
DISTRIBUTION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Yuxiong Ji

Graduate Program in Civil Engineering

The Ohio State University

2011

Dissertation Committee:

Associate professor Rabi G. Mishalani, Co-advisor

Professor Mark R. McCord, Co-advisor

Professor Prem K. Goel
Abstract

Transit route-level passenger Origin-Destination (OD) matrices are important inputs for ridership forecasting, service design and operations. Collecting transit passenger OD flows has typically been costly and time-consuming. The emerging popularity of Automatic Passenger Counter (APC) technologies in public transportation presents new opportunities for transit agencies to obtain passenger flow OD information.

A methodology is developed to estimate a route-level probability OD flow matrix for a time-of-day period that takes advantage of the increasing availability of large quantities of boarding and alighting data collected via APC technologies. Unlike previous transit OD estimation approaches, the problem formulation considers the distribution, rather than only the means, of the APC data. A Heuristic Expectation Maximization (HEM) algorithm is developed to provide an approximate solution to the estimation formulation while achieving computational feasibility for applications to realistically long bus routes.

The performance of the HEM algorithm is evaluated numerically and empirically by comparing OD flow estimates to estimates produced by three other methods: the Expectation Maximization (EM) method, the Iterative Proportion Fitting (IPF) method and a Conditional Maximization (CM) method. The EM method provides a solution to the formulation, but it is computationally prohibitive for practical use. The IPF method
represents traditional methods, where only the means of the APC data are considered. The CM method is developed in this study as another computationally feasible method that considers the distribution of APC data, but it is seen to extract information on the probability OD flows less efficiently than the HEM method.

The numerical study demonstrates that the HEM algorithm provides a good approximation to the solution produced by the EM algorithm. In addition, the numerical and empirical studies demonstrate that the HEM algorithm produces more accurate estimates than the IPF and CM methods when large quantities of APC data are available. Moreover, the numerical and empirical studies quantify the effects of three pertinent factors on the accuracy of OD flow estimates: the number of bus trips with APC data, the sample size of an OD flow survey used with APC data to estimate OD flows, and the magnitude of measurement errors in APC counts. The HEM algorithm does not work particularly well when only a few bus trips have APC data. However, the improvement in the quality of the estimates produced with increased quantities of APC data when using the HEM method is greater than when using the CM method and is much greater than when using the IPF method.

The numerical studies further demonstrate that the HEM method outperforms the IPF and CM methods if the number of bus trips with APC counts is larger than a certain threshold. This threshold increases with the sample size of an OD flow onboard survey if such a survey is used in the estimation. The threshold also increases with increasing magnitude of measurement errors in APC counts. Nevertheless, the threshold is relatively small compared to what can be feasibly collected from APC-equipped buses.
Additional properties of probability OD flow estimation using the HEM algorithm are also investigated, such as convergence of the algorithm, sensitivity of OD estimates to the underlying OD structure and the distributional assumptions, and performance of OD estimation under congested conditions on buses due to bus bunching and high demand.
Dedication

This document is dedicated to my family.
Acknowledgments

This dissertation will not be possible without the guidance and continuous encouragement from my advisors and mentors, Dr. Rabi G. Mishalani and Dr. Mark R. McCord. I am lucky to have their faith in me. Rabi and Mark are always generous with their time and give me valuable advice and help on everything including study, research and life. Working with them is an enjoyable and rewarding experience. Their unselfish help deserves more than thanks.

I am also indebted to Professor Prem K. Goel for his invaluable contributions to this dissertation. Prem never hesitates to offer his help whenever I need it. He is generous to share his creative ideas with me and always guides me on the right direction. Numerous meetings with him broaden my horizon and inspire many ideas in the dissertation. His suggestions directly result in Sections 5.5 through 5.7 in Chapter 5. I also like to thank him for checking the assumptions and formulations in my methodology and giving me advices on statistical terminology and methodology.

Thanks to Sarah Blouch, director of OSU’s Transportation and Parking Services, and Chris Kovitya of OSU’s Campus Area Bus Service (CABS) for their instrumental role in the development of OSU’s Campus Transit Lab (CTL), which made the collection of empirical APC data and field passenger OD flow data on the buses possible.
Thanks for the company of my close friends and colleagues, Seoungbum and Cheng. No road is long with good company. Also thanks the teammates in Transportation Student Group for their help in the past three years: Xi, Fanyu, Dawei, Qianqian, Brandon, Xudong and Xiaofei.

Thanks my family for their unselfish support and encouragement. They always believe in me and encourage me to realize my dream.

Lastly, I would like to acknowledge the support of the Federal Transit Administration (FTA), U.S. Department of Transportation (US DOT), which funded the presented research. In addition, the field data used in this study were collected through the financial support of the Region V University Transportation Center funded by the Research and Innovative Technologies Administration (RITA), US DOT. The views, opinions, findings, and conclusions presented are the responsibility of the author and do not represent the official policy or position of FTA, RITA, or US DOT in general.
Vita

September 28 1978 ...................................... Born – Fujian, China
2001........................................................... B.S., Civil Engineering (Transportation),
                              Tongji University, China
2004........................................................... M.S., Civil Engineering (Transportation),
                              Tongji University, China
2006........................................................... M.S., Civil Engineering (Transportation),
                              The Ohio State University
2006 to 2008 .................................................. Transportation Analyst, TEMS inc.,
                              Maryland
2009........................................................... Master of Applied Statistics, Statistics, The
                              Ohio State University
2008 to present ............................................... Graduate Research Associate, Department
                              of Civil Engineering, The Ohio State
                              University

Publications

Periods for Bus Route Origin-Destination Passenger Flow Patterns Based on Automatic
Passenger Count Data. CD Proceedings of The 90th Transportation Research Board
Annual Meeting. Washington D.C.

Onboard Survey Sample Size on Transit Bus Route Passenger OD Flow Matrix
Estimation Using APC Data. CD Proceedings of The 90th Transportation Research Board
Annual Meeting. Washington D.C.

of the Effect of Bus Drivers’ Reactions to Schedules on Transit Operations Reliability.
Selected Proceedings of The 12th World Conference on Transport Research. Lisbon,
Portugal, pp c2-03075.

viii


Fields of Study

Major Field: Civil Engineering
# Table of Contents

Abstract ........................................................................................................................................ ii

Dedication ..................................................................................................................................... v

Acknowledgments .......................................................................................................................... vi

Vita ................................................................................................................................................ viii

List of Tables .................................................................................................................................. xvii

List of Figures ............................................................................................................................... xviii

Chapter 1  Introduction .................................................................................................................. 27

1.1 Motivation for Transit Origin-Destination Matrix Estimation .................................................. 27

1.2 Obtaining OD Flow Information from AFC or APC Data ....................................................... 29

1.3 Dissertation Overview .......................................................................................................... 31

Chapter 2  Literature Review and Motivation for the Current Study ........................................... 36

2.1 Introduction ............................................................................................................................. 36

2.2 General Framework of Static OD Estimation Methods ......................................................... 38

2.3 Literature Relevant to Transit Route-level Passenger OD Estimation ..................................... 41

2.3.1 Non-statistical Approaches ................................................................................................. 41

2.3.2 Statistical Approaches ....................................................................................................... 48
2.4 Relationships among Transit Route-level Passenger OD Estimation Methods. 64

2.5 Motivation for the Current Study ........................................................................ 72

Chapter 3 Frameork of the Proposed Approach ...................................................... 78

3.1 Introduction ........................................................................................................ 78

3.2 Model Structure and Assumptions ..................................................................... 79

3.3 Formulation: Posterior Distributions and Point Estimates ................................. 90

Chapter 4 Conditional Maximization and Heuristic Expectation Maximization

Algorithms ............................................................................................................. 100

4.1 Introduction ........................................................................................................ 100

4.2 Conditional Maximization (CM) Algorithm ..................................................... 101

4.3 Expectation Maximization (EM) and Heuristic Expectation Maximization

(HEM) Algorithms .............................................................................................. 105

4.3.1 Implementation of the EM Algorithm to Estimate the Probability OD Flows

.............................................................................................................................. 105

4.3.2 Heuristic Expectation Maximization (HEM) Algorithm ................................ 111

4.3.3 Computation Complexity of the HEM Algorithm ....................................... 118

Chapter 5 Numerical Study ..................................................................................... 119

5.1 Introduction ........................................................................................................ 119

5.2 OD Estimation Methods Considered in the Evaluation .................................... 120
7.2 APC Measurement Error Model ................................................................. 190
7.3 Effects of the Number of Bus Trips with APC Counts and OD Survey Sample Size on the Accuracy of OD Estimates ................................................................. 191
7.4 Effects of OD Survey Sample Size and Measurement Errors in APC Counts on the Accuracy of OD Estimates ................................................................. 205
7.5 Effects of the Number of Bus Trips with APC Counts and Measurement Errors in APC Counts on the Accuracy of OD Estimates ................................................................. 212

Chapter 8 OD Estimation under Congestion on Buses ........................................... 218

8.1 Introduction ............................................................................................... 218
8.2 General Assumptions .................................................................................. 219
8.3 Formulation: Posterior Distributions and Point Estimates .......................... 222
8.4 Applying the CM and HEM Algorithms to Estimate the APM ..................... 224
8.5 Numerical Demonstration ......................................................................... 230
  8.5.1 Overview .............................................................................................. 230
  8.5.2 Performance Measures .......................................................................... 231
  8.5.3 Results and Discussion .......................................................................... 232

Chapter 9 Summary and Future Research ......................................................... 240

9.1 Summary ................................................................................................. 240
9.2 Future Research ......................................................................................... 250
9.2.1 Uncertainty Measures ................................................................. 252
9.2.2 Extension to a Two-Route with One Transfer Network ............... 260

Reference .................................................................................................................. 269

Appendix A Advantage of the Proposed Approach ............................................. 272
A.1 Introduction ............................................................................................................. 272
A.2 Method Comparison .............................................................................................. 274
A.3 Relationship between the Probability OD Flows and the First and Second
Moments of the Trip-level Boarding and Alighting Counts ................................. 280
  A.3.1 First Moments of the Trip-level Boarding and Alighting Counts .............. 280
  A.3.2 Second Moments of the trip-level Boarding and Alighting Counts ....... 282

Appendix B Markov Chain Monte Carlo Algorithms ............................................. 288
B.1 Introduction ............................................................................................................. 288
B.2 MCMC1 Algorithm for the Formulation that Directly Incorporates Measurement
Errors in APC Counts .................................................................................................. 289
  B.2.1 Conditional Posterior Distributions for MCMC1 Algorithm ................... 290
  B.2.2 Implementation of the MCMC1 Algorithm ................................................... 294
B.3 MCMC2 Algorithm for the Formulation that Assumes that APC Counts are Free
of Measurement Errors .......................................................................................... 297
  B.3.1 Foundation for Step 2 of the MCMC2 Algorithm ................................. 298
B.3.2 Conditional Posterior Distributions for the MCMC2 Algorithm .......... 301
B.3.3 Implementation of the MCMC2 Algorithm .................................. 304
B.4 Technical Comments on the MCMC Algorithms ............................. 305
B.5 Numerical Evaluation on a Simplistic, Illustrative Bus Route ............... 306
  B.5.1 Experiment Overview ............................................................... 306
  B.5.2 Results .................................................................................. 308

Appendix C Use of the IPF Method in the CM Algorithm .......................... 312

Appendix D Convergence of the EM Algorithm in the Context of the Proposed Probability OD Estimation Problem .......................................................... 315

Appendix E SSD and Chi2 Results from the Analysis of Chapter 7 ............ 319
  E.1 Effects of the Number of Bus Trips with APC Counts and OD Survey Sample Size on the Accuracy of OD Estimates .................................................. 320
  E.2 Effects of OD Survey Sample Size and Measurement Errors in APC Counts on the Accuracy of OD Estimates .................................................. 330
  E.3 Effects of the Number of Bus Trips with APC Counts and Measurement Errors in APC Counts on the Accuracy of OD Estimates .................................. 336

Appendix F Empirical and Numerical Studies on COTA Route 7 ............... 342
  F.1 Introduction ............................................................................... 342
  F.2 Empirical Study ......................................................................... 343
  F.3 Numerical Study ......................................................................... 346
List of Tables

Table 2.1 Reference to Methods Presented in Figure 2.1 ......................................................... 67
Table 2.2 Reference to Formulation or Algorithms Presented in Figure 2.2 .................. 72
Table 3.1 Descriptions of Notation in Figure 3.1 and the Following Development ....... 80
Table 5.1 Probability OD Matrix 1 (OD1) on the Simple, Illustrative Bus Route ........ 128
Table 5.2 Probability OD Matrix 2 (OD2) on the Simple, Illustrative Bus Route ........ 128
Table 5.3 Number of Observations by Bus trips and Passengers ................................. 139
Table 5.4 Performance Measures and the Corresponding $RP$ Measures of OD Estimation Methods based on Two Sequence Orders ................................................................. 156
Table 6.1 Summary of Three Experiments by the Number of Bus Trips and Measurement Errors in Boarding and Alighting Counts ............................................................... 176
Table 9.1 Structure of the Network OD Matrix for a Two-route with One Transfer Network (Cui, 2006) ........................................................................................................... 262
Table 9.2 Boarding and Alighting Counts and Total Transfer Flows for One Pair of Bus Trips in the Hypothetical Network Presented in Figure 9.1 ................................. 263
Table 9.3 Representation of Boarding and Alighting Counts and the Number of Transfer Passengers as the Marginal Values of an OD Matrix ........................................... 265
Table 9.4 Estimated Route-level Probability OD Matrices ................................................. 267
Table 9.5 Estimate the Transfer Probability OD Matrix from the Origin and Destination Probabilities ................................................................................................. 267
Table A.1 Assumptions of Five Methods under Comparison .................................................. 273
Table B.1 Trip-level OD Matrix for a Hypothetical Bus Route ........................................... 297
Table B.2 Determine an OD Matrix from Boarding and Alighting Counts on a Bus Trip
.................................................................................................................................................. 301
Table B.3 Assumed Probability OD Flow Matrix on an Illustrative Bus Route ........... 307
Table B.4 Performance and RP Measures of Various Probability OD Flow Estimates . 310
Table E.1 Correspondence of the Plots in Appendix D and in Chapter 7 ....................... 319
List of Figures

Figure 2.1 Relationships among Methods Relevant to Transit Route-level Passenger OD Matrix Estimation .......................................................... 66
Figure 2.2 Algorithms for the Vardi’s Formulation .................................................. 72
Figure 2.3 QQ plot of the Empirical Trip-level Total Demands versus Poisson Distribution ........................................................................................................... 75
Figure 3.1 Relationships among Parameters, Variables and Observed Data .................. 79
Figure 5.1 ECDFs of the Probability OD Flow Estimates on the Illustrative Bus Route under OD1 in Table 5.1 ........................................................................... 130
Figure 5.2 ECDFs of the Probability OD Flow Estimates on the Illustrative Bus Route under OD2 in Table 5.2 ........................................................................... 131
Figure 5.3 ECDFs of the Performance and RP Measures of OD Estimation Methods on the Illustrative Bus Route under OD1 in Table 5.1 ........................................ 135
Figure 5.4 ECDFs of the Performance and RP Measures of OD Estimation Methods on the Illustrative Bus Route under OD2 in Table 5.2 ........................................ 136
Figure 5.5 Map of the CLS Bus Route .......................................................................... 138
Figure 5.6 Comparisons of the IPF, CM and HEM Probability OD Flow Estimates to the Assumed True Probability OD Flows on the CLS Bus Route .............................. 141
Figure 5.7 ECDFs of the Probability OD Flow Estimates for Individual OD Pairs on the CLS Bus Route .................................................................................. 142
Figure 5.8 ECDFs of the Performance and RP Measures of OD Estimation Methods on the CLS Bus Route ............................................................... 144
Figure 5.9 ECDFs of Passenger Distance Traveled Derived from Various OD Estimates on the CLS Bus Route ............................................................... 146
Figure 5.10 Performance at Each Iteration of the CM and HEM Algorithms .............. 148
Figure 5.11 Convergence of the CM Algorithm for Individual OD Pairs ..................... 150
Figure 5.12 Convergence of the HEM Algorithm for Individual OD Pairs ................. 151
Figure 5.13 Convergence of the CM Algorithm for Individual OD Pairs in the Scale of Logarithm ........................................................................................................ 152
Figure 5.14 Convergence of the HEM Algorithm for Individual OD Pairs in the Scale of Logarithm ........................................................................................................ 153
Figure 5.15 Comparisons of Probability OD Flows Estimated by the HEM Algorithm Implemented based on Two Different Stop Sequence Orders ......................... 155
Figure 5.16 ECDFs of Performance Measures under Three Probability OD Matrices .... 160
Figure 5.17 ECDFs of RP Measures under Three Probability OD Matrices ............... 161
Figure 5.18 Sample Variance versus Sample Mean of the Trip-level OD Flow on the CLS Bus Route ........................................................................................................ 164
Figure 5.19 ECDFs of the Performance Measures under Two Distributional Assumptions .............................................................................................................. 167
Figure 5.20 ECDFs of the RP Measures under Two Distributional Assumptions ....... 168
Figure 6.1 Summary of Trip-level Total Boarding and Alighting Counts on the CLS Bus Route ........................................................................................................ 172
Figure 6.2 Performance Measures of Various Estimates Resulting from Three Experiments ........................................................................................................... 180

Figure 6.3 RP Measures of Various Estimates Resulting from Three Experiments...... 181

Figure 6.4 CDFs of Passenger Distance Traveled (PDT) Derived from Various OD Estimates Resulting from Experiment 2 .................................................................................................................. 187

Figure 7.1 $HD^2$ Performance Measure versus the Number of Bus Trips for which APC Data were Collected under Different OD Survey Sample Sizes (Given Measurement Error Free APC Counts) ........................................................................................................................................ 193

Figure 7.2 $HD^2$-based RP Measure versus the Number of Bus Trips for which APC Data were Collected under Different OD Survey Sample Sizes (Given Measurement Error Free APC Counts) ........................................................................................................................................ 194

Figure 7.3 $HD^2$ Performance Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data (Given Measurement Error Free APC Counts). 200

Figure 7.4 $HD^2$-based RP Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data (Given Measurement Error Free APC Counts). 201

Figure 7.5 ECDFs of the $HD^2$ Measures under Various Numbers of Bus Trips with APC Data and Various OD Survey Sample Sizes (Given Measurement Error Free APC Counts) ........................................................................................................................................ 204

Figure 7.6 ECDFs of the $HD^2$-based RP Measures under Various Numbers of Bus Trips with APC Data and Various OD Survey Sample Sizes (Given Measurement Error Free APC Counts) ........................................................................................................................................ 204
Figure 7.7 $HD^2$ Performance Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes (Given 200 bus trips with APC Counts) ........................................................................................................................................ 207

Figure 7.8 $HD^2$-based RP Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes (Given 200 bus trips with APC Counts) ........................................................................................................................................ 208

Figure 7.9 ECDFs of the $HD^2$ Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts (Given 200 bus trips with APC Counts) ........................................................................................................................................ 211

Figure 7.10 ECDFs of the $HD^2$-based Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts (Given 200 bus trips with APC Counts) ........................................................................................................................................ 211

Figure 7.11 $HD^2$ Performance Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts (Given a Null Base OD Flow Matrix) ........................................................................................................................................ 214

Figure 7.12 $HD^2$-based RP Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts (Given a Null Base OD Flow Matrix) ........................................................................................................................................ 215

Figure 7.13 ECDFs of the $HD^2$ Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts (Given a Null Base OD Flow Matrix) ........................................................................................................................................ 217
Figure 7.14 ECDFs of the $HD^2$-based $RP$ Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts (Given a Null Base OD Flow Matrix) ....................................................................................................................... 217

Figure 8.1 Number of Passengers who are not Able to Board Buses Due to the Bus Capacity Constraint by Stops in One Simulation Trial............................................................ 235

Figure 8.2 ECDFs of the $HD^2$ and $HD^2$-based $RP$ Measures of the Probability OD Matrix and APM Estimates................................................................................................................. 236

Figure 8.3 ECDFs of the $SSD$ and $SSD$-based $RP$ Measures of the Probability OD Matrix and APM Estimates................................................................................................................. 237

Figure 8.4 ECDFs of the $Chi^2$ and $Chi^2$-based $RP$ Measures of the Probability OD Matrix and APM Estimates................................................................................................................. 237

Figure 9.1 Structure of a Hypothetical Network with Two Bus Routes........................................ 261

Figure B.1 Numerical Posterior Distributions of the Probability OD Flows................................ 308

Figure E.1 $SSD$ Performance Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes ........................................ 320

Figure E.2 $SSD$-based $RP$ Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes .................................................. 321

Figure E.3 $SSD$ Performance Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data ................................................................. 322

Figure E.4 $SSD$-based $RP$ Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data ................................................................. 323
Figure E.5 ECDFs of the SSD Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes ........................................ 324
Figure E.6 ECDFs of the SSD-based RP Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes .............. 324
Figure E.7 Chi2 Performance Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes .................................... 325
Figure E.8 Chi2-based RP Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes ........................................... 326
Figure E.9 Chi2 Performance Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data ......................................................................... 327
Figure E.10 Chi2-based RP Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data ................................................................. 328
Figure E.11 ECDFs of the Chi2 Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes ......................... 329
Figure E.12 ECDFs of the Chi2-based RP Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes ............. 329
Figure E.13 SSD Performance Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes .................................................. 330
Figure E.14 SSD-based RP Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes ......................................................... 331
Figure E.15 ECDFs of the SSD Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts .................................................. 332
Figure E.16 ECDFs of the SSD-based Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts .......................................................... 332

Figure E.17 Chi2 Performance Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes .......................................................... 333

Figure E.18 Chi2-based RP Measure versus the Level of Measurement Errors in APC Counts Assuming under Different OD Survey Sample Sizes .................................................. 334

Figure E.19 ECDFs of the Chi2 Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts ......................................................... 335

Figure E.20 ECDFs of the Chi2-based Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts ......................................................... 335

Figure E.21 SSD Performance Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts ........................................ 336

Figure E.22 SSD-based RP Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts ........................................ 337

Figure E.23 ECDFs of the SSD Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts ........................................ 338

Figure E.24 ECDFs of the SSD-based RP Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts .......... 338

Figure E.25 Chi2 Performance Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts ........................................ 339

Figure E.26 Chi2-based RP Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts ................................................ 340
Figure E.27 ECDFs of the $\text{Chi}^2$ Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts ... 341
Figure E.28 ECDFs of the $\text{Chi}^2$-based $RP$ Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts ... 341
Figure F.1 Schematic Route Map of Bus Route 7 ................................................. 342
Figure F.2 Comparison of the Probability OD Estimates produced by the IPF, CM and HEM methods on Route 7 SB ................................................................. 344
Figure F.3 ECDFs of Passenger Distance Traveled (PDT) Derived from Various Probability OD Flow Estimates on Route 7 SB ....................................................... 346
Figure F.4 Comparisons of the IPF, CM and HEM Probability OD Flow Estimates to the Assumed True Probability OD Flows on Route 7 SB ........................................ 348
Figure F.5 ECDFs of the Probability OD Flow Estimates for Individual OD Pairs on Route 7 SB ................................................................................................................ 349
Figure F.6 ECDFs of the Performance and $RP$ Measures of OD Estimation methods on Route 7 SB ............................................................................................................. 350
Figure F.7 $HD^2$ Performance Measure versus the Number of Bus Trips with APC ...... 352
Figure F.8 $HD^2$-based Performance Measure versus the Number of Bus Trips with APC Counts ........................................................................................................................... 353
Figure F.9 ECDFs of the $HD^2$ Measure under Various Numbers of Bus Trips with APC Counts .................................................................................................................. 354
Figure F.10 ECDFs of the $HD^2$-based $RP$ Measure under Various Numbers of Bus Trips with APC Counts ................................................................................................. 354
Figure F.11 SSD Performance Measure versus the Number of Bus Trips with APC Counts .......................................................... 355

Figure F.12 SSD-based RP Performance Measure versus the Number of Bus Trips with APC Counts .......................................................... 356

Figure F.13 ECDFs of the SSD Measure under Various Numbers of Bus Trips with APC Counts .......................................................................................... 357

Figure F.14 ECDFs of the SSD-based RP Measure under Various Numbers of Bus Trips with APC Counts .......................................................................................... 357

Figure F.15 Chi2 Performance Measure versus the Number of Bus Trips with APC Counts .......................................................................................... 358

Figure F.16 Chi2-based RP Performance Measure versus the Number of Bus Trips with APC Counts .......................................................................................... 359

Figure F.17 ECDFs of the Chi2 Measure under Various Numbers of Bus Trips with APC Counts .......................................................................................... 360

Figure F.18 ECDFs of the Chi2-based RP Measure under Various Numbers of Bus Trips with APC Counts .......................................................................................... 360
Chapter 1    Introduction

1.1 Motivation for Transit Origin-Destination Matrix Estimation

An origin-destination (OD) matrix is one of the most fundamental inputs for transportation planning and management. An OD matrix contains information on the number of vehicles or passengers travelling between different pairs of a region, different ramp pairs of a corridor, different leg pairs of an intersection or different stop pairs on a transit line or route. In bus transit applications, transit passenger OD matrices can be considered at either the network level or the route level. A network level OD matrix summarizes passenger flows between boarding and alighting stops that do not necessarily belong to the same route. As a result, one or more transfers may occur between the origin and destination of the bus trip. A route-level passenger OD matrix, which is the focus of this research, considers passenger flows between boarding and alighting stops on a single bus route.

Network-level OD matrices are essential inputs in applications such as network design, determination of transfer stations, schedule coordination, and passenger demand forecasting. Network-level OD matrices also provide important information for real time schedule coordination and control. For example, a bus may be held at a transfer bus stop if many passengers on an arriving bus need to transfer to this bus. The number of passengers expected to transfer would be deduced from network-level OD matrices.
Route-level OD matrices are also useful in transit design and operation. For example, good estimates of the numbers of passengers travelling between stop pairs on a route are useful when deciding whether to extend or split routes, whether to combine routes or implement a new route, or where to locate stops. In addition, route-level OD matrices are useful for designing timetables, determining headways, and developing short turning, expressing, and holding strategies. Route-level OD matrices can also be used to set fares and quantify the impact of service changes on ridership.

In addition, route-level OD matrices can be used to detect and correct sample biases in OD survey data. For example, passengers complete on-board survey while traveling on a single bus route. Passengers who travel short distances may not have enough time to complete the survey before they alight. As a result, these short trip passengers would tend to be undersampled. If they have different characteristics from the non-short trip passengers, the survey would need to be weighted to avoid producing biased results. A good route-level OD matrix estimate can be used to expand the survey data according to the proportion of passengers travelling different trip distances, rather than according to the proportion of surveys returned.

Moreover, route-level OD matrices can be used as inputs to network-level OD estimation methods. Cui (2006) proposed two methods to estimate transfer OD matrices for pairs of routes. Both methods use route-level OD matrices as inputs. Improving the accuracy of the route-level OD matrix estimates will improve the accuracy of the transfer OD matrix estimate.
Although useful, collecting transit passenger OD flows has typically been costly and time-consuming. The increasing use of new technologies, such as Automatic Fare Card (AFC) and Automatic Passenger Counter (APC) systems, in public transit agencies presents new opportunities for these agencies to estimate OD flow information.

1.2 Obtaining OD Flow Information from AFC or APC Data

Automatic Fare Card (AFC) systems use identifiable cards to track individual passenger trips (Chan, 2007). However, in many bus systems, passengers only need to tap on (have the AFC information recorded) when boarding and do not have to tap off when alighting. As a result, such AFC systems only record information at the bus stops where passengers board and do not record information at the bus stops where passengers alight. Therefore, additional assumptions need to be made to use the AFC data for OD flow estimation (Cui, 2006).

APC technologies are being implemented by many transit agencies. APC systems provide passenger boarding and alighting counts at bus stops on a regular basis. Such counts provide indirect information on OD flows. Specifically, the sum of passengers originating at a stop and destined for downstream bus stops equals the boarding count at the bus stop and the sum of passengers destined for a bus stop who originate from upstream bus stops equals the alighting count at the bus stop. Mathematically, the row and column totals of an OD matrix for a bus trip are the boarding and alighting passenger counts on that bus trip, respectively.

To take advantage of this indirect information about OD flows contained in boarding and alighting counts, an approach to estimate OD flow matrices given the
constraints of boarding and alighting counts must be used. Since the APC systems are installed for other uses, the marginal cost of using APC boarding and alighting counts to estimate OD flow matrix would be low when using such an approach.

