flows $\{v^{rs}\}$ are known. Let the travel time on the shortest path between origin r and destination s be denoted by $=\bar{c}^{rs}=\min_{p,k}\{c_p^{rs}(k)\}$. The above problem of finding $\{v^{rs}\}$ can now be written as follows:

$$\min_{\{v^{rs}\}\in\Omega_v} z(\mathbf{v})=\sum_{rs}\left(\bar{c}^{rs}+\frac{1}{\zeta}\ln q^{rs}\right)v^{rs} \tag{17A3.6}$$

where Ω_v is delineated by the following three constraints, that is, trip production, trip attraction and non-negativity constraints:

$$\sum_{s}v^{rs}=\bar{q}^{r}\ \ \forall r \tag{17A3.7}$$

$$\sum_{r}v^{rs}=\bar{q}^{s}\ \ \forall s \tag{17A3.8}$$

$$v^{rs}\geq 0\ \ \forall r,s. \tag{17A3.9}$$

Denoting the total number of origins by R and destinations by S, the Simplex method for linear programming problem is described as follows:

Step 1: Select an initial feasible solution with $R+S-1$ flow-carrying links, known as 'basis' in the sense of linear programming.
Step 2: Check whether the solution can be improved by using a currently empty link. If not, stop; if yes, continue.
Step 3: Determine the amount of flow that can be assigned to the new link without violating any constraint.
Step 4: Adjust the flow on all other flow-carrying links and update the network. Go to Step 2.

In trip distribution, the Hitchcock transportation problem can be represented by a bipartite graph and solved by a linear programming technique if each centroid denotes either origin or destination, but not both. However, when a certain centroid denotes both origin and destination, the problem of accruing intra-trips, a degeneracy, might occur. Figure 17A3.1 illustrates this in the form of a

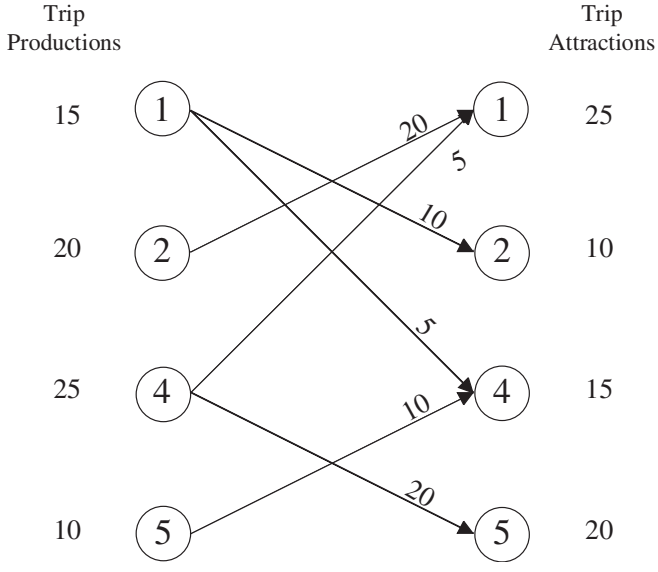


Figure 17A3.1 Degenerate initial solution

bipartite network with four nodes (numbered 1, 2, 3 and 4) representing both origins and destinations. An important characteristic of the Hitchcock problem is that of 'at optimality', by which the number of links carrying flow equals the minimum number of links that can connect R supply nodes to S demand nodes. In other words, there should be $(R+S-1)$ links for which $v^{rs} \geq 0$, on all other links $v^{rs}=0$ (Sheffi 1985). The reason for subtracting 1 from $(R+S)$ is that the required condition of $\sum_r \bar{q}^r = \sum_s \bar{q}^s$ causes one of the existing constraints to be redundant, which leads to the loss of 1 degree of freedom. Therefore, the total number of 'links' with positive flow in a 'basis' should be 7 $(=4+4-1)$. However, without allowing intra-trips, the result obtained by employing the northwest corner method indicates that the total number of links with positive flow is only six, implying a phenomenon of degeneracy. As a consequence, a spanning tree cannot be formed and the subsequent flow augmenting procedure fails.

Evans Algorithm

The Evans algorithm, which may be regarded as a generalization of the FW method, converts the original problem into a series of subproblems by using a partial linearization technique. When

applied to the TD–JEDA problem, only the first term of the objective is linearized and the second term is kept in its original form, as follows:

$$\min_{(\mathbf{g},\mathbf{v})\in\Omega_{DUE-DC-OD-D-R}} \sum_{rs}\sum_{p}\sum_{k}\bar{c}_p^{rs}(k)g_p^{rs}(k)+\frac{1}{\zeta}\sum_{rs}\left(v^{rs}\ln v^{rs}-v^{rs}\right). \tag{17A3.10}$$

Define $\bar{c}^{rs}=\min_{p,k}\left\{c_p^{rs}(k)\right\}$ and, by the relationship of $\sum_p\sum_k g_p^{rs}(k)=v^{rs}$, the above problem of finding $\{v^{rs}\}$ can now be written as follows:

$$\min_{\{v^{rs}\}\in\Omega_v} \sum_{rs}\bar{c}^{rs}v^{rs}+\frac{1}{\zeta}\sum_{rs}\left(v^{rs}\ln v^{rs}-v^{rs}\right). \tag{17A3.11}$$

Dualizing the trip production and trip attraction constraints with dual variables $\{\pi^r\}$ and $\{\pi^r\}$, respectively, and adding these dualities into the above objective function, yields the following Lagrangian:

$$L(v,\pi)=\sum_{rs}\bar{c}^{rs}v^{rs}+\frac{1}{\zeta}\sum_{rs}\left(v^{rs}\ln v^{rs}-v^{rs}\right). \tag{17A3.12}$$

$$+\sum_r\pi^r\left(\bar{q}^s-\sum_s v^{rs}\right)+\sum_s\pi^s\left(\bar{q}^s-\sum_r v^{rs}\right).$$

Taking the derivative with respect to O–D flow v^{rs} results in:

$$\bar{c}_p^{rs}+\frac{1}{\zeta}\ln v^{rs}-\pi^r-\pi^s=0\ \ \forall r,s. \tag{17A3.13}$$

By arithmetic manipulation, the subproblem O–D flow can be expressed as follows:

$$v^{rs}=e^{-\zeta(\bar{c}^{rs}-\pi^r-\pi^s)}\ \ \forall r,s. \tag{17A3.14}$$

Summing over r and s, respectively, we have:

$$\bar{q}^r=e^{\zeta\pi^r}\sum_s e^{-\zeta(\bar{c}^{rs}-\pi^s)}\ \ \forall r \tag{17A3.15}$$

$$\bar{q}^s=e^{\zeta\pi^s}\sum_r e^{-\zeta(\bar{c}^{rs}-\pi^r)}\ \ \forall s. \tag{17A3.16}$$

By arithmetic manipulation, we obtain:

$$e^{\zeta\pi^r}=\frac{\bar{q}^r}{\sum_s e^{-\zeta(\bar{c}^{rs}-\pi^s)}}\ \ \forall r \tag{17A3.17}$$

$$e^{\zeta\pi^s}=\frac{\bar{q}^s}{\sum_r e^{-\zeta(\bar{c}^{rs}-\pi^r)}}\ \ \forall s. \tag{17A3.18}$$

Substituting equations (17A3.17) and (17A3.18) into equation (17A3.14), yields:

$$v^{rs}=e^{-\zeta\bar{c}^{rs}}=\frac{\bar{q}^r}{\sum_s e^{-\zeta(\bar{c}^{rs}-\pi^s)}}\frac{\bar{q}^s}{\sum_r e^{-\zeta(\bar{c}^{rs}-\pi^r)}}\ \ \forall r,s. \tag{17A3.19}$$

Let:

$$A^r=\frac{1}{\sum_s e^{-\zeta[c_p^{rs}(k)-\pi^s]}}\ \ \forall r,s,p,k \tag{17A3.20}$$

$$B^s=\frac{1}{\sum_r e^{-\zeta[c_p^{rs}(k)-\pi^r]}}\ \ \forall r,s,p,k. \tag{17A3.21}$$

Equation (17A3.19) can be simplified as:

$$v^{rs} = A^r \bar{q}^r B^s \bar{q}^s e^{-\zeta \tilde{c}^{rs}} \quad \forall r, s, \tag{17A3.22}$$

where:

$$A^r = \left(\sum_s B^s \bar{q}^s e^{-\zeta \tilde{c}^{rs}} \right)^{-1} \quad \forall r \tag{17A3.23}$$

$$B^s = \left(\sum_r A^r \bar{q}^r e^{-\zeta \tilde{c}^{rs}} \right)^{-1} \quad \forall s. \tag{17A3.24}$$

In fact, equation (17A3.22) is the optimality condition of the subproblem (17A3.1). The coefficients $\{A^r\}$ and $\{B^s\}$ defined by equations (17A3.23) and (17A3.24) are associated with trip productions (R rows) and trip attractions (S columns). A matrix balancing method called the RAS algorithm is commonly adopted for solving this nonlinear subproblem within the Evans algorithm. The complete Evans algorithm is decomposed into two stages. In the first stage, the RAS algorithm is applied for estimating O–D demands based on the estimated minimum O–D travel times, and the obtained O–D demands are, in turn, used to determine link flow solutions. The double-stage procedure may be better illustrated by the following graphs. Consider the 4-node 3-link network shown in Figure 17A3.2a. The first stage of the Evans method involves finding a set of feasible O–D demands based on the current O–D travel times, as shown in Figure 17A3.2b. Once all O–D demands are estimated, traffic assignment is applied to a time–space network as shown in Figure 17A3.2c.

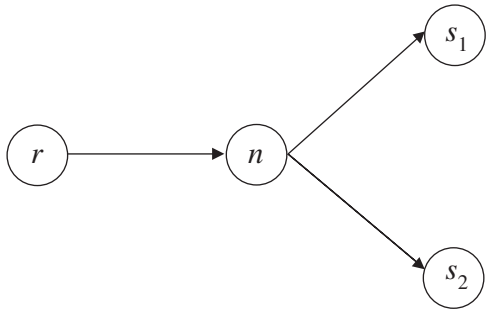


Figure 17A3.2a Static network

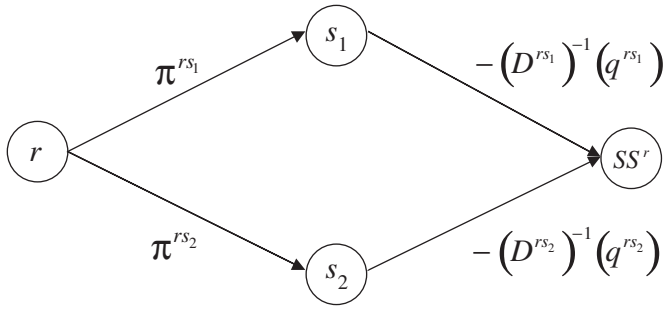


Figure 17A3.2b Network representation for the first stage of the Evans algorithm

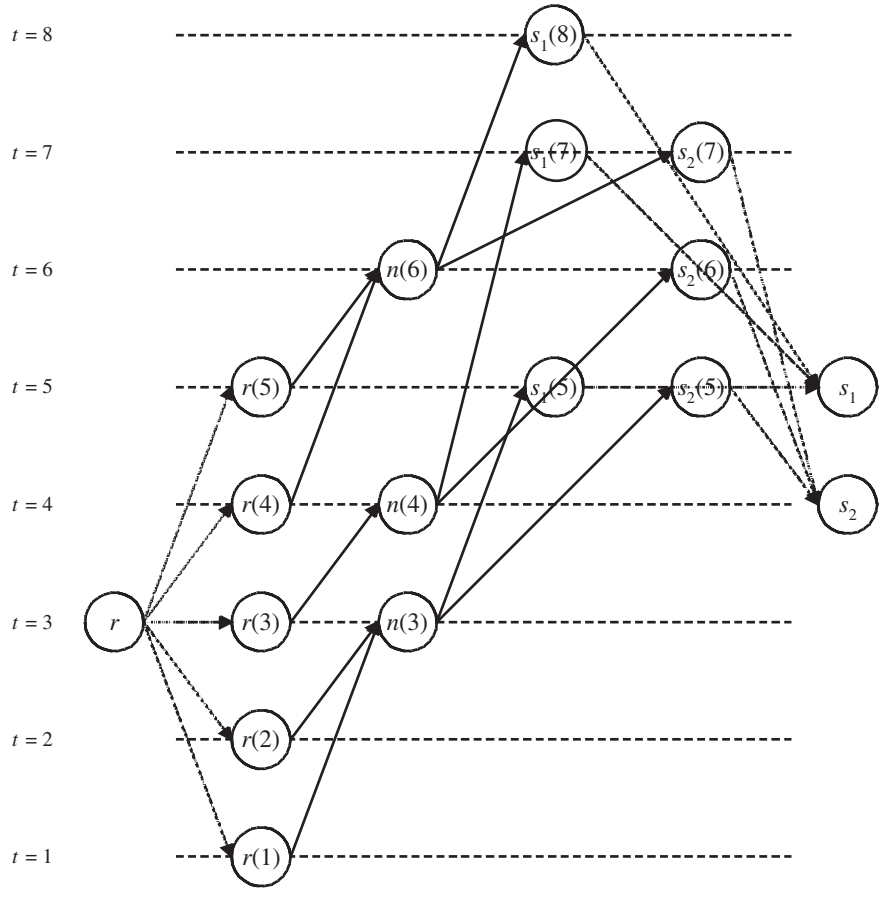


Figure 17A3.2c Network representation for the second stage of the Evans algorithm

Given thise conceptualization of the Evans algorithm, Step 3 of the ND method can be described as follows:

Step 3: Third Loop Operation Let $l=1$. Set an initial solution $\{q^{rsl}\}$, $\{q^{rsn}\}$, $\{u_a^l(t)\}=\{u_a^n(t)\}$, $\{c_a^l(t)\}=\{c_a^n(t)\}$, and compute the associated link travel time $\{c_a^l(t)\}$. Define $\bar{c}^{rsl}=\min_{p,k}\{\bar{c}_p^{rsl}(k)\}$.

Step 3.1: First stage of the solution algorithm Find subproblem O–D trips $\{v^{rsl}\}$ using the RAS algorithm.

Step 3.1.1: Initialization Let $i=0$ and $\{A^{ri}\}=\{1\}$.

Step 3.1.2: Balancing trip attractions (columns) Let $i=i+1$,

$$B^{si}-\left(\sum_r A^{ri}\bar{q}^r e^{-\xi\bar{c}rsl}\right)^{-1} \quad \forall s. \tag{17A3.25}$$

Step 3.1.3: Balancing trip productions (rows)

$$A^{si} = \left(\sum_{s} B^{si} \bar{q}^{s} e^{-\zeta \bar{c}^{rsl}} \right)^{-1} \quad \forall r. \tag{17A3.26}$$

Step 3.1.4: Convergence check If

$$\operatorname{Max}_{r} \left| \frac{A^{ri} - A^{ri-1}}{A^{ri}} \right| \leq \varepsilon, \text{ and } \operatorname{Max}_{s} \left| \frac{B^{si} - B^{si-1}}{B^{si}} \right| \leq \varepsilon,$$

go to Step 3.1.5; otherwise, set $i = i+1$ and return to Step 3.1.2.

Step 3.1.5: Estimate O–D trips

$$v^{rsl} = A^{r*} \bar{q}^{r} B^{s*} \bar{q}^{s} e^{-\zeta \bar{c}^{rsl}} \; \forall r, s. \tag{17A3.27}$$

Step 3.2: Second stage of the solution algorithm Assign O–D trips v^{rsl} to the path with minimum route travel time, yielding link flows $\{p_a^l(t)\}$.

Step 3.3: Define search direction by $[p_a^l(t) - u_a^l(t)]$, $\forall a, t$ and $(v^{rsl} - q^{rsl})$, $\forall r, s$.

Step 3.4: Solve the following linear combination problem for move size α by the bisection method.

$$\min_{\lambda} \sum_{a} \sum_{t} \int_{0}^{u_a^l(t) + \alpha^l [p_a^l(t) - u_a^l(t)]} c_a \left[\bar{\mathbf{u}}^n \backslash \bar{u}_a^n(t), \omega \right] d\omega \tag{17A3.28}$$

$$+ \frac{1}{\zeta} \sum_{rs} (\omega \ln \bar{\omega} - \bar{\omega}) \Big|_{q^{rsl} + \alpha^l (v^{rsl} - q^{rsl})}.$$

Step 3.5: Update link flows and O–D demands as follows:

$$u_a^{l+1}(t) = u_a^l(t) + \alpha [p_a^l(t) - u_a^l(t)], \;\; \forall a, t \tag{17A3.29}$$

$$q^{rsl+1} = q^{rsl} + \alpha^l (v^{rsl} - q^{rsl}), \;\; \forall r, s. \tag{17A3.30}$$

Compute link travel times $\{c_a^{l+1}(t)\}$.

Step 3.6: Convergence check for the third loop If

$$\max_{a,t} \left| \frac{u_a^{l+1}(t) - u_a^l(t)}{u_a^{l+1}(t)} \right| \leq \varepsilon \text{ and } \max_{r,s} \left| \frac{q^{rsl+1} - q^{rsl}}{q^{rsl+1}} \right| \leq \varepsilon,$$

let $\{q^{rsn+1}\} = \{q^{rsl+1}\}$, $\{u_a^{n+1}(t)\} = \{u_a^{l+1}(t)\}$, $\{c_a^{n+1}(t)\} = \{c_a^{l+1}(t)\}$ and go to Step 4. Otherwise, let $l = l+1$, go to Step 3.1.

As previously discussed, intra-trips for each 'centroid' are externally treated and not allowed to accrue in the solution procedure. To take this into account, the RAS algorithm needs to be modified. The modified RAS (mRAS) algorithm is identical to the original RAS algorithm, except that equations (17A3.25)–(17A3.28) are replaced by the following four formulae:

$$B^{si} = \left(\sum_{r \neq s} A^{ri} \bar{q}^{r} e^{-\zeta \bar{c}^{rsl}} \right)^{-1} \quad \forall \; s \tag{17A3.31}$$

$$A^{ri} = \left(\sum_{s \neq r} B^{si} \bar{q}^s e^{-\zeta \bar{c}^{rsl}} \right)^{-1} \quad \forall\, r \tag{17A3.32}$$

$$v^{rsl}|_{r \neq s} = A^{r*}\, \bar{q}^r\, B^{s*}\, \bar{q}^s e^{-\zeta \bar{c}^{rsl}}\ \forall r, s \tag{17A3.33}$$

$$\min_{\lambda} \sum_a \sum_t \int_0^{u_a^l(t) + \alpha^l [p_a^l(t) - u_a^l(t)]} c_a\left[\bar{\mathbf{u}}^n \setminus \bar{u}_a^n(t), \omega\right] d\omega \tag{17A3.34}$$

$$+ \frac{1}{\zeta} \sum_r \sum_{s \neq r} (\omega \ln \bar{\omega} - \bar{\omega}) \Big|_{q^{rsl} + \alpha^l (v^{rsl} - q^{rsl})}.$$

NOTE

* We wish to thank Professor David Boyce for guiding the first author during his PhD study into the colorful field of transportation planning and networks. We also express our gratitude to the National Science Council, Taiwan, for its financial support.

REFERENCES

Bachem, A. and B. Korte (1979), 'On the RAS-algorithm', *Computing*, **25**, 189–98.

Chabini, I., O. Drissi-Kaitouni and M. Florian (1994), 'Parallel implementations of primal and dual algorithms for matrix balancing', in D.A. Belslev (ed.), *Computational Techniques for Econometrics and Economic Analysis*, Dordrecht: Kluwer Academic, pp. 173–85.

Chabini, I. and M. Florian (1995), 'An entropy based primal-dual algorithm for convex and linear cost transportation problems', Report CRT-95-17, University of Montreal.

Chen, H.K. (1999), *Dynamic Travel Choice Models: A Variational Inequality Approach*, Berlin: Springer-Verlag.

Chen, H.K., C.W. Chang and M.S. Chang (2001), 'A comparison of link-based versus route-based algorithms in the dynamic user-optimal route choice problem', *Transportation Research Record* **1667**, 114–20.

Chen, H.K. and Y.C. Chen (1999), 'Comparisons of the Frank–Wolfe and Evans methods for the doubly constrained entropy distribution/assignment problem', *Journal of the Eastern Asia Society for Transportation Studies*, **3** (1), 261–76.

Evans, S.P. (1976), 'Derivation and analysis of some models for combining trip distribution and assignment', *Transportation Research*, (10), 37–57.

Jayakrishnan, R., W.K. Tsai, J.N. Prashker and S. Rajadhyaksha (1994), 'A faster path-based algorithm for traffic assignment', Presented at the Transportation Research Board 73rd Annual Meeting, Washington, DC.

Luenberger, D.G. (1984), *Linear and Nonlinear Programming*, 2nd edn, Reading, MA: Addison-Wesley.

Schneider, M.H. and S.A. Zenios (1990), 'A comparative study of algorithms for matrix balancing', *Operations Research*, (38), 439–53.

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

Tatineni, M., H. Edwards and D. Boyce (1998), 'Comparison of disaggregate simplicial decomposition and Frank–Wolfe algorithms for user-optimal route choice', *Transportation Research Record*, **1617**, 157–62.

Tobin, R.L. and T. Friesz (1988), 'Sensitivity analysis for equilibrium network flow', *Transportation Science*, **22** (4), 242–50.

18. Numerical experiments with a decision support methodology for strategic traffic management

Torbjörn Larsson, Jan T. Lundgren, Michael Patriksson and Clas Rydergren

1. INTRODUCTION

One of the essential aspects of strategic traffic management is to determine the modifications which need to be made in the traffic system in order to improve its long-term functionality. The functionality of a traffic network can typically be expressed in terms of traffic flows and functions thereof. Therefore, functionality wishes and requirements can often be formulated as goals in terms of the link flows. These goals may be influenced by a number of circumstances such as environmental, safety, practical, political, and economic considerations. Examples of such goals are maximal traffic flows allowed on links, maximal travel time between two locations and maximal exhaust fume emissions in a certain area. In order to reach the goals set by the traffic manager, the behavior of the travelers has to change. Their behavior can be altered by modifications in the network.

Each link in the traffic network is associated with a generalized travel time function. These functions may include factors such as the actual travel time, queuing delays, and monetary outlay. The effect of an action in the network is in our setting reflected in a change in these functions; the resulting flow shifts in the traffic equilibrium solution then describes the change in the travelers' behavior. Adjustments in the generalized travel times can be achieved, for example, by the setting of traffic signal green times, adjustments in speed limits, modifications in the street network, and road pricing, or any combination thereof.

The traffic management process is today often based on repeated scenario analyses. In Larsson et al. (2000) an alternative methodology is proposed. This methodology supports the direct derivation of adjustments to the generalized travel times, and is based on a mathematical model which integrates the traffic management decisions with a traffic equilibrium model. The methodology can be seen as a two-stage procedure where a sequence of convex optimization problems are solved. In this chapter, we give examples of how the procedure can be used in a realistic setting and highlight both its advantages and shortcomings. The numerical properties and results of the two-stage management procedure are compared with results from a direct search procedure on a bilevel formulation of the traffic management problem.

The proposed methodology produces tentative adjustments of the generalized travel times. However, it does not prescribe which actions, or combinations of actions, should

be used to achieve these adjustments. Further, in practice it is likely that the generalized travel time adjustments can be implemented only approximately. The implementation of the tentative travel time adjustments through action in the traffic network is a task for the traffic engineer, and may of course also be made subject to additional considerations. These issues, however, are beyond the scope of this chapter.

The chapter is organized as follows. A short summary of the two-stage traffic management procedure is given in Section 2. In Section 3 the traffic management procedure is exemplified on the well-known small-scale Sioux Falls network and on the larger Linköping network. In Section 4, results from a Hooke and Jeeves direct search technique applied to a bilevel formulation of the traffic management problem are presented and compared to the solutions obtained from the two-stage procedure. Section 5 concludes the chapter.

2. SUMMARY OF THE STRATEGIC TRAFFIC MANAGEMENT PROCEDURE

The proposed management procedure is based on solving a sequence of optimization problems. In this section the necessary notation is introduced and the optimization problems are stated.

Let the traffic network be defined by a set of nodes $\mathcal{N}$ and a set of directed links $\mathcal{A}$. Let d_{pq} denote a fixed travel demand from the origin p to the destination q, and let C be the set of origin–destination (O–D) pairs (p, q). Let $\mathcal{R}_{pq}$ denote the (non-empty) set of routes in pair (p, q) and h_{pqr} denote the flow on route $r \in \mathcal{R}_{pq}$. The (compact) set of feasible route flows, $\mathcal{H}$, is then given by the solutions, h, to the system:

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \quad \forall (p, q) \in C, \tag{18.1a}$$

$$h_{pqr} \geq 0, \quad \forall r \in \mathcal{R}_{pq}, \quad \forall (p, q) \in C. \tag{18.16b}$$

Let the flow on the links in the network be described by a vector, $f \in \Re^{|\mathcal{A}|}$, where each element represents the flow (for example, in cars per hour) on a specific link. Let $\delta_{pqra} = 1$ if route r from p to q include link a, and 0 otherwise. The link flow on link a is then calculated as:

$$f_a = \sum_{(p,q) \in C} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr}, \quad \forall a \in C, \tag{18.2}$$

Let $\mathcal{F}$ be the set of link flows that might result from the route flows in $\mathcal{H}$, that is, the link flows f that satisfy constraints (18.1a) and (18.1b).

Next, we introduce the notation necessary for modeling the goals of the traffic manager. The manager's goals are reflected by the constraints $g_k(f) \leq 0$, $k \in \mathcal{K}$, where $\mathcal{K}$ is a finite index set. The functions g_k are assumed to be convex, and typically model flow capacities on individual links or on combinations of links, or restrictions on travel times on links or sequences of links. Define the set:

$$\mathcal{G} = \{ f \in \Re^{|\mathcal{A}|} \mid g_k(f) \leq 0, \quad k \in \mathcal{K} \}.$$

Let the generalized link travel times be given by $t_a(f_a)$, for each link $a \in \mathcal{A}$, where the functions t_a are assumed to be non-negative, continuous and strictly increasing. Further, let the vector $\rho \in \Re^{|\mathcal{A}|}$ represent the manager's decision variables, that is, the adjustment of the generalized travel time on each link. Then the adjusted generalized travel time is $t_a(f_a) + \rho_a$.

The adjustments to the generalized travel times that it is possible to achieve, using some actions in the network, are defined by restrictions on the manager's decision variables. Let $\mathcal{P} \subseteq \Re^{|\mathcal{A}|}$ denote the set of the feasible values of the decision variables $\rho \in \Re^{|\mathcal{R}|}$, that is, the modifications of the generalized travel times that can be made. We assume that:

$$\mathcal{P} = \{\rho \in \Re^{|\mathcal{A}|} \mid w_l(\rho) \leq 0, \quad l \in \mathcal{L}\}.$$

where the functions w_l, $l \in \mathcal{L}$, are assumed to be linear and $\mathcal{L}$ is a finite index set. The set $\mathcal{P}$ is typically used for modeling restrictions on which links it is possible to make adjustments to, and to model restrictions of the size of the adjustment allowed on individual links. The feasible adjustments are, in turn, dependent of the set of actions that may be considered in the traffic network or on specific links.

The traffic management problem can then be formulated as the bilevel model (for example, Migdalas 1995; Larsson and Patriksson 1998):

$$\min \varphi(f, \rho) \tag{18.3a}$$

subject to:

$$f \in \mathcal{G}, \tag{18.3b}$$

$$\rho \in \mathcal{P}, \tag{18.3c}$$

$$f \text{ is an equilibrium link flow solution given } \rho. \tag{18.3d}$$

The objective function φ can, for example, be chosen such that the total adjustment or the total generalized travel time is minimized.

The interaction between the manager and the travelers is here described as a Stackelberg game, where the manager is the leader and the travelers are the followers, that is, they reevaluate their route choices after any change in the traffic network.

The bilevel problem (18.3a) is in general non-convex and therefore very difficult to solve. Furthermore, the function $\varphi(f, \rho)$ is typically not differentiable since f as a function of ρ through (18.3d) is not differentiable (for example, Patriksson and Rockafellar 2003). Several heuristics have been proposed for the solution of similar bilevel problems, for example genetic algorithms in Cree et al. (1998), simulated annealing in Huang and Bell (1998), and by approximating the bilevel problem by a nonlinear non-convex problem in Chen et al. (1998). Marcotte (1986) and Marcotte and Marquis (1992) describe a heuristic for finding a solution to the bilevel problem by solving a convex optimization problem, and present a worst-case analysis of their approach. Recent advances in the field of transport bilevel problems can be found in Yang and Bell (2001).

Compared to most bilevel problems, ours has special characteristics. Our overall decision problem includes both primary and secondary goals of which the former are taken

into account in the lower level, while the latter appear at the upper level. Because of this, the lower level is the most significant one, while the upper is of less importance. This characteristic motivates our solution approach, which is a sequential procedure where we first solve a problem related to the lower level, and then a problem related to the upper level.

Our solution approach has two main stages. In the first stage, equilibrium flows satisfying the manager's goals are computed. In the second stage, travel time adjustments for achieving the goals are computed. First, we need to establish whether the goal constraints are too restrictive, that is, whether the goals are consistent with the travel demand. An inconsistency corresponds to management goals which are too ambitious given the travel demand, and which must therefore be revised. When the management goals and the travel demand have been found to be consistent, we solve the goal-constrained traffic equilibrium problem,

$$\min_{f \in \mathcal{F} \cap \mathcal{G}} \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(s)ds. \tag{18.4}$$

Solution algorithms for this problem are presented in Larsson and Patriksson (1998) and Larsson et al. (2004a). Let the vector $\bar{f}$ denote the optimal solution to problem (18.4). Since we will utilize Lagrange multipliers for the constraints defining $\mathcal{G}$ in our development, we must impose a constraint qualification on the collection of nonlinear functions among g_k; for example, the Slater condition (for example, Bertsekas 1995, Assumption 5.3.2) is sufficient; in the case that all functions g_k are linear (as indeed they are in our examples) a constraint qualification is fulfilled automatically.

Second, we establish whether the goals and the set of feasible travel time adjustments are consistent, that is, if the goals can be attained by travel time adjustments in $\mathcal{P}$. Let $\mathcal{Q}(\bar{f})$ denote the set of travel time adjustments which, when added to the original travel time functions, turn the link flow solution $\bar{f}$ into an equilibrium link flow solution. The set $\mathcal{Q}(\bar{f})$ can be defined by the set of ρ such that the equilibrium condition:

$$h_{pqr} > 0 \;\Rightarrow\; \sum_{a \in \mathcal{A}} \delta_{pqra}[t_a(\bar{f}_a) + \rho_a] = \pi_{pq}, \quad (p, q) \in C,\ r \in \mathcal{R}_{pq},$$

$$h_{pqr} = 0 \;\Rightarrow\; \sum_{a \in \mathcal{A}} \delta_{pqra}[t_a(\bar{f}_a) + \rho_a] \geq \pi_{pq}, \quad (p, q) \in C,\ r \in \mathcal{R}_{pq},$$

is satisfied for some values of π. The set $\mathcal{Q}(\bar{f})$ can alternatively be described by the linear constraints:

$$\sum_{a \in \mathcal{A}} \delta_{pqra}[t_a(\bar{f}_a) + \rho_a] - \pi_{pq} \geq 0, \quad \forall r \in \mathcal{R}_{pq}, \quad \forall (p, q) \in C, \tag{18.5a}$$

$$\sum_{a \in \mathcal{A}} [t_a(\bar{f}_a) + \rho_a]\bar{f}_a - \sum_{(p, q) \in C} \pi_{pq} d_{pq} = 0. \tag{18.5b}$$

This reformulation (see Larsson et al. 2004b) makes it possible to formulate a problem for finding adjustments to the travel times, such that $\bar{f}$ become an equilibrium solution, as a linear optimization problem.

To check whether it is possible to find adjustments $\rho \in \mathcal{P}$ to the generalized travel times such that the formulated goals are satisfied, we need to check whether the set $\mathcal{Q}(\bar{f}) \cap \mathcal{P}$ is non-empty. This consistency check can be formulated as the optimization problem:

$$\min_{\rho \in \mathcal{Q}(\bar{f}), \mu \in \mathcal{P}} \sum_{a \in \mathcal{A}} \bar{w}_a |\rho_a - \mu_a|, \tag{18.6}$$

where $\bar{w}_a$, $a \in \mathcal{A}$ are positive weights. (These weights can be used to reflect a preference on which links we would like to have travel time adjustments μ_a close to ρ_a.) In the case of inconsistency, the objective value of an optimal solution to (18.6) is positive and either the goals or the set of feasible adjustments have to be revised.

When the goals and the feasible adjustments are verified as consistent, travel time adjustments that are optimal with respect to the objective function φ are found by solving a *target flow equilibrium pricing* problem,

$$\min_{\rho \in \mathcal{Q}(\bar{f}) \cap \mathcal{P}} \varphi(\rho). \tag{18.7}$$

This is an inverse nonlinear multicommodity network flow problem. (In the case where φ is linear, this problem is linear.) An algorithm for solving the problems (18.6) and (18.7) is presented in Larsson et al. (2004b).

Given travel time adjustment, ρ_a, for all links $a \in \mathcal{A}$, an equilibrium link flow solution can be found by solving the traffic equilibrium problem:

$$\min_{f \in \mathcal{F}} \sum_{a \in \mathcal{A}} \int_0^{f_a} [t_a(s) + \rho_a] ds, \tag{18.8}$$

where the vector ρ acts as a fixed addition to the link travel time. See, for example, Patriksson (1994, Chapter 2.2), for details on the traffic equilibrium problem. The problem (18.8) is identical to the lower-level problem (18.3d). It can be solved with the disaggregate simplicial decomposition algorithm of Larsson and Patriksson (1992).

The problem (18.8) can be used for finding information about how the goals may be revised in the case of inconsistency in the problem (18.6). Given a solution μ to the problem (18.6), let $\rho = \mu$ and solve the traffic equilibrium problem (18.8). Given the link flow solution from solving (18.8), we may evaluate the resulting violation of the goal constraints. This provides information about the level of violation of the management goals.

The model (18.3) is formulated for the case of a traffic equilibrium with fixed travel demand. This model and the proposed methodology for traffic management can, with small modifications, be extended to the cases of elastic demand and modal split. Such extensions are of importance for the application of the proposed methodology to real-world applications. They are discussed in Larsson et al. (2000) but no numerical experiments have been performed.

The case where the travel time functions are non-separable complicates stage one of our procedure considerably. In this case, it is typically not possible to find the goal-constrained traffic equilibrium problem by solving a convex optimization problem. Instead the stage one problem needs to be stated and solved as a variational inequality problem with side constraints. However, stage two, is unchanged.

3. NUMERICAL ILLUSTRATIONS OF THE TRAFFIC MANAGEMENT PROCEDURE

In this section, management scenarios in the networks of Sioux Falls, SD, and Linköping, Sweden, are used to illustrate our procedure. The order of the computations is illustrated and the advantages and shortcomings of the procedure are pointed out. The computations have been performed on a Sun Ultrasparc 10 with a 300Mhz CPU and 320MB of memory.

3.1 The Sioux Falls Network: An Example Scenario

In the first example, the two-stage traffic management procedure is illustrated on the classic small-scale network of Sioux Falls. The network has 24 nodes, 76 links and 528 O–D pairs for which the travel demand is positive. The management goal is to reduce the inflow into a specified node to 80 per cent of the inflow in the current scenario. The aim is to achieve this goal by imposing adjustments to the link travel times on a small number of selected links. The example is carried out in the eight steps shown in Figure 18.1.

Step 1 The current link flow scenario is computed by solving the traffic equilibrium problem (18.8) for $\rho_a = 0$, $a \in \mathcal{A}$, given the travel demand and travel time functions for each of the links in the network. The result of this network assignment is shown in Figure 18.2. In the figure, each road segment is made up of two links, one in each direction. The travelers use the right-hand-side link and the link flow is given to the right of the corresponding link. The link widths in the figure are proportional to the flow on the links.

Step 2 The management goals are to reduce the inflow into node 15 (in the standard coding of the Sioux Falls network, see LeBlanc et al. 1975) to 80 per cent of the current inflow. The inflow into the node is 69.68 units in the current scenario. The management goals are formulated as two linear constraints. The first is a single link flow capacity constraint to reduce the flow on the link entering node 15 from the East to 80 per cent of the flow in the current scenario. This is achieved by reducing the flow from 19.12 units to 15.29 units on this link. An upper bound on the link flow of 15.29 is imposed on the link. The restriction is shown in Figure 18.2. The second restriction includes the flow on the three links entering node 15 from the North, West and South, shown in Figure 18.2. In the current flow scenario, the total flow towards node 15 is 50.56 units on these links. We formulate a constraint that restricts the sum of the flow on the three links to be less than 40.44 units.

Step 3 The side-constrained traffic equilibrium problem (18.4) is solved for finding an equilibrium solution which satisfies the management goals formulated in Step 2. Let $\bar{f}$ denote the optimal link flow solution to problem (18.4). The difference between the current scenario flows and the link flow solution $\bar{f}$ is shown in Figure 18.3. The numbers in the figure indicate the flow difference of the inflow to node 15 on the four links, respectively, and the link widths are proportional to the difference in flow. The lighter links indicate a reduction in flow and the darker indicate an increase compared to the flow in the initial scenario.