Many approaches in the literature can be used to estimate OD matrices from passenger boarding and alighting counts. Most approaches were proposed before APC systems were widely adopted. Since manually collecting boarding and alighting counts is costly and expensive, these approaches were developed considering that boarding and alighting counts would be available on a limited number of bus trips. Large amounts of boarding and alighting data can now be collected by APC systems. The availability of large quantities of such data motivates the development of an approach to efficiently use the large set of APC passenger count data to estimate OD matrices on a bus route.

Besides APC boarding and alighting counts, onboard survey OD data are sometimes available for some bus routes. Onboard survey OD data provide direct information on OD flows, and are valuable for inferring the underlying OD flows. Therefore, the proposed approach should have the ability to combine information in both onboard survey OD data and APC data to estimate OD matrices on a bus route.

This study develops a statistically-based approach to combine the information in the APC boarding and alighting counts and, possibly, onboard survey OD data to estimate a period-level probability OD flow matrix on a bus route in a homogeneous time-of-day period. The probability OD matrix represents the probabilities that a randomly selected passenger from the population in the given time-of-day period travels from the specified origins to the specified destinations. The approach developed is
computationally expensive. Therefore, an efficient solution procedure must be developed if the approach is to be used in practice. A Heuristic Expectation Maximization (HEM) algorithm is developed to solve the proposed formulation. The advantage of the proposed approach, compared to traditional transit route-level OD estimation methods, lies in its ability to take better advantage of the large quantities of boarding and alighting data that are now available from APC technologies implemented on many bus systems. The superior performance of the HEM algorithm is demonstrated in empirical and numerical studies by comparing OD estimates produced with estimates produced from a method representative of what could presently be considered the default method for practice.

1.3 Dissertation Overview

In Chapter 2, methods relevant to transit route-level passenger OD estimation are reviewed. Relationships among these methods are developed and discussed. The review shows that most methods are related to one another and produce the same or similar results. Among these methods, the Iterative Proportion Fitting (IPF) method is identified as being the method used for comparison purposes in this dissertation. With this background on OD estimation methods, the motivation for the current methodological approach is then presented.

In Chapter 3, the model structure and general assumptions of the proposed approach are presented. Two formulations are then developed. Both formulations consider passenger demand variation across bus trips. However, the first formulation assumes APC counts are free of measurement errors while the second formulation
considers measurement errors in APC counts. The difficulties in estimating the probability OD flow matrices based on both formulations are discussed.

In Chapter 4, algorithms are developed for the formulation that assumes that APC counts are free of measurement errors. The Heuristic Expectation Maximization (HEM) algorithm is developed to produce a point estimate of the period-level probability OD matrix by maximizing the marginal posterior likelihood of the probability OD matrix. The HEM algorithm is a variant of the traditional Expectation Maximization (EM) algorithm, which could be used to produce a solution to the proposed formulation. The EM algorithm is computationally expensive for applications to realistically long bus routes, while the HEM algorithm is computationally feasible. The computational time of the EM algorithm increases exponentially when the number of bus stops increases. Nevertheless, the computational time of the HEM algorithm increases by the square of the number of bus stops. For comparison purposes, a Conditional Maximization (CM) algorithm is also developed to produce a point estimate of the period-level probability OD matrix by maximizing the joint posterior likelihood of the probability OD matrix and the volume OD matrices on bus trips for which APC counts were collected.

In Chapter 5, the performance of the HEM algorithm is evaluated numerically. The performance is first evaluated on an illustrative, short bus route and then on an operational bus route. The EM algorithm is only considered on the illustrative, short bus route for comparison purposes. Simulation results on the illustrative bus route demonstrate that the HEM algorithm is able to arrive at a good approximate solution of the EM algorithm. Simulation results on both the illustrative bus route and the operational
bus route show that the HEM algorithm produces better probability OD flow estimates than the IPF and CM methods. The performance of the CM and HEM estimates are shown not to depend on the values used to initialize the algorithms. The sensitivity of OD estimates to the structure of the probability OD matrix and the distributional assumptions are investigated. It is shown that the HEM algorithm is robust with respect to OD structure, while the IPF and CM methods are sensitive to OD structure. It is also shown that the IPF, CM and HEM methods are robust with respect to the distributional assumptions and that the HEM algorithm outperforms the IPF and CM methods even when the distributional assumptions based on which the HEM algorithm is developed are not satisfied.

In Chapter 6, the performance of the HEM algorithm is evaluated empirically based on large quantities of APC data and large scale manually collected true OD data on an operational bus route. The empirical APC counts are prone to large measurement errors. The performance of the IPF, CM and HEM estimates is evaluated by comparing the corresponding estimates to the probability OD matrix estimated from the large scale manually collected true OD data. The empirical study shows that the HEM algorithm does not work particularly well when the amount of APC data is small. However, when APC counts on all available bus trips are considered for OD estimation, the HEM algorithm outperforms the CM and IPF methods in the absence of survey OD data. The HEM and IPF methods perform similarly and outperform the CM algorithm in the presence of survey OD data.
In addition, the empirical study quantifies the effect of three pertinent factors on the accuracy of OD estimates: measurement errors in APC counts, the number of bus trips for which the APC counts were collected, and the availability of survey OD data. It is shown that measurement errors in APC counts deteriorate the performance of OD estimates. Nevertheless, the impact of measurement errors in APC counts is greater on the CM and HEM estimates than on the IPF estimates. In addition, it is found that increasing the number of bus trips with APC counts would improve the performance of the IPF, CM and HEM estimates. Nevertheless, the performance of the CM and HEM estimates improves faster than that of the IPF method. Moreover, it is shown that incorporating survey OD data in OD estimation improves the accuracy of OD estimates. When a small amount of APC counts are used in OD estimation, the performance improvement due to the incorporation of survey OD data is similar for the IPF, CM and HEM estimates. However, the performance improvement due to the incorporation of survey OD data is larger for the IPF estimates than for the CM and HEM estimates when APC counts on all available bus trips are used in OD estimation.

In Chapter 7, a numerical study is conducted to further investigate the effect of the three pertinent factors on the accuracy of OD estimates. The simulation is based on the large scale manually collected true OD data on the same operational bus route as that in Chapter 6. The results obtained in the numerical study are consistent with what have been seen in the empirical study. Furthermore, it is shown that more APC data are needed if the HEM algorithm is to outperform the IPF and CM methods when the OD survey sample size is larger and when the magnitude of measurement errors in APC counts is
larger. Nevertheless, the amount of APC data that are needed such that the HEM algorithm can outperform the IPF and CM methods is relatively small compared with the amount of APC data that can be collected if buses are equipped with APC systems.

In Chapters 3 through 7, it is assumed that the probability OD matrix is stable across bus trips in a homogenous time-of-day period. Chapter 8 points out that this assumption may not be reasonable under congestion on buses due to bus bunching and high demand. Chapter 8 proposes to estimate an Alighting Probability Matrix (APM), which provides the probabilities that a passenger alights at downstream bus stops conditional on the bus stop at which he or she boarded, under congestion on buses. The general assumptions and a formulation for estimating the APM are presented first. Then, it is shown that the CM and HEM algorithms developed in Chapter 4 can also be used to estimate the APM. Finally, a numerical example is presented. The numerical results show that the HEM algorithm produces better APM estimates than the IPF and CM methods in the presence and in the absence of congestion on buses.

In Chapter 9, the research is briefly summarized, and future work is suggested along several dimensions, such as evaluation of the proposed methodology under a wider set of condition and longer bus routes, development of efficient algorithms for the formulation that directly incorporates measurement errors in APC counts, uncertainty measures of the probability OD flow estimates, and extension of the proposed methodology to network level OD estimation.
Chapter 2    Literature Review and Motivation for the Current Study

2.1 Introduction

This chapter presents a review of literature relevant to transit route-level passenger OD estimation and the motivation for the development of a new approach to estimate period route-level OD matrix from APC data. Since the focus of this research is on a methodology for bus transit route-level passenger OD estimation, only methods used in transit route-level passenger OD estimation and methods that are used in OD estimation for highway or communication networks that appear to be applicable to bus transit route-level passenger OD estimation are included in the review. The extensive literature that focused on highway network-level vehicle OD estimation and the relatively sparse literature that devoted to transit network-level passenger OD estimation are not included in this review.

OD estimation methods can be classified as being static or dynamic. Static methods estimate an OD matrix for a relatively long time-of-day period and assume travelers begin and end their travel in the same period. Trips travelling across periods are considered negligible. Therefore, static methods use vehicle or passenger counts in the same time-of-day period to estimate OD matrices. Dynamic methods estimate OD matrices for a series of consecutive short time-of-day periods and assume that travelers may begin their travel in one time-of-day period and end their travel in some subsequent time-of-day period. Therefore, dynamic methods use vehicle or passenger counts in
multiple consecutive short time-of-day periods to estimate time-varying OD matrices, in which interactions among OD flows across different time-of-day periods are usually taken into account. This study focuses on static OD estimation methods.

The OD flow estimation problem consists of determining an OD flow matrix that matches (broadly defined) link counts. In highway applications, link counts usually come from loop detectors installed on a limited number of road segments in the network. These detectors can provide 24-hour continuous count data. In transit applications, link counts are usually in the form of either bus loads or, more practically, boarding and alighting counts, which can be obtained from APC technologies.

For the purpose of the discussion in this chapter, most OD estimation methods are discussed in the context of transit route-level passenger OD estimation. Link counts are assumed to be in the form of boarding and alighting counts, since this study focuses on using APC boarding and alighting counts as inputs for OD estimation. A set of boarding and alighting counts on one bus trip in a predefined time-of-day period in the transit route-level passenger OD estimation application is analogous to a set of link flows in a predefined time-of-day period on a specific day in highway vehicle OD estimation application.

This chapter is organized as follows: in Section 2.2, a general framework that summarizes static OD estimation methods is introduced. In Section 2.3, studies relevant to transit route-level passenger OD estimation are reviewed. In Section 2.4, the relationships among methods relevant to transit route-level passenger OD estimation are presented and discussed. In Section 2.5, the motivation for the current study is presented.
### 2.2 General Framework of Static OD Estimation Methods

Information about OD flows contained in traffic link counts is represented by an assignment mapping, which can be presented as:

\[ x = \tau \times T \]  

(2.1)

where:

- \( x \) = true link count vector,
- \( T \) = true OD flow vector,
- \( \tau \) = assignment mapping matrix, whose entry \( \tau(i, j) \) represents the proportion of flows of OD pair \( j \) that use link \( i \).

The true OD matrix is rearranged in the form of a vector and presented in Equation 2.1 for convenience. An OD matrix and OD vector are used interchangeably in this study. The true OD matrix may represent the actual number or the expected number of passengers travelling along each OD pair on a bus trip.

In practice, the observed link counts could be different from the true link counts because of measurement errors in counts. In addition, if the true OD matrix describes the expected passenger flows on a bus trip in a given time-of-day period, the observed link counts are obviously different from the true link counts obtained by assigning the true OD vector to the network (i.e., \( \tau \times T \)), since passenger demand varies across bus trips and link counts are collected at bus trip level. The relationship between the observed link counts and the true link counts is given by:
\[ \hat{x} = x + \varepsilon = \tau \times T + \varepsilon \]  
(2.2)

where

\[ \hat{x} = \text{observed link count vector}, \]
\[ \varepsilon = \text{vector of deviations between observed and true link counts.} \]

Usually, Equation 2.2 is not sufficient to determine a unique OD matrix, since the number of unknown variables (number of OD pair flows) tends to be larger than the number of independent equations. Therefore, many existing methods incorporate an OD matrix formed from exogenous data (i.e., other than link flow data). The exogenous data could come from a survey or from a transportation model for example. The collection of exogenous data is often referred to as base OD matrix in the literature. If no exogenous information is available to form a base OD matrix, a common practice in some methods is to use a “null” base OD matrix. In the null matrix, it is assumed that passengers are equally likely to travel along any feasible OD pair (Furth and Navick, 1992)

Cascetta and Nguyen (1988) provided a unified framework for static OD estimation. This framework presents static OD estimation methods as a general optimization problem. The objective function considers the correspondence of the true OD matrix to a base OD matrix and the true link flows to the observed link flows. The formulation is represented by:

Program 2.1:

\[ \min_{T \geq 0} : f_1(T, T_0) + f_2(x, \hat{x}) \]  
(2.3)

where:
$f_1(T,T_0)$ function measuring the “distance” between the true OD matrix $T$ and a base OD matrix $T_0$, and

$f_2(x,\hat{x})$ function measuring the “distance” between the observed link flows $\hat{x}$ and the true link flows $x$ obtained by assigning the true OD matrix to the network (i.e., $x = \tau \times T$).

In some methods, it is assumed that the true OD matrix is able to reproduce the observed link flows. In this case, $f_2(x,\hat{x}) = 0$ and Program 2.1 becomes:

Program 2.2:

\[
\begin{align*}
\min_{T \geq 0} & : f_1(T,T_0) \\
\text{st} & : \tau \times T = \hat{x}
\end{align*}
\]  

(2.4a)

(2.4b)

Programs 2.1 and 2.2 use the base OD matrix and link counts as inputs to estimate the true OD matrix.

Hazelton (2001) pointed out that the definition of the true OD matrix is different in the problem of OD estimation and OD reconstruction, although both problems are generally referred to as OD estimation problem. In the problem of OD estimation, the true OD matrix summarizes the expected number of vehicle or passenger trips for each OD pair in a time-of-day period or on a bus trip. In the problem of OD reconstruction, the true OD matrix summarizes the actual number of vehicle or passenger trips for each OD pair in one time-of-day period or on one bus trip. The distinction between the problems of OD estimation and OD reconstruction is important in terms of explaining model assumptions and model outputs. Under certain circumstances, Hazelton (2001)
demonstrated that the solutions to these two problems could be quite different given the same input data.

The next section presents a review of methodologies related to transit route-level passenger OD estimation. The review will be presented mainly in the context of transit route-level passenger OD estimation, although the original papers being reviewed may have dealt with OD estimation for highway or other types of networks. Link counts are assumed to be in the form of boarding and alighting counts at the bus trip level.

2.3 Literature Relevant to Transit Route-level Passenger OD Estimation

In this section, OD estimation methods are classified into two categories: non-statistical approaches and statistical approaches. Non-statistical approaches are based on optimization and iterative techniques, or assumptions about passenger travel patterns. Statistical approaches are based on the distributional assumptions about the observed data and the prior information of the true OD matrix.

2.3.1 Non-statistical Approaches

In what follows, the various non-statistical methods are described. They are referred to by the authors’ method.

Tsycgalnitzky method

Tsycgalnitzky (1977) proposed a method to reconstruct an OD matrix for a bus trip from boarding and alighting counts at each bus stop along a bus route. The basic assumption of the Tsycgalnitzky method is that each ‘qualified’ onboard passenger is equally likely to alight at a given stop, regardless of the stop at which he or she boarded. A passenger is ‘qualified’ if she or he has traveled on board at least some minimum
distance. The minimum distance assumption is made because passengers would prefer walking or biking if travel distance is short. Therefore, passengers usually will not board at a bus stop and alight at a close downstream stop. The assumption of the Tsygalnitzky method leads to a straightforward computation of a unique OD matrix. The Tsygalnitzky method was tested by Simon and Furth (1985) using empirical survey OD data.

**Li and Cassidy method**

Li and Cassidy (2007) extended the work of Tsygalnitsky (1977) and proposed an algorithm to reconstruct OD matrices for multiple bus trips. The algorithm produces an OD matrix that satisfies the boarding and alighting counts for each stop on each bus trip first, and then produces an Alighting Probability Matrix (APM) based on an OD matrix that is obtained by aggregating the reconstructed OD matrices for each OD pair over all bus trips.

The Li and Cassidy method and the Tsygalnitsky method are similar in that both methods are based on boarding and alighting counts only. Base OD information, which is used in other methods, is not considered in both methods. In addition, both methods use the idea of “qualified” passengers, where a qualified passenger is a passenger who is “qualified” to alight at a given bus stop because she or he is still on board and has traveled at least some minimum distance.

The Li and Cassidy method recognized that the equally likely alighting assumption made in the Tsygalnitsky method might be unrealistic for bus routes serving activity centers, such as train stations and shopping centers, and therefore relaxed this assumption. In the Li and Cassidy method, bus stops are classified as either major or
minor stops. Major stops are considered to be stops serving activity centers, such as commuter train stations or large business centers, along the route. The other stops are treated as minor stops. The distinction is important since it allows the alighting probabilities for ‘qualified’ passengers to be different, depending on whether the alighting stop under consideration is a major or minor stop and whether the passenger boarded at a major or minor stop.

Although the Li and Cassidy method addresses the equally likely alighting problem, the designation of major and minor stops appears subjective and different designations may lead to quite different results.

*Iterative Proportional Fitting (IPF) method*

The IPF method is a popular method to reconstruct a volume OD matrix from boarding and alighting counts because it is simple and easy to implement. The IPF method is also referred to as biproportional method, the Furness or Fratar iterative procedure, the Kruithof algorithm, or the Bregman’s balancing method (Bacharach, 1970; Ben-Akiva et al., 1985).

The inputs to the IPF method are boarding and alighting counts for every bus stop along a bus route and a base OD matrix. Beginning with the two-dimensional base OD matrix, the IPF method iteratively adjusts the matrix by row and column until the marginal row and column totals of the reconstructed OD matrix converge to the given boarding and alighting counts, respectively. When the algorithm converges, the relationship between the reconstructed OD flow matrix and the base OD matrix can be shown to be:
\[ \hat{T}(i, j) = R_i \times T_0(i, j) \times C_j \]  

(2.5)

where \( \hat{T}(i, j) \) and \( T_0(i, j) \) are the reconstructed and the base passenger flows travelling from stop \( i \) to stop \( j \), respectively. The variables \( R_i \) and \( C_j \) are factors for row \( i \) and column \( j \), respectively. The degree to which the marginal row and column totals of the reconstructed OD flow matrix match the given boarding and alighting counts depends on the convergence criterion used in the solution procedure.

When onboard survey data are available, they can be used to form the base OD matrix used as input to the IPF method. However, if the OD survey sample size is relatively small compared to the dimension of the OD matrix, it is likely that travelling along OD pairs with low likelihood of travel will not be observed, and the survey will produce entries of zero for the corresponding OD pairs in the base OD matrix. The zero cell entries lead to the well-known “non-structural zeros” problem (Ben-Akiva et al., 1985). Non-structural zeros cause reconstruction difficulties when applying the IPF method, since the reconstructed passenger flows for OD pairs will be zero if the corresponding flows are zero in the base OD flow matrix, even though it may be feasible for passengers to travel along the given OD pairs. Ben-Akiva et al. (1985) introduced several methods to overcome the “nonstructural zeros” problem. One popular method is to add a small number in all feasible cells of the base OD matrix. However, the appropriate size of the small number to be added is subjective.

In the absence of onboard survey data, the null base OD matrix can be used in the IPF method. The null base matrix assigns equal values (e.g., ones) to all feasible OD pairs, assuming that passengers are equally likely to travel along any feasible OD pairs.
Furth and Navick (1992) proved that the IPF method using the null base OD matrix is equivalent to the Tsygalnitsky method.

The connection between the IPF method using the null base and the Tsygalnitsky method helps indicate situations in which the IPF method using the null base may or may not perform well. Li and Cassidy (2007) have pointed out that the Tsygalnitsky method may not work well on bus routes serving activity centers, such as shopping centers and metro stations, since the equally likely alighting assumption may not be reasonable in this case. Therefore, the IPF method using the null base may not produce good results on such bus routes either.

In addition, as discussed before, if passengers prefer walking or biking for short travel distances, at a given bus stop, onboard passengers who boarded at a close upstream bus stop would be less likely to alight than onboard passengers who have travelled on the bus for a certain distance. As a result, the equally likely alighting assumption of the Tsygalnitsky method results in overestimating the short travel trips. The Tsygalnitsky method tried to use the “minimum distance” assumption to address this issue (the same concept can also be applied to the IPF method). However, this assumption is too deterministic and may not be appropriate for all bus stops. Since the average travel length for all passengers is determined uniquely by the boarding and alighting counts (Furth and Navick, 1992), the overestimation of short travel trips leads to the overestimation of long travel trips by the Tsygalnitsky method. Therefore, the IPF method using the null base also tends to overestimate both the short and long travel trips, which has been demonstrated empirically by Simon and Furth (1985).
Kikuchi and Perincherry method

Kikuchi and Perincherry (1992) developed a method to reconstruct an OD matrix from boarding and alighting counts and a base OD matrix. Different from a traditional base OD matrix, the base OD matrix they used incorporates information on OD flows in the form of ranges. For example, the passenger flow from stops A to B may be considered to be in the range of 5 to 12. In their methodology, the authors proposed that the ranges are provided by an expert and that the expert specifies bounds for all OD pairs or a subset of OD pairs. When only a subset of OD pairs is specified, a method is presented to derive the ranges for the remaining OD pairs. The reconstructed OD matrix satisfies the given boarding and alighting counts with OD flows falling within the ranges specified in the base OD matrix.

An expert who is very familiar with the route under study is required in the Kikuchi and Perincherry method. In the proposed procedure, the expert specifies the ranges and then uses them as inputs to a developed algorithm to produce an OD matrix. The expert inspects the resulting OD matrix, identifies the problematic OD pairs (i.e., the pairs for which reconstructed passenger flows do not appear reasonable) and then revises the ranges. The revised ranges are input to the developed algorithm to produce a new reconstructed OD matrix. The whole process is repeated until the expert is satisfied with the resulting OD matrix. Although it is not discussed in the paper, inspecting all feasible OD pairs and finding problematic OD values can be burdensome for the expert, since a typical bus route has a very large number of feasible OD pairs.
*Gur and Ben-Shabat method*

Gur and Ben-Shabat (1997) formulated an optimization problem to reconstruct an Alighting Probability Matrix (APM) from boarding and alighting counts of multiple bus trips and a base APM matrix. An APM provides the probabilities that passengers alight at downstream bus stops, given the bus stop at which they boarded. The APM is assumed to be constant across bus trips. Coupled with the aggregated boarding counts across bus trips for each bus stop, the estimated APM reproduces the aggregated alighting counts across all bus trips for each bus stop.

The objective function in the Gur and Ben-Shabat formulation is a weighted sum of two components. The first component is the sum of the squared differences between the observed and estimated alighting counts over all stops and all bus trips. The second component is a measure of divergence between the base and estimated alighting probabilities based on the concept of entropy content (Vanzuylen and Willumsen, 1980). This measure is a function of observed boarding counts, estimated alighting probabilities and the base alighting probabilities.

Based on the results presented in the paper, it seems that the performance of the proposed formulation is sensitive to the weights for the two components in the objective functions. Since it appears that the combination of the two components in the objective function does not have a physical or behavioral interpretation, it would be difficult to choose the optimal weights for the two components in the objective function.
2.3.2 Statistical Approaches

This section reviews the approaches for OD estimation that are based on distributional assumptions about the observed data and the prior information of the true OD matrix. The “statistical” approaches presented in this section include Maximum Likelihood, Bayesian, Generalized Linear Squares (GLS) and Markov Chain methods. As will be seen, the GLS method does not rely on distributional assumptions. However, it is included in statistical approaches because it has the same formulation as Maximum Likelihood or Bayesian methods under certain distributional assumptions. In addition, a formulation proposed by Vardi (1996) and related studies conducted by other researchers are reviewed. Vardi’s formulation has similar properties as the approach proposed in this study for transit route-level passenger OD estimation.

**Maximum Likelihood methods**

Maximum Likelihood (ML) methods make distributional assumptions about the base OD matrix and boarding and alighting counts. In the ML methods, the base OD matrix is assumed to be obtained from OD surveys. The ML estimates are obtained by maximizing the joint likelihood of the survey derived OD data and boarding and alighting counts or the logarithm of their joint likelihood. Usually, it is assumed that survey OD data and boarding and alighting counts are independent. Under this assumption, the ML estimates can be obtained by an optimization problem that is equivalent to Program 2.1 in Equation 2.3, which is given by:
\[
\max_{T \geq 0} \log(p(T_0 | T)) + \log(p(\hat{x} | x)) \tag{2.6}
\]

where:

\[\log(p(T_0 | T)) = \text{logarithm of the probability of observing the base OD flows } T_0,\]
conditional on the true OD flows \( T \) (e.g., expected OD flows on a bus trip in a given time-of-day period),

\[\log(p(\hat{x} | x)) = \text{logarithm of the probability of observing boarding and alighting counts } \hat{x},\]
conditional on the true boarding and alighting counts \( x \) with \( x = \tau \times T \).

Under the assumption that the observed trip-level boarding and alighting counts are unbiased (i.e., the expectations of the observed trip-level boarding and alighting counts equal the actual trip-level boarding and alighting counts), if boarding and alighting counts on a sufficient number of bus trips are available, the sample means of the observed trip-level boarding and alighting counts across bus trips are good approximations of the true trip-level boarding and alighting counts obtained by assigning the true OD matrix to the network. The true OD matrix describes the expected OD flows on a bus trip in a given time-of-day period. As a model approximation, the ML estimates may be obtained by an optimization problem that is equivalent to Program 2.2 in Equation 2.4, which is given by:

\[
\max_{T \geq 0} \log(p(T_0 | T)) \tag{2.7a}
\]

\[st: \tau \times T = \bar{x} \tag{2.7b}\]
where $\bar{x}$ are sample means of the observed trip-level boarding and alighting counts, the means are taken by stop across bus trips, and the rest of the notations has been defined above.

In the literature, some papers sought to find an OD matrix that reproduces either boarding and alighting counts on a bus trip or the boarding and alighting counts aggregated by stop across multiple bus trips, assuming that measurement errors in boarding and alighting counts are zero (Ben-Akiva et al., 1985; Geva et al., 1983). These formulations could also be summarized by an optimization formulation equivalent to Program 2.2 in Equation 2.4, which is given by:

$$\max_{T \geq 0} : \log(p(T_0 | T)) \quad (2.8a)$$

$$st : \tau \times T = x \quad (2.8b)$$

where $x$ are boarding and alighting counts on a bus trip or aggregated by stop across multiple bus trips.

Given the same survey OD data and boarding and alighting counts on multiple bus trips, Equations 2.7 and 2.8 produce different volume OD flow estimates, since the true OD matrix is constrained by the sample means of the trip-level boarding and alighting counts in Equation 2.7 and by the aggregated boarding and alighting counts in Equation 2.8. However, both equations can be shown to produce the same normalized OD flow estimates under certain assumptions about the base OD flow matrix. That is, the normalized estimated OD flow matrices determined by Equations 2.7 and 2.8 are the same, where the normalization is performed by dividing each cell in the estimated volume OD flow matrix by the total boarding (or alighting) volume.
Even though Equations 2.7 and 2.8 may result in the same normalized OD flow estimates, they represent different OD estimation problems (see Section 2.2). Equation 2.7 represents the OD estimation problem in that the true OD matrix $T$ summarizes the expected number of passenger trips for each OD pair on a bus trip in a given time-of-day period, while Equation 2.8 represents the OD reconstruction problem in that the true OD matrix $T$ summarizes the actual number of passenger trips for each OD pair on one bus trip or multiple bus trips.

Equations 2.7 and 2.8 assume that survey OD data and boarding and alighting counts are conditional on the same true OD matrix. However, they were developed based on some different assumptions. As discussed above, the ML methods assume that survey OD data and boarding and alighting counts are independent. This assumption implies that survey OD data and boarding and alighting counts were collected on different bus trips. Equation 2.7 was developed based on the assumptions above. However, Equation 2.8 was developed based on the assumption that survey OD data were collected on bus trips on which boarding and alighting counts were collected. Equation 2.8 reconstructs an OD matrix for bus trips on which boarding and alighting counts were observed. If survey OD data and boarding and alighting counts were collected on different bus trips, the true OD flows for bus trips on which boarding and alighting counts were collected should be different from those for bus trips on which survey OD data were collected, since passenger demand varies across bus trips.

As discussed before, the variation of the trip-level boarding and alighting counts results from passenger demand variation across bus trips and measurement errors in
boarding and alighting counts. Incorporating both demand variation and measurement errors in one formulation should be straightforward. However, developing an algorithm to solve the estimation problem under such conditions is likely to be challenging. Therefore, usually only demand variation, measurement errors, or the compound effect of both is considered in the formulation in the literature. Poisson and Multivariate Normal distributions are commonly used to capture the variation of the trip-level boarding and alighting counts.