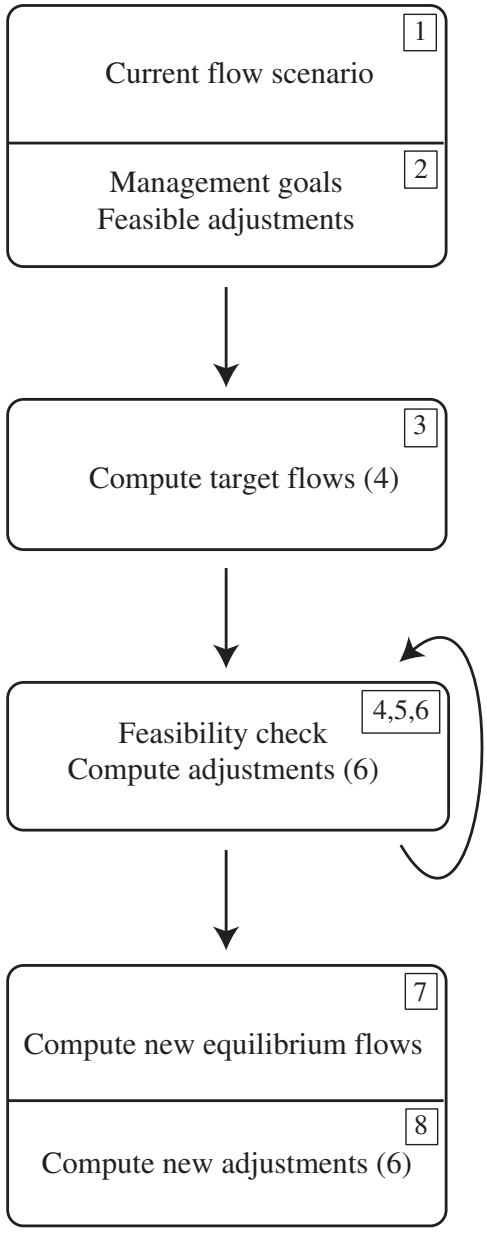


Figure 18.1 Order of computations

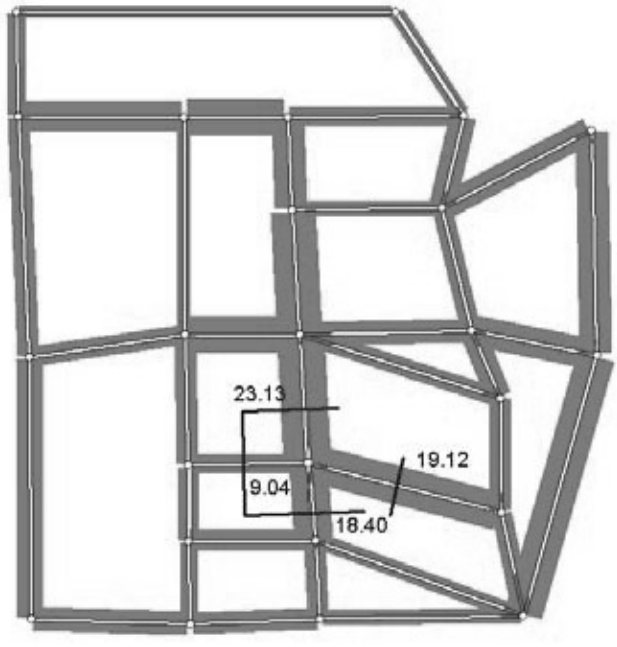


Figure 18.2 Initial traffic equilibrium scenario and management goals

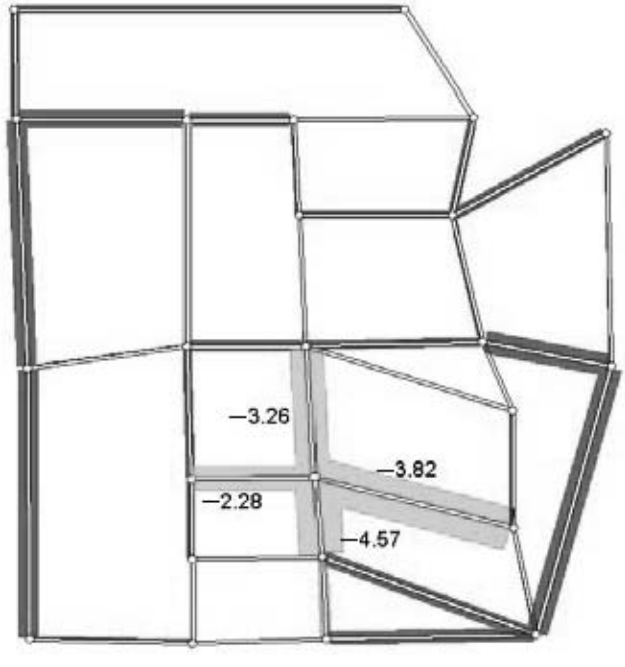


Figure 18.3 Difference between initial scenario and target flow solution

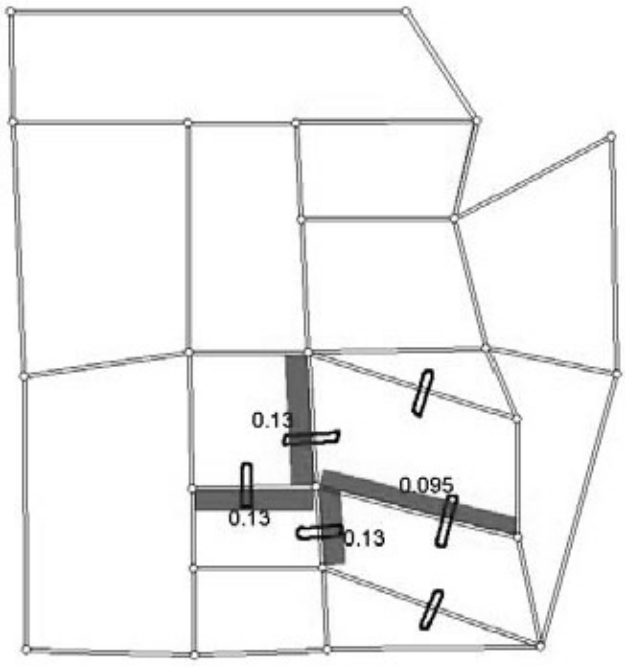


Figure 18.4 Minimal adjustments to reach management goals

Step 4 Given the target link flow vector $\bar{f}$, we solve the target flow equilibrium pricing problem for finding adjustments to the link travel times which, when implemented, turn $\bar{f}$ into an equilibrium solution. The target flow equilibrium pricing problem (18.7) is solved with the objective of finding a set of travel time adjustments which minimize the total adjustment made in the network. This is achieved by choosing the objective function:

$$\varphi(\rho) = \sum_{a \in \mathbb{A}} \bar{v}_a |\rho_a|, \tag{18.9}$$

where $\bar{v}_a = 1$, $a \in \mathcal{A}$. The problem is solved with $\mathcal{P} \equiv \Re^{|\mathcal{A}|}$. The solution is shown in Figure 18.4.

Step 5 With a preference for solutions where we do not make any adjustments the travel time functions on the links entering and exiting node 15 from the East, we re-solve the target flow equilibrium pricing problem with new link weights $\bar{v}_a$, $a \in \mathcal{A}$. We set $\bar{v}_a = 1$ for all links, except for the two links entering and exiting node 15 from the East where $\bar{v}_a = 1000$. Restrictions are put on the adjustments such that adjustments are only allowed on links entering and exiting from node 15 and on the two road segments parallel to the link entering and exiting node 15 from the East. The roads segments where adjustments are *allowed* are marked in Figure 18.4. The solution to the target flow pricing problem for the new weights is given in Figure 18.5. The adjustments on the two links entering and exiting from the East are reduced with this choice of weights, but are not both zero.

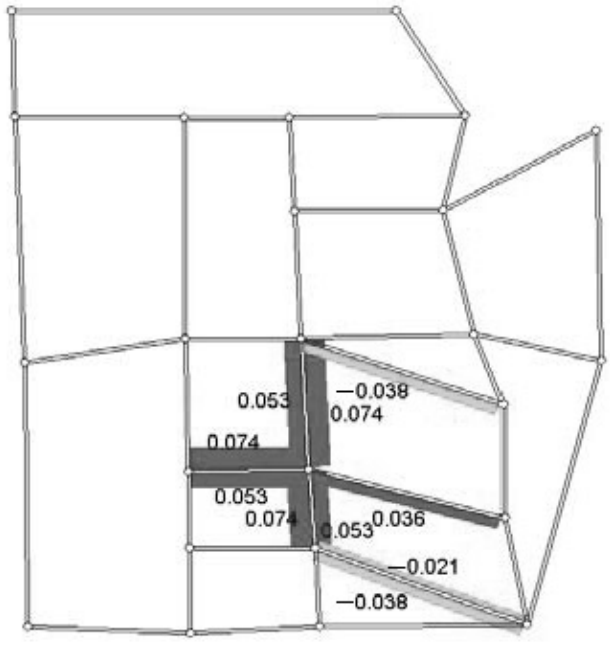


Figure 18.5 Restricted minimal weighted adjustments to reach management goals

Step 6 From Figure 18.5 we can observe that the travel time adjustments on links to and from the East is non-zero. With the preference toward zero adjustments on these links, we impose restrictions by restricting the adjustments on the links entering and exiting node 15 from the East to be zero. The feasibility check problem (18.6) is solved with $\bar{w}_a = \bar{v}_a$, $a \in \mathcal{A}$, and the solution indicates that the target flow equilibrium pricing problem is infeasible. The μ-part of the solution to the problem (18.6) is a vector of travel time adjustments, feasible with respect to the restrictions imposed on the adjustments, and the solution closest to a solution in $\mathcal{Q}(\bar{f})$ measured in a weighted ℓ_1-norm. This set of adjustments is shown in Figure 18.6. However, when this set of adjustments is implemented, the adjustments typically do not result in an equilibrium solution which satisfies the management goals. Since we are not able to find a solution for the case where we have restricted the set of links where it is viable to impose adjustments, we need to revise the goals. This can be done in two ways. If the link flow capacities are treated as 'hard' constraints, we have to be satisfied with a solution where adjustments are also imposed on other links, that is, the solution shown in Figure 18.5 is the best we can do. If the flow restrictions are assumed to be somewhat more 'soft' than the restrictions on the adjustments, we may proceed with the following steps.

Step 7 The adjustments from the feasibility detection problem (18.6) are used to compute new link flows by performing an equilibrium assignment with modified link travel times. The link travel times are modified to be $t_a(f_a) + \hat{\rho}_a$, $a \in \mathcal{A}$, where $\hat{\rho}_a$ are the adjustments shown in Figure 18.6. The link flows which result from the equilibrium

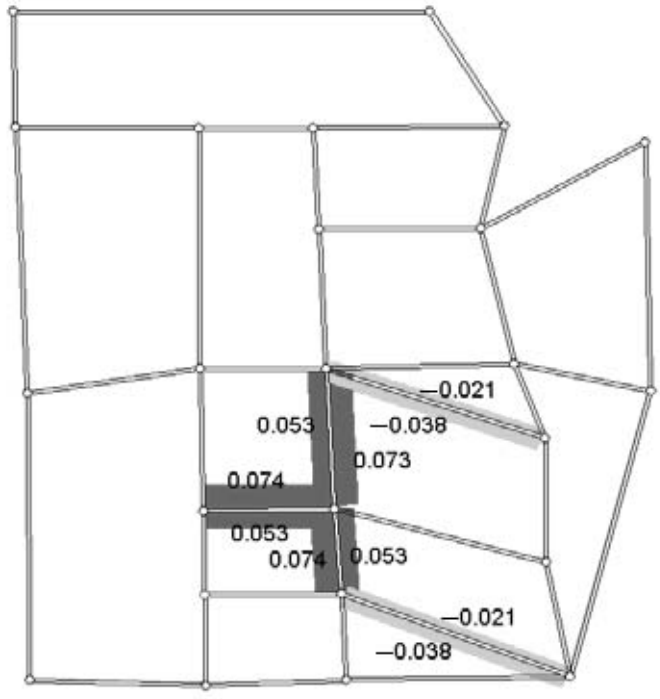


Figure 18.6 Adjustments from feasibility check

assignment are shown in Figure 18.7. The original traffic management goals are not satisfied, but instead of the required flow of 15.29 of eastern link, the flow is 16.28 and the total inflow into the intersection on the other three links is 41.95 instead of the required 40.44.

Step 8 We solve the target flow equilibrium pricing problem with the same restrictions on the adjustments as imposed in Step 5 and for the target flow vector computed in Step 7. The solution coincides with the solution shown in Figure 18.6.

From Step 7 it is clear that the procedure has not provided an equilibrium solution that satisfies the goals. Not being satisfied with the solution to the management problem, the modified link travel time functions computed in Step 7 can be used as the initial travel time functions and the two-stage management procedure can be re-applied on the modified network. We re-apply the management procedure on the equilibrium solution in Figure 18.7 and the travel times modified by the adjustments given in Figure 18.6. A side-constrained traffic equilibrium problem is solved where the side constraints given in Step 2 are used. As a result we obtain a link flow solution satisfying the management goals. The next step is to try to find minimal adjustments of the travel times in order to turn these link flows into equilibrium link flows. We apply the restrictions on the adjustments shown in Step 6. From solving the feasibility check problem we can conclude that the target flow

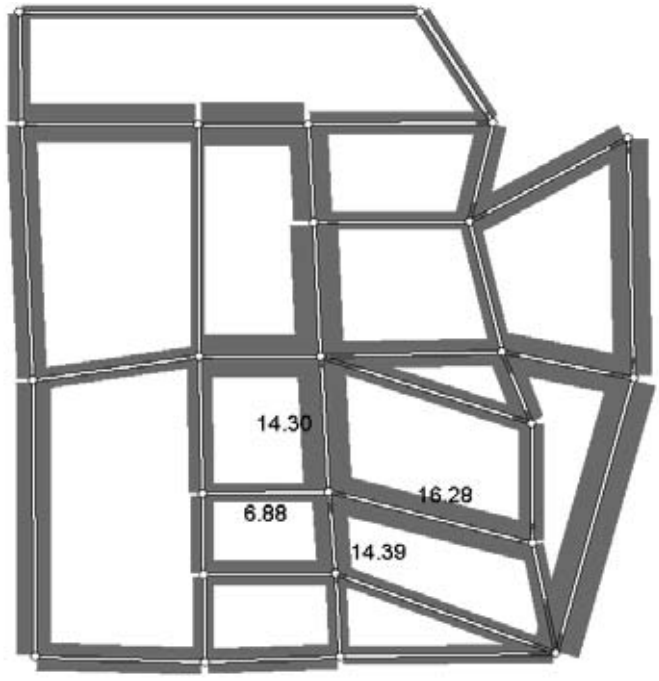


Figure 18.7 Equilibrium solution for adjusted travel times

equilibrium pricing problem is infeasible. Further, we obtain a set of feasible adjustments from the feasibility check problem. The travel time functions are modified by these adjustments and a set of new link flows is computed by solving a traffic equilibrium problem. This additional run of the procedure results in a new set of adjustments for which the flow on the link entering node 15 is 15.76 and the sum of the link flows on the links entering node 15 from North, West and South is 40.60. For cases when repeated application of the two-stage procedure does converge to a solution, this solution satisfies both the manager's goals and the restrictions on the adjustments.

From the illustration we observe that if the formulated restrictions are 'hard', the management procedure, at least in this example, is not capable of finding a solution by applying the two-stage procedure once. However, for the case where the travel time adjustments are 'soft', the procedure provides adjustments for satisfying the flow goals. If the management link flow goals are 'soft', the management procedure is capable of finding a management scheme for which the flow goals either are satisfied or close to being satisfied. In this example, re-applying the two-stage procedure to the solution of the first run, resulted in feasible adjustments for which the corresponding link flows are close to satisfying the management goals.

3.2 The Linköping Network: Two Example Scenarios

In this section, the management procedure is illustrated on a network model of the city of Linköping, Sweden. The network model has 717 nodes, 882 links and 12372 O–D pairs.

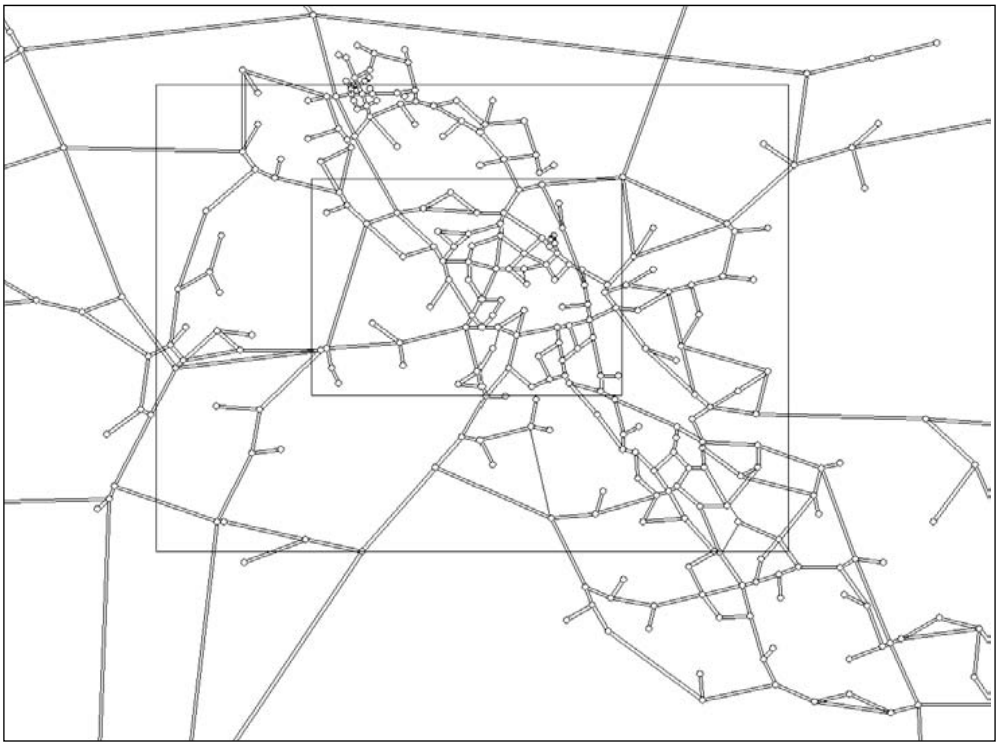


Figure 18.8 Network of the city of Linköping

Two scenarios are presented. In the first, the aim is to decrease the traffic through the central part of the network by increasing the traffic on a bypass route. In the second scenario, the aim is to reduce the flow on one of the central road segments. The central part of Linköping is shown in the network model in Figure 18.8.

Scenario 1 The aim is to find which actions need to be taken in the network such that the flow passing the central part of the city is reduced. Since the travel demand between origins and destinations is modeled as fixed, this can be achieved by increasing the traffic on a bypass route. We introduce goal constraints which state that the flow is increased by 20 per cent on the bypass route. These management goals are formulated as four linear flow constraints of the type $f_{a1} + f_{a2} \geq 1.2(\hat{f}_{a1} + \hat{f}_{a2})$ where links $a1$ and $a2$ are links going in the two opposite directions of a road segment a, and where $\hat{f}$ represents the link flows in the initial scenario. In Figure 18.9, the equilibrium solution characterizing the current traffic situation and the placements of the link flow restrictions are shown. The side-constrained traffic equilibrium problem (18.4) is solved to find the target flow vector $\bar{f}$. The difference between the initial link flow solution $\hat{f}$ and the target flow $\bar{f}$ is shown in Figure 18.10. In the figure, the link widths are proportional to the difference in traffic flow, the darker links indicate an increase and the lighter links indicate a decrease in the link flow. Given the optimal dual solution to the side-constrained traffic equilibrium problem, one

Figure 18.9 Equilibrium link flows in the central part of Linköping

set of travel time adjustments can be computed based on the values of the dual multipliers to the side constraints (see Larsson and Patriksson 1999). The locations of these adjustments are on the links covered by the side constraints.

We have restricted the adjustments to be zero on a sequence of three road segments on and around the topmost flow constraint shown in Figure 18.9. Any adjustment is allowed on the other links. The objective function in the target flow equilibrium pricing problem is chosen to minimize the perceived travel time adjustment, that is, the objective is chosen as in (18.9) with $\bar{v}_a = \bar{f}_a$, $a \in \mathcal{A}$. The resulting travel time adjustments are shown in Figure 18.11. From Figure 18.11, it can be observed that the target flows can be achieved by decreasing the travel time on parts of the bypass route together with an increase in travel times on links directed from the bypass route to the central part of the network.

Scenario 2 The aim is to reduce the flow on one of the central road segments in the Linköping network. The equilibrium solution characterizing the current traffic situation and the placement of the link flow restrictions is shown in Figure 18.12. The restrictions are formulated as two linear constraints. The first one restricts the sum of the flow in both directions to be 80 per cent of the links' flow in the current scenario. This is the flow restriction shown to the right in Figure 18.12. The second restriction aims at reducing the sum of flow from two of the connecting links to be 80 per cent of the inflow in the current scenario. This is the leftmost restriction shown in Figure 18.12.

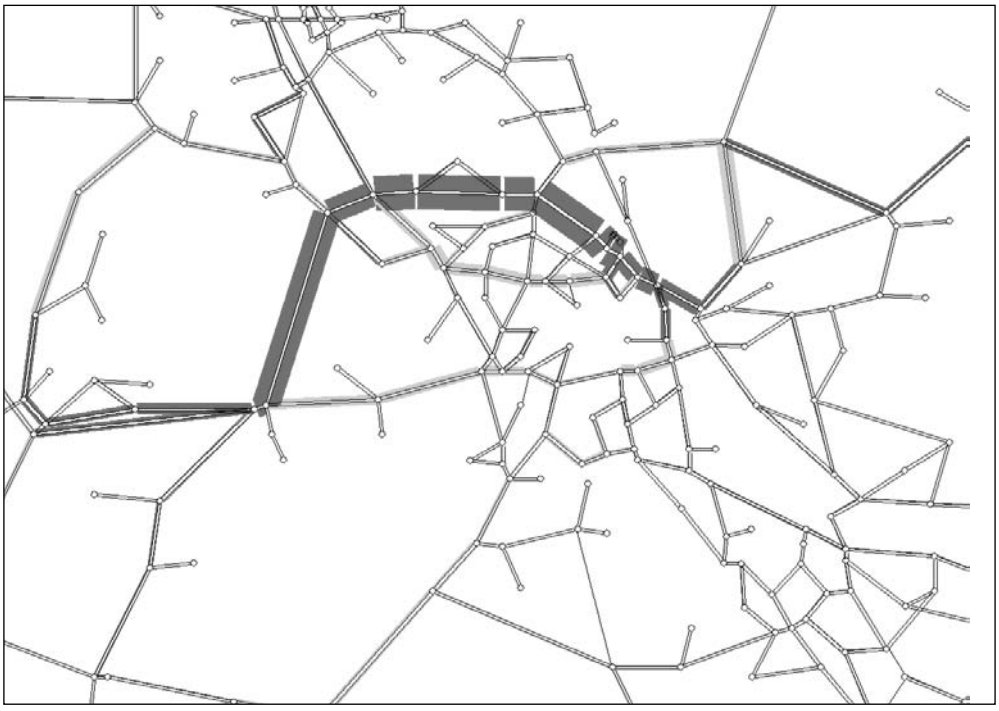


Figure 18.10 Resulting difference between equilibrium solution with imposed adjustments and initial scenario

We solve the side-constrained traffic equilibrium problem (18.4), where the formulated constraints are introduced as the side constraints. The solution to this problem is a link flow satisfying the management goals. Denote this solution by $\bar{f}$. The difference between the initial link flow solution and $\bar{f}$ is shown in Figure 18.13. In the figure, the link widths are proportional to the difference in flow, and the darker links indicate a decrease and the lighter links indicate an increase in link flow.

Given the target link flow solution $\bar{f}$, an instance of the target flow equilibrium pricing problem is constructed. The objective function is chosen as in (18.9). The link weight $\bar{v}_a$ on each link is chosen as one of the three values 1, 10 and 100. The value 100 is set on links where adjustments are undesirable. Links close to where the flow goals were introduced have a weight of 1 and links in between these sets have a link weight of 10. The link weights are shown in Figure 18.14, where each link width is proportional to the weight chosen. No additional restrictions are imposed on the travel time adjustments. The solution to the target flow equilibrium pricing problem is shown in Figure 18.15. We observe that the adjustment on two of the links is around 53 seconds each.

Additional restrictions on individual link travel time adjustments are now introduced to restrict the travel time adjustments to be less than 40 seconds. The target flow equilibrium pricing problem is re-solved and the resulting travel time adjustments are shown in Figure 18.16. Adjustments are present also in some parts of the network not shown in the figure.

Figure 18.11 Travel time adjustments under constraints

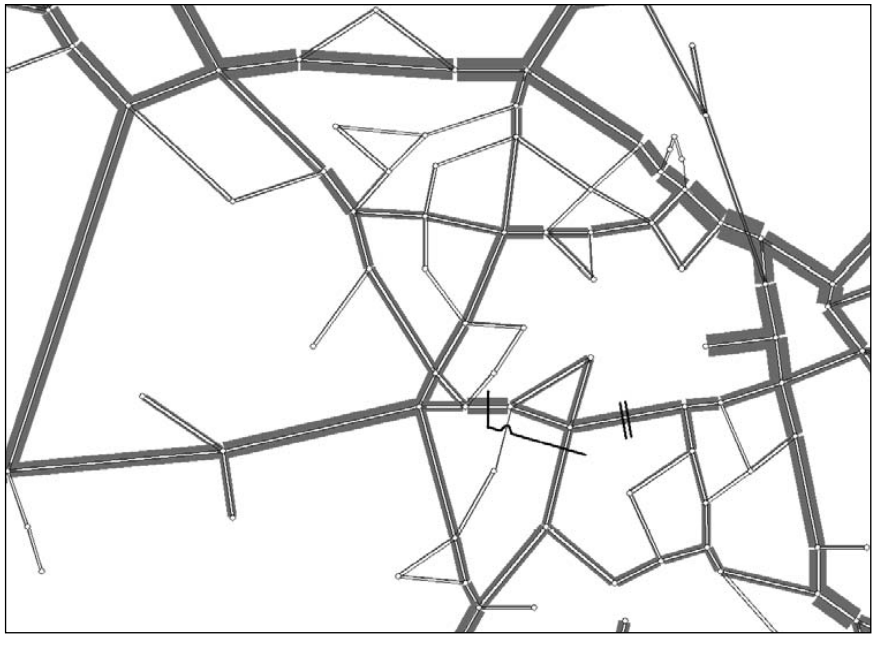


Figure 18.12 Equilibrium link flows in the central part of Linköping

Figure 18.13 Difference between equilibrium solution and side-constrained equilibrium solution satisfying the management goals

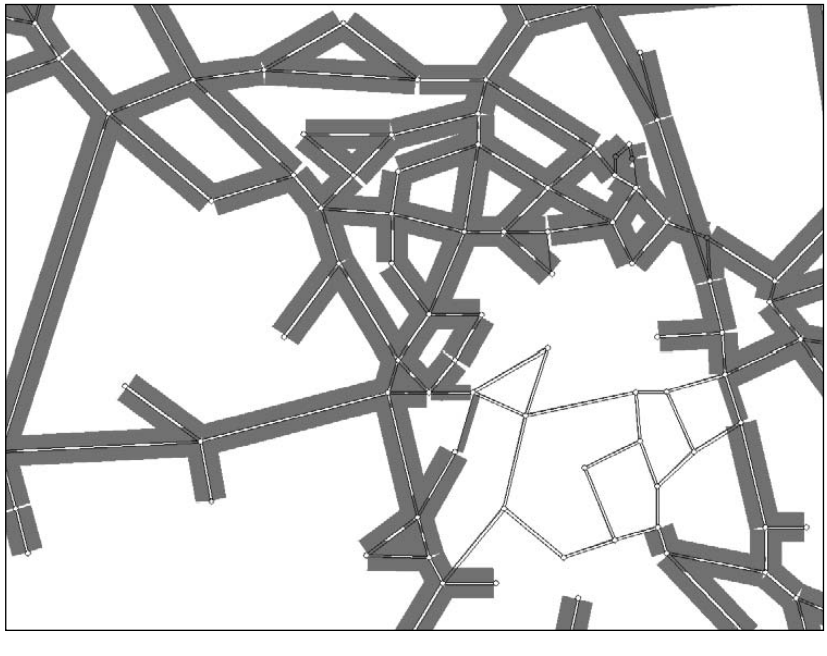


Figure 18.14 Link weights for the target flow equilibrium pricing problem

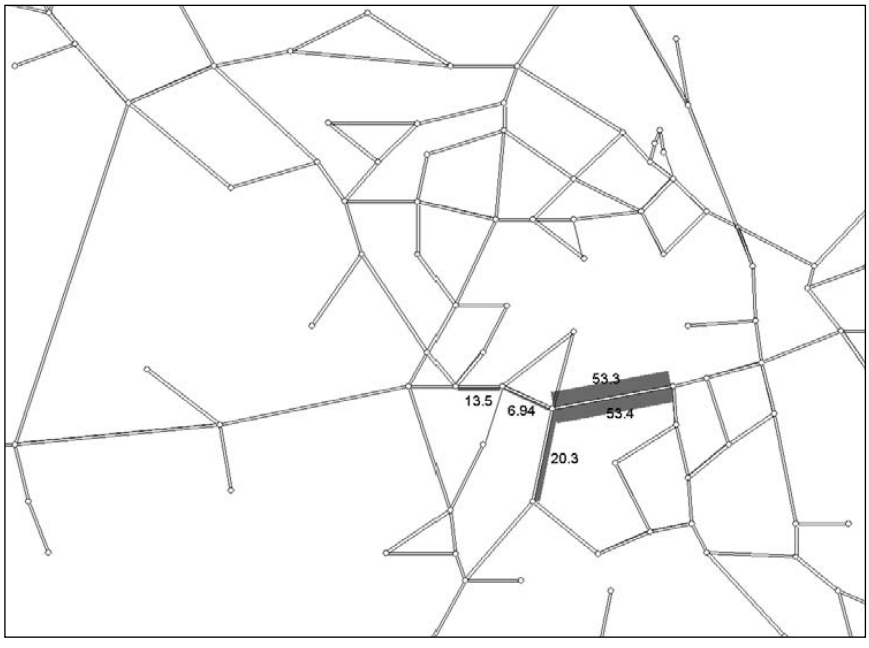


Figure 18.15 Set of minimal unconstrained travel time adjustments

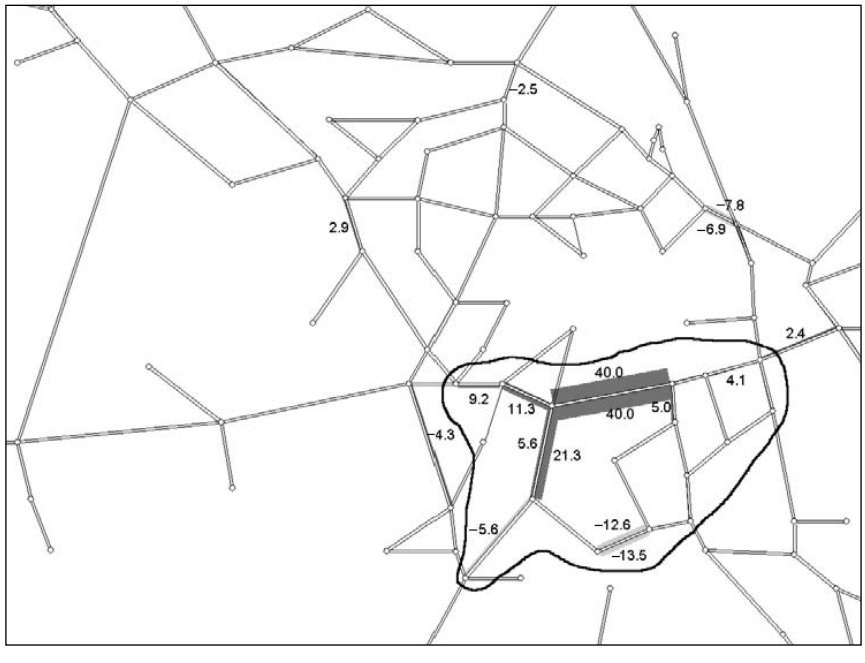


Figure 18.16 Travel time adjustments under constraints

Now, we impose further restrictions on the travel time adjustments such that adjustments on individual links are less than 40 seconds, and such that it is only possible to impose adjustments on the links which are inside the region marked in Figure 18.16. By solving the feasibility check problem (18.6) with the additional restrictions on the travel time adjustments, the corresponding target flow equilibrium pricing problem is shown to be infeasible. The link weights $\bar{w}_a = 1$, $a \in \mathcal{A}$, were used. A set of feasible travel time adjustments $\bar{\mu}$ obtained from solving the problem (18.6) are shown in Figure 18.17. We compute an equilibrium solution based on the feasible travel time adjustment from the problem (18.6). We modify the generalized travel times by adding the adjustments shown in Figure 18.17 and solve the resulting traffic equilibrium problem to obtain the link flow solution $\bar{f}^2$. The difference between $\bar{f}^2$ and the link flow solution in the initial scenario is shown in Figure 18.18. When comparing the link flows on the road segment where the aim was to reduce the total flow to 80 per cent of the flow in the initial scenario with the flows when the adjustments shown in Figure 18.17, we observe that the flow is reduced to 82 per cent of the flow in the initial scenario. If we consider the total flow on the other two links, where the aim was to reduce the inflow to 80 per cent we observe that the flow is reduced to 81 per cent. From the solution to the target flow equilibrium pricing problem where the target flow vector is chosen as $\bar{f}^2$, we conclude that the minimal perceived travel time adjustment to achieve the flow differences shown in Figure 18.18 are to be obtained by using the travel time adjustments given in Figure 18.17.

From the numerical illustration, we observe that the management procedure is not capable of finding a traffic flow which satisfies the goals by travel time adjustments which

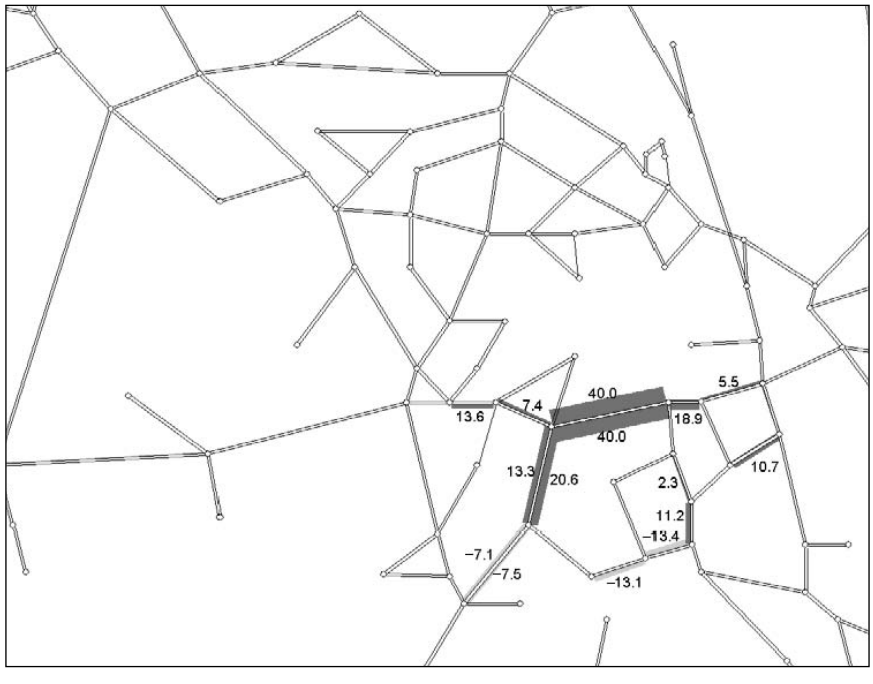


Figure 18.17 Feasible set of travel time adjustments

Figure 18.18 Resulting difference between equilibrium solution with imposed adjustments and initial scenario

are feasible with respect to the imposed restrictions. However, this does not necessarily imply that this instance of problem (18.3) does not have a solution. However, the procedure may be guided to find a solution which satisfies the formulated restrictions on the adjustments used and for which the link flow solution differs from the flow goals by a few per cent only. To find a set of adjustments such that the equilibrium link flows satisfy the original flow goals with the given procedure, it is necessary to augment the set of possible adjustments, either to allow larger adjustments or to allow adjustments on a larger number of links. The total computational time spent in Scenario 2 is about 30 minutes.

4. NUMERICAL RESULTS FROM A DIRECT SEARCH PROCEDURE

The two-stage traffic management procedure, illustrated in Section 3, can be seen as a heuristic for finding a solution to the bilevel problem formulation (18.3) of the traffic management problem. In this section we apply the Hooke and Jeeves (1961) direct search method with discrete steps to the bilevel problem, and compare the results to ours.

An overview of properties of, and solution procedures for, bilevel problems can be found in, for example, Luo et al. (1996) and Bard (1998). A reformulation of the bilevel model is made before the direct search method is applied. The constraints that define the

set $\mathcal{G}$, that is, $g_k(f) \leq 0$, $k \in \mathcal{K}$, are penalized using an exact penalty in the objective function, giving an equivalent bilevel problem:

$$\min \varphi(\rho) + M \sum_{k \in \mathcal{K}} \max [0, g_k(f)] \tag{18.10a}$$

subject to

$$\rho \in \mathcal{P} \tag{18.10b}$$

$$f \textit{ is} \text{ an equilibrium link flow solution given } \rho, \tag{18.10c}$$

where M is a sufficiently large constant. In the numerical examples we have chosen the objective function:

$$\varphi(f,\rho) = \sum_{a \in \mathcal{A}} \bar{v}_a |\rho_a|.$$

The weights $\bar{v}_a = 1$, $a \in \mathcal{A}$, are used in the tests. The objective function in the bilevel problem (18.10) is in general not convex due to the dependency of the lower level with respect to ρ. We exemplify the effects of the nonconvexity on the Sioux Falls problem. A flow capacity constraint is introduced on one link in the network, requiring that the link flow is reduced by approximately 95 per cent of the equilibrium link flow solution. One adjustment variable ρ is used to adjust the travel time on the selected link. In Figure 18.19, the objective function value $\varphi[f(\rho), \rho]$ shown for travel time adjustments, ρ, in the interval $[0.2, 0.7]$ for $M = 1$ is shown. Two local minima can be observed. At the local minimum with $\rho = 0.32$, the flow capacity constraint is not satisfied. As can be expected, the existence of this local minimum, corresponding to an infeasible solution, cannot be circumvented by increasing the penalty parameter M further.

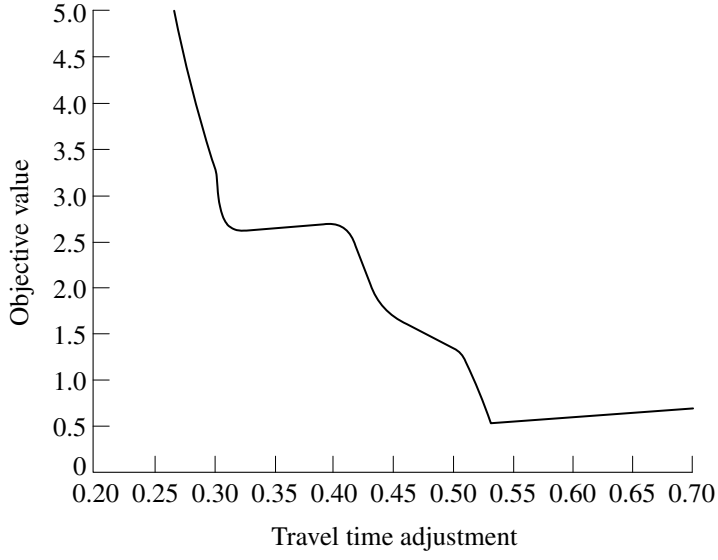


Figure 18.19 Objective values for one link travel time adjustment

The Hooke and Jeeves direct search method with discrete steps is applied to the problem (18.10). The search method is implemented according to the scheme given in Bazaraa et al. (1993, Section 8.5). In the search procedure, the objective function is evaluated for a systematic choice of values on the ρ variables. In each objective function evaluation, a traffic equilibrium problem is solved where the link travel times in the traffic equilibrium problem are given by $t_a(f_a)+\rho_a$ for $a \in \mathcal{A}$. Since the traffic equilibrium problems differ from iteration to iteration by changes in link travel times only, we have chosen to solve them using the disaggregate simplicial decomposition (DSD) method of Larsson and Patriksson (1992), due to its good reoptimization facilities. As a result of the nonconvexity of problem (18.10), the search procedure will be sensitive to the choice of the starting solution. In the numerical experiments, several start solutions are tested. Numerical evaluations of the direct search method of Hooke and Jeeves, sensitivity analysis based techniques and a genetic algorithm for solving this type of bilevel problem are earlier presented in Huang and Bell (1998).

4.1 The Sioux Falls Example

In this section, numerical results from the direct search procedure applied to a traffic management scenario in the Sioux Falls network are shown. The numerical results are presented for the scenario used in Section 3.

In each evaluation of the objective function, the traffic equilibrium problem is reoptimized to an accuracy such that the relative error between the upper and lower bound on the objective function is smaller than 10^{-4} per cent. The initial step size in the Hooke and Jeeves algorithm is set to $\Delta = 0.001$ and the acceleration factor is $\alpha = 1$. The method is terminated when $\Delta < 10^{-6}$.

In the initial experiment we choose the flow goals according to Figure 18.2, above. The restrictions on the travel time adjustments are chosen such that adjustments are allowed only on links in and out of node 15 in the network. The search procedure is initialized with a number of different values on the ρ variables. The first starting solution is chosen as the set of travel time adjustments obtained from the two-stage procedure (see Figure 18.4, above). The direct search procedure identifies this starting solution as a potential local minimum and the search method is terminated with the starting solution as the best solution found. This solution has an objective value of 0.476. The results for two other starting solutions are given in Figures 18.20 and 18.21. The objective values for these three solutions are 0.421 and 0.533, respectively. For the solution shown in Figure 18.20, the flow constraints formulated from the management goals are all binding and for the solution in Figure 18.21 the flow constraint restricting the inflow from the North, West and South, is not binding.

In a second experiment, we modified the restrictions such that travel time adjustments are allowed on the links marked in Figure 18.4 above, except on the link entering and exiting node 15 from the East. Results from two different starting solutions are given in Figures 18.22 and 18.23. The objective values of these two solutions are 0.739 and 0.499, respectively. The computational time for the solution given in Figure 18.22 is around 30 minutes and 176 iterations and 2872 objective evaluations were performed.