Usually, boarding and alighting counts on a bus trip are assumed to be random variables with means equal to the true boarding and alighting counts $\tau \times T$. The true OD matrix $T$ summarizes the expected OD flows on a bus trip in a given time-of-day period. Under the assumption that OD flows on a bus trip are independently Poisson distributed conditional on the true OD flows, it can be shown that the trip-level boarding and alighting counts are Multivariate Poisson distributed (Karlis and Meligkotsidou, 2005). If the correlations among the trip-level boarding and alighting counts across bus stops are assumed to be zero, the trip-level boarding and alighting counts may be assumed to be independently Poisson distributed (Spiess, 1987).

Ben-Akiva (1987) also assumed that boarding and alighting counts are independently Poisson distributed. However, in this case, the Poisson assumption for boarding and alighting counts was used to capture measurement errors in boarding and alighting counts instead of passenger demand variation across bus trips.

In a Multivariate Normal model, a variance-covariance matrix is used to quantify the variation of passenger demand across bus trips. If it is assumed that OD flows on a
bus trip are Multivariate Normal distributed conditional on the true OD flows for a bus trip and the variance-covariance matrix, it can be shown that the trip-level boarding and alighting counts are also Multivariate Normal distributed (Maher, 1983).

The Multivariate Normal distribution is often used to model the variation of OD flows in the same time-of-day period across days in highway OD estimation. However, it may not be appropriate in transit applications, since passenger OD flows at the bus trip level are usually very small. Although the Possion distribution seems appropriate for capturing the variation of OD flow across bus trips, the equality of the mean and variance of OD flow imbedded in the Possion assumption may be too restrictive for transit applications.

Besides the distributional assumptions made for the trip-level OD flows and the observed boarding and alighting counts, distributional assumptions are also made for survey OD data. The true OD flows are treated as the parameters of the assumed distribution of the survey OD data. Commonly used distributions for survey OD data include Multivariate Hypergeometric, Multinomial and Poisson (Cascetta and Nguyen, 1988; Geva et al., 1983). The distribution of survey OD data depends on the sampling procedure (e.g., simple random, stratified or cluster sampling) and the true OD matrix to be estimated. The simple random sampling, which is discussed in the following, is most often assumed in the literature. That is, onboard passengers are randomly sampled to ask for their travel origins and destinations. Similar approaches could be developed for other sampling procedure. The Multivariate Hypergeometric distribution is appropriate only for the problem of OD reconstruction (Equation 2.8). The Multivariate Hypergeometric
distribution assumes that passengers are sampled without replacement from a finite population. The population is assumed to be finite in the problem of OD reconstruction, while not in the problem of OD estimation.

The Multinomial and Poisson distributions are appropriate for both the problem of OD reconstruction (Equation 2.8) and OD estimation (Equations 2.6 and 2.7). Recall that the population is finite in the problem of OD reconstruction. If the OD survey sample size is much smaller than the population size, the Multinomial distribution is a good approximation of the Multivariate Hypergeometric distribution. In addition, in both the OD reconstruction and OD estimation problems, if the total number of sampled passengers and the number of feasible OD pairs are sufficiently large, the Poisson distribution is a good approximation of the Multinomial distribution (Cascetta and Nguyen, 1988).

The Multivariate Normal distribution is rarely used for the survey OD data because the number of surveyed passengers in many OD pairs would be very low. As a result, the Multivariate Normal distribution may not fit the data well. However, if the number of sampled passengers along each OD pair is sufficiently large, the Multivariate Normal distribution could be a good approximation of the Poisson distribution.

It is worth noting that the distributions of survey OD data discussed above assume that the same passenger will not be sampled twice. Actually, the same passenger may be sampled more than once if the survey was carried out in the same time-of-day period on multiple days, since the majority of the transit users travel on a bus route on a regular
basis. However, the OD survey sample size is usually small and, therefore, the number of passengers who are sampled more than once is low and can be considered negligible.

Bayesian methods

Bayesian methods are different from the ML methods in that Bayesian methods consider incorporating prior information about the true OD flows and make distributional assumption about the prior information. The prior distribution \( \pi(T) \) of the true OD flows, the likelihood \( p(\hat{x} \mid T) \) of observing boarding and alighting counts and the likelihood \( p(T_0 \mid T) \) of observing survey OD data are combined to form the posterior distribution of the true OD flows. The posterior distribution represents the total information about the true OD flows after boarding and alighting counts and survey OD data have been observed. Like ML methods, Bayesian methods also assume that boarding and alighting counts and survey OD data are independent. The posterior likelihood of the true OD flows is proportional to the product of the prior likelihood of the true OD flows, the likelihood of observing boarding and alighting counts and the likelihood of observing survey OD data. Using notation defined above, their relationship can be written by:

\[
\pi(T \mid \hat{x}, T_0) \propto \pi(T) \times p(\hat{x} \mid T) \times p(T_0 \mid T)
\] (2.9)

The Bayesian estimates can be obtained by maximizing the posterior likelihood of the true OD flows given in Equation 2.9 or the logarithm of their likelihood:

\[
\max_{T \geq 0} \log(\pi(T)) + \log(p(\hat{x} \mid x)) + \log(p(T_0 \mid T))
\] (2.10)

Equation 2.10 is equivalent to Program 2.1 in Equation 2.3. If it is assumed that the sample means of the trip-level boarding and alighting counts are good approximations
of the true trip-level boarding and alighting counts, as a model approximation, the Bayesian estimates may be obtained by an optimization problem equivalent to Program 2.2 in Equation 2.4, which is given by:

$$\max_{T \geq 0} \log(\pi(T)) + \log(p(T_0 | T)) \quad (2.11a)$$

$$st : \tau \times T = \hat{x} \quad (2.11b)$$

In the literature, some papers sought to find an OD matrix that reproduces boarding and alighting counts on a bus trip or the aggregated boarding and alighting counts across multiple bus trips, assuming that measurement errors in boarding and alighting counts are zero (Vanzuylen and Willumsen, 1980). It can be shown that these formulations can also be summarized by an optimization formulation given by:

$$\max_{T \geq 0} \log(\pi(T)) + \log(p(T_0 | T)) \quad (2.12a)$$

$$st : \tau \times T = x \quad (2.12b)$$

where $x$ are boarding and alighting counts on a bus trip or aggregated boarding and alighting counts by stop across multiple bus trips.

Given the same prior information, boarding and alighting counts and survey OD data, it can be shown that Equations 2.11 and 2.12 could result in the same OD flow estimates in the form of probabilities under certain prior distribution of the true OD matrix $T$ and distribution of the survey OD data $T_0$. However, Equation 2.11 represents the problem of OD estimation, while Equation 2.12 represents the problem of OD reconstruction as discussed in ML approaches.
The distributional assumptions about boarding and alighting counts and survey OD data in Bayesian methods are similar to those in ML approaches. In addition, Bayesian methods make assumptions about the prior distribution of the true OD matrix. The prior information is incorporated through the parameters of the prior distribution. Commonly used prior distributions include Multinomial, Poisson, and Multivariate Normal distributions (Cascetta and Nguyen, 1988; Maher, 1983). In Equation 2.12, if the prior distribution of the true OD matrix is Multinomial, Equation 2.12 is also referred to as Entropy maximization method (Cascetta and Nguyen, 1988; Vanzuylen and Willumsen, 1980).

It is worth noting that a null base is often referred to as non-informative prior in Bayesian approaches. The non-informative prior could be seen as the most naïve “guess” about an OD pattern on a bus route before observing any data. That is, the non-informative prior implies that passengers are equally likely to travel along any feasible OD pairs (Vanzuylen and Willumsen, 1980).

**Generalized Least Squares (GLS) method**

Similar to the ML methods that assume the trip-level boarding and alighting counts and the base OD flows are Multivariate Normal distributed, the GLS method also uses the variance-covariance matrix to quantify the variation of the trip-level boarding and alighting counts and the variation of the base OD flows. The inputs of the GLS method are the sample means of the trip-level boarding and alighting counts $\bar{x}$, a base OD matrix $T_0$, a variance-covariance matrix for the sample means of the trip-level boarding and alighting counts $W$, and a variance-covariance matrix for the base OD
matrix $Z$. The base OD matrix is assumed to be an unbiased sample estimate of the true OD matrix. The variance-covariance matrix $Z$ for the base OD matrix is used to capture the variation of the sample estimate of the true OD matrix $T_0$, which results from sample error. Similarly, the variance-covariance matrix for boarding and alighting counts $W$ is used to capture the variation of the sample means of the trip-level boarding and alighting counts $\bar{x}$, which results from passenger demand variation across bus trips and measurement errors in boarding and alighting counts. The GLS estimates can be obtained by minimizing the sum of “distance” between the true OD matrix and the base OD matrix and “distance” between the true boarding and alighting counts and observed boarding and alighting counts. This can be written as:

$$\min_{T \geq 0} (T_0 - T)' Z^{-1} (T_0 - T) + (\bar{x} - \tau T)' W^{-1} (\bar{x} - \tau T)$$  \hspace{1cm} (2.13)

Equation 2.13 is equivalent to Program 2.1 in Equation 2.3. If it is assumed that the sample means of the trip-level boarding and alighting counts are good approximations of the true trip-level boarding and alighting counts, the variances of the sample means of the trip-level boarding and alighting counts are close to zero. As a model approximation, the GLS estimates can be obtained by an optimization problem equivalent to Program 2.2 in Equation 2.4. This can be written as:

$$\min_{T \geq 0} (T_0 - T)' Z^{-1} (T_0 - T)$$ \hspace{1cm} (2.14 a)

$$st : \tau \times T = \bar{x}$$ \hspace{1cm} (2.14 b)

Equations 2.14 are also referred to as Constrained Generalized Least Squares (CGLS) estimator.
Li (2009) developed a Markov chain model to estimate an Alighting Probability Matrix (APM) from boarding and alighting counts collected on multiple bus trips. Li described the first-order and second-order Markov models to estimate the APM. In the fist-order Markov model, the first-order transition probability that a passenger alights at a given stop given that he or she is onboard immediately upstream of the given stop is crucial. In the second-order Markov model, besides the first-order transition probability, the second-order transition probability that a passenger alights at a given stop given that he or she is onboard immediately upstream of the given stop and the previous stop is also crucial. The discussion in the paper emphasized the fist-order Markov models. The properties of the second-order Markov model were not investigated.

In the first-order Markov model, the transition probability at a given bus stop can be estimated uniquely by boarding and alighting counts. Specifically, the ratio of alighting counts at the given bus stop to the number of onboard passengers immediately upstream of the given bus stop yields the estimate of the transition probability. It is shown that the APM is determined uniquely by the transition probabilities for all bus stops. In the second-order Markov model, the first and second-order transition probabilities are estimated by simulation. It is also shown that the APM is determined uniquely by the first and second-order transition probabilities for all bus stops.

Li showed that the posterior distributions of the transition probabilities for one bus trip can be used as the prior distributions of the transit probabilities for the next bus trip in the first-order Markov model. However, it can be shown that doing so is
equivalent to aggregating boarding and alighting counts over two bus trips first, and then applying the Markov model on the aggregated boarding and alighting counts. Similar properties were not discussed for the second-order Markov model.

Unlike the other methods, the Markov chain model incorporates prior information about the transition probabilities instead of the APM or the OD matrix. If survey OD data are available, transforming the survey OD data into the structure required by the Markov model may result in loss of information, since survey OD data contains richer information about the underlying APM than transition probabilities. That is, the target is to estimate the APM instead of the transit probabilities. A set of transition probability estimates can be derived from survey OD data. However, the reverse is not true.

_Vardi’s formulation_

Vardi (1996) proposed a formulation to estimate the true OD flows in a given time-of-day period in communication networks. It was assumed that OD flows in a given observation period are independently Poisson distributed conditional on the true OD flows in that period. The inputs to Vardi’s formulation are observed link flows in the same time-of-day period across multiple days. A set of link flows in one observation period in communication network OD estimation is analogous to a set of boarding and alighting counts on one bus trip in transit route-level passenger OD estimation. Survey OD data is not considered in Vardi’s formulation.

Vardi used a simple example to demonstrate the difficulty of maximizing the full likelihood of link flows. Specifically, the maximization requires enumerating all possible OD matrices that reproduce link flows in each observation period. Vardi recognized that
the first and second moments of link flows are also functions of the expected OD flows. As a result, both the first and second moments of link flows contain information that can be used to estimate the expected OD flows. To avoid OD enumeration, he developed an Expectation Maximization (EM) algorithm based on the first and second moments of link flows to produce a solution of OD estimation. A small simulation study was carried out to demonstrate that good estimates can be obtained using both the first and second moments of link flows. However, the proposed EM algorithm depends heavily on the independent Poisson assumption of OD flows. If this assumption is not valid, it can be shown the proposed EM algorithm may produce unreliable estimates.

Tebaldi and West (1998) developed a Markov Chain Monte Carlo (MCMC) algorithm for Vardi’s formulation to simulate the posterior distributions of the true OD flows for a given time-of-day period in moderately large highway networks. Tebaldi and West only considered link flows in one observation period for OD estimation. Therefore, information in higher moments of link flows were not able to be utilized. As a result, the MCMC algorithm developed by Tebaldi and West requires survey OD data to achieve good OD estimates. In addition, it can be seen in the paper that the proposed MCMC algorithm is computationally expensive and therefore not computationally feasible when the number of OD pairs is large.

Hazelton (2000) also applied Vardi’s formulation to estimate the true OD flows for a given time-of-day period in highway networks. The inputs are survey OD data, prior information about the true OD flows and link flows in multiple time-of-day observation periods. Due to the difficulty of maximizing the full likelihood of the link flows,
Hazelton proposed a Multivariate Normal (MVN) approximation and a Generalized Least Square (GLS) approximation to the distribution of link flows conditional on the underlying true OD flows. The parameters in both approximations are the true link flows (linear functions of the true OD flows) and the variance-covariance matrix of link flows. The difference between the MVN and GLS approximations is in their treatment of the variance-covariance matrix. In the MVN approximation, the variance-covariance matrix is a function of the true OD flows, whereas in the GLS approximation, the variance-covariance matrix is estimated exogenously by the sample variance-covariance matrix using link flows in multiple observation periods. It was shown that the GLS approximation is equivalent to the GLS method developed by Cascetta (1984).

Hazelton extended the work of Vardi (1996) by incorporating additional OD information and measurement errors in link flows in the formulation of OD estimation. However, algorithms were not provided to produce a solution for the MVN approximation. Later, Hazelton (2003) claimed that the algorithm to produce the estimates was computationally time consuming and numerically unstable for even moderately sized networks.

Li (2005) proposed an Expectation Maximization (EM) algorithm for Vardi’s formulation to estimate the true OD flows for a given time-of-day period in highway networks. The inputs are prior information about the true OD flows and observed link flows in multiple time-of-day observation periods.

The EM algorithm is an iterative method (Meng and vanDyk, 1997) used to find the marginal mode from the joint distribution. One iteration of the EM algorithm includes
two steps: the E-step and the M-step. The E-step determines the expected OD flows for each observation period, conditional on the underlying true OD flows estimated from previous iteration and observed link flows in the corresponding observation period. The M-step determines the mode of the distribution of the true OD flows conditional on the expected OD flows for all observation periods determined in the E-step and the prior information about the true OD flows.

It was shown that the M-step is relatively easy based on model assumptions. However, determination of the expected OD flows for each observation period in the E-step is difficult. Direct determination of the expected OD flows requires enumerating OD matrices that satisfy link flows for each observation period. Therefore, in the E-step, Li proposed a Multivariate Normal approximation to the Poisson distribution, conditional on the true OD flows (OD flows in a given observation period are assumed to be independently Poisson distributed.). Specifically, the mean and variance of the OD flow are derived from the Poisson assumption first, and then the derived mean and variance of the OD flow are used as the parameters of the Multivariate Normal distribution. It was shown that the conditional expected OD flows for each observation period can be obtained based on the Multivariate Normal approximation.

However, it can be shown that the EM algorithm proposed by Li (2005) requires informative prior information as input. Otherwise, the EM algorithm is not able to converge. That is, different starting values of the EM algorithm may lead to dispersed OD estimates.
Many of the methods reviewed above could be used to estimate transit route-level passenger OD matrices. All these methods reflect the optimization concept summarized by Programs 2.1 and 2.2. Although these methods were developed based on different assumptions and have different “distance” functions in Equations 2.3 and 2.4, most methods are related to each other in some way. The next section discusses the relationships among methods that could be used for transit route-level passenger OD matrix estimation. This discussion provides additional insights for developing a new transit route-level passenger OD matrix estimation method.

2.4 Relationships among Transit Route-level Passenger OD Estimation Methods

Figure 2.1 presents a map developed to describe the relationships among methods relevant to transit route-level passenger OD estimation methods that were reviewed in Section 2.3. Each box in Figure 2.1 represents a method, which is named according to distributional assumptions for the statistical approaches, and by the authors for the non-statistical approaches. The distribution names refer to the distribution of survey OD data in the column of Maximum Likelihood (ML) approaches, and to the prior distribution of the true OD flows in the column of Bayesian (Bayesian) approaches. For example, the box corresponding to the Multinomial in the column of Bayesian approaches represents the Bayesian approaches that assume the prior distribution of the true OD flows are multinomially distributed. The number in the box refers to the number in the first column of Table 2.1, which provides the references for each method.

For the convenience of comparison, it is assumed that the methods in Figure 2.1 estimate (reconstruct) an OD matrix that reproduces the boarding and alighting counts
This assumption was already made for methods other than ML, Bayesian and GLS methods. Making such an assumption does not lose generality for the ML, Bayesian and GLS methods. In the general framework of the ML, Bayesian and GLS approaches, the true OD flow matrix does not reproduce the boarding and alighting counts due to passenger demand variation across bus trips and measurement errors in boarding and alighting counts. However, as discussed in Section 2.3, if boarding and alighting counts on a large number of bus trips are available, the sample means of the trip-level boarding and alighting counts are good approximations of the true trip-level boarding and alighting counts obtained by assigning the true OD flows (which represent the expected OD flows on a bus trip of a given time-of-day period) to the route. As a result, the general formulation of the ML, Bayesian or GLS approaches could be approximated by a special formulation in which the estimated OD matrix reproduces the sample means of the trip-level boarding and alighting counts.

Three basic relationships are defined in Figure 2.1: equivalent, conditionally equivalent and approximate. Their definitions are as follows:

- **Equivalent**: Methods A and B are equivalent if they produce exactly the same OD flow estimates (in the form of volume or probability), given the same input data (e.g., boarding and alighting counts and the base OD matrix). This relationship is represented by a two directional solid arrow between A and B in Figure 2.1.

- **Conditionally equivalent**: Methods A and B are conditionally equivalent if they produce exactly the same OD flow estimates, under certain circumstances (e.g., the
null base OD matrix is used in OD estimation). This relationship is represented by a two directional dotted arrow between A and B in Figure 2.1.

- **Approximate**: Method B is an approximation of Method A if the distributional assumptions in Method B could approximate those in Method A and both methods produce similar OD flow estimates. This relationship is represented by a solid arrow from A to B in Figure 2.1.

![Figure 2.1 Relationships among Methods Relevant to Transit Route-level Passenger OD Matrix Estimation](image)
<table>
<thead>
<tr>
<th>Index</th>
<th>Method</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Markov</td>
<td>Li (2009)</td>
</tr>
<tr>
<td>2</td>
<td>Li and Cassidy</td>
<td>Li and Cassidy (2007)</td>
</tr>
<tr>
<td>3</td>
<td>Tsygalnitzky</td>
<td>Simon and Furth (1985); Ben-Akiva et al. (1985); Furth and Navick (1992);</td>
</tr>
<tr>
<td>4</td>
<td>IPF</td>
<td>Bacharach (1970); Furth and Navick (1992)</td>
</tr>
<tr>
<td>5</td>
<td>Entropy</td>
<td>Vanzuylen and Willumsen (1980)</td>
</tr>
<tr>
<td>6</td>
<td>Gur and Ben-Shabat</td>
<td>Gur and Ben-Shabat (1997)</td>
</tr>
<tr>
<td>7</td>
<td>CGLS</td>
<td>Hendrickson and McNeil (1984); Ben-Akiva et al. (1985); Cassetta (1984); Cascetta and Nguyen (1988);</td>
</tr>
<tr>
<td>8</td>
<td>Multinomial Bayesian</td>
<td>Cascetta and Nguyen (1988)</td>
</tr>
<tr>
<td>9</td>
<td>Poisson Bayesian</td>
<td>Cascetta and Nguyen (1988)</td>
</tr>
<tr>
<td>10</td>
<td>Normal Bayesian</td>
<td>Maher (1983); Cascetta and Nguyen (1988)</td>
</tr>
<tr>
<td>11</td>
<td>Hypergeometric</td>
<td>Geva et al. (1983)</td>
</tr>
<tr>
<td>12</td>
<td>Multinomial</td>
<td>Geva, Hauer &amp; Landau (1983); Ben-Akiva et al. (1985); Cascetta and Nguyen (1988);</td>
</tr>
<tr>
<td>13</td>
<td>Poisson</td>
<td>Ben-Akiva et al. (1985); Spiess (1987); Cascetta and Nguyen (1988)</td>
</tr>
<tr>
<td>14</td>
<td>Normal</td>
<td>Cascetta and Nguyen (1988)</td>
</tr>
</tbody>
</table>

Table 2.1 Reference to Methods Presented in Figure 2.1

There are two groups of equivalent OD estimation methods. The first group includes the Bayesian approaches that use a Multinomial prior distribution on the true OD matrix (i.e., number 8 in Figure 2.1), the entropy method (i.e., number 5 in Figure 2.1), and the IPF method (i.e., number 4 in Figure 2.1). Cascetta and Nguyen (1988) have shown that the entropy method produces point estimates of OD flows that are the same as those produced by the Bayesian approaches that use a Multinomial prior distribution on the true OD matrix. It can also be shown that the IPF method produces the same OD flow estimates as point estimates produced by the Bayesian approaches that use a Multinomial prior distribution on the true OD matrix.
The second group of equivalent OD estimation methods includes the Bayesian approaches that use a Multivariate Normal prior distribution on the true OD matrix (i.e., number 10 in Figure 2.1), the ML approaches that assume survey OD data are Multivariate Normal distributed (i.e., number 14 in Figure 2.1), and the Constrained Generalized Least Squares (CGLS) method (i.e., number 7 in Figure 2.1). Cascetta and Nguyen (1988) have pointed out that methods in the second group produce exactly the same OD flow estimates.

As discussed above, conditionally equivalent methods produce exactly the same OD flow estimates under certain circumstances. There are two groups of conditionally equivalent OD estimation methods. The first group includes the Tsygalnitsky method (i.e., number 3 in Figure 2.1), the IPF method (i.e., number 4 in Figure 2.1), the Li and Cassidy method (i.e., number 2 in Figure 2.1) and the Markov model (i.e., number 1 in Figure 2.1). Furth and Navick (1992) proved that the Tsygalnitsky method is equivalent to the IPF method using the null base OD matrix. Li and Cassidy (2007) showed that the Tsygalnitsky method is a special case of their algorithm. It could also be shown that the Tsygalnitsky method produces the same estimates as the Markov model when the non-informative prior is used in the Markov model.

The second group of conditionally equivalent methods includes the Entropy method (i.e., number 5 in Figure 2.1) and the Gur and Ben-Shabat method (i.e., number 6 in Figure 2.1). Recall that the objective function in the Gur and Ben-Shabat method is a weighted sum of two components. The first component is the sum of the squared differences between the observed and estimated alighting counts over all stops and all bus
trips. The second component is a measure of divergence between the base and estimated APM based on the concept of entropy content. If the weight for the first component in the objective function is set to be zero, Gur and Ben-Shabat pointed out that their method produces the same estimates as the Entropy method.

As discussed above, one method is considered as an approximation of the other if the distributional assumptions in one method could approximate those in the other and both methods result in similar estimates. The Bayesian and ML approaches are considered approximations of each other. In the literature, the Bayesian approaches are typically used with one of three types of prior distribution on the true OD matrix: the Multinomial (i.e., number 8 in Figure 2.1), Poisson (i.e., number 9 in Figure 2.1), or Multivariate Normal (i.e., number 10 in Figure 2.1) distribution. The Poisson distribution is a good approximation of the Multinomial distribution if the total demand (i.e., the sum of true OD flows) and the number of feasible OD pairs are sufficiently large. The Normal distribution is a good approximation of the Poisson distribution if the true passenger flow along each OD pair is sufficiently large.

The ML approaches are typically used with one of four types of distribution of survey OD data: the Multivariate Hypergeometric (i.e., number 11 in Figure 2.1), Multinomial (i.e., number 12 in Figure 2.1), Poisson (i.e., number 13 in Figure 2.1), or Multivariate Normal (i.e., number 14 in Figure 2.1) distribution. As discussed in Section 2.3, the Multinomial distribution is a good approximation of the Multivariate Hypergeometric distribution if the total demand (i.e., the sum of true OD flows) is very large compared with the sample size of OD survey. The Poisson distribution is a good
approximation of the Multinomial distribution if the total number of sampled passengers and the number of feasible OD pairs are sufficiently large. The Normal distribution is a good approximation of the Poisson distribution if the number of sampled passengers along each OD pair is sufficiently large.

Given the relationships shown in Figure 2.1, it is not surprising to see that many different methods produce similar results, as presented by Cascetta and Nguyen (1988) and Ben-Akiva et al. (1985).

Methods presented in Figure 2.1 can be applied to boarding and alighting counts on one bus trip or multiple bus trips. Given boarding and alighting counts on multiple bus trips, methods in Figure 2.1 may treat boarding and alighting counts differently. For example, some methods aggregate boarding and alighting counts by stop across bus trips first, and then estimate an OD matrix from the aggregated counts, such as the IPF, ML, Bayesian and GLS methods. Some methods estimate an OD matrix for each bus trip first, and then aggregate the estimated OD matrices across bus trips to produce the final OD flow estimates (e.g., the Tsygalnitzky method and the Li and Cassidy method). It is clear that the former approach uses information in the sample means (i.e., the first moments) of the trip-level boarding and alighting counts as inputs for OD estimation (see Appendix A for a mathematical illustration). Although it is not mathematically clear that the latter approach is strictly based on the first moments of the trip-level APC data, given the close similarity in the estimates the two approaches produce, it is believed that the latter approach is at least predominantly based on the first moments of the APC data.
Notice that Figure 2.1 does not include Vardi’s formulation and the related work. Vardi’s formulation considered the full likelihood and not just the first moments, of link flows. Vardi’s formulation aimed to estimate OD matrices from large amounts of link flow data. The complexity of Vardi’s formulation motivates the development of four algorithms, which are summarized in Figure 2.2.

In Figure 2.2, the solid box represents the formulation, and the dashed box represents the algorithm. Each dashed box in Figure 2.2 is named according to an algorithm or a combination of algorithm and author. The number in the box refers to the number in the first column of Table 2.2, which provides the references for the formulation or algorithm.

Specifically, Vardi (1996) proposed an EM algorithm based on the method of moments (i.e., number 2 in Figure 2.2) to provide point estimates of the expected OD flows. Li (2005) developed another version of the EM algorithm (i.e., number 3 in Figure 2.2) to provide point estimates of the expected OD flows. Tebaldi and West (1998) developed the Markov Chain Monte Carlo (MCMC) algorithm (i.e., number 4 in Figure 2.2) to simulate the posterior distributions of the true OD flows. Hazelton (2000) proposed to use the GLS method (i.e., number 5 in Figure 2.2) to provide an approximate solution for Vardi’s formulation after introducing additional assumptions.
2.5 Motivation for the Current Study

In the transit literature, various methods were proposed for estimating route-level passenger OD flows from boarding and alighting counts before APC systems were prevalent, when data were collected on an infrequent basis and on a limited number of bus trips. As discussed above, methods in the transit literature use information in the sample means (i.e., the first moments) of the trip-level boarding and alighting counts for OD estimation. Statistically, the distribution of the trip-level boarding and alighting counts is more informative than the first moments of the trip-level boarding and alighting counts. Therefore, better OD estimates may be obtained by taking advantage of the
distribution rather the first moments only. For example, since passengers travelling from stop \(i\) to stop \(j\) by definition board at stop \(i\) and alight at stop \(j\), the boarding count at stop \(i\) and the alighting count at stop \(j\) are correlated with one another. This correlation may contain useful information for inferring the OD flow from stop \(i\) to stop \(j\). Such information is reflected in the distribution of the trip-level boarding and alighting counts, but not in the first moments of the trip-level boarding and alighting counts (See Appendix A for a mathematical illustration). Naturally the entire distribution contains potentially useful information to infer the OD flows.