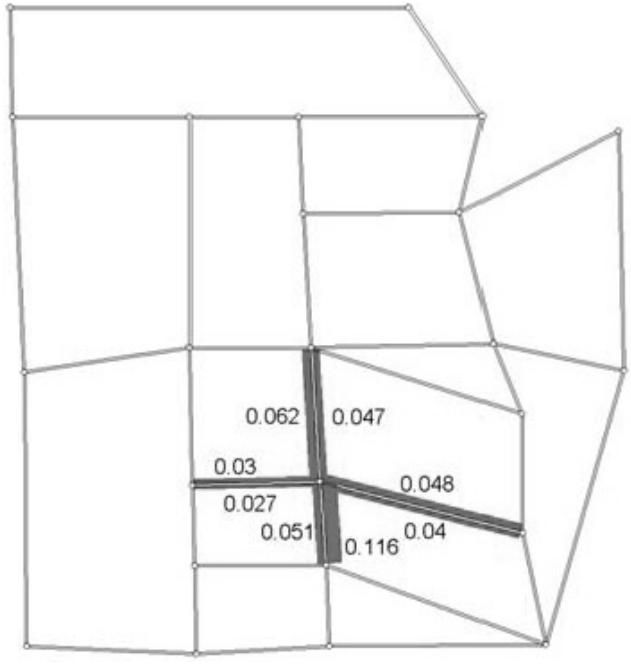


Figure 18.20 Adjustments resulting from first start solution in scenario 1

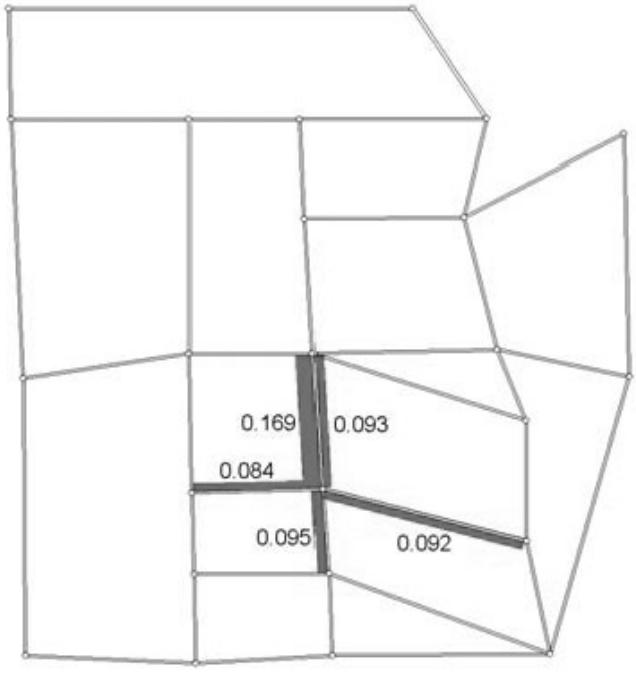


Figure 18.21 Adjustments resulting from second start solution in scenario 2

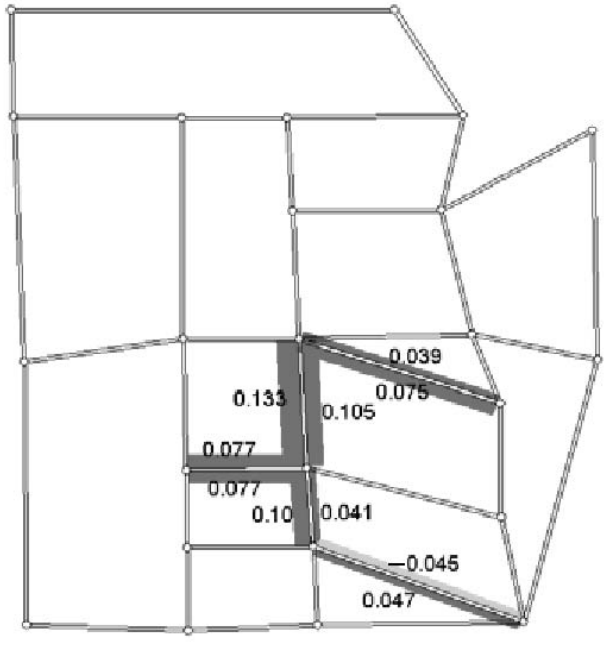


Figure 18.22 Adjustments resulting from first start solution in scenario 2

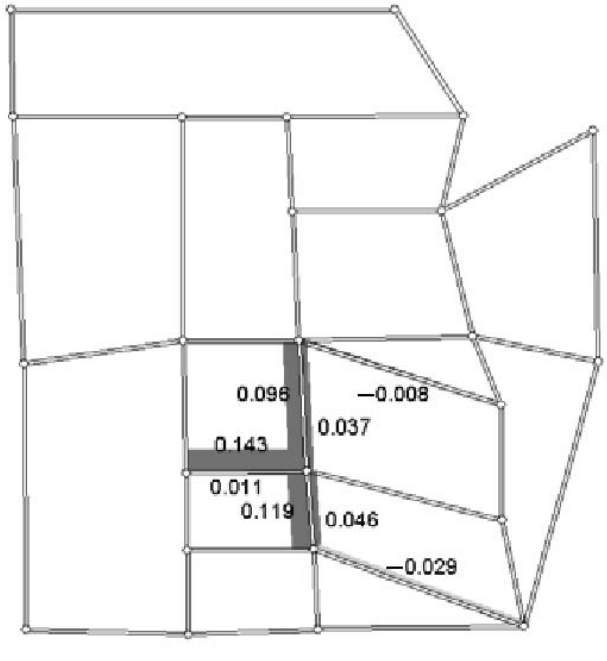


Figure 18.23 Adjustments resulting from second start solution in scenario 2

4.2 The Linköping Example

In this section, results from the direct search procedure applied to an instance of the bilevel problem (18.10) for the Linköping network are presented. The numerical results are presented for the second of the two scenarios that were used to illustrate the two-stage traffic management procedure in Section 3. The flow goals are to reduce to inflow from two links and on a road segment to 80 per cent of the flow in the current scenario. The flow constraints are shown in Figure 18.9, above. The travel time adjustments are restricted to zero on all links outside the zone marked in Figure 18.11, above, and are bounded to be below 40 on all links inside the zone.

In each iteration of the direct search procedure, the traffic equilibrium problem is re-optimized such that the relative error between the upper and lower bound generated by the DSD algorithm is smaller than 10^{-4} per cent. Further, the initial step size in the Hooke and Jeeves method is set to $\Delta = 5.0$ and the acceleration factor $\alpha = 1$. The method is terminated when $\Delta 10^{-5}$. The search method is initialized with a number of different values of the ρ variables.

In the first experiment, the adjustments resulting from the traffic management procedure used in Section 3.2 are used as the starting solution, that is, the direct search is initialized with the adjustments shown in Figure 18.17, above. The search procedure terminated with the solution shown in Figure 18.24. Two solutions which result from two other initial values of the ρ variables are shown in Figures 18.25 and 18.26. The three solutions correspond to feasible management actions, both with respect to the management

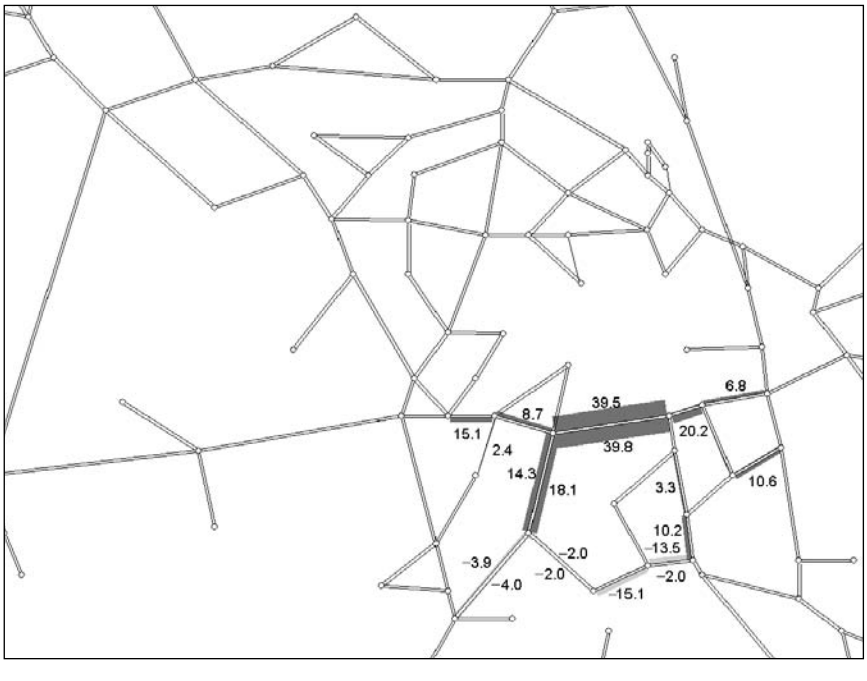


Figure 18.24 Travel time adjustments from first start solution

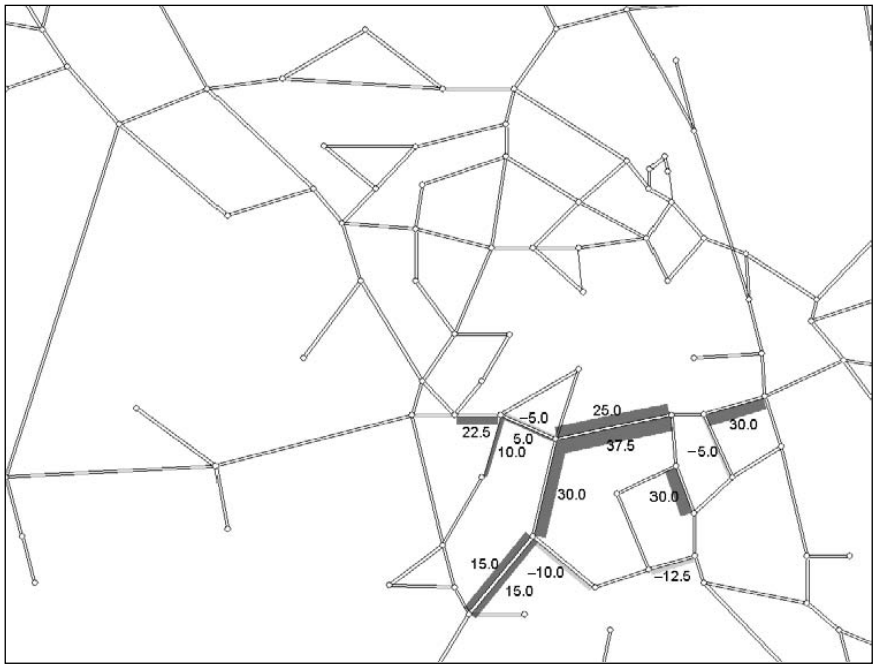


Figure 18.25 Travel time adjustments from second start solution

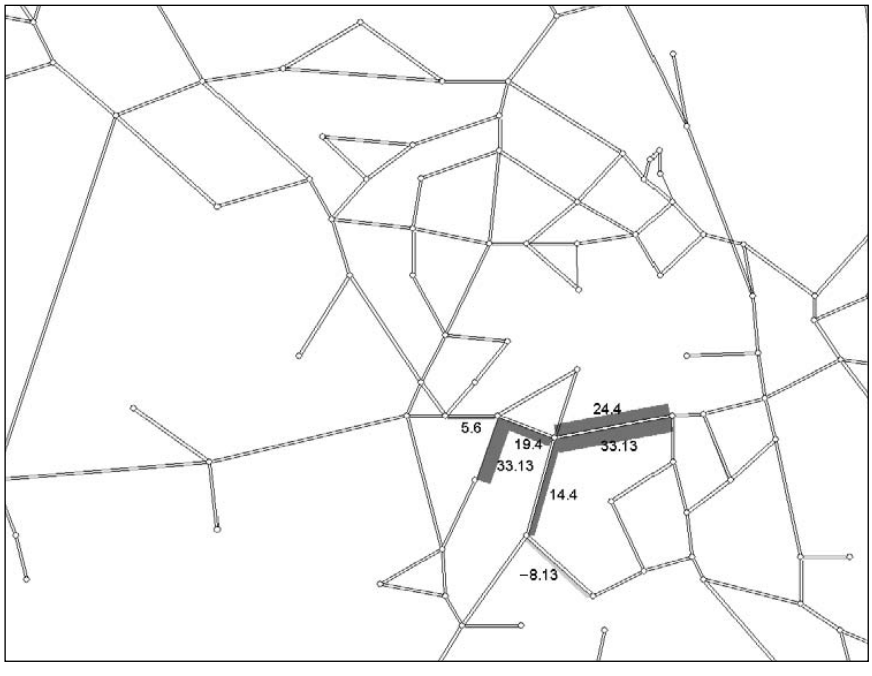


Figure 18.26 Travel time adjustments from third start solution

goals and the feasible actions. For the solution shown in Figure 18.25 the flow constraints formulated from the management goals are both binding. In the solutions in Figures 18.24 and 18.26 the leftmost flow constraint shown in Figure 18.12, above, is not binding.

The computational time for one instance of the Linköping scenario is approximately eight hours. The corresponding computational time for the two-stage procedure is around 30 minutes. From the results, we observe that the direct search procedure does generate significantly different solutions at termination for different initial values of the travel times adjustment variables.

Three observations can be made from the numerical results from the direct search procedure. First, the procedure is capable of finding travel time adjustments for which the traffic flows satisfy the flow goals, in problem scenarios where the two-stage procedure failed. Second, the solution at termination of the direct search procedure is highly dependent on the initial solution. Further, the direct search procedure is computationally very demanding. Instances of the Linköping problem where travel time adjustments are allowed on all links in the network, the computational time will probably be excessive. The computational time for the instances tested here, where adjustments are allowed only on a small subset of the links, is still orders of magnitude larger than the computational time for the two-stage procedure on the same scenarios.

5. CONCLUSIONS

Using numerical examples, we have shown how the proposed traffic management procedure can be applied to small- and medium-scale network scenarios. The order of the computations is highlighted and the flexibility and shortcomings of the methodology are illustrated. Examples show that the procedure can be guided such that approximate solutions to the management problem can be found in a reasonable computational time. Numerical experiments with a direct search method indicate that such procedures are computationally very demanding; it was capable of finding feasible solutions to the bilevel formulation of the management problem, but the quality of the solution at termination is highly dependent on the initialization of the procedure. In a comparison between the two-stage procedure and the direct search procedure, we observe that the direct search procedure has an advantage in finding feasible solutions for cases where the two-stage procedure, without guidance, does not find a solution; however, the two-stage procedure has the advantage that it can provide a rough solution to the traffic management problem much more quickly.

REFERENCES

Bard, J.F. (1998), *Practical Bilevel Optimization: Algorithms and Applications*, Dordrecht: Kluwer Academic.

Bazaraa, M.S., H.D. Sherali and C.M. Shetty (1993), *Nonlinear Programming: Theory and Algorithms*, 2nd edn, New York: John Wiley & Sons.

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

Chen, M., D.H. Bernstein, S.I.J. Chien and K.C. Mouskos (1998), 'A simplified formulation of the toll design problem', paper submitted for presentation and publication to the Transportation Research Board, National Research Council, July.

Cree, N.D., M.J. Maher and B. Paechter (1998), 'The continuous equilibrium optimal network design problem: a genetic approach', in M.G.H. Bell (ed.), *Transportation Networks: Recent Methodological Advances*, Amsterdam: Pergamon, pp. 163–74.

Hooke, R. and T.A. Jeeves (1961), 'Direct search solution of numerical and statistical problems', *Journal of Association Computer Machinery*, **8**, 212–29.

Huang, H.J. and M.G.H. Bell (1998), 'Continuous equilibrium network design problem with elastic demand: derivative-free solution methods', in M.G.H. Bell (ed.), *Transportation Networks: Recent Methodological Advances*, Amsterdam: Pergamon, pp. 175–93.

Larsson, T., J.T. Lundgren, M. Patriksson and C. Rydergren (2000), 'A decision support methodology for strategic traffic management', in P. Marcotte (ed.), *Current Trends in Transportation and Network Analysis – Papers in Honor of Michael Florian*, Dordrecht: Kluwer Academic, pp. 147–64.

Larsson, T. and M. Patriksson (1992), 'Simplicial decomposition with disaggregate representation for the traffic assignment problem', *Transportation Science*, **26**, 4–17.

Larsson, T. and M. Patriksson (1998), 'Side constrained traffic equilibrium models – traffic management through link tolls', in P. Marcotte and S. Nguyen (eds), *Equilibrium and Advanced Transportation Modelling*, New York: Kluwer Academic, pp. 125–51.

Larsson, T. and M. Patriksson (1999), 'Side constrained traffic equilibrium models – analysis, computation and applications', *Transportation Research*, **33B**, 233–64.

Larsson, T., M. Patriksson and C. Rydergren (2004a), 'A column generation procedure for the side constrained traffic equilibrium problem', to appear in *Transportation Research B*.

Larsson, T., M. Patriksson and C. Rydergren (2004b), 'Inverse nonlinear multicommodity network flow optimization by column generation', to appear in *Optimization Methods & Software*.

LeBlanc, L.J., E.K. Morlok and W.P. Pierskalla (1975), 'An efficient approach to solving the road network equilibrium traffic assignment problem', *Transportation Research*, **9**, 309–18.

Luo, Z.Q., J.S. Pang and D. Ralph (1996), *Mathematical Programs with Equilibrium Constraints*, Cambridge: Cambridge University Press.

Marcotte, P. (1986). 'Network design problem with congestion effects: a case of bilevel programming', *Mathematical Programming,* **34**, 142–62.

Marcotte, P. and G. Marquis (1992), 'Efficient implementation of heuristics for the continuous network design problem', *Annals of Operations Research*, **34**, 163–76.

Migdalas, A. (1995), 'Bilevel programming in traffic planning: models, methods and challenge', *Journal of Global Optimization*, **7**, 381–405.

Patriksson, M. (1994), *The Traffic Assignment Problem – Models and Methods*, Utrecht: VSP.

Patriksson, M. and R.T. Rockafellar (2003), 'Sensitivity analysis of variational inequalities over aggregated polyhedra, with application to traffic equilibria', to appear in *Transportation Science*.

Yang H. and M.G.H. Bell (2001), 'Transport bilevel programming problems: recent methodological advances', *Transportation Research*, **35B**, 1–4.

19. Free trade and transportation infrastructure in Brazil: towards an integrated approach

Paulo Resende, Joaquim J.M. Guilhoto and Geoffrey J.D. Hewings

1. INTRODUCTION

In modeling the welfare gains associated with the development of free trade agreements, little attention has been paid to the mechanism by which the projected increased flows of goods and services will be moved between the countries. For the most part, the models have implicitly assumed an acceptable transportation infrastructure with enough capacity to absorb increased demand on the highway systems. However, in developing economies, these assumptions are less tenable; in the euphoria surrounding the creation of MERCOSUL, the free trade agreement between Brazil, Argentina, Paraguay and Uruguay, the promise of enhanced trade was not measured against some fundamental realities. For example, it was noted that three-quarters of all terrestrial trade between Brazil and Argentina (with some additional trade with Chile) uses a single bridge across the Uruguay River (*The Economist* 1996). In fact there are only two other bridges linking these two countries; obviously, increased trade will face significant transportation and transfer costs, and these have not been prominent features of most general equilibrium models that have explored MERCOSUL to date.

This chapter offers a modest first step in an attempt to place the gains from trade onto a highway transportation network in the hope that some of the major sources of capacity limitations and bottlenecks can be identified. The Brazilian transport configuration is currently characterized by highly concentrated flows of products and services on its highway system, whereas the railroad and inland waterway modes are not considered to be of primary importance. Notwithstanding this concentration, where more than 60 per cent of the general freight is carried by heavy highway vehicles, there have been few efforts directed to continuous and endurable planning for maintenance, operation, safety, and even modest expansion of the highway network, especially within the last two decades. The other modes, such as railways, inland waterways, and coastal navigation, have offered non-competitive levels of productivity and efficiency. For example, inland waterways account for just 1 per cent of the general freight, notwithstanding the fact that the country has approximately 20 000 miles of navigable rivers. The railroad system, once considered a critical element in the process of territorial integration, has faced a tremendous deterioration in its equipment and operational network and in the morale of the personnel, so that the Brazilian rail transportation has come perilously close to financial disaster; recent

privatization efforts (many involving international investors) may offer some hope for a rebound in its fortunes.

In this chapter, a set of simulation exercises was conducted to examine the likely consequences of the trade agreements, especially MERCOSUL, on the efficiency, capacity, and development of the highway system in Brazil. The main reason to consider MERCOSUL as the main source of trade impacts and changes in the traffic volumes is exactly because this trade agreement is the major factor to the increase of demands for better and more competitive transportation facilities in Brazil. Accordingly, capacity problems, which cannot be solved in the short run, are evaluated in terms of their potential costs to the economy – that is, lost output or increased highway transportation costs that reduce the competitiveness of the sectors. In this case, it is assumed that highway bottlenecks result in higher costs when compared to efficient and productive transportation links. One of the major contributions of this analysis is the identification of the specific role highway transportation plays in the economy, based on which some suggestions for priorities in investments are presented, for the existing facilities should be upgraded in terms of capacity levels. Naturally, due to space limitations, many of the details of the original study could not be presented; however, the general findings of the impacts of MERCOSUL upon the current Brazilian highway transportation system still offer some important insights.