Using information in link flows beyond the first moments for OD estimation has been recognized by Vardi (1996) for communication networks since it considered the full likelihood of link flows. Due to the complexity of the full likelihood, an EM algorithm based on the first and second moments of link flows was developed. It was demonstrated numerically that good OD estimates could be obtained by the EM algorithm. Vardi’s formulation was also applied to highway networks by several researchers (Hazelton, 2000; Li, 2005; Tebaldi and West, 1998).

It seems natural to extend Vardi’s formulation to transit networks for OD estimation in the present environment. Vardi’s formulation requires large amounts of link flow data. With the increasing deployment of APC systems, transit agencies have access to passenger boarding and alighting counts on an on-going basis. The availability of large amounts of APC boarding and alighting counts at the bus trip-level makes it possible to apply Vardi’s formulation to transit networks.
However, Vardi’s formulation may not be appropriate for transit route-level passenger OD estimation. In the context of the transit route-level passenger OD estimation, Vardi’s formulation would imply that the trip-level OD flows are independently Poisson distributed conditional on the expected trip-level OD flows. This assumption may be too restrictive in transit applications since the Poisson assumption restricts the variance of the trip-level OD flow to be equal to the mean of the trip-level OD flow. The reasonableness of the independent Poisson assumption could be evaluated by testing whether the trip-level total demand is Poisson distributed, since the trip-level total demand would be Poisson distributed if the trip-level OD flows are independently Poisson distributed. The trip-level total demand equals the sum of boarding or alighting counts on a bus trip, and therefore this demand can be easily obtained from APC data.

Empirical analysis demonstrates that the null hypothesis that the trip-level total demand is Poisson distributed is rejected based on empirical data and, therefore, the Poisson assumption does not appear reasonable for transit applications. Specifically, the trip-level total demands collected on 1077 APC-equipped bus trips in the same time-of-day period are used to test whether the trip-level total demand is Poisson distributed. The description of these APC data is presented in Section 6.2 of Chapter 6. The sample mean and variance of the trip-level total demand is 51.3 and 524.4, respectively. The variance of the trip-level total demand is more than nine times larger than the mean, indicating that the Poisson distribution does not appear to summarize the empirical trip-level total demand well. Figure 2.3 presents the QQ plot of the empirical trip-level total demand versus the Poisson distribution. If the Poisson distribution is able to fit the empirical data
well, the crosses points in the QQ plot would approximately lie on the dashed line. As can be seen in Figure 2.3, data in both the left and right tails of the empirical distribution of the trip-level total demand deviate systematically from the dashed line, indicating that Poisson is not a good distributional assumption for the trip-level total demand. The p-value of the Chi-Square goodness-of-fit test (Rice, 2006) is nearly zero for these data. Based on the discussion above, Vardi’s formulation does not appear suitable for transit route-level passenger OD estimation.

![Figure 2.3 QQ plot of the Empirical Trip-level Total Demands versus Poisson Distribution](image)

Figure 2.3 QQ plot of the Empirical Trip-level Total Demands versus Poisson Distribution
In addition, Vardi’s formulation assumes that the expected volume OD flow matrix is stable across bus trips in a given time-of-day period. This assumption may not be reasonable. The expected volume OD flow matrix for a bus route can be represented in the form of two components. One component is the expected trip-level total demand for the route. This demand describes the expected number of passengers travelling on a bus trip of the given bus route in a given time-of-day period. The other component is the probability OD flow matrix, which represents the probabilities that a randomly selected passenger from the population in the given time-of-day period travels from the specified origins to the specified destinations. The time-of-day variations of the expected trip-level total demand and of the probability OD flow matrix are likely to be different, which would shorten the time period over which the expected volume OD flow matrix could be considered stable. This difference has been demonstrated empirically (Ji et al., 2011). This difference also motivates the separate estimation of the expected trip-level total demand and the probability OD matrix.

Therefore, the objective of the study presented in this dissertation is to develop a method for estimating bus transit route-level passenger OD flows that takes into account the distribution of the trip-level boarding and alighting counts and makes assumptions appropriate for the bus transit context. As discussed before, it is more reasonable to estimate the expected trip-level total demand and the probability OD matrix separately. Therefore, this study focuses on estimating the probability OD matrix. The probability OD matrix is more informative than the expected trip-level total demand in describing passenger travel patterns. In addition, the probability OD matrix is more difficult to
estimate than the expected trip-level total demand. The trip-level total demand can be directly obtained from APC data.

To effectively taking advantage of the full distribution of the trip-level boarding and alighting counts, large datasets are needed. As a result of the adoption of APC technologies by many transit agencies, such datasets are currently readily available, and more agencies are expected to implement these technologies in the near future. As such, the alignment of a methodological and a technological opportunity motivate the objective of the research.
Chapter 3  Framework of the Proposed Approach

3.1 Introduction

Chapter 2 investigated the relationships among methods relevant to transit route-level passenger OD estimation. This investigation provides additional insights for developing a new transit route-level passenger OD estimation method. This chapter introduces a new approach to estimate a period-level OD matrix based on APC boarding and alighting counts and survey OD data collected in the same homogeneous time-of-day period.

As discussed in Chapter 2, it is more appropriate to estimate the expected trip-level total demand and the probability OD matrix separately. This study focuses on the estimation of the probability OD matrix since the probability OD matrix is more informative than the expected trip-level total demand in describing passenger travel patterns. The probability OD matrix is for the most parts assumed to be stable across bus trips in the given time-of-day period. However, this assumption is relaxed in Chapter 8, but is assumed in effect through Chapters 3-7.

This chapter is organized as follows: Section 3.2 presents the model structure and assumptions of the proposed approach. Section 3.3 presents the posterior distributions and point estimates of the probability OD matrix. The difficulties of estimating the probability OD flow matrix are discussed.
3.2 Model Structure and Assumptions

Figure 3.1 illustrates the hierarchical structure that describes the relationships among the parameter (represented by a Greek letter), variables (represented by upper case letters), and observed data (represented by lower case letters) assuming an analysis for a given homogeneous time-of-day period. The direction of the arrow indicates which parameter or variables (at the tail end of the arrow) determine which variable or data (at the head end of the arrow). The notation used in Figure 3.1 and the following development is summarized in Table 3.1 in addition to being presented subsequently where appropriate.

Figure 3.1 Relationships among Parameters, Variables and Observed Data
<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_s$</td>
<td>Number of bus stops in the given bus route.</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of feasible OD pairs in the given bus route. $N = N_s \times (N_s - 1) / 2$.</td>
</tr>
<tr>
<td>$L$</td>
<td>Number of bus trips for which APC counts were collected in the specified period.</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Hyper-parameter matrix in the prior distribution of the probability OD matrix. $\mu(r,q)$ is the parameter for OD pair from stop $r$ to stop $q$.</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Unknown parameter to be estimated $(N_s \times N_s)$. Probability OD matrix in the specified period. $\alpha(r,q)$ represents the probability that passenger travels from stop $r$ to stop $q$. In addition, $\alpha(r,:)$ and $\alpha(:,q)$ represents the sum of the probabilities in row $r$ and the sum of the probabilities in column $q$, respectively.</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Deterministic matrix $((2 \times (N_s - 1)) \times N_s)$, $\tau(s,t)$ is the cell value in $s^{th}$ row and $t^{th}$ column, which equals 1 if passenger flow of OD pair $t$ contributes to passenger counts in the $s^{th}$ cell of APC count vector $x_l$, otherwise, $\tau(s,t)$ equals zero.</td>
</tr>
<tr>
<td>$T_l$</td>
<td>Volume OD matrix $(N_s \times N_s)$ on bus trip $l$. $T_l(r,q)$ represents OD flow from stop $r$ to stop $q$ on bus trip $l$.</td>
</tr>
<tr>
<td>$T^c$</td>
<td>Collection of all trip-level OD matrices $(N_s \times N_s \times L)$ across all bus trips in the specified period.</td>
</tr>
<tr>
<td>$\text{tot}_l$</td>
<td>Trip-level total demand: number of passengers travelling on bus trip $l$.</td>
</tr>
<tr>
<td>$\text{tos}$</td>
<td>Total OD survey sample size.</td>
</tr>
<tr>
<td>$x_l$</td>
<td>True APC count vector $((2 \times (N_s - 1)) \times 1)$ at all bus stops on bus trip $l$ of the specified period. The vector entries include the boarding counts followed by the alighting counts.</td>
</tr>
<tr>
<td>$x^c$</td>
<td>Collection of true APC count vector $((2 \times (N_s - 1)) \times L)$ across all bus trips in the specified period.</td>
</tr>
<tr>
<td>$\hat{x}_l$</td>
<td>Observed APC count vector $((2 \times (N_s - 1)) \times 1)$ at all bus stops on bus trip $l$ of the specified period.</td>
</tr>
<tr>
<td>$\hat{x}^c$</td>
<td>Collection of observed APC counts $((2 \times (N_s - 1)) \times L)$ across bus trips in the specified period.</td>
</tr>
<tr>
<td>$z$</td>
<td>Aggregated OD survey flow matrix $(N_s \times N_s)$ across surveyed bus trips in the specified period.</td>
</tr>
<tr>
<td>$b_l$</td>
<td>True boarding count vector $(N_s - 1) \times 1$ at all stops on bus trip $l$ of the specified period. $b_l(r)$ is the true boarding count at stop $r$ on bus trip $l$. $b_l$ is a subset of $\hat{x}_l$.</td>
</tr>
</tbody>
</table>

Table 3.1 Descriptions of Notation in Figure 3.1 and the Following Development
Table 3.1 continued

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{b}_l )</td>
<td>Observed boarding count vector (((N_s - 1) \times 1)) at all stops on bus trip ( l ) of the specified period. ( \hat{b}_l(r) ) is the observed boarding count at stop ( r ) on bus trip ( l ). ( b_l ) is a subset of ( \hat{x}_l ).</td>
</tr>
<tr>
<td>( a_l )</td>
<td>True alighting count vector (((N_s - 1) \times 1)) at all stops on bus trip ( l ) of the specified period. ( a_l(q) ) is the true alighting count at stop ( q ) on bus trip ( l ). ( a_l ) is a subset of ( \hat{x}_l ).</td>
</tr>
<tr>
<td>( \hat{a}_l )</td>
<td>Observed alighting count vector (((N_s - 1) \times 1)) at all stops on bus trip ( l ) of the specified period. ( \hat{a}_l(q) ) is the observed alighting count at stop ( q ) on bus trip ( l ). ( a_l ) is a subset of ( \hat{x}_l ).</td>
</tr>
</tbody>
</table>

To illustrate this structure, consider first the trip-level volume OD flow matrix \( T_l \) in Figure 3.1. This matrix represents passenger flows along feasible OD pairs on bus trip \( l \) in the given time-of-day period. For the problem of interest, the trip-level volume OD matrix is not observed directly. Instead, the marginal values of the trip-level OD matrix are observed in the form of the boarding and alighting counts. For instance, on bus trip \( l \), the marginal values of the trip-level volume OD flow matrix \( T_l \) are boarding and alighting counts \( x_l \). Due to measurement errors, the observed boarding and alighting counts \( \hat{x}_l \) collected by APC systems are different from the true boarding and alighting counts \( x_l \). For completeness, the collection of volume OD flow matrices, the collection of boarding and alighting counts and the collection of observed boarding and alighting counts for all bus trips within the given time-of-day period of interest across different days is represented by \( T^c, x^c \) and \( \hat{x}^c \), respectively. In the above and the subsequent development, the direction of the arrow in Figure 3.1 flows from the variables or parameters that underly an outcome to the variables denoting the outcome.
The trip-level volume OD flow matrix $T_i$ is determined by both the underlying probability OD matrix $\alpha$ for the time-of-day period of interest and the trip-level total demand $tot_i$. The trip-level total demand $tot_i$ represents the total number of passengers travelling on bus trip $l$, which equals the total true boarding (alighting) passenger counts.

In addition, an OD flow survey may be conducted on different bus trips for the given time-of-day period. The survey OD flow data $z$ are obtained by aggregating survey OD flow data for each OD pair across bus trips in the given time-of-day period. The survey OD flow data $z$ are determined by the probability OD matrix $\alpha$ and the total survey sample size $tos$.

As in traditional OD estimation methods, this study assumes that OD survey data and trip-level APC counts are independent. This independence implies that OD survey data and trip-level APC counts were collected on different bus trips. If survey OD data and APC counts were collected on the same bus trip, overlapping information between the two sources of data results. A traditional OD survey consists of randomly interviewing passengers at bus stops or onboard a bus to collect the origins and destinations of their travels. Usually, the OD survey sample size is small. When OD survey is conducted at some time in the past prior to the availability of APC data, currently a common scenario, OD survey data and APC data collected subsequently would have no overlap and, therefore, would be independent from one another. In the event of concurrent survey and APC data collection, only a small proportion of passengers on bus trips for which APC counts were collected may be surveyed, and therefore, the overlapped information in both data sets is negligible. Another rare type of
transit passenger OD survey is one where the origins and destinations of travels of all or majority of passengers on a bus trip are collected. In this case, OD survey data contain almost complete travel information about passengers on that bus trip. As a result, the APC data on that bus trip are redundant and could be omitted from OD estimation.

An OD matrix (either probability or volume OD matrices) may be rearranged into a vector, in which case it would be referred to as an OD vector. Sometimes it is more convenient to consider a vector than a matrix structure. In this study, these two terms are used interchangeably even if one structure is favored over the other in some discussion. In addition, sometimes it is more convenient to split the true count vector \( x_l \) into a vector of true boarding counts \( b_l \) and a vector of true alighting counts \( a_l \). The vector of true boarding counts \( b_l \) represents the number of true boarding passengers at each bus stop on bus trip \( l \) and the vector of true alighting counts \( a_l \) represents the number of true alighting passengers at each bus stop on bus trip \( l \). The observed count vector \( \hat{x}_l \) may be split into a vector of observed boarding counts \( \hat{b}_l \) and a vector of observed alighting counts \( \hat{a}_l \).

Generally, it is assumed that no passengers alight at the first (starting terminal) bus stop and no passengers board at the last (ending terminal) bus stop of the given bus route. It is also assumed that no passengers board and alight at the same bus stop (i.e., passenger flows along the diagonal of the OD matrix are zero). And, an OD pair is feasible if the origin stop is upstream of the destination stop. For the remainder of this dissertation, the OD flow related parameters (i.e., \( \alpha \)), variables (i.e., \( T_l \)) or data (i.e., \( z \)) reflect only the feasible OD pairs. Assuming there are \( N_s \) bus stops on a given bus route, the size of the trip-level true (observed) boarding \( b_l (\hat{b}_l) \) or alighting \( a_l (\hat{a}_l) \) count vector
is $N_s - 1$ and the size of the trip-level true (observed) counts $x_l (\hat{x}_l)$ is $2 \times (N_s - 1)$, and the number of feasible OD pairs in an OD matrix is $N_s \times (N_s - 1) / 2$.

The structure presented in Figure 3.1 could be better understood in the context of assumed distributions on parameters, variables, and observed data. The numbers next to some of the arrows in Figure 3.1 correspond to the following distributional assumptions and structural constraints:

(1) Conditional on the trip-level total demand and the probability OD matrix $\alpha$, trip-level OD matrices across bus trips in the specified time-of-day period are assumed to be independent. For a given bus trip $l$, conditional on the trip-level total demand $\text{tot}_l$ and the probability OD matrix $\alpha$, the trip-level volume OD matrix $T_l$ is assumed to be multinomially distributed:

$$f(T_l | \alpha) = f(T_l | \text{tot}_l, \alpha) \sim \text{Multinomial}(\text{tot}_l, \alpha)$$

$$= \frac{\text{tot}_l!}{\prod_{i=1}^{N_s-1} \prod_{j=i+1}^{N_s} T_l(i,j)!} \times \prod_{i=1}^{N_s-1} \prod_{j=i+1}^{N_s} \alpha(i,j)^{T_l(i,j)}$$

(3.1)

The dependence on the trip-level total demand $\text{tot}_l$ is suppressed in Equation 3.1 because the trip-level total demand $\text{tot}_l$ is assumed to be known and fixed for bus trip $l$ in the model structure (because it is directly observed from measurement error free APC data). As a result, subsequent discussion about the trip-level OD matrix $T_l$ is assumed to be conditional on this known trip-level total demand $\text{tot}_l$.

(2) Conditional on the trip-level OD matrix $T_l$, the true boarding count $b_l(r)$ at stop $r$ and alighting count $a_l(q)$ at stop $q$ on bus trip $l$ are determined uniquely, which are given by:
\[ b_i(r) = \sum_{q=r+1}^{N_s} T_i(r,q) \quad r = 1,\ldots,N_s - 1 \]  
(3.2a)

\[ a_i(q) = \sum_{r=1}^{q-1} T_i(r,q) \quad q = 2,\ldots,N_s \]  
(3.2b)

Equations 3.2a and 3.2b could also be represented in matrix form as follows:

\[ x_l = \tau \times T_l \]  
(3.3)

where:

\[ \tau = \text{the assignment mapping matrix}. \]

The vector \( x_l \) includes the boarding counts followed by the alighting counts. In Equation 3.3, the trip-level OD matrix \( T_l \) is rearranged in the form of vector. The relationship between the trip-level OD vector \( T_l \) and the trip-level true counts \( x_l \) is captured by the assignment mapping matrix, which is constant across bus trips. The cell value in \( s^{th} \) row and \( t^{th} \) column \( \tau(s, t) \) equals 1 if passenger flow of OD pair \( t \) contributes to passenger counts in the \( s^{th} \) cell of passenger count vector \( x_l \), otherwise, \( \tau(s, t) \) equals zero.

(3) The observed passenger counts \( \hat{x}_l \) are assumed to be independently distributed across bus stops and bus trips conditional on the true passenger counts \( x_l \). The conditional distribution of the observed passenger counts is represented by \( f(\hat{x}_l | x_l) \).

(4) Conditional on the total sample size of the OD survey \( t_{os} \) and the probability OD matrix \( \alpha \), the survey OD data \( z \) is assumed to be multinomially distributed:
The dependence on the survey sample size \( tos \) is suppressed in Equation 3.4 because the survey sample size \( tos \) is assumed to be known and fixed in the model structure. Hence, all subsequent discussion about the survey OD data \( z \) is assumed to be conditional on the OD survey sample size \( tos \).

The subsequent model development is based on the Bayesian framework given the prior nature of some information that could be considered in the model. The prior information represents knowledge or belief about an unknown quantity before observing any data. The analyst can specify the uncertainty associated with the prior information subjectively based on his or her own expert opinion. The prior knowledge and the likelihood of observing the available data are combined to form the inferred (posterior) knowledge about the parameters of interest (Berger, 1985).

This study focuses on using the trip-level APC boarding and alighting counts to estimate the probability OD matrix \( \alpha \). Survey OD flow data, if available, are also considered in the estimation. Considering an informative prior, such as a model derived OD matrix, is not within the scope of this research. Such consideration of prior information is discussed as part of future research in Chapter 9.

The prior distribution of the probability OD matrix \( \alpha \) is assumed to be Dirichlet, which is given by:
where,
\[ \Gamma(\cdot) = \text{the Gamma function and } \Gamma(n) = (n-1)! \text{ if } n \text{ is a positive integer}, \]
\[ \alpha(r, q) = \text{probability OD flows such that } \alpha(r, q) \geq 0 \text{ and } \sum_{r=1}^{N_s-1} \sum_{q=r+1}^{N_s} \alpha(r, q) = 1, \text{ and} \]
\[ \mu(r, q) = \text{hyper-parameter of the prior distribution of the probability OD matrix where the} \]
\[ \text{hyper-paramete } \mu(r, q) \text{ is positive for any feasible OD pair } (r, q). \]

The Dirichlet prior distribution is adopted because it is a conjugate prior distribution of the Multinomial distribution, which the trip-level OD matrices are assumed to follow. The conjugacy property means that the posterior distribution follows the same parametric form as the prior distribution, and only the hyper-parameters need to be updated in determining the posterior distribution from the prior distribution and the available data.

Any prior information about the probability OD matrix, such as a model derived OD matrix, could be represented by the hyper-parameter \( \mu \). When no prior information is available, a uniform prior distribution can be used. The uniform prior distribution is obtained by setting \( \mu(r,q) = 1 \) for all feasible OD pairs \((r,q)\). The uniform distribution assigns equal probability density function value to any matrix \( \alpha \) whose cell entries sum to one. Therefore, the hyper-parameter matrix \( \mu \) is analogous to the null base OD matrix in the transportation OD estimation literature in the sense that both represent the situation
where a random passenger is equally likely to travel between any feasible OD pair. The uniform Dirichlet prior distribution is also referred to as a non-informative prior since the posterior mode in this case is the maximum likelihood estimate (Gelman et al., 2004).

Assuming the analyst only knows the total demand for a bus trip, the most naïve guess about the OD matrix for that trip is that any feasible OD matrix, where the sum of its cell entries equals the given trip-level total demand, is equally likely. The uniform prior reflects this intuition. That is, the predictive distribution of the trip-level OD flows (given the trip-level total demand) is uniform. To show this, under the uniform Dirichlet prior and based on Equations 3.1 and 3.5, the predictive distribution for trip-level OD flows $T_i$ is given by:

$$f(T_i | tot_i) = \int f(T_i | \alpha, tot_i) f(\alpha) d\alpha$$

$$= \int \frac{(tot_i)!}{\prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} T_i(r,q)!} \times \prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} \alpha(r,q)^{T_i(r,q)} \times (N-1)! \times \prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} \alpha(r,q)^{1-1} d\alpha$$

$$= \frac{(tot_i)!}{\prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} T_i(r,q)!} \times (N-1)! \times \frac{\prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} T_i(r,q)!}{(tot_i + N - 1)!}$$

$$\times \int \frac{(tot_i + N - 1)!}{\prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} T_i(r,q)!} \times \prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} \alpha(r,q)^{T_i(r,q)+1-1} d\alpha$$

(3.6)

Terms A and B in Equation 3.6 are the Multinomial distribution and the uniform Dirichlet distribution, respectively. Term C in Equation 3.6 equals one since the
component inside the integration is the density function of a Dirichlet distribution. Therefore, Equation 3.6 is given by:

\[ f(T_i | tot_i) = \frac{(tot_i)!(N-1)!}{(tot_i + N - 1)!} \]

(3.7)

As implied by Equation 3.7, any OD matrix \( T_i \) that satisfies \( \sum_{r=1}^{N_s-1} \sum_{q=r+1}^{N_s} T_i(r,q) = tot_i \) is equally likely.

As in Vardi’s formulation (see Section 2.3.2), the proposed approach in this study is based on the distribution of the trip-level boarding and alighting counts. The difference between the proposed approach and existing OD estimation methods is discussed in detail in Appendix A, where the proposed approach is compared with four existing methods: the Iterative Proportional Fitting (IPF) method (Ben-Akiva, 1987; Ben-Akiva et al., 1985), Poisson Maximum Likelihood Estimator (PMLE) (Spiess, 1987), Vardi’s formulation (Vardi, 1996), and Normal Maximum Likelihood Estimator (NMLE) (Cascetta and Nguyen, 1988). In summary, Appendix A shows that the IPF, PMLE and NMLE methods strictly use the sample means (i.e., the first moments) of the trip-level boarding and alighting counts as inputs for OD estimation while Vardi’s formulation and the proposed approach use the distribution of the trip-level boarding and alighting counts for OD estimation. In general, the distribution is more informative than the first moments. Since the proposed approach allows for the use of more information in APC boarding and alighting counts than is used in current state-of-the-art transit OD estimation methods, it
is expected that the proposed approach would produce better estimates, at least when the distributional assumptions are valid.

The proposed approach is less restrictive than Vardi’s formulation. If the trip-level OD flows are multinomially distributed and the trip-level total demand is Poisson distributed, it can be shown that Vardi’s formulation and the proposed approach result in the same estimate of the probability OD matrix. As Vardi’s formulation assumes that the trip-level total demand is Poisson distributed, the final OD estimates based on Vardi’s formulation clearly depend on this assumption. By contrast, in the proposed approach, the trip-level total demand is assumed to be known and given in the model and, therefore, the estimate of the probability OD matrix does not depend on the distribution of the trip-level total demand, thus, offering more flexibility than Vardi’s formulation.

3.3 Formulation: Posterior Distributions and Point Estimates

This section derives both the joint posterior distribution of the probability OD matrix and the trip-level OD matrices and the marginal posterior distribution of the probability OD matrix. The marginal posterior distribution of the probability OD matrix can be used to describe the uncertainty in the probability OD matrix. For example, the posterior variance could reflect the variation of the probability OD matrix.

In practice, a point estimate of the probability OD matrix may be required by many applications. Usually, the mean, median or mode of the posterior distribution is used as the point estimate. In this study, the mode of the marginal posterior distribution is chosen as the point estimate because the mode is easier to obtain than the mean and the median given that the marginal posterior distribution is not known in closed-form.
As a less desirable alternate to use the marginal posterior distribution to estimate the probability OD matrix, when doing so is subject to computational or theoretical difficulties, the joint posterior distribution can also be used. This study develops approaches based on the marginal posterior distribution. However, as discussed in the next chapter, a joint posterior distribution approach is also presented for comparative purposes as part of the numerical and empirical aspects of this study presented in Chapters 5 through 7.

If the true trip-level OD matrix $T_l$ is known, the true trip-level APC counts $x_l$ are determined uniquely. Therefore, the following constraints apply to all the distributions developed in the following:

\[
\sum_{q=r+1}^{N_s} T_l(r,q) = b_l(r) \quad r = 1,\ldots,N_s - 1; l = 1,\ldots,L
\]

(3.8a)

\[
\sum_{r=1}^{q-1} T_l(r,q) = a_l(q) \quad q = 2,\ldots,N_s; l = 1,\ldots,L
\]

(3.8b)

Condition on the survey OD data $z$ and the observed trip-level APC counts $\hat{x}^c$, the joint posterior distribution of the probability OD matrix $\alpha$, the trip-level OD matrices $T^c$ and the true trip-level APC counts $x^c$ is given by:

\[
f(\alpha,T^c,x^c | \hat{x}^c,z) = \frac{f(\alpha,T^c,x^c,\hat{x}^c,z)}{f(\hat{x}^c,z)}
\]

\[
\propto f(\alpha,T^c,x^c,\hat{x}^c,z)
\]

\[
\propto f(\hat{x}^c | x^c,T^c,\alpha) \times f(x^c | T^c,\alpha) \times f(T^c | \alpha) \times f(z | \alpha) \times f(\alpha)
\]

(3.9)

In Equation 3.9, the marginal distribution of the observed trip-level APC counts and the survey OD data $f(\hat{x}^c,z)$ can be obtained by integrating the joint likelihood density
\( f(\alpha, T^c, x^c, \hat{x}^c, z) \) with respect to the probability OD matrix \( \alpha \), the true trip-level OD matrix \( T^c \) and the true trip-level APC counts \( x^c \). Therefore, \( f(\hat{x}^c, z) \) is not a function of the probability OD matrix \( \alpha \), the trip-level OD matrices \( T^c \) and the true trip-level APC counts \( x^c \). Consequently, \( f(\hat{x}^c, z) \) can be included in the proportionality constant as is done in the second and third lines of Equation 3.9. In addition, Equation 3.9 also reflects the assumption that the survey OD data \( z \) and the true trip-level APC counts \( x^c \) are independent.

The conditional distribution \( f(\hat{x}^c \mid x^c, T^c, \alpha) \) of the observed trip-level APC counts \( \hat{x}^c \) depends only on the true trip-level APC counts \( x^c \) since the probability OD matrix \( \alpha \) and the trip-level OD matrices \( T^c \) affect \( \hat{x}^c \) only through \( x^c \). Similarly, the conditional distribution \( f(x^c \mid T^c, \alpha) \) of the true trip-level APC counts \( x^c \) depends only on the trip-level OD matrices \( T^c \) since the probability OD matrix \( \alpha \) affects \( x^c \) only through \( T^c \). Based on the discussed above, Equation 3.9 reduces to:

\[
\begin{align*}
f(\alpha, T^c, x^c, \hat{x}^c, z) \propto f(\hat{x}^c \mid x^c) \times f(x^c \mid T^c) \times f(T^c \mid \alpha) \times f(z \mid \alpha) \times f(\alpha) 
\end{align*}
\]

(3.10)

Based on the assumption that the trip-level OD flows are independent across bus trips conditional on the probability OD matrix \( \alpha \), the joint conditional distribution of the trip-level OD matrices \( f(T^c \mid \alpha) \) can be obtained by the product of the conditional distributions of the trip-level OD matrices across bus trips. In addition, the true trip-level APC counts \( x^c \) are determined uniquely by the trip-level OD matrices \( T^c \) as captured by Equations 3.8a and 3.8b, which apply at the bus trip level. Moreover, based on the assumption that the observed trip-level APC counts \( \hat{x}^c \) are independently distributed conditional on the true
trip-level APC counts $x^c$, the joint conditional distribution of the observed trip-level APC counts $\hat{x}^c$ can be obtained by the product of the conditional distributions of the observed trip-level APC counts across bus trips. Therefore, Equation 3.10 reduces to:

$$f(\alpha, T^c, x^c | \hat{x}^c, z) \propto \prod_{l=1}^{L} f(\hat{x}_l | x_l) \times f(x_l | T_l) \times f(T_l | \alpha) \times f(z | \alpha) \times f(\alpha)$$

(3.11)

In Equation 3.11, $f(x_l | T_l)$ equals one if $x_l$ is determined by $T_l$ as given by Equations 8a and 8b and zero otherwise. The other conditional distributions $f(T_l | \alpha)$ and $f(z | \alpha)$ and the prior distribution of $\alpha$ in Equation 3.11 are given by Equations 3.1, 3.4 and 3.5, respectively. Therefore, Equation 3.11 can be expressed by:

$$f(\alpha, T^c, x^c | \hat{x}^c, z) \propto \prod_{l=1}^{L} \left[ f(\hat{x}_l | x_l) \times f(x_l | T_l) \times \prod_{r=1}^{N_l} \prod_{q=r+1}^{N_l} \alpha(r,q)^{\frac{\text{tot}_l \times f_T(r,q)}{T_l(r,q)!}} \times \prod_{r=1}^{N_l} \prod_{q=r+1}^{N_l} \alpha(r,q)^{\alpha(r,q) - 1} \right]$$

(3.12)

In Equation 3.12, Term A represents the likelihood of observing APC counts $\hat{x}_l$ on bus trip $l$ conditional on the true APC counts $x_l$ that are determined by OD matrices $T_l$ as given by Equations 3.8a and 3.8b. Term B represents the likelihood of observing an OD matrix $T_l$ conditional on the probability OD matrix $\alpha$. Term C represents the likelihood of observing the survey OD data $z$ in the given time-of-day period conditional on the probability OD matrix $\alpha$, and Term D represents the prior likelihood of the probability OD matrix $\alpha$. 
The marginal posterior distribution of the probability OD matrix $\alpha$ can be derived from the joint posterior distribution conditional on the survey OD data $z$ and the observed trip-level APC counts $\hat{x}^c$ of Equation 3.12 by summing the joint posterior distribution over feasible OD matrices $T_l$ for each bus trip. Recognizing that $x^c$ is uniquely determined by $T^c$ based on Equations 3.8a and 3.8b, the marginal posterior distribution of $\alpha$ is given by:

$$f(\alpha | \hat{x}^c, z) \propto \prod_{l=1}^{L} \left[ \sum_{T_l} f(\hat{x}_l | x_l) \times f(x_l | T_l) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r,q)^{z(r,q)} \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r,q)^{\mu(r,q)-1} \right] \times \frac{\prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r,q)^{\mu(r,q)-1} \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r,q)^{\mu(r,q)-1}}{T_l(r,q)!}$$

(3.13)

In Equation 3.13, Term A represents the distribution of observed APC data on bus trip $l$ conditional on the probability OD matrix $\alpha$. As is already discussed, taking advantage of this distribution is an important distinguishing aspect of the developed formulation and approach to estimate the probably OD matrix $\alpha$. Terms B and C in Equation 3.13 are the same as Terms C and D in Equation 3.12, respectively.

The marginal posterior distribution of the probability OD matrix in Equation 3.13 is not in closed-form. The numerical posterior distribution of the probability OD matrix $\alpha$ may be simulated by the Monte Carlo Markov Chain (MCMC) algorithm, which is developed and demonstrated on a simple, illustrative short bus route in Appendix B. However, the MCMC algorithm is computationally expensive for practical use on a
realistically long bus route given that the dimension of the OD matrix and the number of bus trips with APC counts are usually large.

A point estimate of the probability OD matrix can be given by the mode of the marginal posterior distribution of the probability OD matrix $\alpha$, which can be obtained by maximizing the marginal posterior likelihood of Equation 3.13 or the natural logarithm of the marginal posterior likelihood. What follows is maximization formulation of the latter, referred to as Program 3.1:

\[
\max_{\alpha \geq 0} : \sum_{l=1}^{L} \log \left[ \sum_{T_l} f(\hat{x}_l | x_l) \times f(x_l | T_l) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\mu(r,q)^{T_r(r,q)} T_r(r,q)!}{\alpha(r,q)} \right]
\]
\[
+ \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \left( z(r,q) + \mu(r,q) - 1 \right) \log(\alpha(r,q))
\]
\[\text{s.t.}
\sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \alpha(r,q) = 1
\]  

(3.14a)  

In Equation 3.14a, the true trip-level APC counts $x_l$ are determined uniquely by the trip-level OD matrix $T_l$, as given by Equations 3.8a and 3.8b.

Program 3.1 is analogous to Program 2.1 of Chapter 2 (see Section 2.2). However, Program 3.1 cannot be solved by traditional optimization solvers. Notice that the decision variables in Program 3.1 are the probability OD flows $\alpha$. The true trip-level APC counts $x^c$ and the trip-level OD matrices $T^c$ are latent variables, where, as already discussed, the true trip-level APC counts $x_l$ are determined uniquely by the true trip-level OD matrix $T_l$ on bus trip $l$. Therefore, the evaluation of the objective function of Equation 3.14a requires enumerating all possible OD matrices $T_l$ for each bus trip, which would be very
time consuming if the dimension of the OD matrix and the number of bus trips with APC counts are large.

It is worth mentioning that within the framework of maximum likelihood estimation, under the uniform Dirichlet distribution (i.e., \( \mu(r,q) = 1 \)), Equation 3.13 is the likelihood of observing the trip-level APC counts and the survey OD data. Therefore, under the uniform Dirichlet distribution, the solution of Program 3.1 is also the maximum likelihood estimate.

The OD estimation problem is relatively simpler if it is assumed that APC counts are measurement error free. In this case, the observed trip-level APC counts \( \hat{x}^c \) equal the true trip-level APC counts \( x^c \). Based on Equation 3.12, assuming APC counts are measurement error free, the joint posterior distribution of the probability OD matrix \( \alpha \) and the trip-level OD matrices \( T^c \), condition on the survey OD data \( z \) and the true trip-level APC data \( x^c \), is given by:

\[
f(\alpha, T^c | x^c, z) \propto \prod_{l=1}^{L} f(x_l | T_l) \times \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q)^{T_l(r,q)}}{T_l(r,q)!} \times \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \alpha(r,q)^{z(r,q)} \times \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \alpha(r,q)^{\mu(r,q)-1}
\]

(3.15)

\( f(\hat{x}_l | x_l) \) in Equation 3.12 equals one in Equation 3.15 since \( \hat{x}_l \) equals \( x_l \) under the assumption that APC counts are free of measurement errors. In addition, the factorial of the trip-level total demand \( \text{tot}_l! \) in Equation 3.12 is reflected in the proportionality constant of the conditional distribution of the probability OD matrix and trip-level OD matrices in Equation 3.15 since the trip-level total demand \( \text{tot}_l \) is known and given by APC data if APC counts are free of measurement errors.
The marginal posterior distribution of the probability OD matrix $\alpha$ can be derived from the joint posterior distribution of Equation 3.15 condition on the survey OD data $z$ and the trip-level APC data $x^c$ by summing the joint distribution of Equation 3.15 over the feasible OD matrix $T_i$ for each bus trip, where $T_i$ satisfies the APC counts $x_i$ as captured by Equations 3.8a and 3.8b, which is given by:

$$f(\alpha | x^c, z) \propto \prod_{l=1}^{L} \sum_{T_l \in S_l} \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\alpha(r, q)^{T_l(r, q)}}{T_l(r, q)!} \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q)^{x_l(r, q)} \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q)^{\mu(r, q)-1}$$

(3.16)

In Equation 3.16, Term A represents the distribution of the APC counts (in this case assumed to be the true boarding and alighting counts) on bus trip $l$ conditional on the probability OD matrix $\alpha$. As is already discussed and as is the case regarding the corresponding element in Equation 3.13, taking advantage of this distribution is an important distinguishing aspect of the developed formulation and approach to estimate the probably OD matrix $\alpha$. Term B represents the likelihood of observing the survey OD data $z$ in the given time-of-day period conditional on the probability OD matrix $\alpha$, and Term C represents the prior likelihood of the probability OD matrix $\alpha$. $T_i \in S_i$ in Equation 3.16 reflects the function $f(x_l | T_i)$ in Equation 3.15 and represents all OD matrices $T_i$ that satisfy the APC counts $x_i$ as given by Equations 3.8a and 3.8b.

A point estimate of the probability OD matrix could be given by the mode of the marginal posterior distribution, which can be obtained by maximizing the marginal
posterior likelihood of Equation 3.16 or the natural logarithm of the marginal posterior likelihood. What follows is maximization formulation of the latter, referred to as Program 3.2:

$$\max_{\alpha \geq 0} \sum_{l=1}^{L} \log \left[ \sum_{T \in S} \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q)T_l(r,q)}{T_l(r,q)!} \right] + \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} (z(r,q) + \mu(r,q) - 1) \times \log(\alpha(r,q)) \tag{3.17a}$$

s.t.

$$\sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \alpha(r,q) = 1 \tag{3.17b}$$

Program 3.2 is analogous to Program 2.1 of Chapter 2 and Program 3.1. As in Program 3.1, under the uniform Dirichlet distribution, the solution of Program 3.2 is also the maximum likelihood estimate. Program 3.2 is relatively simpler than Program 3.1 since the number of feasible OD matrices for each bus trip given by $S_l$ is bounded by the constraints of the APC boarding and alighting counts. This is not the case in Program 3.1 because the true counts and the observed counts are not equal, but rather related to one another through $f(\hat{x}_j | x_i)$. Nevertheless, because of the requirement of enumerating all feasible OD matrices that satisfy the given APC boarding and alighting counts for each bus trip in Equation 3.17a (as given by Equations 3.8a and 3.8b), traditional optimization solvers are computationally prohibitive to estimate the probability OD flows based on Program 3.2 for realistically long bus routes.

Although Program 3.2 is relatively simpler than Program 3.1, estimating the probability OD matrix for Program 3.2 is still not trivial given the enumeration issue already discussed. The next chapter focuses on Program 3.2 and develops
computationally efficient algorithms to determine a point estimate of the probability OD matrix.
Chapter 4  Conditional Maximization and Heuristic Expectation Maximization Algorithms

4.1  Introduction

This chapter develops a computationally implementable algorithm, referred to as the Heuristic Expectation Maximization (HEM) algorithm, to provide a good approximate estimate of the period-level probability OD flow matrix. The HEM algorithm maximizes the marginal posterior likelihood of probability OD matrix, assuming that APC counts are free of measurement errors. The HEM algorithm is a variant of the traditional Expectation Maximization (EM) algorithm (Gelman et al., 2004). The EM algorithm could be used to determine the solution to the estimation problem for Program 3.2 of Chapter 3. However, applying the EM algorithm requires enumeration of all OD flow matrices consistent with APC counts for each bus trip, while the HEM algorithm avoids this and obtains good approximations to the EM estimates based on the marginal posterior mode.

As discussed in Chapter 3, the joint posterior mode has often been used as point estimate when it is difficult to determine the marginal posterior mode (O'Hagan, 1976). It is worthwhile to develop a solution for this approach for comparison purposes. In this approach, instead of determining the mode of the marginal posterior distribution of the probability OD matrix as the point estimate, a mode of the joint posterior distribution of the probability OD matrix and the trip-level OD matrices is determined. The resulting
probability OD matrix is considered as the point estimate. The Conditional Maximization (CM) algorithm is applied for this purpose. In the remainder of this chapter, the CM algorithm is introduced in Section 4.2 and the EM and HEM algorithms are presented in Section 4.3.

4.2 **Conditional Maximization (CM) Algorithm**

The Conditional Maximization (CM) algorithm is an iterative method for finding values of a set of variables that maximize an objective function (Gelman et al., 2004). The CM algorithm starts with some initial values for one set of dimensions of the function and finds the maxima of the function with respect to the complementary set of dimensions conditional on the initial values of the original set. The maxima with respect to the original set are then determined conditional on the most recent maxima of the complementary set of dimensions. The above two steps are then repeated. The value of the objective function increases over successive iterations, eventually converging to a local mode of the objective function.

The CM algorithm could be used to find a mode of the joint posterior distribution of the probability OD matrix and the trip-level OD matrices. To begin with, the joint posterior distribution of the probability OD matrix and the trip-level OD matrices of Equation 3.15 is reproduced here for convenience:

\[
f(\alpha, T^c \mid x^c, z) \propto \prod_{i=1}^{L} \left[ f(x_i \mid T_i) \times \prod_{r=1, q=r+1}^{N_r} \frac{\alpha(r, q)^{T_i(r, q)}}{T_i(r, q)!} \right] \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q)^{z_{(r, q)}} (r, q)^{-1} \tag{4.1}
\]
In addition, Equations 3.8a and 3.8b capturing the relationship between the trip-level OD matrices and the trip-level APC boarding and alighting counts are reproduced for convenience:

\[ \sum_{q=r+1}^{N_s} T_l(r,q) = b_l(r) \quad r = 1, \ldots, N_s - 1; l = 1, \ldots, L \]  

(4.2a)

\[ \sum_{r=1}^{q-1} T_l(r,q) = a_l(q) \quad q = 2, \ldots, N_s; l = 1, \ldots, L \]  

(4.2b)

Recall, in Equation 4.1, \( f(x_l | T_l) \) equals one if the OD matrix \( T_l \) satisfies the APC counts \( x_l \) on bus trip \( l \) as given by Equations 4.2a and 4.2b and zero otherwise.

To find a mode of the joint posterior distribution presented in Equation 4.1, the CM algorithm is carried out as follows:

1. Start with some initial (likely crude) estimate of the probability OD matrix, \( \alpha^0 \).
2. For \( h=1, 2, \ldots \) apply the following two steps iteratively:
   
   (a) Find a set of OD matrices \( T^c \) that maximizes Equation 4.1 conditional on the current probability OD matrix \( \alpha^{h-1} \).
   
   (b) Conditional on \( T^c \) determined in Step (a), find a value of \( \alpha \) that maximizes the Equation 4.1.

Equation 4.1 reflects the assumption that \( T_l \) are independent across \( l \), and \( T_l \) are independent of \( z \) for all \( l \) conditional on the probability OD matrix \( \alpha \). Therefore, given the current probability OD matrix \( \alpha^{h-1} \), the matrix \( T_l \) that maximizes Equation 4.1 in Step 2a can be determined independently for all \( l \) by maximizing:
Equation 4.3 is obtained by removing the terms in Equation 4.1 that are not related to $T_l$. The value of $T_l$ that maximizes Equation 4.3 also maximizes Equation 4.1, given that $\alpha$ is known. Equation 4.3 can be interpreted as the likelihood of observing OD flow matrix $T_l$ that satisfies the APC counts $x_l$ on bus trip $l$, conditional on the probability OD flow matrix $\alpha$. This likelihood is obtained based on the assumption that the trip-level OD flows are multinomially distributed conditional on the probability OD flows.

The trip-level OD matrix $T_l$ that maximizes Equation 4.3 can be obtained approximately using the IPF method, in which the current probability OD matrix $\alpha^{k-1}$ is treated as the base OD flow matrix and APC counts $x_l$ are treated as the marginal values of the OD matrix to be determined. The argument for this approximation is presented in Appendix C.

In Step 2b, given $T^c$ (i.e., the set of $T_l$ for all $l$) determined in Step 2a, the value of $\alpha$ that maximizes the Equation 4.1 can be determined by maximizing:

$$f(\alpha | T^c, z) \propto \prod_{r=1}^{N_s} \prod_{q=r+1}^{N_s} \alpha(r,q)^{\sum_{i=1}^{L} T_l(r,q)} \times f(x_l | T_l)$$  \hspace{1cm} (4.4)
\[ \alpha(r, q) = \frac{\sum_{l=1}^{L} T_l(r, q) + z(r, q) + \mu(r, q) - 1}{\sum_{i=1}^{N_r-1} \sum_{j=1}^{N_q} \sum_{l=1}^{L} T_l(i, j) + z(i, j) + \mu(i, j) - N} \] (4.5)

where,

\( N \) is the number of feasible OD pairs.

The mode of Equation 4.5 exists if:

\[ \sum_{l=1}^{L} T_l(r, q) + z(r, q) + \mu(r, q) > 1 \quad \forall \; r, q \] (4.6)

The CM algorithm introduced above is easy to implement since it involves iteratively applying the IPF method to boarding and alighting counts on each bus trip along with Equation 4.5, which consists of aggregating the trip-level OD matrices determined by the IPF method across bus trips, until the algorithm converges. The aggregation reflected in Equation 4.5 is the normalized average in the absence of survey OD data, however, when survey OD data are available, the survey OD data are included in this aggregation step. Whether the CM algorithm is better than applying the IPF method on its own is investigated numerically and empirically in the following chapters.

The CM algorithm approximately finds a mode of the joint posterior distribution of the probability OD matrix and the volume OD matrices for bus trips on which the APC counts are collected. Therefore, the CM algorithm also considers information in the distribution of the trip-level APC data. However, when the number of bus trips with APC counts is large, the joint mode may not provide a good summary of the posterior distribution since the dimension of the joint mode increases as the number of bus trips.
with APC counts increases. The purpose of this research is to estimate the period-level probability OD matrix. The trip-level OD matrices may be considered as ‘nuisance’ variables. Therefore, estimating the probability OD matrix from the marginal posterior distribution of the probability OD matrix is expected to be superior to the estimate based on the joint posterior distribution (i.e., the CM algorithm). Although the dimension of the marginal mode is also large, the dimension is constant and does not increase as the number of bus trips with APC counts increases. Based on the discussion above, it is necessary to develop an operational algorithm to find the mode of the marginal posterior distribution. This is the subject of the next section.

4.3 Expectation Maximization (EM) and Heuristic Expectation Maximization (HEM) Algorithms

4.3.1 Implementation of the EM Algorithm to Estimate the Probability OD Flows

The Expectation Maximization (EM) algorithm is an iterative method for finding the mode of the marginal posterior distribution from the joint posterior distribution (Meng and vanDyk, 1997). In the context of the OD estimation problem of interest, the EM algorithm finds the mode of the marginal posterior distribution of the probability OD matrix from the joint posterior distribution of the probability OD matrix and the trip-level OD matrices. The convergence of the EM algorithm has been proven in the literature (Meng and vanDyk, 1997). The discussion of the convergence of the EM algorithm in the context of the probability OD estimation problem of interest is presented in Appendix D.

The EM algorithm, as it applies to the probability OD estimation problem at hand, can be described as follows:

1. Start with some (likely crude) estimate of the probability OD matrix, \( \alpha^0 \).
2. For \( h=1, 2, \ldots \) apply the following two steps iteratively:

(a) E-step: Determine the expected log joint posterior density function:

\[
E(\log(f(T^c, \alpha | x^c, z)) | \alpha^{h-1}, x^c, z) = \sum_{c} \log(f(T^c, \alpha | x^c, z)) \times f(T^c | \alpha^{h-1}, x^c, z) \quad (4.7)
\]

(b) M-step: Find the value \( \alpha = \alpha^h \) that maximizes:

\[
E(\log(f(T^c, \alpha | x^c, z)) | \alpha^{h-1}, x^c, z)
\]

The EM algorithm relies on the joint posterior distribution of the probability OD matrix and the trip-level OD matrices. In addition, the distribution of the trip-level OD matrices conditional on the probability OD matrix, the APC counts and survey OD data is required in the E-step. These distributions are presented in the following.

The joint posterior distribution of the probability OD matrix \( \alpha \) and the trip-level OD matrices \( T^c \) has been presented in Equation 3.15, which is reproduced here for convenience:

\[
f(\alpha, T^c | x^c, z) \propto \prod_{l=1}^{L} \left[ f(x_l | T_l) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q)^{T_l(r,q)}}{T_l(r,q)!} \right] \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r,q)^{z(r,q)+\mu(r,q)-1} \quad (4.9)
\]

In Equation 4.9, \( f(x_l | T_l) \) equals one if the OD matrix \( T_l \) satisfies APC counts \( x_l \) on bus trip \( l \) and zero otherwise. The logarithm of the joint posterior distribution of Equation 4.9 is given by:
log \left [ f \left ( \alpha, T^c \mid x^c, z \right ) \right ]
= \sum_{l=1}^{L} \sum_{r=1}^{N-1} \sum_{q=r+1}^{N} \left [ T_l(r,q) \times \log(\alpha(r,q)) - \log(T_l(r,q)!) \right ]
+ \sum_{r=1}^{N-1} \sum_{q=r+1}^{N} \left [ z(r,q) + \mu(r,q) - 1 \right ] \times \log(\alpha(r,q)) + \text{constant}
= \sum_{r=1}^{N-1} \sum_{q=r+1}^{N} \left [ \sum_{l=1}^{L} T_l(r,q) + z(r,q) + \mu(r,q) - 1 \right ] \times \log(\alpha(r,q))
- \sum_{l=1}^{L} \sum_{r=1}^{N-1} \sum_{q=r+1}^{N} \log(T_l(r,q)!) + \text{constant}
(4.10)

where \( T_l \) in Equation 4.10 satisfies the APC counts \( x_l \) on bus trip \( l \). In what follows, the E-step in its \( h^{th} \) iteration is discussed in detail followed by the same for the M-step.

**E-step**

The E-step determines the expectation of Equation 4.10 with respect to the trip-level OD matrices \( T^c \). The expectation is performed based on the distribution \( f(T^c \mid \alpha^{h-1}, x^c, z) \). Based on Equation 4.10, \( E(\log (f(\alpha, T^c \mid x^c, z)) \mid \alpha^{h-1}, x^c, z) \) is given by:

\[
E(\log (f(\alpha, T^c \mid x^c, z)) \mid \alpha^{h-1}, x^c, z)
= \sum_{r=1}^{N-1} \sum_{q=r+1}^{N} \left ( \sum_{l=1}^{L} E(T_l(r,q) \mid \alpha^{h-1}, x^c, z) + z(r,q) + \mu(r,q) - 1 \right ) \times \log(\alpha(r,q))
- \sum_{l=1}^{L} \sum_{r=1}^{N-1} \sum_{q=r+1}^{N} E(\log(T_l(r,q)!) \mid \alpha^{h-1}, x^c, z) + \text{constant}
(4.11)
\]

For a complete evaluation of Equation 4.11, both \( E(T_l(r,q) \mid \alpha^{h-1}, x^c, z) \) and \( E(\log(T_l(r,q)!) \mid \alpha^{h-1}, x^c, z) \) must be determined. However, for the purpose of executing the M-step, only the first expression needs to be determined since the second expression does not affect the M-step (i.e., the determination of a new probability OD matrix \( \alpha \) in the
M-step does not rely on the value of $E(T_i(r,q) | \alpha^{h-1}, x^c, z)$). The conditional expected OD flow $E(T_i(r,q) | \alpha^{h-1}, x^c, z)$ is given by:

$$E(T_i(r,q) | \alpha^{h-1}, x^c, z) = \sum_{T_l \in S_l} T_l(r,q) \times f(T_i | \alpha^{h-1}, x^c, z)$$  \hspace{1cm} (4.12)$$

In Equation 4.12, $T_l \in S_l$ represents all OD matrices $T_l$ that satisfy the APC counts $x_l$. The conditional distribution of OD matrix on bus trip $l$, $f(T_l | \alpha^{h-1}, x^c, z)$, is developed in the following.

The conditional distribution of interest $f(T_l | \alpha^{h-1}, x^c, z)$ can be derived from the joint posterior distribution of Equation 4.9. Equation 4.9 reflects the discussed assumption in Chapter 3 that $T_l$ are independent across $l$, and $T_l$ are independent of $z$ for all $l$ conditional on $\alpha$. Therefore, the conditional distribution $f(T_l | \alpha^{h-1}, x^c, z)$ can be obtained from Equation 4.9 by removing terms that are not related to $T_l$, which is given by:

$$f(T_l | \alpha^{h-1}, x^c, z) = f(T_l | \alpha^{h-1}, x_l) \propto f(x_l | T_l) \times \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{(\alpha^{h-1}(r,q))^{T_l(r,q)}}{T_l(r,q)!}$$  \hspace{1cm} (4.13)$$

Substituting Equation 4.13 into Equation 4.12 leads to:

$$E(T_i(r,q) | \alpha^{h-1}, x^c, z) = E(T_i(r,q) | \alpha^{h-1}, x_l)$$

$$= \sum_{T_l \in S_l} T_l(r,q) \times c_l \times \prod_{i=1}^{N_i} \prod_{j=1}^{N_j} \frac{(\alpha^{h-1}(i,j))^{T_l(i,j)}}{T_l(i,j)!}$$  \hspace{1cm} (4.14)$$

In Equation 4.14, $E(T_i(r,q) | \alpha^{h-1}, x^c, z) = E(T_i(r,q) | \alpha^{h-1}, x_l)$ because

$$f(T_l | \alpha^{h-1}, x^c, z) = f(T_l | \alpha^{h-1}, x_l)$$ as presented in Equation 4.13. In addition, $f(x_l | T_l)$ in Equation 4.13 is reflected in the collection of $T_l \in S_l$, which represents all OD matrices $T_l$.
that satisfy the APC counts \( x_l \). Moreover, \( c_l \) in Equation 4.14 is the proportionality constant in Equation 4.13 that satisfies the totality axiom for the distribution \( f(T_l \mid \alpha^{h-1}, x_l) \):

\[
c_l \sum_{T_l \in S_l} \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{(\alpha^{h-1}(r,q))^{T_l(r,q)}}{T_l(r,q)!} = 1
\]  

(4.15)

As is evident from the summation in Equations 4.14 and 4.15, obtaining the expected trip-level OD flows requires the enumeration of all OD flow matrices \( T_l \) that satisfy APC counts for each bus trip. Therefore, the E-step of the EM algorithm requires the enumeration of OD flow matrices for each bus trip, which is not feasible for realistic applications as already discussed when commenting on the direct optimization of Equation 3.17 of Chapter 3. This matter is addressed by developing a heuristic alternative resulting in the Heuristic Expectation Maximization (HEM) method presented in Section 4.3.2.