2. THE OVERALL MODELING SYSTEM

A myriad of economic models have been employed to project the impacts of free trade agreements among countries, or regions within countries, but there have rarely been initiatives to measure the effects, or impacts, of such agreements upon the efficiency of highway transportation networks even to the extent of measuring efficiency as the ratios between traffic volumes and operational capacities. When a transportation system is already built and planned to accommodate the increases in demand, the impacts of the economic changes are relatively and proportionally spread out along the spatial and temporal arrangement of the traffic volumes. Such a dynamic process becomes even more perceptible when a highway transportation system exists and operates under highly efficient standards of travel time, operational costs, and safety. However, when the existing transportation infrastructure is deficient and lacks the basic elements to properly move the products along the networks, the economic impacts of any proposed increases in demand are likely to affect the routes negatively, leading to additional bottlenecks or critical segments. In most developing countries, this latter option is the one that is found more often; it arises in large part because the transportation infrastructure needs massive investment to avoid the transportation facilities becoming a barrier to the global competitiveness of the free trade member countries. Accordingly, it is necessary to evaluate and analyse the impacts of the free trade agreements on transportation facilities through a methodological approach where the links between trade and transport can be quantified and applied to the parameters that measure the efficiency of moving goods and services.

In this chapter, MERCOSUL in South America is taken as the paramount element to generating economic changes in Brazil, where its impacts on the Brazilian highway transportation infrastructure may lead to additional bottlenecks and, consequently, to opportunities to help realize the gains from international trade. To facilitate this exploration,

two economic models and one transportation model are combined and analysed together, so that the economic impacts can be transposed to effects on efficiently moving goods and services within the highway network. The first economic model is a macroeconometric model providing annual forecasts for the Brazilian economy at the macro level up to the year 2014 (for a description of this model, see Fonseca 1991). The results of this model are then fed into a five-region interregional model of the Brazilian economy (see Figure 19.1 for the location of the macro regions). The forecasts generated by the macroeconometric model are used as a guide to determine how the five regions and their associated economic sectors in the interregional model would grow during the period, as well as the flows of goods among the regions. The transportation model is based on the relationship between traffic volume and operational capacity (V/C) for each highway segment. The parameter V/C can be stratified into ranges of levels of service, where each level of service is a qualifying endowment of the V/C ratio.

Any sort of mathematical approach to analysing trade–transportation data and, by any means, to numerically linking these two concepts, demands a type of theoretical background that is able to support the assumptions and lines of thoughts of the mathematical analysis itself. In this study, the theoretical foundation behind the mathematical exercises is constructed through the specification of regional parameters. Within each region, each highway segment links a production pole to a consumption pole. Accordingly, the goods and services are transported along the existing highway facilities, leading to traffic volumes that can easily be quantified. Each segment, however, has an operational capacity limited to geometric conditions, percentage of heavy vehicles in the traffic mix, terrain, drivers' composition, and so on. Therefore, several adjustments need to be made in the development of the relationship between volume and capacity, and thus the determination of the level of service for each highway segment.

With the assumption that each highway segment will move the production of a certain pole in the direction of the consumption center, an expansion in consumption or production in any of these poles will lead to an increase in the traffic volume of the segment linking the two poles, thereby altering the volume-to-capacity ratio. This is exactly the principle that will direct all the analyses in this study, where MERCOSUL is the main generator of trade factors that will change the production/consumption levels of the Brazilian poles.

The chapter provides a mechanism to link the trade impacts resulting from the economic model and the highway transportation parameters that identify the operational conditions of the highway facilities. From the economic model, given the implementation of MERCOSUL, a series of impact coefficients among the regions within Brazil can be generated. These coefficients must affect the highway transportation facilities in such a way that the increases in transportation demands reflect the magnitude of the economic forecasts. In other words, the changes in traffic volumes, given the free trade agreements, must mirror the economic impacts, for they are quantified by the economic coefficients. Based on the nature of the trade–transportation linkages, the main hypothesis is that within a certain period of time, MERCOSUL would generate percentage change in interregional and intraregional trade in Brazil, and that these changes, depicted by the impact coefficients, would affect the highway system, with increases in the traffic volumes, which will result in increases of the V/C ratios, thereby leading to decreases in the levels of service of the highway segments.

When the two models are aggregated and the impacts of MERCOSUL in the Brazilian highway transportation systems are evaluated, the operational efficiency of the highway segments can be investigated by answering the following questions:

- Are the goods and services being moved along the transportation network through the most efficient highway routes?
- Are there any problems of under- or overestimation of the network capacity?
- Is the distribution of flows across the highway being used in its most efficient way, within an optimized format as far as the economic context demands, that is, without creating bottlenecks?

By answering these questions, it would be possible to build a planning strategy that considers not only the current demands for a better way to transport goods and services, but also the projected effects of the free trade agreements on higher demands for a highway transportation system where competitiveness is taken as the economic target.

3. THE ECONOMIC MODEL

To obtain the results for the impact of MERCOSUL over the flow of the goods and services inside and among the Brazilian regions, an interregional model for the Brazilian economy was constructed at the level of five regions (South, Southeast, North, Northeast, and Central West, as shown in Figure 19.1), 26 sectors and 43 categories of goods and services.

Figure 19.1 Macro regions of Brazil

The forecasts at the macro level, to the year 2014, are derived from an econometric model that was used as a guide to determine how the intra- and interregional relations would change in the economy. In addition, as attention is directed essentially to international transactions, forecasts were also made for the growth of exports and imports of the five regions.

As the main focus of this study is the role of trade into the highway transportation infrastructure, the results from the economic models were used to create indexes of growth of the intra- and interregional flows of those goods that have a more intensive use of transportation systems. For the international flows, attention was directed to all the goods.

The results are summarized in Figures 19.2–14. What we can see is that the interregional flows grew more in the Central West, followed by the South, Southeast, North, and Northeast regions. For exports, the larger growth rates are found in the Central West, followed by the North, Northeast, South and Southeast regions, while imports grew more vigorously in the North, followed by the Central West, South, Southeast and Northeast. The higher growth rates of the external sector of the Central West and North are due to the fact that these regions have a low value of imports and exports in the base year, as compared to the other regions. In such a way, greater trade liberalization has a tendency to increase the external flow (in percentage terms) from these regions, more so than in the other regions.

When attention is directed to the interregional flows, what we find is an increase in the relations among the North, Northeast, and Central West regions. The exports of the North to the Central West and Southeast grow more than those to the South and Northeast regions. The imports from the Northeast and Central West increase considerably more than those from the other regions. The exports of the Northeast increase to the Central West while showing about the same growth to the other regions. The growth of imports from the other regions to the Northeast was at about the same level.

The exports of the Central West grew more to the North and Southeast and less to the Northeast and South. The imports grew more from the North and Northeast and less from the Southeast and South. The exports of the Southeast expanded more to the Central West, South, and Northeast than to the North. The imports grew more from the North and Central West and less from the South and Northeast. Finally, the exports of the South increased more to the Southeast, Central West, and Northeast, and less to the North; imports from the North and Southeast had higher growth rates than those from the Northeast and Central West.

4. THE TRADE–TRANSPORTATION LINKS

The aggregation of the trade–transportation models was conducted through several steps, where the primary concern was the spatial, or regional, standardization of the coefficients to be applied to the highway traffic volumes. Figure 19.15 shows a schematic sequence of these steps. The spatial partitioning used in the economic model, as related to MERCOSUL and Brazil, was defined by five macro regions shown in Figure 19.1. The transportation model grouped six corridors, based on the main routes of goods and services around the country. So, the transportation model can be characterized as a user-optimizing network production model taking into consideration the level of congestion in the critical

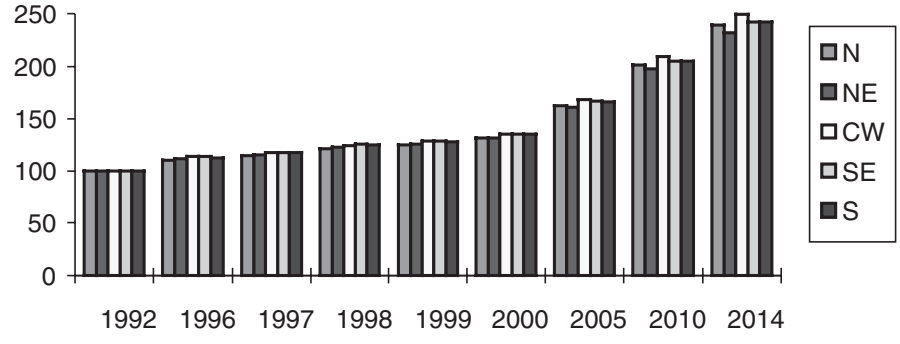


Figure 19.2 Index of growth of transport-demanding goods (1992 = 100)

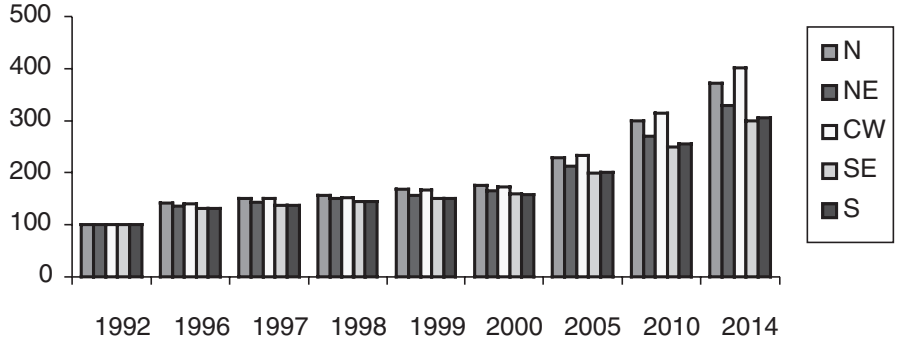


Figure 19.3 Index of growth of exports to the external market, by region (1992 = 100)

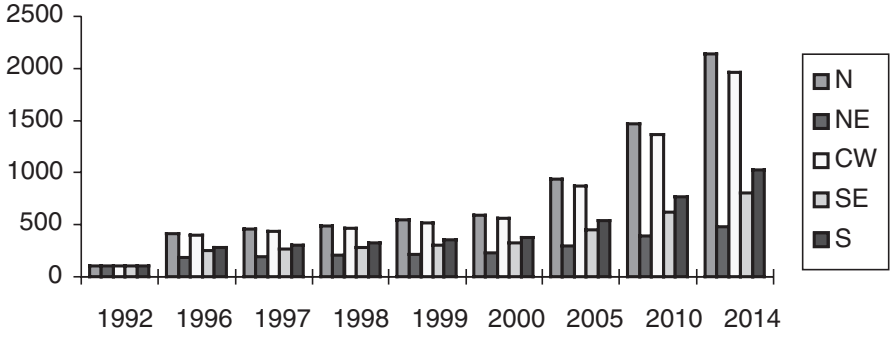


Figure 19.4 Index of growth of imports from the external market, by region (1992 = 100)

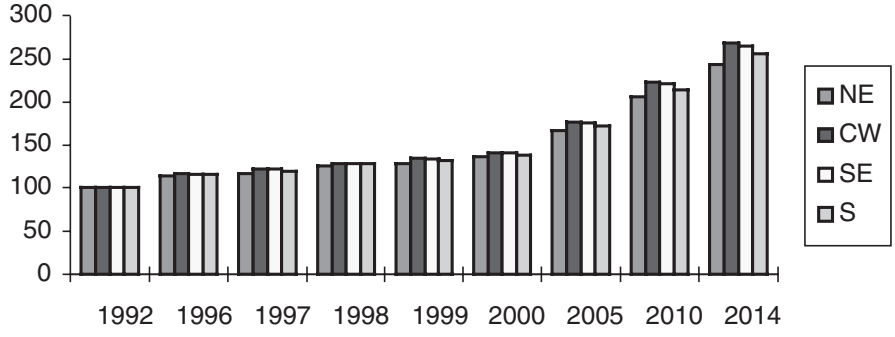


Figure 19.5 Index of growth of regional exports – North region transport-demanding goods (1992 = 100)

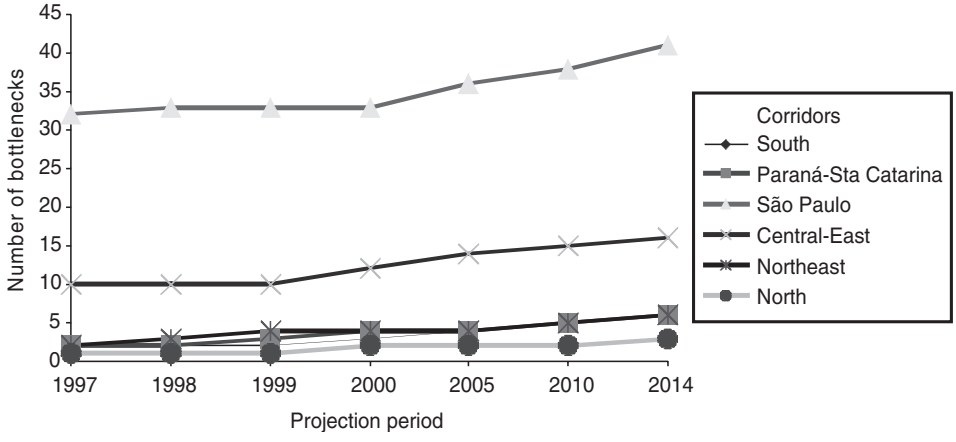


Figure 19.6 Index of growth of regional imports – North region transport-demanding goods (1992 = 100)

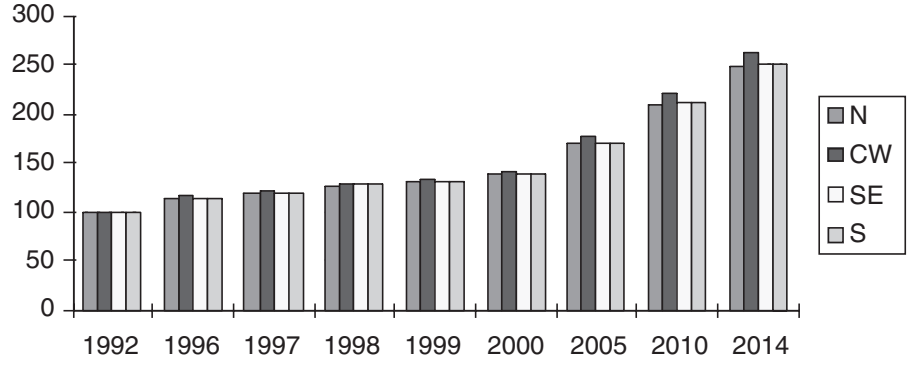


Figure 19.7 Index of growth of regional exports – Northeast region transport-demanding goods (1992 = 100)

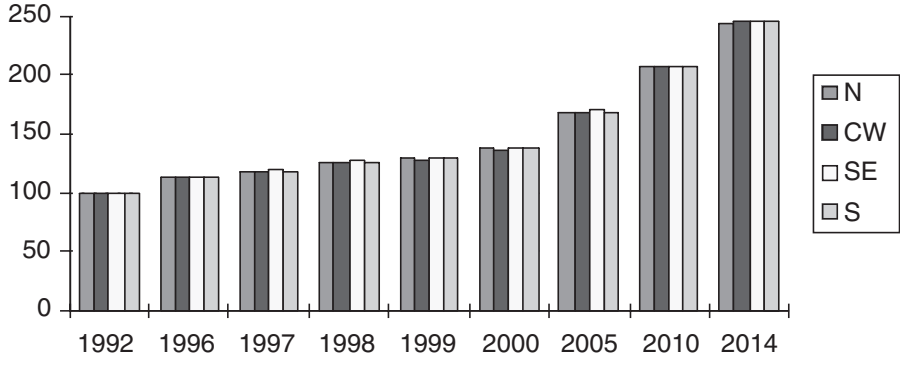


Figure 19.8 Index of growth of regional imports – Northeast region transport-demanding goods (1992 = 100)

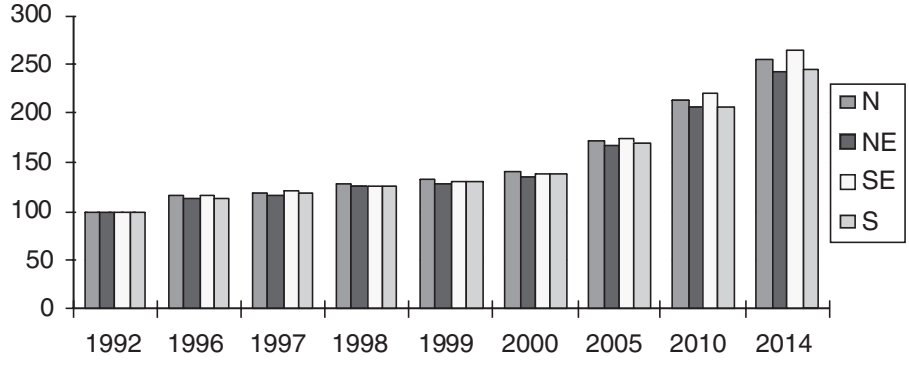


Figure 19.9 Index of growth of regional exports – Central West region transport-demanding goods (1992 = 100)

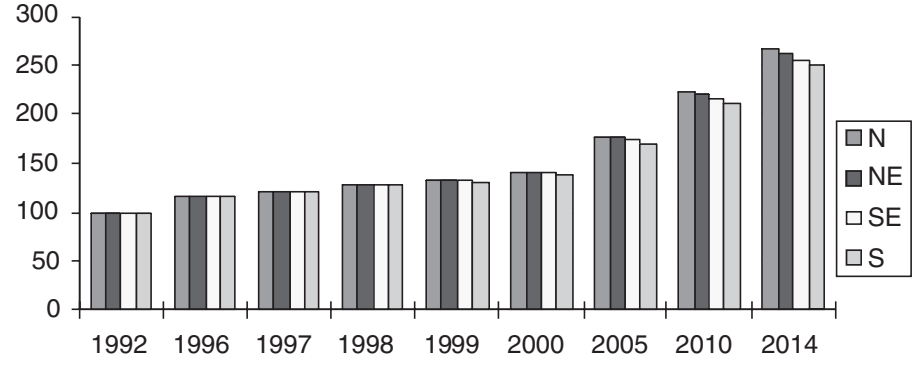


Figure 19.10 Index of growth of regional imports – Central West region transport-demanding goods (1992 = 100)

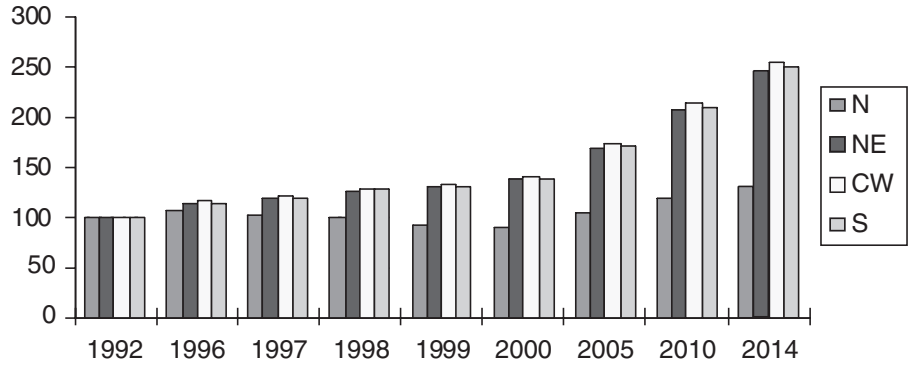


Figure 19.11 Index of growth of regional exports – Southeast region transport-demanding goods (1992 = 100)

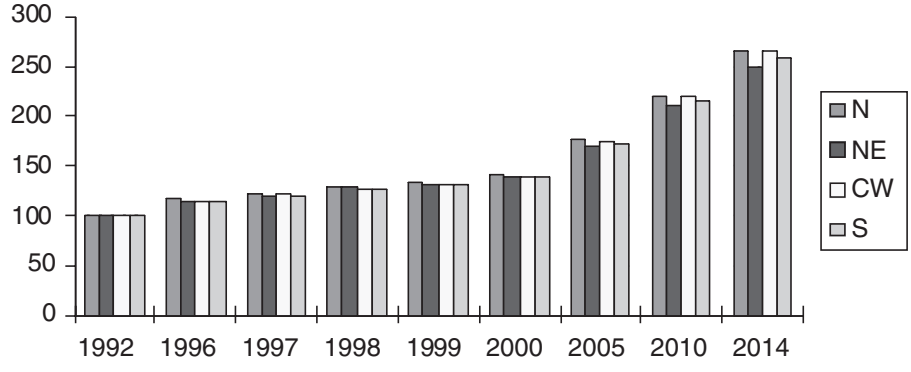


Figure 19.12 Index of growth of regional imports – Southeast region transport-demanding goods (1992 = 100)

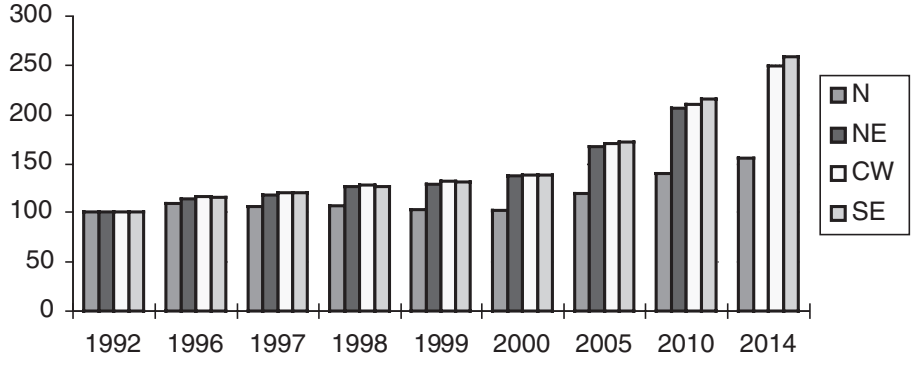


Figure 19.13 Index of growth of regional exports – South region transport-demanding goods (1992 = 100)

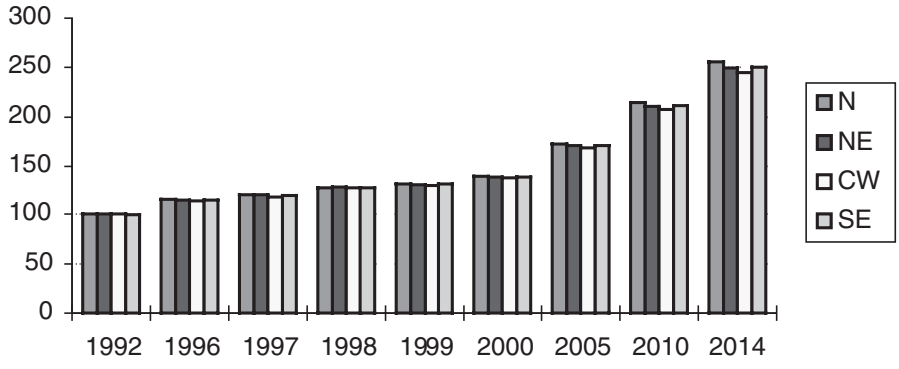


Figure 19.14 Index of growth of regional imports – South region transport-demanding goods (1992 = 100)

points of the system. From the economic models considered there had to be a statistically significant representation of products within each economic sector in the O–D (origin–destination) matrix.