**M-step**

The M-step finds a new probability OD matrix \( \alpha^h \) that maximizes the expected log posterior density function in Equation 4.11. As discussed under the E-step, only the first term of Equation 4.11 is pertinent as the second term is independent from the probability OD matrix \( \alpha \), the optimal value of which is being determined in the M-step. Also recall that \( E(T_l(r,q) \mid \alpha^{h-1}, x_l) \) has been determined by Equation 4.14. Therefore Equation 4.11 for the purpose of the M-step can be reduced to:

\[
E(\log(f(T,\alpha \mid x,z)) \mid \alpha^{h-1}, x^c, z) \\
= \sum_{r=1}^{N_r} \sum_{q=r+1}^{N_q} \left( \sum_{l=1}^{L} E(T_l(r,q) \mid \alpha^{h-1}, x^c, z) + z(r,q) + \mu(r,q) - 1 \right) \times \log(\alpha(r,q))
\]  

+ constant

(4.16)

109
Maximizing Equation 4.16 is straightforward because Equation 4.16 has the form of a Dirichlet log posterior density function. The new probability OD matrix $\alpha^h$ that maximizes Equation 4.16 can be obtained by the mode of the Dirichlet posterior density function, which is given by (Gelman et al., 2004):

$$\alpha^h(r, q) = \frac{g(r, q) - 1}{\sum_{i=1}^{N_s-1} \sum_{j=i+1}^{N_s} g(i, j) - N}$$

(4.17)

where $N$ is the number of feasible OD pairs and

$$g(i, j) = \sum_{l=1}^{L} E(T_l(i, j) | \alpha^{h-1}, \mathbf{x}, \mathbf{z}) + z(i, j) + \mu(i, j)$$

(4.18)

Equation 4.18 holds if:

$$g(r, q) > 1 \ \forall \ r, q$$

(4.19)

The situation where $g(r, q)$ is less than or equal to one suggests that more data are needed for probability OD flow estimation. As evident from the elements of Equation 4.18, additional data could include trip-level APC counts, survey OD data or prior information. When $g(r, q)$ equals one, removing OD pair $(r, q)$ from the set of feasible OD pairs or not does not affect the final estimates. Most importantly, as can be seen in Equation 4.18, the sum of the conditional expected OD flow for OD pair $(r, q)$ over bus trips contributes to the value of $g(r, q)$. Therefore, the value of $g(r, q)$ is a monotone function of the number of bus trips with APC counts. Therefore, either way the situation where $g(r, q)$ is less than or equal to one could be resolved in the presence of APC data on a large number of bus trips.
Based on the above development, the implementation of EM algorithm in the current probability OD flow estimation problem can be simplified as follows:

1. Start with some (likely crude) estimate of the probability OD matrix, $\alpha^0$.

2. For $h=1, 2, \ldots$ apply the following two steps iteratively:
   
   (a) E-step: for each bus trip, determine the conditional expected values of OD flows
   $$E(T_l(r,q) | \alpha^{h-1}, x_l)$$
   based on Equation 4.13.

   (b) M-step: Determine $\alpha^h$ that maximizes
   $$E(\log( f(T, \alpha | x, z)) | \alpha^{h-1}, x^c, z)$$
   of Equation 4.16. Equation 4.16 has the form of a Dirichlet log posterior density function. The value of $\alpha$ that maximizes the Dirichlet distribution in Equation 4.16 is obtained through Equation 4.17, which represents the mode of the Dirichlet distribution.

4.3.2 Heuristic Expectation Maximization (HEM) Algorithm

As discussed in the previous section, the E-step of the EM algorithm requires OD enumeration for each bus trip (recall the summation in Equations 4.14 and 4.15). The suggested HEM algorithm avoids this enumeration problem by replacing the E-step of the EM algorithm with a heuristic resulting in an approximate determination of the conditional expected OD flows.

To do so, define $D^h_l(n|r)$ as the number of passengers onboard the bus on trip $l$ immediately upstream of stop $n$ who boarded at stop $r$. The expected value of this random variable, $E(D^h_l(n|r))$, is determined by subtracting the expected number of passengers
who boarded at stop $r$ and alighted at stops upstream of stop $n$ from the total number of passengers who boarded at stop $r$ on bus trip $l$, which is given by:

$$E(D_l^h(n\mid r)) = b_l(r) - \sum_{q=r+1}^{n-1} E(T_l^h(r,q\mid x_q,\alpha^{h-1})) \quad n = 3, ..., N_s; r < n$$

(4.20a)

and

$$E(D_l^h(n\mid r)) = b_l(r) \quad n = 2; r < n$$

(4.20b)

Recall that $b_l(r)$ is the boarding count at stop $r$ on bus trip $l$. Given Equation 4.20, at the $h$th iteration of the HEM algorithm, for bus trip $l$, the expected OD flows conditional on the probability OD matrix $\alpha$ can be determined recursively, beginning at the first alighting stop on the route and continuing downstream until the last alighting stop. At bus stop $n$, the conditional expected OD flow from stop $s$ ($\forall s: s < n$) to stop $n$ is determined approximately by:

$$E(T_l^h(s,n\mid \alpha^{h-1},x_q)) \approx a_l(n) \times \frac{E(D_l^h(n\mid s)) \cdot \alpha^{h-1}(s,n)}{\sum_{r=1}^{n-1} E(D_l^h(n\mid r)) \cdot \alpha^{h-1}(r,n)}$$

(4.21)

The second term of the product on the left hand side of Equation 4.21 is an approximation of the conditional probability of a random passenger boarded at stop $s$ given that she alighted at stop $n$.

Equation 4.21 is used to determine the conditional expected OD flows $E(T_l^h(s,n\mid \alpha^{h-1},x_q))$ for each bus trip. The conditional expected OD flows are determined recursively from the first alighting stop (i.e., $n = 2$) to the last alighting stop (i.e., $n = N_s$) on the bus route. At bus stop 2, the expected onboard passenger is actually boarding
count at stop 1 (Equation 4.20b). At a given bus stop \( n (n > 2) \), the conditional expected OD flows between stops upstream of stop \( n \) would have already been determined and are used in the summation term of Equation 4.20a to calculate the conditional expected OD flows destined for bus stop \( n \). Notice that this recursive nature of the E-step of the HEM algorithm circumvents the need for enumerating all possible OD flow matrices in the E-step of the EM algorithm.

What follows describes the argument supporting the development of the heuristic E-step in Equation 4.21. Since what follows applies to a general bus trip in the specified time-of-day period for any iteration of the HEM algorithm, the subscript \( l \) and the superscript \( h \) are dropped for convenience without any loss of completeness.

Suppose that there are \( d(n) \) onboard passengers immediately upstream of bus stop \( n \), among which \( D(n \mid s) \) passengers boarded at stop \( s \) \((s < n)\). The number of onboard passengers \( d(n) \) can be obtained directly from boarding and alighting counts, which is given by:

\[
d(n) = \sum_{q=1}^{n-1} b(q) - \sum_{q=2}^{n-1} a(q)
\]

(4.22)

The number of onboard passengers who boarded at stop \( s \) \( D(n \mid s) \) can be obtained by subtracting the number of passengers who boarded at stop \( s \) and alighted at bus stops upstream of bus stop \( n \) from the total boarding count at stop \( s \), which is given by:

\[
D(n \mid s) = b(s) - \sum_{q=s+1}^{n-1} T(s,q)
\]

(4.23)

The probability OD flow \( \alpha(s, n) \) represents the probability that a randomly selected passenger boards at stop \( s \) and alights at stop \( n \). Given that an onboard passenger
just upstream of stop $n$ boarded at stop $s$ (i.e., she has not alighted upstream of bus stop $n$), the probability that she will alight at bus stop $n$ is given by the conditional alighting probability:

$$P_n(A_n | B_s) = \frac{\alpha(s,n)}{\sum_{q=n}^{N_s} \alpha(s,q)} \tag{4.24}$$

where $A_n$ represents alighting at stop $n$ and $B_s$ represents boarding at stop $s$.

According to the Bayes’ law (Rice, 2006), the probability that an onboard passenger upstream of stop $n$ boarded at stop $s$ given that he or she alights at stop $n$ is given by:

$$P_n(B_s | A_n) = \frac{P_n(A_n | B_s) \times P_n(B_s)}{P_n(A_n)} = \frac{P_n(A_n | B_s) \times P_n(B_s)}{\sum_{r=1}^{n-1} P_n(A_n | B_r) \times P_n(B_r)} \tag{4.25}$$

where $P_n(B_r) = \text{probability that an onboard passenger immediately upstream of stop } n \text{ actually boarded at stop } r$. This probability is estimated by the ratio of the number of onboard passengers who actually boarded at stop $r$ to the total number of onboard passengers, which can be obtained from Equations 4.22 and 4.23:

$$P_n(B_r) = \frac{D(n | s)}{d(n)} = \frac{b(r) - \sum_{q=r+1}^{n-1} T(s,q)}{\sum_{q=1}^{n-1} b(q) - \sum_{q=2}^{n-1} a(q)} \tag{4.26}$$

Substituting Equations 4.24 and 4.26 into Equation 4.25 leads to the following:
The collection of the probabilities that an onboard passenger immediately upstream of bus stop \( n \) boarded at stops upstream of stop \( n \) given that she alights at stop \( n \) is represented by:

\[
P_n(B_s \mid A_n) = \left( P_n(B_1 \mid A_n), \ldots, P_n(B_{n-1} \mid A_n) \right)
\]  

(4.28)

where \( B_{1:(n-1)} \) represents the event of boarding at any of the stops 1 through \( (n - 1) \).

Conditional on the alighting passenger count \( a(n) \) at stop \( n \) and \( P_n(B_{1:(n-1)} \mid A_n) \), it is assumed that OD flows destined for stop \( n \) are multinomially distributed. This assumption is represented by:

\[
f(T(1:(n-1),n) \mid a(n), P_n(B_{1:(n-1)} \mid A_n)) \sim Multinomial(a(n), P_n(B_{1:(n-1)} \mid A_n))
\]  

(4.29)

where:

\[
T(1:(n-1),n) = (T(1,n), \ldots, T(n-1,n))
\]  

(4.30)

The conditional probability of Equation 4.27 depends on the probability OD flow matrix \( \alpha \), APC counts \( x \), and OD flows \( U^n \) between bus stops upstream of bus stop \( n \). OD flows \( U^n \) are random variables, which are not uniquely determined by APC counts. Therefore, the distribution of Equation 4.29 is conditional on the probability OD flow matrix \( \alpha \), APC counts \( x \), and OD flows \( U^n \) between bus stops upstream of bus stop \( n \), which is expressed as follows:
\[ f(T(1:(n-1),n) | U^n, x, \alpha) \sim \text{Multinomial}(a(n), P_n(B_{1(n-1)} | A_n)) \]  

(4.31)

Based on Equations 4.27 and 4.31, the expected OD flows between stop \( s \) and stop \( n \) (\( \forall s: s < n \)) conditional on the probability OD flow matrix \( \alpha \), APC counts \( x \), and OD flows \( U^n \) between bus stops upstream of bus stop \( n \) is given by:

\[
E(T(s,n) | U^n, x, \alpha) = a(n) \times \frac{\sum_{q=n}^{n-1} \frac{\alpha(s,n)}{\sum_{q=n}^{N} \alpha(s,q)} \times b(s) - \sum_{q=n+1}^{n-1} T(s,q)}{\sum_{q=n}^{n-1} \sum_{r=1}^{N} \alpha(r,q) \times b(r) - \sum_{q=r+1}^{n-1} T(r,q)}
\]

(4.32)

Conditional on the probability OD flow matrix \( \alpha \) and APC counts \( x \), the expected OD flow from stop \( s \) to stop \( n \) can be determined approximately by applying the Taylor series first order approximation and the law of total expectation – leading to the approximation \( E(g(x)) \approx g(E(x)) \) (Rice, 2006) – to Equation 4.32. That is:

\[
E(T(s,n) | x, \alpha) = E_{U^n|x,\alpha}(E(T(s,n) | U^n, x, \alpha))
\]

\[
\approx a(n) \times \frac{\sum_{q=n}^{n-1} \frac{\alpha(s,n)}{\sum_{q=n}^{N} \alpha(s,q)} \times b(s) - \sum_{q=n+1}^{n-1} E(T(s,q) | x, \alpha)}{\sum_{q=n}^{n-1} \sum_{r=1}^{N} \alpha(r,q) \times b(r) - \sum_{q=r+1}^{n-1} E(T(r,q) | x, \alpha)}
\]

\[
= a(n) \times \frac{\sum_{q=n}^{n-1} \frac{\alpha(s,n)}{\sum_{q=n}^{N} \alpha(s,q)} \times E(D(n | s))}{\sum_{r=1}^{n-1} \sum_{q=n}^{N} \alpha(r,q) \times E(D(n | r))}
\]

(4.33)
where \( E(D(n \mid r)) \) is defined in Equations 4.20. Equation 4.33 is the result shown in Equation 4.21 rearranged for convenient presentation.

In the E-step of the HEM algorithm, it is possible that the conditional expected OD flow from a boarding bus stop \( s \) to a given alighting stop \( n \), \( E(T(s, n) \mid x, \alpha) \) determined by Equation 4.20 exceeds the expected number of onboard passengers who boarded at stop \( s \) \( E(D(n \mid s)) \), which results in a negative conditional expected OD flow from stops \( s \) to \( n + 1 \) when applying the next recursive calculation. To avoid this situation, when for a given stop \( s = k \) the value of \( E(T(k, n) \mid x, \alpha) \) is calculated to be greater than \( E(D(n \mid k)) \), the value of \( E(T(k, n) \mid x, \alpha) \) is reset to be equal to that of \( E(D(n \mid k)) \). In this case, the conditional expected OD flow from a remaining bus stop \( s = 1, 2, \ldots, k - 1, k + 1, \ldots, n - 1 \) to stop \( n \) is the product of the remaining alighting count \( a(n) - E(D(n \mid k)) \) and the probability that an onboard passenger boarded at stop \( s \) given that she alights at stop \( n \), which is given by:

\[
E(T(s,n) \mid x,\alpha) = (a(n) - E(D(n \mid k))) \times \frac{E(D(n \mid k)) \cdot \alpha(s,n)}{\sum_{q=n}^{N} \alpha(s,q)} \times \sum_{r=1, r \neq k}^{n-1} \left( E(D(n \mid r)) \cdot \alpha(r,n) / \sum_{q=n}^{N} \alpha(r,q) \right)
\]

If, as a result, \( E(T(s, n) \mid x, \alpha) \) is found to be greater than \( E(D(n \mid s)) \) for additional boarding stops, the same correction and a similar recalculation to that of Equation 4.33 is applied.
4.3.3 Computation Complexity of the HEM Algorithm

The computation complexity quantifies the amount of resources (time or steps) necessary to execute the HEM algorithm. The number of iterations needed between the E- and M-steps is determined by the convergence criterion. This criterion is defined as follows: each probability value in the probability OD flow matrix being estimated changes in absolute value by less than a given threshold from one iteration to the next.

In one iteration, the worst computation case is that where passengers boarded and alighted at all bus stops on all bus trips as can be seen from Equation 4.30. In this worst case, for a bus route with $N_s$ bus stops, the number of feasible OD pairs is $\frac{1}{2} \times N_s \times (N_s - 1)$, and therefore, for $L$ bus trips with APC data, the number of times Equation 4.20 needs to be evaluated is on the order of $O(L \times N_s^2)$. That is, in one iteration, the number of times Equation 4.20 needs to be evaluated increases linearly with the number $L$ of bus trips with APC counts and increases by the square of the number of bus stops $N_s$.

The computation time in actuality could be much smaller than this upper bound since not all bus stops have boarding and alighting counts on a given bus trip. For a bus trip, if no passengers boarded at stop $r$, it is not necessary to evaluate Equation 4.30 for OD pairs originating from stop $r$. Similarly, for a bus trip, if no passengers alighted at stop $q$, it is not necessary to evaluate Equation 4.30 for OD pairs destining for stop $q$. 
Chapter 5  Numerical Study

5.1 Introduction

Chapter 4 developed the HEM algorithm to estimate the period-level probability OD matrix on a transit bus route. For comparison purposes, the CM algorithm was also developed in Chapter 4. This chapter evaluates the performance of the HEM algorithm based on simulated APC boarding and alighting data. Some other methods that are considered in the evaluation for comparison purposes are introduced in Section 5.2. The overall performance of various estimates could be assessed by different measures. These measures are introduced in Section 5.3.

In Section 5.4, the HEM algorithm is evaluated on a simple, illustrative short bus route and on an operational bus route. The comparison of the solutions produced by the EM and HEM algorithms is the main reason the hypothetical simple short bus route is developed. As already discussed, the EM algorithm can be applied only on a short bus route because it is computationally prohibitive to apply to realistic long bus routes. The numerical study on an operational bus route demonstrates that it is feasible to apply the HEM algorithm to a practical bus route.

Since both the CM and HEM algorithms are iterative procedures and require an initial probability OD flow matrix estimate to start the iteration, Section 5.5 investigated whether the final estimates of the CM and HEM algorithms depend on the starting values of the iterations. In Section 5.6, a variant of the HEM algorithm is proposed and
evaluated. In Sections 5.7 and 5.8, the sensitivity of the OD estimates to the structure of the underlying probability OD matrix and the distributional assumptions made in developing the formulation are evaluated, respectively.

The investigations in this chapter are based on APC boarding and alighting counts on an assumed given number of bus trips only. APC counts are assumed to be measurement error free. Survey OD data are not considered in the investigations. The effects of the number of bus trips with APC counts, OD survey data and measurement errors in APC counts are investigated empirically in Chapter 6 and numerically in Chapter 7.

Both the CM and HEM algorithms are iterative procedures. The algorithms stop when a convergence criterion is met. The criterion is defined in Chapter 4 as follows: each probability value in the probability OD flow matrix being estimated changes in absolute value by less than a given threshold from one iteration to the next. In this study, the threshold for both the CM and HEM algorithms is chosen as 0.01 / Ns, where Ns is the number of bus stops on the given bus route. This specification applies to all analysis presented in this and subsequent chapters.

5.2 OD Estimation Methods Considered in the Evaluation

In this study, the probability OD estimates produced by the HEM algorithm are compared with those produced by some other methods. These methods are introduced in the following.

The null base OD matrix represents the most naïve “estimates” (where the term “estimate” is used loosely in this context) regarding OD flows in the absence of any data
or information. The normalized null base OD matrix, where each cell contains the reciprocal of the total number of feasible cells, represents the non-informative (normalized) OD flows to which the estimated matrices are compared to in the evaluation process.

OD flow estimates produced by the IPF method are also considered in the evaluation. As already discussed, the IPF method is an easy-to-implement procedure for transit route-level passenger OD flow estimation that can combine information from an onboard survey and from APC counts, that has been shown to produce estimates that are similar to many other OD estimation methods, and whose performance has been assessed empirically (McCord et al., 2010; Mishalani et al., 2011). Results produced from the IPF method can be considered representative of what could readily be obtained in practice.

When boarding and alighting counts on multiple bus trips are available, the IPF method traditionally aggregates boarding and alighting counts by stop across bus trips, and then reconstruct an OD matrix from the aggregated boarding and alighting counts (Ben-Akiva et al., 1985). As a result, the IPF method relies on the aggregated boarding and alighting counts or the first moments of the trip-level boarding and alighting counts to estimate the probability OD matrix. Appendix A presents the reason behind this characteristic of the IPF method in mathematical terms. McCord et al. (2010) applied the IPF method slightly differently to estimate the probability OD matrix. They used the IPF method to reconstruct a volume OD matrix that is consistent with APC counts for each bus trip, and then aggregated the reconstructed OD matrices across bus trips. The aggregated OD matrix was normalized by the total volume to produce a point estimate of
the probability OD matrix for the period during which the bus trips took place. The procedure developed by McCord et al. (2010) could produce slightly better probability OD flow estimates than the traditional procedure when APC counts are aggregated first because OD flows for some OD pairs could be uniquely determined by boarding and alighting counts at bus trip level. While it is not mathematically clear that this variation in implementing the IPF method is strictly based on the first moments of the trip-level APC data, given the close similarity in the estimates the two approaches produce, it is believed that the latter approach is at least predominantly based on the first moments of the APC data.

More specifically, the procedure of determining the probability OD matrix at the period-level from the trip-level APC counts by stop using the IPF method in this study is as follows. First, for each bus trip for which APC counts are available, volume stop-to-stop OD flows are reconstructed by the IPF method from the trip-level APC counts. Second, the reconstructed trip-level volume OD flow matrices are summed to produce an aggregated volume OD matrix. Third, this aggregated volume OD flow matrix is normalized by dividing each cell entry by the total number of boarding passengers. The normalized OD flow matrix is considered to be the estimate of the probability OD flow matrix as produced by the IPF method.

As discussed before, the EM algorithm could be used to provide a solution to the proposed formulation. However, the EM algorithm is computationally prohibitive for applications to realistically long bus routes. The EM algorithm is considered only on an
illustrative short bus route in Section 5.4 to evaluate whether the HEM algorithm could provide a good approximate solution.

As discussed in Chapter 4, the CM algorithm consists of iteratively applying the IPF method to boarding and alighting counts on each bus trip and aggregating the trip-level OD matrices reconstructed by the IPF method across bus trips. In such a way, the CM algorithm approximates finds a mode of the joint posterior distribution of the probability OD matrix and the trip-level OD matrices. The CM algorithm considers the distribution of APC data albeit not as effectively as the EM algorithm, which is based on the marginal posterior distribution of the probability OD matrix. Given this and the fact that the HEM algorithm provides an approximate solution to that produced by the EM algorithm, it is also worthwhile to compare the HEM algorithm to the CM algorithm.

Finally, in the simulation, a volume OD flow matrix $T_l$ is generated for each bus trip $l$. If these generated matrices could be observed, they could be used to produce another estimate of the period-level probability OD flow matrix. Since trip-level OD flow matrices are assumed to be multinomially distributed conditional on the underlying probability OD flow matrix and the trip-level total demand, and volume OD flow matrices for different bus trips are assumed to be independent, the maximum likelihood estimate of the probability OD flow matrix, conditional on observing the generated trip level volume OD flow matrices, is given by:

$$\hat{\alpha}_{\text{gen}}(r,q) = \frac{\sum_{l=1}^{L} T_l(r,q)}{\sum_{l=1}^{L} \text{tot}_l}$$

(5.1)
These estimates are used in the simulation-based evaluation as well, representing the best that could be possibly arrived at through observation.

Since the evaluation in this chapter is based on APC counts only, the null base OD matrix is used in the IPF method, and the uniform prior distribution is used in the CM, HEM and EM algorithms to produce the probability OD flow estimates. As discussed in Chapter 3, the uniform prior is analogous to the null base in that both imply that passengers are equally likely to travel along any OD pair. Therefore, the null base in this study also refers to the uniform prior for the CM, HEM and EM algorithms.

5.3 Performance Measures

The performance of the various estimates could be assessed by different measures. The main measure used in this study is the squared Hellinger Distance, which is often used to measure the similarity between two probability distributions (Yang et al., 2000). This measure consists of the sum of the squared differences between the square root of the estimated probability OD flows and the square root of the underlying true probability OD flows:

\[
HD^2 = \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \left( \sqrt{\hat{\alpha}(r,q)} - \sqrt{\alpha(r,q)} \right)^2
\]  

(5.2)

where, \(\hat{\alpha}(r,q)\) is the probability OD flow from stop \(r\) to stop \(q\) “estimated” from one of the methods considered, and \(\alpha(r,q)\) is the underlying, true probability OD flow from stop \(r\) to stop \(q\). Naturally, small values of \(HD^2\) represent a higher degree of similarity between the two matrices in question.
Some other measures, such as the Sum of Squared Differences (SSD) and the Chi-square (Chi2) statistic, are also considered in this study. The SSD is defined as the sum of the squared differences between the estimated probability OD flows and the underlying true probability OD flows:

$$SSD = \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} (\hat{\alpha}(r,q) - \alpha(r,q))^2$$

(5.3)

The Chi2 statistic is defined as the sum of the ratios of the square difference between the estimated probability OD flows and the underlying true probability OD flows to the underlying true probability OD flows:

$$Chi2 = \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \frac{(\hat{\alpha}(r,q) - \alpha(r,q))^2}{\alpha(r,q)}$$

(5.4)

The measures presented above do not indicate the performance of the estimates in relative terms. Therefore, a measure of relative performance $RP$ (McCord et al., 2010) is defined to quantify the relative performance of the estimated matrix with respect to the null base OD matrix. The $RP$ metric quantifies the improvement of the estimated OD flow matrix from the null base OD matrix as a proportion of the measure of dissimilarity between the null and the underlying true OD flows, which is given by:

$$RP = \frac{P_{null} - P}{P_{null}}$$

(5.5)

where $P_{null}$ is the performance measure ($HD^2$, SSD or Chi2) for the null base as the estimate and $P$ is the performance measure for the various estimates. The $RP$ metric would be zero if the null base OD flow matrix is used as an “estimate”, and it would be one if the underlying true probability OD matrix is used as an estimate. Therefore, low
RP values indicate little improvement in performance, and RP values close to one indicate great improvement in performance. (RP values could be less than zero if the estimated OD matrix was worse than the null matrix.)

5.4 Numerical Evaluation of OD Estimation Methods

5.4.1 Experiment Overview

This section evaluates the CM and HEM algorithms on a simple, illustrative bus route and on an operational bus route. The distribution of the underlying, “true” bus trip-level total demand and the probability OD flow matrix for the illustrative bus route are assumed arbitrarily, while the distribution of the trip-level total demand and the underlying probability OD flow matrix for the operational bus route are constructed based on OD flow survey data collected on the operational route of interest. The trip-level volume OD flow matrices and the trip-level APC counts are simulated based on the model structure presented in Figure 3.1 of Chapter 3 and the assumed underlying true probability OD flow matrix and the assumed trip-level total demand distribution.

The illustrative bus route is short enough that the complete enumeration of the OD flow matrices satisfying Equations 4.2a and 4.2b required to obtain the solution by applying the EM algorithm is possible. Therefore, the estimated solutions produced by the HEM algorithm are compared to the solutions produced by the EM algorithm, and both sets of solutions are also compared to the assumed underlying true probability OD flows. The operational bus route, on the other hand, is long enough such that executing the complete enumeration required by the EM algorithm is not feasible. Therefore, the
HEM estimates are only compared to the assumed underlying probability OD flows when considering the performance of the HEM algorithm for the operational bus route.

5.4.2 Numerical Evaluation on a Simple Bus Route

The simple bus route is used to illustrate the comparisons among the different methods and, especially, to investigate the ability of the proposed HEM algorithm to produce results that approximate those produced by the EM algorithm.

There are four bus stops on the illustrative bus route beginning at stop 1 and terminating at stop 4. The trip-level total demand is assumed to be negative binomially $NB(r,p)$ distributed. The Negative Binomial distribution is traditionally used to model the number of failures occurring in a series of independent Bernoulli trials before a specified number of successes $r$ is reached, where the probability of a success in a single trial is $p$ (Rice, 2006). (The negative binomial distribution is also known as the Gamma-Poisson (mixture) distribution (Gelman et al., 2004). That is, the $NB (r, p)$ could be seen as a Poisson ($\lambda$) distribution, in which $\lambda$ is distributed according to Gamma ($r, p / (1 – p)$).)

The Negative Binomial distribution is used to model the case where the variance of the trip-level total demand is larger than the mean of the trip-level total demand. This case is desirable in this analysis because, as discussed in Chapter 2, the variance of the trip-level total demand is found to be much larger than the mean of the trip-level total demand in the empirical data. In this analysis, the number of successes $r$ is set to be 30 and the probability of success $p$ is set to be 0.5. The resulting mean and variance of the trip-level total demand is 30 and 60, respectively.
Two probability OD matrices, OD1 and OD2, are considered to be the underlying true probability OD matrix for generating the trip-level APC boarding and alighting counts. These two probability OD matrices are shown in Tables 5.1 and 5.2, respectively. Notice that OD1 and OD2 have different cell values but the same marginal values. Therefore, the final probability OD flow estimates would be similar under OD1 and OD2 for methods based on the first moments of APC counts. In addition, notice that OD1 and OD2 are almost the same matrices, but labeled differently. For example, OD flow from stop 1 to stop 3 is 0.09 under OD1 and OD flow from stop 1 to stop 4 is also 0.09 under OD2. As a result, the performance measures would not depend on the two underlying true probability OD matrices.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Boarding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.07</td>
<td>0.09</td>
<td>0.30</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.35</td>
<td>0.12</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Alighting</strong></td>
<td>0</td>
<td>0.07</td>
<td>0.44</td>
<td>0.49</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.1 Probability OD Matrix 1 (OD1) on the Simple, Illustrative Bus Route

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Boarding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.07</td>
<td>0.30</td>
<td>0.09</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.33</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Alighting</strong></td>
<td>0</td>
<td>0.07</td>
<td>0.44</td>
<td>0.49</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.2 Probability OD Matrix 2 (OD2) on the Simple, Illustrative Bus Route
Based on an assumed probability OD matrix and the assumed distribution of the trip-level total demand, the volume OD flow matrix for each of 100 assumed bus trips is generated and the resulting trip-level APC boarding and alighting counts are derived. The period-level probability OD estimates are produced by the various methods and the performance measures are calculated for each of the period-level probability OD estimates. The null base OD matrix is used in the IPF, CM, HEM and EM methods when producing the probability OD flow estimates. This process is repeated 100 times, once for each generated set of APC data on 100 bus trips, for each assumed probability OD matrix to produce the Empirical Cumulative Density Functions (ECDFs) of the IPF estimate, the CM estimate, the HEM estimate, the EM estimate, and the estimate based on the generated “true” trip-level OD flows using Equation 5.1 (OD_{gen}) for each cell of the OD flow matrix. The resulting ECDFs are presented in Figures 5.1 and 5.2. The underlying true probability OD flows of Tables 5.1 and 5.2 are indicated by vertical lines in Figures 5.1 and 5.2.
Figure 5.1 ECDFs of the Probability OD Flow Estimates on the Illustrative Bus Route under OD1 in Table 5.1
Figure 5.2 ECDFs of the Probability OD Flow Estimates on the Illustrative Bus Route under OD2 in Table 5.2

In both Figures 5.1 and 5.2, the ECDFs of the probability OD flow estimates for OD pairs 1-2 and 3-4 are identical across the five methods since the trip-level OD flows for these two OD pairs are uniquely determined by APC boarding and alighting counts (i.e., passengers alighting at stop 2 must have boarded at stop 1, and passengers boarding at stop 3 must alight at stop 4). For the other four OD pairs, the OD\textsuperscript{gen} estimates are the best; namely the distributions of the probability OD flow estimates derived directly from the generated volume OD flows are centered on the true values and the variances are small. This result is expected, since the true trip-level volume OD flow matrices contain
richer information about the probability OD flows than do the marginal values of the true trip-level OD matrices as captured by the trip-level APC counts.

In addition, the ECDFs of the EM and HEM estimates are seen to be similar in Figures 5.1 and 5.2. This result indicates that the HEM algorithm is able to provide good approximations to the solutions produced by the EM algorithm, which produces the solution to the estimation problem. The ECDFs of the probability OD flows estimated by the HEM and EM algorithms are almost centered on the true values, but the variances are larger than those of the OD_{gen} estimates. The larger variances of the HEM and EM estimates, compared to the variance of the OD_{gen} estimates, reflect the additional uncertainty of the probability OD flow estimates that result from the indirect observations of the trip-level OD matrices in the form of the trip-level APC boarding and alighting counts.

Moreover, Figures 5.1 and 5.2 show that the IPF estimates have large biases and small variances. More specifically, the true probability OD flows are out of the range of the ECDFs of the IPF estimates. The IPF method as implemented for the most part if not entirely uses information in the first moments of APC counts. The first moments of APC counts do not provide sufficient information on the probability OD flow matrix. Instead, the first moments of APC counts only provide information on the marginal values of the probability OD flow matrix in the IPF method. The structure of the estimated probability OD matrix is predominantly determined by the null base OD matrix, which results in the small variations among the estimates. In addition, the fact that the null base OD flow
matrix does not accurately reflect the actual passenger travel patterns results in the large biases in the final IPF OD estimates.

Furthermore, the true probability OD flows are also out of the range of the ECDFs of the CM estimates. These results indicate that the CM estimates are influenced by the null base, however, the influence of the null base on the CM estimates is smaller than that on the IPF method.

Lastly, it would be interesting to examine whether the IPF, CM or HEM estimates can distinguish the difference in the underlying probability OD matrices. Recall that the two underlying probability OD matrices have the same marginal values but different cell values. As can be seen, for any given OD pair, the IPF estimates in Figures 5.1 and 5.2 are similar, indicating that the IPF method cannot distinguish between the underlying probability OD matrices. This result is consistent with the fact that the IPF method is based on the first moments of the trip-level APC boarding and alighting counts. By contrast, the EM and HEM algorithms take advantage of the distribution of the trip-level APC boarding and alighting counts and they successfully distinguish between the underlying probability OD matrices. As can be seen, the EM and HEM estimates are close to the underlying truth in Figures 5.1 and 5.2 for each OD pair. The CM estimates are different in Figures 5.1 and 5.2 and are better than the IPF estimates. However, the CM algorithm fails to infer the underlying probability OD matrices as well as the EM and HEM algorithms.
Figures 5.1 and 5.2 compare the probability OD flow estimates to the underlying true values for individual OD pairs. The overall performance of various estimates could be summarized by the performance and relative performance measures discussed in Section 5.3. Figures 5.3 and 5.4 present the ECDFs of the $HD^2$, SSD and Chi2 measures and their corresponding RP measures for various estimates. The underlying true probability OD flows for the results in Figures 5.3 and 5.4 are the ones presented in Tables 5.1 and 5.2, respectively. As can be seen, the EM algorithm performs the best among the methods that are based on APC boarding and alighting counts (i.e., the IPF, CM, HEM and EM methods). However, as already discussed, the enumeration required in the EM algorithm is computationally expensive and therefore it is prohibitive to apply to realistic bus routes. The HEM algorithm is computationally efficient such that it can be applied on realistic long bus routes and, according to the results obtained with the illustrative bus route, can provide similar OD estimates to those produced by the EM algorithm. The CM algorithm, as seen in Figures 5.3 and 5.4, is worse than the HEM algorithm but better than the IPF method.
Figure 5.3 ECDFs of the Performance and RP Measures of OD Estimation Methods on the Illustrative Bus Route under OD1 in Table 5.1
5.4.3 Numerical Evaluation on an Operational Bus Route

It was shown numerically on a illustrative short bus route in the previous section that the HEM algorithm produces similar probability OD flow estimates to those produced by the EM algorithm and that the EM and HEM algorithms perform better than the IPF and CM algorithms. This section evaluates the accuracy of various estimates on an operational bus route by comparing the estimates to the assumed underlying probability OD flows. Because the EM algorithm is computationally expensive, it is
infeasible to execute the EM algorithm on an operational bus route to estimate the probability OD matrix.

The performance of various methods is assessed on an operational bus route based on a probability OD matrix that is constructed from survey OD data. The data were collected through The Ohio State University’s Campus Transit Lab (CTL). The CTL is based on the Campus Area Bus Service (CABS), a transit service owned and operated by The Ohio State University (OSU) that serves approximately four million passengers annually on seven routes in the vicinity of the OSU campus.

This study focuses on the Campus Loop South (CLS) bus route, which is 8.38 km long and contains 20 bus stops. The CLS route follows a loop pattern transporting passengers between the west campus park-and-ride facility, where four stops are located, and the main campus of OSU, as presented in Figure 5.5. The park-and-ride stops are aggregated into a “pseudo” stop considered to be the terminus of the CLS route where bus trips begin and end. Furthermore, the beginning and ending terminus are treated as two separate stops. The OD pair starting and ending at the terminus is considered infeasible since passengers rarely travel from the west campus park-and-ride facility back to the same facility on the same bus trip. Given the stop aggregation at the park-and-ride facility, CLS bus trips are considered to serve 18 stops and 152 feasible OD pairs for the purpose of OD flow estimation.
Large-scale onboard surveys were carried out on the CLS bus route between 8 and 10 a.m. on weekdays of three academic quarters in 2009. Passenger travel patterns are assumed to be homogeneous in the given time-of-day period. Table 5.3 summarizes the numbers of bus trips and passengers for which the field OD data were collected. To collect the empirical field OD data, two data collectors rode CLS buses, with one person stationed near the front door and one stationed near the rear door. They distributed cards indicating the boarding stop to passengers as they boarded the bus and collected the cards as the passengers alighted. By filing the cards according to the alighting stop and bus trip, the cards could be used to determine both the OD flows and the corresponding boarding and alighting volumes for the various bus trips. More than 95% of the passengers travelling on the targeted trips were successfully sampled in this way.
Trip-level OD flow matrices are aggregated for each OD pair across bus trips, and then the aggregated OD flow matrix is normalized by the total boarding passengers. This normalized OD flow matrix is used as the underlying, true probability OD flow matrix in this numerical study. In addition, the sample mean and sample variance of the trip-level total demand is 62.3 and 566.6, respectively. The Negative Binomial distribution fits the trip-level total demand data well (namely, the p-value of the Chi-Square goodness-of-fit test is 0.31). Assuming a Negative Binomial distribution for the trip-level total demand is used to generate the trip-level total demand in the simulation-based analysis, the Maximum Likelihood Estimate of the specified number of successes \( r \) and the probability of a success \( p \) is 7.24 and 0.1, respectively. This estimation is based on the total number of boarding counts on the 36 surveyed bus trips. Based on the assumed probability OD matrix and the distribution of the trip-level total demand, OD matrices and the resulting true APC counts can be generated based on the model structure presented in Figure 3.1 in Chapter 3.

The probability OD flow estimates by various methods are produced from the simulated trip-level APC counts on 500 generated bus trips. The null base OD matrix is used in the IPF, CM and HEM methods when producing the probability OD flow estimates. The computational time for producing the period-level probability OD flow
estimates by the CM or HEM algorithm is less than 2 seconds on a 2.80 GHz Intel CPU, demonstrating that it is feasible to apply the CM and HEM algorithms on an operational bus route.

Figure 5.6 compares the estimates of the probability OD flows by the IPF, CM and HEM methods to the underlying true values based on one set of 500 generated bus trips (falling within the period the underlying probability OD matrix corresponds to, namely 8-10 am). The horizontal and vertical values of one data point in Figure 5.6 represent the assumed underlying true and estimated probability OD flows (by the IPF, CM or HEM method) for one feasible OD pair. Data points that are close to the diagonal line indicate that the estimated values are close to the assumed underlying true values for the corresponding OD pairs. By visual inspection, the IPF method is worse than the CM and HEM algorithms since the IPF method tends to overestimate the low probability OD flows and underestimate the high probability OD flows. In addition, the HEM algorithm is better than the CM algorithm since the HEM estimates are closer to the underlying true values for most OD pairs.
Figure 5.6 Comparisons of the IPF, CM and HEM Probability OD Flow Estimates to the Assumed True Probability OD Flows on the CLS Bus Route

Thus far, only one set of 500 generated bus trips has been considered. As in the case of the analysis for the simple illustrative route, the Empirical Cumulative Distribution Functions (ECDFs) of the IPF estimate, the CM estimate, the HEM estimate, and the estimate based on the generated “true” trip-level OD flows (OD\text{gen}) for each OD pair can be derived from the estimates produced from many simulation trials. Therefore, different trip-level OD matrices and the resulting APC counts are generated. A total of 100 simulation trials are carried out, where in each trial APC boarding and alighting counts for 500 bus trips are generated from a corresponding generated trip-level OD matrix. The null base OD matrix is used in the IPF, CM and HEM methods when
producing the probability OD flow estimates.

There are 152 feasible OD pairs on the CLS bus route. The ECDFs of the probability OD flow estimates for a sample of 16 OD pairs are presented in Figure 5.7 for illustration purposes. The subplots in Figure 5.7 are sorted in descending order by the magnitude of the underlying true probability OD flows from the left to the right and from the upper to the bottom of the figure. In Figure 5.7, the underlying true probability OD flows are presented in the bracket of the subplot title and are also plotted as vertical lines in the subplots. Since the underlying true probability OD matrix is constructed based on the sample survey OD matrix, the probability OD flows for some feasible OD pairs happen to be zero (the last row of Figure 5.7 depicts such cases).

Figure 5.7 ECDFs of the Probability OD Flow Estimates for Individual OD Pairs on the CLS Bus Route
As expected, the performance of the OD\textsuperscript{gen} estimates is the best; that is, the distributions of the OD\textsuperscript{gen} estimates are centered on the assumed underlying true probability OD flows and the variances are small. In addition, the assumed underlying true probability OD flows are within the range of the ECDFs of the HEM estimates for OD pairs whose probability OD flows are not zero (see Rows 1, 2 and 3 in Figure 5.7). For OD pairs whose probability OD flows are zero, the HEM algorithm is more likely to produce very small probability OD flow estimates than the CM and IPF methods. Moreover, the variances of the HEM estimates are larger than those of the OD\textsuperscript{gen} estimates, reflecting the additional uncertainty of the probability OD flow estimates that results from the indirect observations of the trip-level OD matrices in the form of the APC boarding and alighting counts. Furthermore, the ECDFs of the CM estimates are farther away from the underlying true values than those of the HEM estimates. Lastly, the underlying true probability OD flows are beyond the range of the ECDFs of the probability OD flows estimated by the IPF method. The results are consistent with those of the previous section for a simple, illustrative bus route.

The overall performance of various estimates could be summarized by the performance and relative performance measures, which is convenient given the large number of feasible OD pairs. Figures 5.8 presents the ECDFs of the $HD^2$, $SSD$ and $Chi2$ measures and their corresponding $RP$ measures for various estimates. As can be seen, The HEM algorithm performs the best among the methods that are based on APC counts (i.e., the IPF, CM and HEM methods). The CM algorithm is better than the IPF method,
but is worse than the HEM algorithm. The results are consistent with those of the previous section for a simple, illustrative bus route.

Figure 5.8 ECDFs of the Performance and $RP$ Measures of OD Estimation Methods on the CLS Bus Route

Since the distribution of Passenger Distance Travelled (PDT) can be easily derived from the probability OD matrix and the distances between OD pairs (McCord et al., 2010), the performance of various methods could also be evaluated in terms of the distribution of PDT. Doing so allows for evaluating the overall performance rather than comparing estimates at the individual OD pair level. Figure 5.9 presents the ECDFs of the PDTs derived from various probability OD flow estimates, which are produced from
the simulated trip-level APC counts on one set of 500 bus trips (the same set used in producing Figure 5.6). The null base OD matrix is used in the IPF, CM and HEM methods when producing the probability OD flow estimates. As can be seen, the IPF method overestimates both the short and long travel trips (e.g., travel distance less than 0.8 km and larger than 4.8 km). This result is consistent with the discussion in Chapter 2, in which it is pointed out that the IPF method using the null base tends to overestimate both the short and long travel trips. By contrast, the ECDF of the PDT derived from the HEM estimates is very close to the one derived from the underlying true probability OD matrix. The ECDF of the PDT derived from the CM estimates is between the ones derived from the IPF estimates and the HEM estimates.
5.5 Investigating the Convergence of the CM and HEM Algorithms

Both the CM and HEM algorithms are iterative procedures and require an initial probability OD flow matrix estimate to start the iteration. The default initial matrix adopted in this study is the base OD matrix (the null base or a base constructed from survey OD data). However, theoretically, any probability OD matrix is eligible to serve as the initial matrix (in the case of the IPF method, however, the only possible initial matrix is the base OD matrix). This section investigates whether the CM or HEM algorithm is able to converge to the same or similar estimate of the probability OD matrix with overdispersed starting matrices. The inability to converge to similar estimates may
indicate that the objective function has multiple local optima rather than a unique mode or that the algorithm is not well designed.

This numerical investigation is carried out using the structure of the CLS bus route. OD matrices and the resulting APC boarding and alighting counts on 500 assumed bus trips are generated based on the assumed underlying probability OD matrix and the assumed distribution of the trip-level total demand discussed in Section 5.4. The CM and HEM algorithms are applied to the generated trip-level APC counts to estimate the period-level probability OD flow matrix. The null base is used for OD estimation in the CM and HEM algorithms in this analysis. Therefore, the CM and HEM algorithms start from a probability OD flow matrix generated randomly from the uniform Dirichlet distribution, as presented in Equation 3.5 where the parameter $\mu(r, q) = 1$ for all feasible OD pairs. Doing so produces a starting matrix that is equally likely to fall anywhere in the feasible space. The determined probability OD matrix in each iteration of each of the two algorithms is recorded. Based on the same simulated trip-level APC counts, the OD estimation is repeated 200 times, where with each replication a new starting probability OD matrix is generated for each of the two algorithms.

Since the dimension of the probability OD matrix is large, the $HD^2$, SSD and $Chi^2$ measures are first used to summarize the difference between the determined probability OD matrix and the underlying true probability OD matrix in each iteration of each of the two algorithms. Figure 5.10 visualizes the measures at each iteration of the CM and HEM algorithms for 200 replications. For comparison, the measures of the final IPF estimates are also presented in Figure 5.10. The different measures are indicated in the titles and
the labels of the y axis of the subplots. Since the patterns based on different measures are similar, the following discussion focuses on the $HD^2$ measure.

Figure 5.10 Performance at Each Iteration of the CM and HEM Algorithms

As can be seen in Figure 5.10, after about 25 iterations, the $HD^2$ values of all CM and HEM replications are smaller than that of the IPF method. The $HD^2$ values decrease monotonically and eventually all replications converge for both the CM and HEM algorithms. These results demonstrate that the overall performance of the estimates arrived at by the CM and HEM algorithms does not depend on the starting matrix initiating the iterations. Figure 5.10 also shows that the performance of the HEM algorithm is superior to that of the CM algorithm regardless of the starting matrix.
Figure 5.10 summarizes the convergence of the CM and HEM algorithms in terms of the performance of the whole OD matrix. This does not imply that the same probability OD flow matrix is arrived at in each replication. Given the high dimension of the OD matrix, the convergence of the CM and HEM algorithms may vary across OD pairs. Figures 5.11 and 5.12 visualize the determined probability OD flow for individual OD pair at each iteration of the CM and HEM algorithms, respectively. There are 152 feasible OD pairs on the CLS bus route. Of those, 16 OD pairs are presented in Figures 5.11 and 5.12 for illustration. The subplots in Figures 5.11 and 5.12 are sorted in descending order by the magnitude of the underlying probability OD flows from the left to the right and from the top to the bottom of the figure. For comparison purposes, the underlying probability OD flows and the probability OD flows estimated by the IPF method are also shown. The values of the underlying probability OD flows are also presented in parenthesis in the subplot title.
Figure 5.11 Convergence of the CM Algorithm for Individual OD Pairs
As can be seen in Figures 5.11 and 5.12, in general, it seems that all replications converge to the same or similar estimates for both the CM and HEM algorithms and that the final estimates are closer to the underlying true values than the estimates produced by the IPF method. Moreover, in the case of relatively large probability OD flows (i.e., 1st two rows in Figures 5.11 and 5.12), the replications of the HEM algorithm converges to estimates superior to those of the CM algorithm regardless of the starting matrix. However, for OD pairs with zero or very low probability OD flows (i.e., subplots in the third and fourth rows), it is hard to distinguish the replications from each other. Therefore, the results in Figures 5.11 and 5.12 are reproduced by plotting the logarithm.
of the determined probabilities for the 16 OD pairs, as shown in Figures 5.13 and 5.14, respectively.

Figure 5.13 Convergence of the CM Algorithm for Individual OD Pairs in the Scale of Logarithm
As can be seen in Figures 5.13 and 5.14, for OD pairs with relatively high probability OD flows (i.e., the first two rows), all replications converge to the same or similar values for both the CM and HEM algorithms. However, the convergence may be poor for some OD pairs with zero or very low probability OD flows (for example, see the subplot in the third row and the third column for OD pair from stop 3 to stop 14 in Figures 5.13 and 5.14).

The above investigation shows that the overall performance of the CM and HEM algorithms in terms of the quality of the arrived at estimates does not depend on the matrix initiating the iterations and that the HEM algorithm outperforms the IPF and CM.
methods regardless of the starting matrix. However, the convergence of the CM and HEM algorithms depends on the values of the underlying true probability OD flows. For OD pairs with relatively high probability OD flows, the CM and HEM algorithms converge, while for OD pairs with very low probability OD flows, the CM and HEM algorithms may not converge. However, the CM and HEM algorithms still produce very low estimates for these OD pairs. The inability to converge for OD pairs with very low probability OD flows may not be an issue if the transit applications are more concerned with OD pairs with high probability OD flows than with OD pairs with very low probability OD flows.

5.6 A Variant of the HEM Algorithm

In the E-step of the HEM algorithm, the conditional expected OD flows are determined recursively from the first alighting stop on the route and continuing downstream until the last alighting stop. A variant is to determine the conditional expected OD flows recursively from the last boarding stop on a bus route and continuing upstream until the first boarding stop. This section investigates numerically whether both implementations of the HEM algorithm produce similar probability OD estimates. The numerical investigation is carried out using the structure of the CLS bus route. Estimation is based on simulated APC counts on 500 bus trips and the null base.

Figure 5.15 compares the probability OD flows estimated by the two variations of the HEM algorithm. The sequence orders in the E-step are different in the two HEM algorithms. The original sequence represents the HEM algorithm in which the E-step determines the conditional expected OD flows from the first alighting stop on the route.
and continuing downstream until the last alighting stop. The reverse sequence represents the HEM algorithm in which the E-step determines the conditional expected OD flows recursively from the last boarding stop on a bus route and continuing upstream until the first boarding stop. As can be seen in Figure 5.15 where the estimates for all feasible OD pairs are considered, the two variants of the HEM algorithm result in similar probability OD estimates. The maximum absolute difference between the probability OD estimates produced by the two variants of the HEM algorithm is less than 0.005.

Figure 5.15 Comparisons of Probability OD Flows Estimated by the HEM Algorithm Implemented based on Two Different Stop Sequence Orders
The overall performance of the two variants of the HEM algorithm could be summarized by the performance measures and the corresponding $RP$ measures as presented in Table 5.4. For comparison, the performance measures and the corresponding $RP$ measures of the IPF method are also presented. As can be seen, the performance of the two variants of the HEM algorithm is similar and is much better than that of the IPF method.

<table>
<thead>
<tr>
<th></th>
<th>IPF</th>
<th>HEM</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original sequence</td>
<td>Reverse sequence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$HD^2$</td>
<td>0.1560</td>
<td>0.0256</td>
<td>0.0238</td>
<td></td>
</tr>
<tr>
<td>$HD^2$-based $RP$</td>
<td>0.732</td>
<td>0.956</td>
<td>0.959</td>
<td></td>
</tr>
<tr>
<td>$SSD$</td>
<td>0.0048</td>
<td>3.71e-4</td>
<td>3.04e-4</td>
<td></td>
</tr>
<tr>
<td>$SSD$-based $RP$</td>
<td>0.764</td>
<td>0.982</td>
<td>0.985</td>
<td></td>
</tr>
<tr>
<td>$Chi2$</td>
<td>4.3219</td>
<td>0.1073</td>
<td>0.1040</td>
<td></td>
</tr>
<tr>
<td>$Chi2$-based</td>
<td>0.758</td>
<td>0.994</td>
<td>0.994</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4 Performance Measures and the Corresponding $RP$ Measures of OD Estimation Methods based on Two Sequence Orders

The results presented above show that both variants of the HEM algorithm result in similar estimates and that both variants outperform the IPF method. Therefore, either variant of the HEM algorithm could be used as an approximation of the EM algorithm. In this study, the original HEM algorithm in which the E-step determines the conditional expected OD flows from the first alighting stop on the route and continuing downstream until the last alighting stop is used.
5.7 Sensitivity of OD Estimates to the Underlying OD Structures

As discussed in Chapter 2, the underlying assumption of the IPF method using the null base is that the probability that an onboard passenger alights at a given bus stop is the same, regardless of the bus stop at which he or she boarded. If the underlying true OD matrix reflects this assumption, the IPF method using the null base would produce good OD flow estimates. Otherwise, the IPF method using the null base may not perform well. The CM and HEM algorithms do not make this assumption. However, as presented in Section 5.4.3, the CM estimates are heavily influenced by the null base OD matrix.

The property of the IPF method using the null base motivates the investigation of the sensitivity of OD estimates to the underlying OD structures. Three probability OD matrices are constructed to represent three different OD structures in this investigation. These three probability OD matrices have the same marginal values but different cell values. The first probability OD matrix is constructed from the survey OD data collected on the CLS bus route and is referred to as OD-survey (this matrix is the same as the one used to represent the underlying probability OD matrix in Section 5.4.3). The second probability OD matrix is constructed by applying the IPF method using the null base to the marginal values of the OD-survey and is referred to as OD-equality-of-alighting. Recall that the IPF method using the null base assumes that all onboard passengers are equally likely to alight at a given bus trip regardless of the bus stops they boarded. The OD-EOA satisfies the assumption of the IPF method using the null base and, therefore, it is expected that the IPF method using the null base would produce good OD flow estimates from the trip-level APC boarding and alighting counts generated from OD-
The third probability OD matrix is referred to as OD-extreme and is constructed such that it has the same marginal values as the OD-survey and is more different from the OD-EOA than is from the OD-survey as is described in what follows.

The difference between a probability OD matrix and the OD-EOA is quantified by the $\text{Chi}^2$ distance between the given probability OD matrix and the OD-EOA, which is given by:

$$
\text{Chi}^2 = \sum_{i=1}^{N_s-1} \sum_{j=i+1}^{N_s} \frac{(\alpha(i, j) - \alpha^0(i, j))^2}{\alpha^0(i, j)}
$$

(5.6)

where $\alpha(i, j)$ is the probability OD flow from stop $i$ to stop $j$ for the OD-survey or the OD-extreme, and $\alpha^0(i, j)$ is the probability OD flow from stop $i$ to stop $j$ for the OD-EOA. To construct OD-extreme, the third probability OD matrix, numerous probability OD matrices are generated randomly such that they have the same marginal values as those of OD-survey). The $\text{Chi}^2$ distance is calculated for each of these randomly generated matrices (representing the distance from the OD-EOA matrix). The first generated probability OD matrix whose $\text{Chi}^2$ distance is about twice as large as the $\text{Chi}^2$ distance of OD-survey (representing the distance between OD-survey and OD-EOA) is selected as the OD-extreme matrix. In this analysis, the $\text{Chi}^2$ distances for OD-survey and OD-extreme are 0.49 and 0.96, respectively.

The performance of the various estimates (i.e., the IPF, CM, HEM and $\text{OD}^{\text{gen}}$ estimates) is assessed under the three probability OD matrices. Given an assumed probability OD matrix (OD-EOA, OD-survey, and OD-extreme), a volume OD flow matrix is generated from the underlying true probability OD flow matrix and the
distribution of the trip-level total demand for each of 200 bus trips, the corresponding trip-level boarding and alighting counts are derived, the period-level probability OD flow estimates are produced from the various methods, and the performance measures are calculated for each of the period-level probability OD flow estimates. The null base OD matrix is used in the IPF, CM and HEM methods when producing the probability OD flow estimates. This process is replicated 100 times, where with each replication different volume OD flow matrices are generated for the different bus trips. The entire process is then repeated for each assumed underlying true probability OD matrix.

The ECDFs of the performance measures and the corresponding $RP$ measures are presented in Figures 5.16 and 5.17, respectively. The results for the subplots in the same row are based on the same specific underlying probability OD matrix, which is indicated in the title of the subplot. The results for the subplots in the same column are based on the same specific performance measure, which is indicted in the label of the x axis. Since the patterns based on different measures are similar and the patterns based on the performance measure and the corresponding $RP$ measure are similar, the following discussion focuses on the $HD^2$-based $RP$ measure in Figure 5.17.
Figure 5.16 ECDFs of Performance Measures under Three Probability OD Matrices
As expected, under OD-EOA, the performance of the IPF estimates is as good as (even slightly better than) that of the OD$^{gen}$ estimates since the OD-EOA satisfies the assumption of the IPF method using the null base OD matrix. However, the performance of the IPF estimates deteriorates considerably under OD-survey or OD-extreme. In addition, based on the $HD^2$-based $RP$ measure, the performance of the IPF estimates is worse under OD-extreme than under OD-survey. For example, the average $HD^2$-based $RP$ measure of the IPF estimates is 0.998, 0.726 and 0.688 under OD-EOA, OD-survey and OD-extreme, respectively.
In addition, as can be seen in Figure 5.17, The CM algorithm also performs better under OD-EOA than under OD-survey and OD-extreme. However, the performance of the CM algorithm is less sensitive to the OD structure than the IPF method. For example, the average $HD^2$-based $RP$ measure of the CM estimates is 0.957, 0.896 and 0.862 under OD-EOA, OD-survey and OD-extreme, respectively. This result is not surprising given that the CM method does capture information in the distribution of the APC counts while the IPF method does not.

Moreover, the HEM algorithm is robust with respect to the OD structure. The performance of the HEM estimates is similar under the three different underlying true probability OD matrices. For example, the average $HD^2$-based $RP$ measure of the HM estimates is 0.918, 0.928, and 0.892 under OD-EOA, OD-survey and OD-extreme, respectively.

Lastly, Figure 5.17 shows that the performance of the IPF method is better than that of the CM and HEM algorithms under OD-EOA, whereas the performance of the HEM algorithm is better than that of the IPF and CM methods under OD-survey and OD-extreme. However, the advantage of the IPF method over the HEM algorithm under OD-EOA is much smaller than the advantage of the HEM algorithm over the IPF method under OD-survey and OD-extreme. For example, the average $HD^2$-based $RP$ measure of the HEM estimates is lower than that of the IPF estimates by 0.07 under OD-EOA, whereas the average $RP$ measure of the HEM estimates is higher than that of the IPF estimates by 0.202 or 0.204 under OD-survey or OD-extreme, respectively. Based on the
investigation in this section, the HEM algorithm would tend to be preferred over the IPF and CM methods since the underlying OD structure is unknown in practice.