Accordingly, data were collected on the origin and destination of 53 products, totaling approximately 300 million tons, and encompassing 23 out of 27 states in Brazil; the remaining four states are located in the Amazon region where production and consumption are not significant enough to create large highway flows. The sample size then represented 55 per cent of the total traffic volume on the Brazilian highways in an average year of the 1990s. Considering that the most important products would be those whose aggregate value is relatively low, but with high growth rates, the sample size showed may be claimed to be fully representative of the total O–D flow pattern in Brazil. Notwithstanding, several tests were conducted to evaluate the representation of the regional distribution of the products, so that a lack of equilibrium among the several production/consumption areas could be avoided.

The stepwise analysis to standardize the trade coefficients to be applied to the highway traffic volumes had the following sequence:

1. Within a certain forecasting period, there exists only one trade coefficient from, say, macro-region A to macro-region B, from macro-region A to macro-region C, and so on.
2. Since every corridor encompassed one or more macro regions, a simple average was calculated to represent only one trade coefficient to each corridor. It is important to emphasize that one region may be part of more than one corridor; that is why an average of the coefficient had to be considered. However, it is also important to note that, since the regional differences between the two models were not that critical, and since the nature of trade is naturally concentrated (gravity theory), the simple averaging procedure did not lead to significant distortions.
3. The flows among the corridors were identified by an O–D matrix. Since this matrix was composed of micro-regional production/consumption poles, it was possible to aggregate those poles into macro-regional groups within the corridors. In this way,

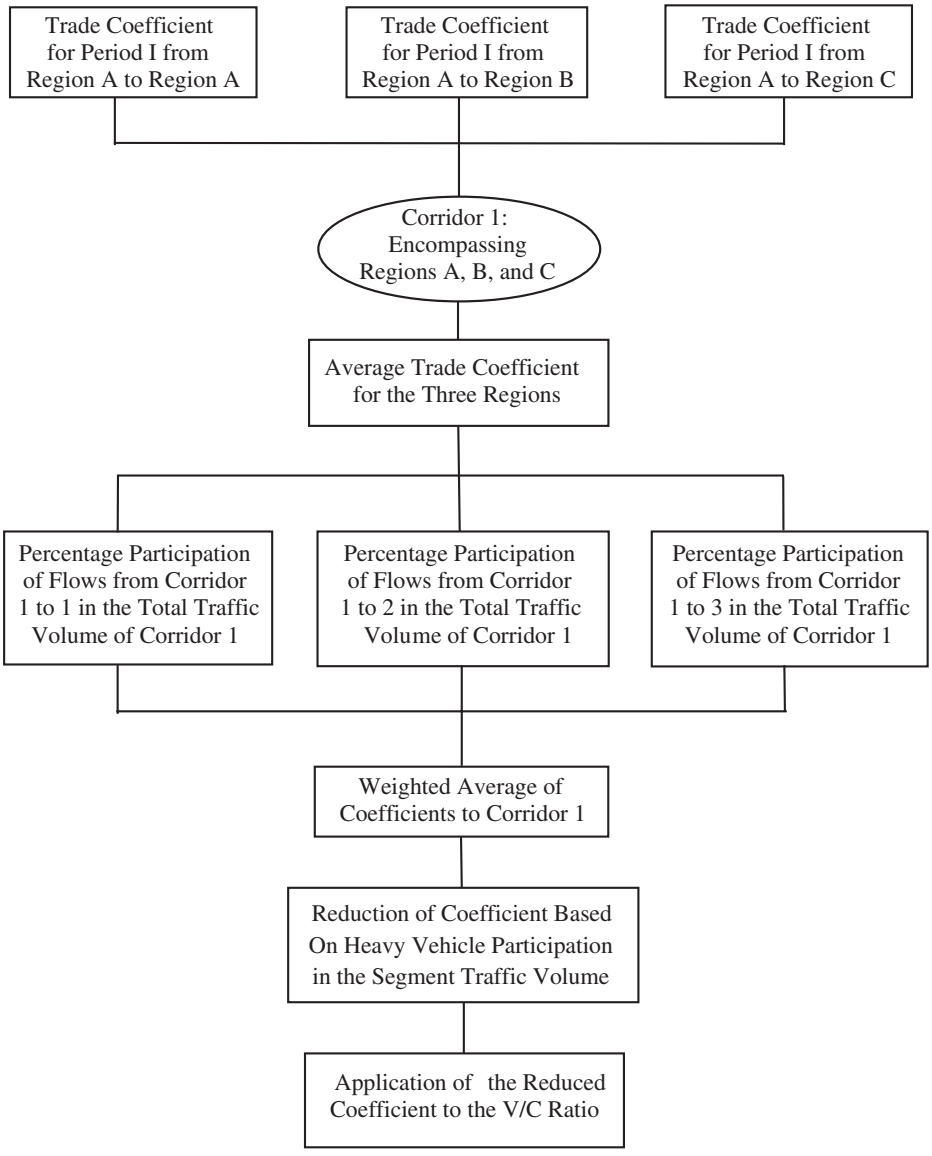


Figure 19.15 Stepwise analysis of the economic impacts of MERCOSUL: a three-region, three-corridor example

the percentage participation of each corridor-to-corridor flow as related to the total traffic volume of a certain corridor could be achieved.

4. By using these percentage participations as weights, a single weighted trade coefficient was created for each corridor, within a certain forecasting period.
5. Assuming that none of the products considered in the O–D matrix was subjected to significant transformations from the origin to the destination (that is, low aggregate

value and high growth weight), it was feasible to assume that a certain percentage change in trade could be directly and linearly applied to the highway traffic volumes. However, this linear effect would be valid only for the heavy goods traffic; we cannot assume that the changes in trade will linearly affect the use and thus the volume of passenger cars. Since the operational capacity takes into account the heavy traffic factor resulting from truck–passenger car equivalent values, the heavy goods traffic participation in the total traffic flow was taken as the major reduction factor in the trade coefficients.

6. After achieving a reduced trade coefficient for each highway segment, the V/C ratios were indexed by the trade coefficients to reflect the impacts of MERCOSUL in the highway transportation network of the six corridors.

Based on this stepwise analysis, it was possible to identify one trade coefficient for each forecasting period, from 1997 to 2014, within each corridor. The following equation shows the mathematical standardization of the trade coefficients for each corridor:

$$I_{c1i} = \left[I_{11} * W_1 + I_{12} * W_2 + I_{13} * W_3) / 100 \right] * R_{\%hv1}$$

where,

I_{c1i} weighted impact coefficient for corridor 1 in the i period for a certain highway segment;
I_{11} averaged coefficient resulting from the individual coefficients from regions of corridor 1 to corridor 1;
W_1 percentage participation in total flows of corridor 1 to corridor 1;
I_{12} averaged coefficient resulting from the individual coefficients from regions of corridor 2 to corridor 1;
W_2 percentage participation in total flows of corridor 2 to corridor 1;
I_{13} averaged coefficient resulting from the individual coefficients from regions of corridor 3 to corridor 1;
W_3 percentage participation in total flows of corridor 3 to corridor 1;
$R_{\%hv1}$ reduction coefficient based on the heavy vehicle percentage for a certain highway segment.

Given this stepwise analysis, it was possible to apply the trade coefficients to each V/C ratio from 1997 to 2014, thereby quantifying the deteriorations in the level of service of each highway segment. From this analysis, it was possible to identify the main bottlenecks within the highway system.

5. MAIN FINDINGS

Among the several analyses that have resulted from the mathematical linkage between the trade and the transportation models, one that has been of utmost importance is the trade effect that generates bottlenecks in the main highway segments, given the already deteriorated conditions of the Brazilian highway system. This study has shown that one of the

main consequences of the impacts of the MERCOSUL free trade agreements on the highway network is the increase in the number of critical points, or bottlenecks, in most of the corridors. Of course, the assumption was made that investments in transportation infrastructure that would be made over the forecast period would be allocated primarily for maintenance rather than for system expansion. According to Figure 19.6, there is an increasing pattern concerning the number of bottlenecks, generating negative consequences, mainly in those corridors where the primary routes cross largely industrialized regions, such as São Paulo, the biggest metropolitan area of South America, and in the Central-East, where the Rio de Janeiro and Belo Horizonte metropolitan areas are located. Almost all the main Brazilian industrial plants are concentrated in these two corridors, such as automobile, steel, fabricated metals, paper and cellulose, food processing and services. These three metropolitan areas account for more than 60 per cent of GDP.

Figure 19.6 shows, for example, that the impact of MERCOSUL on the highway system of the São Paulo corridor accounts for an increase of, approximately, 25 per cent in the number of bottlenecks within the projection period. Moreover, in the Central-East corridor, this increase reaches 60 per cent, going from 10 critical points to 16 in the year 2014. The same pattern can be realized for the other four corridors, the major difference being the existence of a smaller number of bottlenecks, but the deterioration patterns are similar to those observed in the São Paulo and Central-East corridors.

The influence of the free trade agreements in the highway systems shows the potentially negative consequences of not being prepared to handle the increases in demands. It is important to note that the deterioration of the level of service in one highway segment has a ripple effect that spreads out the operational condition of that bottleneck to a large area along the main route. In other words, once the quality of service on a highway segment has deteriorated, the effects in the main operational parameters are multiplied along other segments of the main and adjacent routes, leading to higher transportation costs, increases in travel time, decreases in safety and, by any means, reductions in competitiveness. In turn, this may translate into increases in the final prices of the products, potentially undermining the gains from trade. Therefore, this investigation has achieved its main objective, namely the detection of the influence of the implementation of MERCOSUL upon the main corridors of the Brazilian highway system. These negative effects will continue not only because of MERCOSUL, but also because of the increasing globalization of the economy that also affects the internal flows of goods and services.

6. CONCLUSIONS

In considering the impacts of MERCOSUL, together with the globalization and stabilization of the Brazilian economy, the highway transportation systems must be analysed through a regional perspective, where the trade characteristics of each area can be investigated and updated in response to improvements in the transportation infrastructure. The spatial distribution of activity is not uniform and the consequences of international trade agreements are unlikely to yield similar benefits to all regions of the country (see Haddad and Hewings 2001a, b). In central Brazil, what has been noticed is a limitation in this region's competitiveness due to a transportation system with high levels of deterioration, which directly reflects on travel time, transportation costs, safety and so on.

Based on these findings, a certain number of initiatives could be taken to improve the transportation system. First, transportation development has ignored, until recently, opportunities to expand the development of multimodal systems, to exploit the efficient competitive contributions of each mode and to relieve the overdependence on highway transportation exclusively. Second, regulatory reform has proceeded, albeit slowly, to reduce bureaucratic interference in the development of efficient transportation systems and to encourage foreign direct investment. An acceleration of the privatization processes of highways and railroads would certainly offer opportunities for much needed investment, investment that is unlikely to be forthcoming from state and federal governments facing significant fiscal problems. Third, there is a pressing need to enhance the participation of the railroad and inland waterway systems in the transportation system; at present, less than 1 per cent of interstate trade uses the extensive waterway system. Fourth, there is still a significant economic basis for capacity expansion of some highway segments, where the distance between nodes makes other forms of transportation less competitive.

Given this overall outlook for the Brazilian transportation system, there must be a planning program that provides a detailed inventory of the main investment targets, according to a spatial and temporal schedule of investments. If this program becomes a reality, then the transportation sector can be adjusted to enhance Brazilian competitiveness both in the internal and external markets. Nevertheless, it is important that the past and still current mode of thought of decision makers be changed to a more comprehensive analysis of the role to be played by infrastructure systems, so that the financial resources will not be applied to inefficient transportation segments.

While Brazil-specific investment needs are not known, it was estimated (*The Economist* 1996) that the southern South American countries would need $20 billion in infrastructure investment just to keep pace with demand growing at 4–5 per cent per year. However, this estimate is probably far too low since it includes only investment at the margin and assumes that the existing infrastructure is adequate; this is far from the case.

Finally, the analysis conducted here reinforces some recent remarks about international trade research:

> Most international trade research in the past has ignored the geographic dimension. International trade models, whether empirical or theoretical, whether based on small-country or large-country assumptions and whatever else their attributes, tended until recently to have one curious thing in common: they treated countries as disembodied entities that lacked a physical location in geographic space . . . many of the more interesting aspects of regional trading arrangements require the introduction of a geographic dimension. (Frenkel 1998, p. 1)

While it may not be possible to invest in the effort needed to link transportation systems with computable general equilibrium (CGE) models, greater attention will need to be placed in the nested production functions and consumer choice functions that are typically used in these models to reflect variations in transportation and transfer costs in addition to variations in production and tariff costs. The results presented by Kim (1998), Kim and Kim (2002) and Kim et al. (2002) point to the significant insights that can be gained from integration of transportation and CGE models. Without such an effort, the results from separate, unlinked models may not accurately reflect the gains to trade and some of the major source of nontariff impediments to the realization of those gains.

REFERENCES

The Economist (1996), 'Remapping South America', 12 October, p. 26.

Fonseca, M.A.R. (1991), 'Um Modelo Macroecnométrico de Simulação e Previsão', Anais do 13, Encontro Brasileiro de Econometria, Curitiba, PR.

Frenkel, J.A. (1996), 'Introduction', in J.A. Frenkel (ed.), *The Regionalization of the World Economy*, Chicago: NBER/University of Chicago Press, pp. 1–15.

Fundação do Instituto de Pesquisas Econômicas (1996), 'Índice de Desenvolvimento Econômico do Transporte', Universidade de São Paulo / Confederação Nacional dos Transportes, São Paulo, SP.

Haddad, E.A. and G.J.D. Hewings (2001a), 'Trade and regional development: international and interregional competitiveness in Brazil', in B. Johansson, Ch. Karlsson and R.R. Stough (eds), *Theories of Endogenous Regional Growth*, Heidelberg: Springer-Verlag, pp. 181–208.

Haddad, E.A. and G.J.D. Hewings (2001b), 'Transportation costs and regional development: an interregional CGE analysis', in P. Friedrich and S. Jutila (eds), *Policies of Regional Competition*, Baden-Baden: Nomos Verlag, pp. 83–101.

Kim, E. (1998), 'Economic gain and loss of public infrastructure investment: dynamic computable general equilibrium model approach', *Growth and Change*, **29**, 445–68.

Kim, E. and K. Kim (2002), 'Impacts of regional development strategies on growth and equity of Korea: a multiregional CGE model', *Annals of Regional Science*, **36**, 165–89.

Kim, T.J., H. Ham and D.E. Boyce (2002), 'Economic impacts of transportation network changes: implementation of a combined transportation network and input–output model', *Papers in Regional Science*, **81**, 223–46.

20. Accessibility and site rents in the C-economy[1]

Åke E. Andersson and David Emanuel Andersson

1. THE C-ECONOMY – AN INTRODUCTION

Over the past three decades, the economies of most OECD (Organization for Economic Cooperation and Development) countries have been transformed. The countries that had industrialized by 1970 are now emerging as restructured C-economies. In the industrial era, these economies mainly relied on adequate supplies of natural resources, coupled with a transportation system that was suitable for bulk shipments of freight by sea and rail. The producer's choice of location was then a question of finding an optimal location near a natural resource concentration or at a seaport or railroad junction. Estimates by Grübler (1990) show that railroads and waterways accounted for approximately 80 per cent of US transportation capacity in 1900.

During the industrial growth phase of the twentieth century, governments made sizable infrastructural investments in the new and more versatile road and air networks. The share of road and air infrastructure reached more than 80 per cent of total transportation capacity in 1970 in the United States and Western Europe (ibid.). Technological advances and network creation have been even more dramatic in communications, especially with the introduction of the Internet, leading to an almost complete dematerialization and potential globalization of information flows.

In recent years, transportation networks have accommodated increasing numbers of passengers. The growth of travelers has outpaced the growth rate of freight, leading to an increasing share of transported persons rather than goods. Travelers often have the economic function of being embodiments of knowledge. This knowledge function implies that the transportation systems support networks of cognitive and creative flows, while the communications infrastructure distributes increasing amounts of information (note that knowledge is not information, but rather a creative use or organization of information). Moreover, the total amount of cognitive and creative capacity has been increasing with the global expansion of formal education and research and development investments.

The increasing relative importance of stocks and flows (including network flows) of cognitive and creative capital and the ever-increasing importance of computerized communication systems characterize the emerging C-economy (Andersson 2000). This C-economy is more dependent on accessibility to C-resources than on proximity to financial or natural resources and the attendant large-scale use of land. We should therefore expect a corresponding transformation of both location patterns and their reflection in location-specific land rents.

2. A TWO-STAGE DECISION PROCESS

Discussions with managers of large firms in the real estate sector reveal that decision making regarding the location of property, both within these firms and among their corporate customers, is a two-stage decision process. In the first stage, the firm decides in which region it will buy or develop property. In the second stage, the firm chooses among locations within the region selected in the first stage. It also seems to be common for the corporate headquarters to make the first-stage choice of region, while managers of regional offices decide on individual property acquisitions and development strategies within the selected regions. The reason for this division of labor within firms is the need for detailed spatial knowledge in intraregional decisions.

In the early 1980s, the OECD published the results of a survey of the location criteria of high-technology firms in the United States. The survey showed that the most important determinant of location was accessibility to educated labor. This result applied to all regions, regardless of population size. Similarly, a study by Bindemann (1997) concluded that access to knowledge and skilled labor was by far the most important factor in determining the efficiency of international financial activities. Factors such as political stability, the supply of information or government policies were judged as much less important.

In the emerging C-economy, where both high technology and financial services are of increasing importance, accessibility to educated labor should rank highly in the choice of location, thereby influencing the spatial distribution of land values and rents in the commercial property market. It is also likely that accessibility by road and air transportation should be more important than in earlier times in deciding the choice of location.

3. REGIONAL ACCUMULATION OF CAPITAL

In the C-economy the most important productive input is knowledge. The regional availability of knowledge depends both on the cumulative production of knowledge within the region and on the accessibility to knowledge in other regions. With a simplifying Cobb–Douglas assumption we can illustrate the workings of a growing economy. In the following model we further assume that the level of income and the rate of knowledge accumulation determine the regional growth of capital. We assume that knowledge already accumulated in the region as well as 'imported' knowledge from other regions (located on an $n \cdot n$-grid of regions of equal spatial dimension) explain the regional income. For simplicity we assume that all regions have identical (unit) land areas. The amount of knowledge capital is thus equal to the knowledge density of a region. We further assume that the value of extra-regional and thus public knowledge (that is, knowledge accumulated in other regions) declines with the interregional distance. This spatial discounting implies that the value of available knowledge is the same in all regions, as is indicated by the common elasticity α. The following set of equations describes the growth process:

$$\dot{K}_i = \delta_i K_i^{\alpha} \prod_{j \neq i} \left(K_j / \gamma_j d_{ij}^{\beta}\right)^{\alpha} \cdot L_i^{\lambda} \cdot S_i^{\mu}; (i,j = 1, \ldots, n); \tag{20.1}$$

where:

$\mathrm{D} = \{d_{ij}\}$ = matrix of distances from region i to region j;
K_i = stock of knowledge capital in region i;
δ_i = rate of accumulation of knowledge capital in region i;
α = elasticity of regional product with respect to knowledge capital;
$\mathrm{L}_i = \bar{\mathrm{L}}_i$ = given supply of labor in region i;
$\mathrm{S}_i = 1$ = land area of region i;
γ_j = privateness (excludability) of the knowledge in region j.

A higher value of γ thus implies lower accessibility to public knowledge.

$$\text{Regional product} = \mathrm{Q}_i = \mathrm{K}_i^{\alpha} \prod_{j \neq i} (\mathrm{K}_j / \gamma_j d_{ij}^{\beta})^{\alpha} \mathrm{L}_i^{\lambda} \mathrm{S}_i^{\mu}$$

If each region maximizes profit from the use of its own knowledge capital, the privately determined optimal use of capital equals $\mathrm{K}_i^* = \alpha \mathrm{Q}_i / \rho$, where ρ is the required rate of return on knowledge capital.

The set of equations is a nonlinear differential system of equations with a general equilibrium solution. At the general equilibrium of this system the knowledge capital volume grows at the same rate in all regions, but the relative equilibrium knowledge capital volume per unit of land will be different. This equilibrium volume is a function of the accessibility from a region to all other regions. If all regions were to start with the same amount of knowledge, the more accessible regions would have a more rapid accumulation of knowledge and a faster growth of income than the less accessible regions. The following nonlinear eigen-value equation yields the equilibrium conditions.

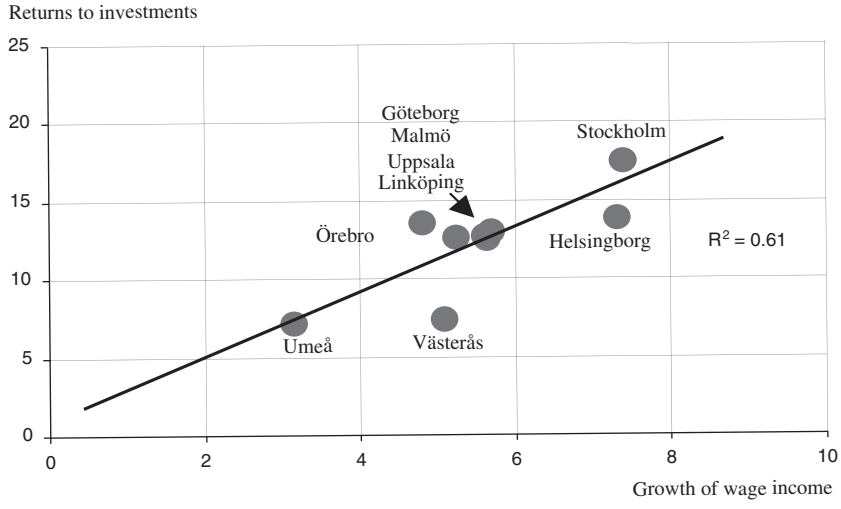
$$\lambda k = \mathrm{Q}(k, \bar{\mathrm{D}}, \bar{\ell}),$$

where:

λ = the common rate of growth of knowledge and per capita product;
k = vector of knowledge in regions;
$\bar{\mathrm{D}}$ = given matrix of distances between regions;
$0 < \bar{\ell}$ = given vector of labor in the n regions.

With normal assumptions about the parameters of the regional economies $\mathrm{Q}(k)$ is strictly positive for all $k > 0$. According to a theorem by Nikaido (1968, pp. 150–52) there exists an equilibrium solution such that $\lambda^* > 0$ and $k^* > 0$, where the asterisk indicates an equilibrium solution. This system is linearizable in the vicinity of the equilibrium and thus we can use the Perron theorem to assert that an improvement in the transport system will increase the equilibrium rate of growth.

This nonlinear eigen-equation has a positive eigen-value. However, the equilibrium rate of growth (λ^*) occurs at different combinations of space per unit of knowledge and regional accessibility. In regions with a low relative accessibility, there will be correspondingly more space per unit of knowledge, and thus the dual relative price of space will have to be lower than in more accessible regions, if the equalized rate of growth is to be sus-



Source: SFI/IPD Svenskt Fastighetsindex, SCB and Temaplan AB.

Figure 20.1 Real estate returns and yearly growth of wage income, 1998–2000

tainable. The equilibrium growth property of this knowledge expansion model is at variance with the disequilibrium assumptions made by Schumpeter (1912). Although disequilibrium at the micro level of the firm might be common, this does not in general imply disequilibrium at the level of a region. Often large corporations are sized and macro organized so as to accommodate disequilibria at lower levels of the organization.

4. REAL ESTATE RETURNS, REGIONAL RATES OF GROWTH AND ACCESSIBILITY TO KNOWLEDGE

A recent empirical study of the Swedish commercial property market showed that there has been a significant statistical association between regional returns to real estate investments and the growth rate of knowledge-oriented regional economies (Andersson et al. 2001).

For this reason, we have estimated regional economic growth (as reflected by growth of regional wage income) as a function of various explanatory variables for a cross-section of Swedish functional urban regions, based on data for the 1998–2000 period. Figure 20.1 is an illustration of the results.

Econometric analyses of per capita incomes and income growth in regions, based on cross-sectional data from functional urban regions, show that accessibility to university-educated labor and regional entrepreneurial activity (as measured by the growth rate of the number of firms with one or more employees) have a positive association with regional per capita income and regional income growth, whereas the level of local and regional taxation is negatively associated with per capita income and income growth (Andersson et al. 2001).

Table 20.1 Hedonic rent function for commercial space in Swedish regions, 2000

Variable	Coefficient	*t*-value
Constant	6.84	
Total property area (ln)	0.93	35.2
Office and shop area (ln)	0.06	4.49
Total property area if located in city of Stockholm (ln)	0.05	3.43
Per cent vacancy	−0.02	8.66
Index of road accessibility	1.04	2.38
Educational density (no. of univ. grad/sq km municipality)	0.00032	2.16
$R^2 = 0.9$; N = 293		

Note: Dependent variable: ln of annual rent in Swedish kronor.

A cross-sectional econometric estimation of commercial property rent levels in 293 Swedish municipalities shows the importance of accessibility and knowledge availability in the determination of the regional rental structure for commercial property (that is, office and retail trade space) (Table 20.1).

The functional form of the hedonic price function is the one with the best statistical properties among the three most common functional forms (that is, linear, log-linear and semi-log). However, all three functional forms lead to specifications that include the same independent variables, which indicates that the specification is robust regarding changes to the functional form or specification of independent variables (variables that were insignificant at the one-tailed 95 per cent confidence level were dropped from the model). An alternative specification with annual rent per square meter as the dependent variable attained robust estimates of the regression parameters but had a lower coefficient of determination.

A number of different accessibility measures were tested. Of the general (national) accessibility measures,[2] only the road accessibility measure is included in the hedonic price equation, because the other accessibility measures had no significant effects on the rent. Moreover, the educational density of the municipality turns out to be statistically superior to more general measures of accessibility to educated labor. One reason for this may be the large average land area of Swedish municipalities (approximately 1400 square kilometers).

The 'City of Stockholm' variable refers to the accessibility advantages of being located in the central business district (CBD) of Stockholm, which is the area where the Swedish stock exchange and the headquarters of most Swedish banks and multinational corporations (MNCs) are located. This variable is measured as the natural logarithm of the floor area if the property is located in the CBD, and is assigned a magnitude of 0 for all properties outside the CBD. This variable reflects accessibility advantages relating to localization economies of the financial services cluster as well as general urbanization economies.

5. OPTIMIZING THE REAL ESTATE PORTFOLIO

Our theoretical analysis as well as our econometric results indicate that investments in new real estate and real estate expansion by acquisitions should be primarily located in C-

regions, especially metropolitan C-regions. However, most profitable property developers and other real estate firms have a diversified regional investment strategy. Real estate and construction investment is by necessity long term, as compared with investments in machinery or stocks. The maps in Figures 20.2 and 20.3 show regional income growth rates and the growth rate volatility (that is, the intertemporal standard deviation of the growth rate).

The metropolitan regions of Malmö, Gothenburg and Stockholm all combine a rapid rate of growth with a greater volatility than the median Swedish region. These regions are also the centers of educational capital and accessibility to other regions in Sweden and abroad. However, there are also some medium-sized C-regions with relatively high levels of knowledge and accessibility. These regions typically combine high-income growth rates with low-to-moderate levels of volatility. Such regions will therefore provide preferable investment opportunities for risk-averse investors in the real estate markets. Real estate investors and developers should be expected to benefit from diversifying their portfolios among a number of regions. Cross-border diversifications should offer particularly attractive combinations of expected growth and risk.

Increasing the regional supply of educated labor or the number of firms or improved inter- and intraregional accessibility should raise expected returns on property investments, while increasing regional or local tax rates should lower expected returns according to the econometric estimates.

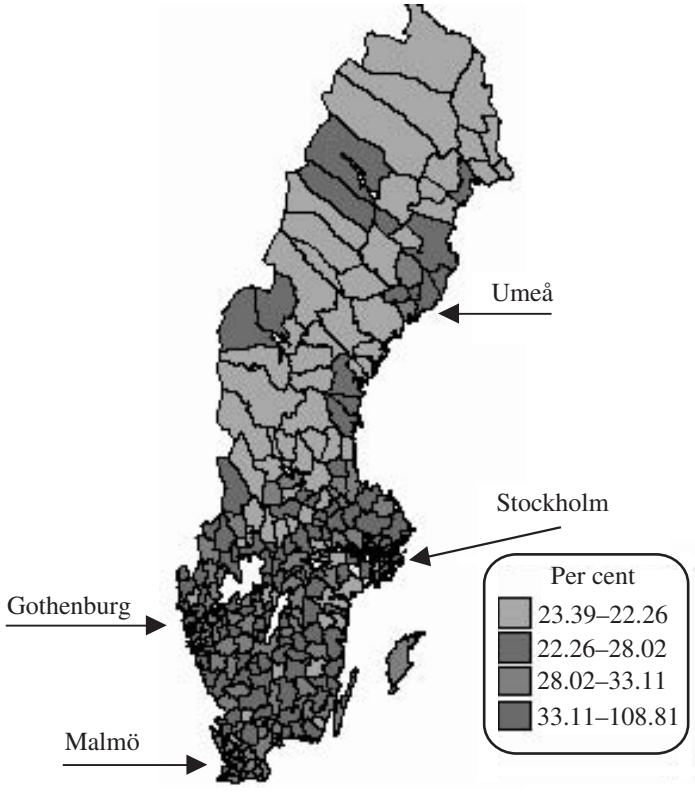


Figure 20.2 Wage income growth in Sweden, 1991–2000

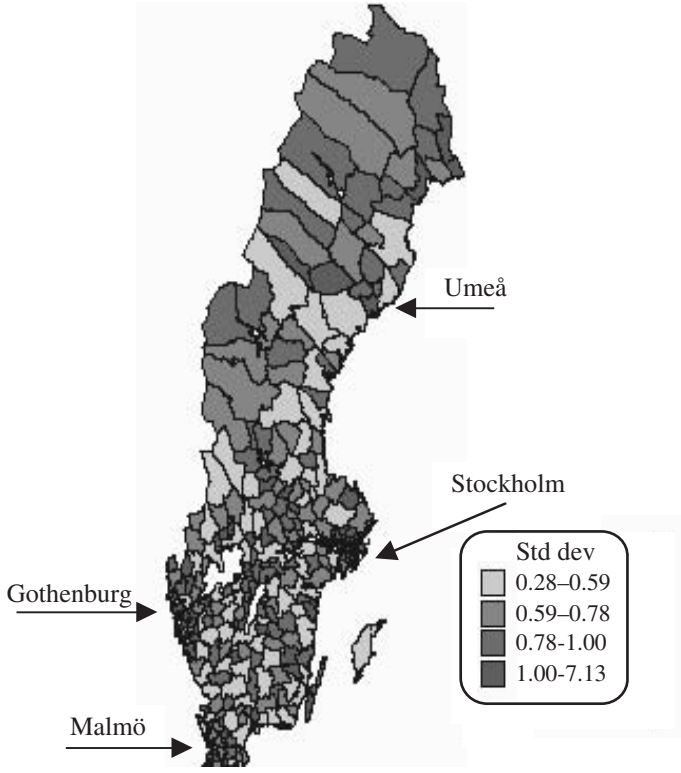


Figure 20.3 Volatility of income growth in different parts of Sweden

The variance–covariance matrix of income-growth statistics for the included regions shows the interdependence of the volatility of each pair of regions. Following Markowitz's (1952) classical model of optimal portfolio selection, adjusted for the more extended time scale of real estate investments (see also Nagurney 1997), the real estate and construction companies would optimize their portfolio strategies by maximizing:

$$\mathrm{H} = w \cdot \mathrm{R} - (1 - w)\mathrm{V},$$

where:

- H = the risk-compensated returns to the real estate portfolio;
- R = vector of expected returns to property investments $= \Sigma \rho_i x_i$; with $\rho_i =$ expected returns to property in region i and x_i the relative proportion of property in region i;
- V = the expected intertemporal and interregional variances and covariances of returns to properties in the n regions; $\mathrm{V} = 2\sum_i \sum_j x_i x_j \sigma_{ij}$;
- $0 \leq w$ = parameter reflecting the risk preference of the real estate firm.

Maximization of H should then be subject to the financial capital constraints, normalized to be:

$$\Sigma x_i = 1 \text{ and } x_i \leq 0.$$

The result of this optimizing procedure is a portfolio of properties in which the benefits of low or even negative correlations between returns in different regions are taken into account. In other words, the optimization possibilities are of the same nature as the portfolio choice problem of a conventional financial investor.

6. THE INTRA-METROPOLITAN PROPERTY RENT STRUCTURE

In the second stage of choosing property investments, the real estate firm has to decide intraregional locations and attendant asking and reservation rents.

Nothing in regional science has been as thoroughly analysed as *intraregional* pricing and use of land (see, for example, Beckmann and Puu 1985; Fujita 1989; Beckmann 1999). For any land-using agent who does not derive any direct utility (or revenue) from the choice of location and who is subject to a binding budget constraint, the following *gradient law* holds:

$$\frac{\partial p_i}{\partial d} . \mathrm{S}_i^* = -\frac{\partial \mathrm{R}_i}{\partial d}$$

where:

p_i = rental bid-price of space of agent *i*;
d = distance from regional center;[3]
S_i^* = optimal use of space by agent *i*;
R_i = cost of transport for agent *i* as a function of distance to the regional center.

The optimal use of space is then determined by the marginal rate of substitution between space and other inputs equaling the relative bid-price of space:

$$\frac{\partial \mathrm{G}_i}{\partial \mathrm{S}_i} \Big/ \frac{\partial \mathrm{G}_i}{\partial \mathrm{B}_i} = p_i(d).$$

This would be realized (implying S_i^* for *i*) if $p_i(d) \geq p_j(d)$ for all *j* agents. Little can be said about the form of the equilibrium rent structure, except that it must be a monotonously decreasing function of distance to the regional center. One possible and simple form is $p\,(d) = 5 p_c d^{\alpha}$; where $\alpha < 0$ and p_c is the rental price at the center of the region. This rental price function for commercial space has been estimated for the Stockholm region on data for the year 2000 and results in the following econometric equation (Table 20.2).

An equation with rent per square meter was also estimated. The distance effect is robust, but the statistical fit improves with the inclusion of the size factor. The estimated results for the intra-metropolitan equation do not differ if time distance data are used instead of metric distances. It is remarkable that the distance to the city center has such a

Table 20.2 Hedonic rent function for commercial space in the Stockholm region, 2000.

Variable	Coefficient	*t*-value
Constant	−6.3	
Total property area (ln)	0.9	26.6
Distance in meters from the CBD (Sergels Torg) (ln)	−0.303	−9.6
$R^2 = 0.89$		

Note: Dependent variable: ln of annual rent in Swedish kronor.
Sources: JM Real Estate Corporation (2000); Wasakronan (2000).

dominant effect on rents. The inclusion of sub-centers as explanatory variables does not improve the estimates. It is clear that the Stockholm region remains a monocentric structure.

7. CONCLUSION

Land-use patterns and spatial distributions of property values have changed considerably in the economically advanced countries and regions. The focus has shifted from natural resource availability and undifferentiated labor supply towards availability of highly educated specialized labor and accessibility to knowledge from other regions. Because knowledge has public good characteristics and knowledge diffusion is subject to spatial friction, there are considerable advantages in agglomerating new economic activities in regions that are accessible and rich in cognitive and creative capacity, that is to say in C-regions.

A simplified growth model shows that a knowledge-oriented economy – consisting of a given number of regions with different accessibilities – can grow at a balanced rate, with increasing amounts of knowledge capital per unit of land and labor as well as increasing per capita incomes. Switching from a standard industrial economy to a knowledge-oriented C-economy would, however, imply increased initial benefits from accessibility and therefore a more rapid growth of knowledge and income per person as well as per unit of land in the most accessible regions.

Our empirical analysis indicates that regions that are accessible by road transportation and that have an initial advantage in terms of the availability of knowledge capital tend to have both higher income growth and higher expected returns to real estate investments, reinforcing the agglomerative tendencies of the C-economy. However, high regional economic growth rates tend to go hand in hand with greater fluctuations in the growth rate, *ceteris paribus*. This is a natural consequence of high rates of growth, because C-regions tend to have higher rates of firm births and firm expansions. The new and growing firms produce mostly new products, which are characterized by having a high income elasticity of demand causing larger fluctuations.

Thus, even in the C-economy, real estate businesses would gain from diversifying their investment portfolios interregionally, combining the higher growth rates of C-regions with the greater stability of more traditional but accessible regions.

NOTES

1. The C in 'C-economy' stands for several typical features of the post-industrial economy, such as creative, cognitive and computer capacities, culture, and communications.
2. Other tested accessibility measures included: population-weighted accessibility in the national air, road and rail networks, accessibility to college-educated labor in all municipalities (with inter-municipal distance effects), accessibility to high-school graduates in all municipalities (with inter-municipal distance effects), and regional air transportation capacity (passenger volume).
3. The regional center is defined as the location with the best overall accessibility.

REFERENCES

Andersson, Å.E. (2000), 'Gateway regions of the world – an Introduction', in Å.E. Andersson and D.E. Andersson (eds), *Gateways to the Global Economy*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, pp. 3–16.

Andersson, Å.E., D.E. Andersson and I. Holmberg (2001), *Grogrund för tillväxt*, Malmö: Sydsvenska Industri- och Handelskammaren.

Beckmann, M. (1999), *Lectures on Location Theory*, Berlin: Springer-Verlag.

Beckmann, M. and T. Puu (1985), *Spatial Economics: Density, Potential, and Flow*, Amsterdam: North-Holland.

Bindemann, K. (1999), *The Future of European Financial Centers*, London: Routledge.

Fujita, M. (1989), *Urban Economic Theory, Land Use and City Size*, Cambridge: Cambridge University Press.

Grübler, A. (1990), *The Rise and Fall of Infrastructure*, Heidelberg: Physica-Verlag.

JM (2002), Annual Report, Stockholm: JM Real Estate Corporation.

Markowitz, H.M. (1952), 'Portfolio selections', *Journal of Finance*, vol 7, 77–91.

Nagurney, A. (1997), *Financial Networks*, Berlin: Springer-Verlag.

Nikaido, H. (1968), *Convex Structures and Economic Theory*, New York: Academic Press.

Schumpeter, J.A. (1912), *Theorie der Wirtschaftlichen Entwicklung* (The Theory of Economic Development), Leipzig.

Wasakronan (2000), Annual Report, Stockholm: Wasakronan.

Index