5.8 Sensitivity of OD estimates to Distributional Assumptions

The simulation results presented in the previous sections compared the HEM, CM and IPF methods under different scenarios and conditions. The comparison results, of course, depend on the distributional assumptions of the volume trip-level OD flows. This section investigates the sensitivity of the OD estimates to the distributional assumption of the trip-level OD flows, based on which the trip-level APC boarding and alighting counts are generated in the simulation. Recall that the distributional assumption behind the volume trip-level OD flows in the developed formulation is the same assumption adopted in generating these flows in the various simulation exercises. In this analysis, the performance of various OD estimates is evaluated based on the trip-level APC boarding and alighting counts generated from a distribution that is different from the one assumed in the developed approach.

As presented in Chapter 3, the volume trip-level OD flows are assumed to be multinomially distributed, conditional on the probability OD flows and the trip-level total demand. In this section, it is assumed that the generated trip-level OD flows are independently Negative Binomial distributed while the Multinomial distribution still holds for the developed approach. In the Negative Binomial distribution, the variance increases with the mean of the trip-level OD flow and the variance is greater than the mean of the trip-level OD flow. These two properties are desirable because they are suggested in the empirical data. Figure 5.18 depicts the relationship between the sample
variance and the sample mean of the trip-level OD flow across bus trips for individual OD pairs on the CLS bus route where data on a total of 36 bus trips are considered. The horizontal and vertical values of one data point are the sample mean and the sample variance of the trip-level OD flow for one OD pair across the 36 bus trips on which the survey OD data were collected. As can be seen in Figure 5.18, generally the variance increases as the mean of the trip-level OD flow increases and the variance is larger than the mean of the trip-level OD flow for most OD pairs, especially for the ones with relatively large mean values.

Figure 5.18 Sample Variance versus Sample Mean of the Trip-level OD Flow on the CLS Bus Route
Based on the Negative Binomial distribution, the expected trip-level OD flow is given by:

\[ E(T(i, j)) = r(i, j) \times \frac{1 - p(i, j)}{p(i, j)} \]  \hspace{1cm} (5.7)

and the variance of the trip-level OD flow is given by:

\[ \text{var}(T(i, j)) = r(i, j) \times \frac{1 - p(i, j)}{p(i, j)^2} \]  \hspace{1cm} (5.8)

where \( r(i, j) \) specifies the number of successes and \( p(i, j) \) specifies the probability of a success in the Negative Binomial distributions for the passenger flow of OD pair \((i, j)\). The ratio of the expectation to the variance of the trip-level OD flow equals the probability of a success \( p(i, j) \):

\[ p(i, j) = \frac{E(T(i, j))}{\text{var}(T(i, j))} \]  \hspace{1cm} (5.9)

It is assumed that the expected trip-level OD flow in Equation 5.7 equals the sample mean of the trip-level OD flow obtained from the survey OD data collected on the CLS bus route. For a given OD pair \((i, j)\), the probability of a success \( p(i, j) \) is assumed to be equal to the ratio of the sample mean to the sample variance of the trip-level OD flow across the 36 bus trips on which OD survey data were collected on the CLS bus route (the probability \( p(i, j) \) is set to be 0.9 if the resulting probability is greater than one). Given the assumed probability \( p(i, j) \) and the assumed expected trip-level OD flow, the specified number of success \( r(i, j) \) can be obtained from Equation 5.7, which is given by:

\[ r(i, j) = E(T(i, j)) \times \frac{p(i, j)}{1 - p(i, j)} \]  \hspace{1cm} (5.10)

By definition, the underlying probability OD flow is given by:
\[ \alpha(i, j) = \frac{E(T(i, j))}{E(tot)} \] (5.11)

The probability OD flow matrix is estimated based on APC boarding and alighting counts on 200 assumed bus trips. The null base OD matrix is used in the IPF, CM and HEM methods when producing the probability OD flow estimates. Given the assumed parameters of the distribution governing the generation of volume trip-level OD flows, OD flow for each OD pair is generated independently to form an OD flow matrix for each of the 200 bus trips and the corresponding trip-level boarding and alighting counts are derived. The period-level probability OD flow estimates are produced from the various methods based on the generated trip-level boarding and alighting counts (while maintaining all the original distributional assumptions), and the performance measures and the corresponding RP measures are calculated for each of the period-level probability OD flow estimates. This process is replicated 100 times, where with different volume OD flow matrices are generated.

The ECDFs of the performance measures and the corresponding RP measures are presented in Figures 5.19 and 5.20, respectively. For comparison purposes, the ECDFs of the performance measures and the corresponding RP measures of various estimates based on the trip-level boarding and alighting counts generated from the Multinomial distribution are also presented (these are the same results produced from the analysis of Section 5.4.3). The results for the subplots in the same row are based on the same specific distributional assumption, which is indicated in the title of the subplot, and the results for the subplots in the same column are based on the same specific performance measure, which is indicted in the label of the x axis. Since the patterns based on different measures
are similar and the patterns based on the performance measure and the corresponding $RP$ measure are similar, the following discussion focuses on the $HD^2$-based $RP$ measure in Figure 5.20.

Figure 5.19 ECDFs of the Performance Measures under Two Distributional Assumptions
As can be seen in Figure 5.20, the ECDFs of the $HD^2$-based $RP$ measures are similar under the two distributional assumptions for each method, indicating that the IPF, CM and HEM methods are insensitive to the distributional assumptions. In addition, the HEM algorithm is better than the IPF and CM methods under both distributional assumptions. For example, the average $HD^2$-based $RP$ measures of the IPF, CM and HEM methods are 0.737, 0.887 and 0.915 respectively under the Negative Binomial distributional assumption. The average $HD^2$-based $RP$ measures of the IPF, CM and HEM methods are 0.721, 0.891 and 0.927 respectively under the Multinomial distributional assumption.
Not surprisingly, the CM and HEM methods perform slightly better when the volume trip-level OD flows are generated by the same distribution as assumed in the respective formulations (the Multinomial distribution). The IPF method does not rely on such an assumption and in this case it performs slightly worse under volume trip-level OD flows generated based on the Multinomial assumption.
Chapter 6  Empirical Study

6.1 Introduction

Based on simulated APC boarding and alighting counts, Chapter 5 demonstrated that, first, the HEM method produces a very good approximation of the solution produced by the EM method and, second, the HEM method is superior to the IPF and CM methods. However, Chapter 5 assumes that APC data are free of measurement errors. The empirical APC data inevitably are subject to measurement errors. In addition, Chapter 5 was strictly based on simulated data even though in the case of the CLS bus route, a field observation derived probability OD matrix was used in the simulation. Therefore, it is necessary to evaluate the HEM method using only empirical APC data.

This chapter evaluates the performance of the IPF, CM and HEM methods using empirical data collected on the Campus Loop South (CLS) bus route operated by Campus Area Bus Service (CABS) at the Ohio State University (OSU). The general information about the CLS bus route has been provided in Chapter 5. The empirical data used in this study include large amounts of APC data and a large-scale manually collected true OD dataset. The empirical data are described in Section 6.2. Measurement errors in APC counts may result in the imbalance of total boarding and total alighting counts. The imbalance of total boarding and total alighting counts may cause the IPF, CM and HEM
method to break down. A balance procedure is introduced in Section 6.3 to deal with this problem.

To quantify the effects of the number of bus trips for which boarding and alighting counts were collected and measurement errors in these counts on the accuracy of OD estimates, three experiments are developed. The effect of using OD survey data as additional inputs in conjunction with this analysis is also considered. The overview of the three empirical experiments and the selection of OD survey data are presented in Section 6.4. The results and discussion are presented in Section 6.5.

6.2 Data Description

APC systems have been installed on all buses of CABS since 2009 as one component of the Campus Transit Lab (CTL). APC boarding and alighting counts collected on 1,119 bus trips of the CLS bus route between 8 and 10 a.m. on weekdays in the winter and spring quarters of 2009 are used in this study. Measurement errors in APC data on the CLS bus route during that time are considered large. Figure 6.1 summarizes the trip-level APC counts on the CLS bus route in terms of the trip-level total boarding count (TTB) and the trip-level total alighting count (TTA). Measurement errors in APC counts would result in a difference between TTB and TTA (a difference between TTB and TTA could in theory result from passengers boarding at a stop on one trip and alighting at a stop on some subsequent trip. However, this case is rare on the CLS bus route due to the definition of a bus trip).

Figure 6.1a presents the ECDF of the difference between TTB and TTA. This difference ranges from -54 to 70. This wide range reflects the large measurement errors.
in APC counts. Figure 6.1b presents the ECDF of the ratio of the absolute difference of between TTB and TTA to the average of TTB and TTA. Transit agencies often use this ratio for data quality control (Boyle 2008). For example, some transit agencies accept the APC counts if this ratio is not greater than 0.1 on at least 90 percent of bus trips. However, based on Figure 6.1b, the ratio is not greater than 0.1 on only 47 percent of bus trips, indicating that measurement errors in APC counts are indeed large.

![Graphs](image)

Figure 6.1 Summary of Trip-level Total Boarding and Alighting Counts on the CLS Bus Route

Figure 6.1c presents the ECDF of the average of TTB and TTA. The average value ranges from 0 to 151 and the median is 45.5. Figure 6.1d compares TTB with TTA.
in a scatter plot. The x and y values of one point are TTB and TTA on one bus trip, respectively. As can be seen, large discrepancy between TTB and TTA exists on some bus trips.

After screening for bus trips with extremely large APC measurement errors, where each bus trip with an absolute value difference between TTB and TTA exceeding 100% of the average of TTB and TTA for that trip is excluded, data for a total of 1,077 bus trips (i.e., 96% of the total number of bus trips) are used in the subsequent analysis. However, TTB and TTA may still not be equal for some trips because of APC measurement errors that were not large enough to result in discarding these trips at the screening stage. The imbalance between TTB and TTA may cause the IPF, CM and HEM methods to break down. A balancing procedure is introduced in Section 6.3 to address this situation.

Large-scale onboard surveys were carried out on the CLS bus route between 8 and 10 a.m. on weekdays in 3 academic quarters of 2009. The survey process has been described in Chapter 5. Briefly, the true origin and destination information of 2,268 passengers on 36 bus trips, where at least 95% of passengers traveling on each bus trip were surveyed, was manually collected.

### 6.3 Procedure to Balance Boarding and Alighting Counts with Measurement Errors

As discussed in Section 6.2, measurement errors in APC counts may result in an imbalance between total boarding and total alighting counts. Measurement errors in APC
counts may also result in the negative load problem, namely, the number of passengers alighting at a given bus stop is larger than the number of onboard passengers.

The IPF, CM and HEM methods assume that APC counts are free of measurement errors. An imbalance in the total boarding and total alighting counts and a negative load could cause the IPF, CM and HEM methods to break down. To overcome this problem, APC counts on each bus trip could be balanced first such that the total boarding count equals the total alighting count and no negative loads exist. Once these conditions are satisfied, the IPF, CM and HEM methods could be applied to the balanced APC counts to estimate the probability OD matrix. The balancing procedure used in this study is described as follows:

1. For each bus trip, if the total boarding count equals total alighting count, go to step 3; otherwise, go to step 2.

2. An average number of total boarding and total alighting counts is calculated as a target trip-level total demand. A boarding factor and an alighting factor are calculated as the ratio of the target trip-level total demand to the total boarding and total alighting counts, respectively. The boarding and alighting counts at each bus stop are then multiplied by the boarding and alighting factors, respectively. Then the adjusted boarding and alighting counts are round off to produce integer values.

3. If a negative load exists immediately downstream of a given bus stop, the absolute value of the negative load is added to the boarding count at the first boarding stop. Step 3 is repeated for all alighting stops sequentially, going from one alighting stop to the next alighting stop. Once the last alighting stop is processed, go to step 4.

4. After Step 3, the total boarding count could be greater than or equal to the total alighting count. If the total boarding count is greater than the total alighting count, the absolute
difference between the total boarding and alighting counts is added to the alighting count at the last alighting stop. Otherwise, stop.

A good discussion regarding such balancing procedures is provided in Furth, et al. (2005), Kikuchi, et al. (2006) and Lu (2008).

6.4 Experiment Overview

To quantify the effects of interest on the accuracy of OD estimates, three experiments are designed based on field OD data and empirical APC data. These three experiments are summarized in Table 6.1 by the number of bus trips with boarding and alighting counts considered in OD estimation and the level of measurement errors in boarding and alighting counts. In Experiment 1, the trip-level boarding and alighting counts are derived from the trip-level OD flows collected on 36 bus trips. The derived boarding and alighting counts are considered measurement error free given the nature of the survey process. In Experiment 3, APC boarding and alighting counts on the selected 1,077 bus trips are used in OD estimation. As discussed before, the empirical APC data are prone to large measurement errors. In Experiment 2, 36 bus trips (the same number of bus trips as in Experiment 1) are randomly drawn without replacement from the selected 1,077 bus trips for which APC counts were collected. APC boarding and alighting counts on the drawn 36 bus trips are used in OD estimation.
The effect of measurement errors in boarding and alighting counts or the number of bus trips with boarding and alighting counts on the accuracy of OD estimates could be quantified by comparing the performance of OD estimates across two selected experiments. For example, the number of bus trips with boarding and alighting counts in Experiments 1 and 2 is the same. However, boarding and alighting counts in Experiment 1 are free of measurement errors while APC counts in Experiment 2 are prone to large measurement errors. The performance difference of the OD estimates resulting from Experiments 1 and 2 reflects the impact of measurement errors in APC counts on the accuracy of OD estimates. In addition, the level of measurement errors in APC counts is the same in Experiments 2 and 3. However, the number of bus trips with APC counts in Experiment 3 is much larger than that in Experiment 2. The performance difference of the OD estimates resulting from Experiments 2 and 3 reflects the impact of the number of bus trips with APC counts on the accuracy of OD estimates.

Moreover, to investigate the effect of using additional survey OD data on OD estimates, estimations based on the null base OD matrix and a base OD flow matrix constructed from randomly sampled passengers of the manually collected true OD data.
are evaluated in each experiment. The sampled data emulate data collected in an onboard survey. Specifically, 100 passengers are drawn from the total of 2,268 passengers without replacement from the manually collected true OD flow data to produce the base OD flow matrix. The OD survey sample size (i.e., 100) considered in this study is relatively large with respect to a typical onboard survey on a bus route since there are only 18 bus stops and 152 feasible OD pairs on the CLS bus route.

The onboard survey OD matrix may reflect zero entries for some feasible OD pairs, especially for long routes (resulting in large OD matrices), due to the combination of the survey sample size and a low likelihood of travel between these OD pairs. Using the onboard survey OD matrix directly as the base OD matrix of the IPF method may cause the well-recognized “non-structural zeros” problem (Ben-Akiva et al., 1985). Non-structural zeros cause estimation difficulties when applying the IPF method, since the estimated passenger flows for OD pairs remain zero if the corresponding flows are zero in the base OD flow matrix, even if it is feasible for passengers to travel between the given OD pairs. To overcome the “non-structural zeros” problem, the onboard survey base OD flow matrix in this study is constructed as the sum of the onboard survey OD matrix and the null base OD matrix where cell values are equal to one for all feasible OD pairs. Under the proposed model structure in Chapter 3, the onboard survey base OD flow matrix, as constructed above, is analogous to the parameter matrix of the prior distribution of the probability OD matrix before observing APC data (where the uniform prior is used) and after observing survey OD data.
The OD flow matrices estimated by the IPF, CM and HEM methods are evaluated by comparing the results to both the directly observed OD flow and the non-informative null base matrices. The former represents the “truth” being estimated, while the latter represents the most naïve “estimate” (where the term “estimate” is used loosely in this context) regarding OD flows in the absence of any data or information. Each matrix is described next. Regarding the observed OD flow matrix, the trip-level OD flows in all quarters are aggregated across bus trips for each OD pair in forming the field OD flow data set. This field OD data set is then normalized by the total sample size. The resulting OD flow matrix represents the best estimate of the probability OD flow matrix for CLS bus route during the 8-10 am period, to which the IPF, CM, and HEM estimated matrices are compared in evaluating their performance. Each cell in the estimate from the null base matrix contains the reciprocal of the total number of feasible cells. The resulting matrix represents the non-informative (normalized) OD flows to which the IPF, CM, and HEM estimated matrices are also compared in the evaluation process.

The $HD^2$, $SSD$ and $Chi2$ performance measures and their corresponding $RP$ measures defined in Equations 5.2 through 5.5 in Chapter 5 are used to evaluate the performance of the IPF, CM, and HEM methods. The $\alpha(r,q)$ in Equations 5.2 through 5.4 represents the assumed underlying probability OD flow from stops $r$ to $q$, while $\alpha(r,q)$ in this chapter represents the normalized field OD flow from stops $r$ to $q$.

### 6.5 Results and Discussion

Figures 6.2 and 6.3 present the performance measures and their corresponding relative performance $RP$ measures, respectively, of various estimates in the three
experiments. The results in the same row correspond to the same experiment and the results in the same column correspond to the same measure. For the scenario where a survey base OD flow matrix is used, passengers are randomly drawn without replacement from the field OD flow data set. To reduce sampling error, the drawings are repeated 100 times. The results shown in Figures 6.2 and 6.3 are average values across these 100 trials. In addition, as discussed above, bus trips in Experiment 2 are randomly sampled without replacement from the empirical APC data. To reduce sampling error, the drawings generating bus trips for Experiment 2 are repeated 100 times (in each trial, a new survey base OD matrix is sampled for the scenario where a survey base OD flow matrix is used). The results for Experiment 2 shown in Figures 6.2 and 6.3 are average values across these 100 trials.
Figure 6.2 Performance Measures of Various Estimates Resulting from Three Experiments
Figure 6.3 $RP$ Measures of Various Estimates Resulting from Three Experiments

Since Figures 6.2 and 6.3 show similar patterns, the following discussion focuses on the relative performance $RP$ metric in Figure 6.3. Recall this metric ranges from zero to one where smaller values indicate poorer performance. Not surprisingly, the comparison of $RP$ measures resulting from Experiment 2 with those resulting from Experiment 1 reveals that measurement errors in APC data deteriorate the performance of the IPF, CM and HEM estimates. Recall that the number of bus trips with boarding and alighting counts in Experiment 1 is the same as that in Experiment 2. However, boarding and alighting counts in Experiment 1 are free of measurement errors and APC boarding and alighting counts in Experiment 2 have large measurement errors. As can be seen in
Figure 6.3, the $RP$ measures of the IPF, CM and HEM estimates resulting from Experiment 2 are lower than the corresponding measures resulting from Experiment 1 (except for the IPF method using the null base evaluated by the $Chi^2$ measure), indicating that measurement errors in APC data would reduce the accuracy of OD estimates. For example, when only APC counts are used for OD estimation (i.e., the null base is used), the $HD^2$-based $RP$ measures resulting from Experiment 2 are lower than the measures resulting from Experiment 1 by 0.08, 0.22 and 0.22 for the IPF, CM and HEM estimates, respectively. When both APC counts and survey OD data are used for OD estimation (i.e., the survey base is used), the $HD^2$-based $RP$ in Experiment 2 is lower than that in Experiment 1 by 0.10, 0.17 and 0.17 for the IPF, CM and HEM estimates, respectively. Notice that the performance of the CM and HEM estimates deteriorates faster than that of the IPF estimates in the presence of large measurement errors in APC counts. Similar results could be seen for the $SSD$-based and $Chi^2$-based $RP$ measures. However, the $Chi^2$-based $RP$ measure resulting from Experiment 2 is higher than the measure resulting from Experiment 1 for the IPF method using the null base. Since the IPF method when the null base is used does not capture structural information about the underlying OD flow matrix, it is possible for this method to accidentally arrive at better results in the presence than in the absence of measurement errors in APC data, which is what is suspected in this particular counter-intuitive result.

Comparison of $RP$ measures resulting from Experiment 2 with those resulting from Experiment 3 reveals that increasing the number of bus trips for which APC counts are collected improves the accuracy of OD estimates. Recall that the magnitude of
measurement errors in APC counts in Experiment 3 is the same as that in Experiment 2. However, the number of bus trips in Experiment 3 is 1,077 and the number of bus trips in Experiment 2 is 36. As can be seen in Figure 6.3, the $RP$ measures of the IPF, CM and HEM estimates resulting from Experiment 3 is higher than the corresponding measures resulting from Experiment 2, indicating that increasing the number of bus trips with APC counts would improve the accuracy of OD estimates. For example, when only APC counts are used for OD estimation (i.e., the null base is used), the $HD^2$-based $RP$ measures resulting from Experiment 3 are higher than the measures resulting from Experiment 2 by 0.03, 0.18 and 0.21 for the IPF, CM and HEM estimates, respectively. When both APC counts and survey OD data are used for OD estimation (i.e., the survey base is used), the $HD^2$-based $RP$ measures resulting from Experiment 3 are higher than the measures resulting from Experiment 2 by 0.03, 0.10 and 0.12 for the IPF, CM and HEM estimates, respectively. Notice that the performance of the CM and HEM estimates improves faster than that of the IPF estimates when the number of bus trips with APC counts increases, especially when only APC counts are used for OD estimation. Similar results could be seen for the $SSD$-based and $Chi2$-based $RP$ measures.

As expected, the performance of the IPF, CM and HEM methods using the survey base OD matrix is always better than that using the null base OD matrix in the three experiments. However, the magnitude of the performance improvement due to the incorporation of survey OD data is not affected by the number of bus trips with APC counts for the IPF method, while the magnitude of the performance improvement is smaller when the number of bus trips with APC counts is larger for the CM and HEM
algorithms. For example, in Experiment 2 (where the number of bus trips with APC counts is 36), the increase in the $H D^2$-based $RP$ measure due to the incorporation of survey OD data is 0.08, 0.08 and 0.10 for the IPF, CM and HEM methods, respectively. In Experiment 3 (where the number of bus trips with APC counts is 1,077), the increase in the $H D^2$-based $RP$ measure due to the incorporation of survey OD data is 0.08, 0.002 and 0.002 for the IPF, CM and HEM methods, respectively. The magnitude of performance improvement is much larger for the IPF method than it is for the CM and HEM methods in Experiment 3, indicating that survey OD data are more important for the IPF method than for the CM and HEM methods in producing good probability OD flow estimates given large amounts of APC data.

The effects of the number of bus trips with boarding and alighting counts, measurement errors in APC counts and survey OD data on the accuracy of OD estimates has been discussed above. The next chapter further investigates the effects of these three factors using simulated data and provides possible explanations of the nature of the results. The following discussion focuses on the comparison of the three methods and the results they produce in the three experiments.

As can seen in Figure 6.3, the HEM algorithm is worse than the IPF or CM method when the number of bus trips with boarding and alighting counts is small (i.e., Experiments 1 and 2). However, when boarding and alighting counts on a large number of bus trips are available (i.e., Experiment 3), based on APC boarding and alighting counts only (i.e., the null base is used), the HEM algorithm is better than the IPF and CM methods, and based on both APC counts and OD survey data (i.e., the survey base is
used), the IPF and HEM methods perform similarly and both methods are better than the CM algorithm.

In practice, boarding and alighting counts on a large number of bus trips could be collected by APC systems. In this case, the results presented above indicate that the HEM algorithm would tend to be preferred over the IPF and CM methods if survey OD data are not available, and that either the IPF or HEM method could be used for estimating the probability OD matrix if survey OD data are available and large measurement errors are presented in the APC data.

It is worthwhile to point out again that the OD survey sample size in this study is considered relatively large given the relatively small number of feasible OD pairs. If the OD survey sample size is relatively small compared with the dimension of the OD matrix, the information about the underlying probability OD matrix contained in the survey OD data is limited. In this case, it is expected that the HEM algorithm would be better than the IPF method since the survey OD data is more important for the IPF method than for the HEM algorithm to produce good OD flow estimates as discussed before.

Lastly, as discussed in Chapter 5, the distribution of Passenger Distance Travelled (PDT) derived from the estimated probability OD matrix could also be used to evaluate the performance of OD estimates. Based on large quantities of simulated APC data, Figure 5.9 shows that the IPF method using the null base tends to overestimate both the short and long travel trips, while the HEM algorithm provides good estimates for both the
short and long travel trips. The distribution of PDT could also be produced from the estimated probability OD matrices based on the empirical APC data.

For example, Figure 6.4 presents the CDFs of the PDT derived from various probability OD flow estimates resulting from Experiment 3 where the number of bus trips with APC counts is large. The probability OD matrix estimates used to produce the CDFs of Figure 6.4 are based on the null base OD matrix. The ‘true’ CDF in Figure 6.4 is for the PDT derived from the normalized field OD data. As can be seen in Figure 6.4, the IPF method using the null base overestimates both the short and long travel trips (e.g., travel distances less than 0.8 km and larger than 4.8 km). The CDF of the PDT derived from the HEM estimates is closer to the “true” CDF than those derived from the IPF and CM estimates. However, because of the presence of large measurement errors in APC counts, the HEM estimates also overestimate both the short and long travel trips, but not as much as the IPF or CM method do.
Figure 6.4 CDFs of Passenger Distance Traveled (PDT) Derived from Various OD Estimates Resulting from Experiment 2
Chapter 7   Numerical Investigation of the Effects of Three Pertinent Factors on
the Accuracy of OD Estimates

7.1 Introduction

The empirical study in Chapter 6 demonstrated that measurement errors in APC
counts, the number of bus trips for which APC counts were collected and survey OD
data are important factors that affect the accuracy of OD estimates. However, the extent of
measurement errors (for example, as captured by the distribution or variance of the
errors) in the empirical APC data is fixed (no errors associated with the field
observations, or errors associated with the APC system in place on the CLS bus route). It
is difficult to obtain empirical APC data with different levels of measurement errors for
evaluating the performance of various OD estimation methods as a function of the level
of measurement errors. In addition, only two levels in the number of bus trips with
boarding and alighting counts are considered (36 or 1,077). Moreover, only two levels of
OD survey data are considered (none, where the null base OD matrix is used, and a base
OD matrix derived from an emulated OD survey based on a sample of 100 passengers). A
numerical approach renders more flexibility in investigating the joint effects of multiple
levels for these three dimensions on the quality of the resulting probability OD flow
estimates.

The numerical study in this chapter is based on the structure of the CLS bus route.
The underlying probability OD matrix and the trip-level total demand distribution are
constructed from the manually collected true OD data as discussed in Section 5.4.3. Based on the assumed probability OD matrix and the assumed distribution of the trip-level total demand, trip-level OD matrices (the same as those adopted in the numerical study of Section 5.4.3), the resulting APC boarding and alighting counts and survey OD data can be generated based on the model structure presented in Figure 3.1 of Chapter 3.

The performance of various estimates (i.e., the IPF, CM, HEM and OD\textsuperscript{gen} estimates) is assessed and discussed based on the $HD^2$ measure and the $HD^2$-based RP measure. Similar results can be obtained based on the SSD and $Chi^2$ measures, which are presented in Appendix E.

This chapter is organized as follows: Section 7.2 introduces the measurement error model that is used to simulate the observed trip-level APC boarding and alighting counts conditional on the true trip-level APC boarding and alighting counts. Sections 7.3 through 7.5 evaluate the effects of the number of bus trips with APC counts, OD survey sample size and measurement errors in APC counts on the accuracy of probability OD flow estimates. Specifically, Section 7.3 evaluates the performance of the various estimates under different combinations of number of bus trips with APC counts and OD survey sample size, assuming that APC counts are free of measurement errors. Section 7.4 evaluates the performance of the various estimates under different combinations of OD survey sample size and level of measurement errors in APC counts, assuming that APC boarding and alighting counts were collected on 200 bus trips. Section 7.5 evaluates the performance of various estimates under different combinations of number of bus trips.
with APC counts and level of measurement errors in APC counts, assuming that the null base OD matrix is used for OD estimation.

7.2 APC Measurement Error Model

A linear measurement error model is adopted to simulate the observed trip-level APC boarding and alighting counts conditional on the true trip-level boarding and alighting counts, which is given by:

\[ \hat{x}_l(r) = \beta \times x_l(r) \]  

(7.1)

where:

\( \beta \) = multiplicative factor, which is assumed to follow symmetric triangular distribution between 1-\( w \) and 1+\( w \). The parameter \( w \) is specified to represent the magnitude of measurement errors,

\( x_l(r) = \) the value in the \( r^{th} \) entry of the true boarding and alighting count vector \( x_l \) on bus trip \( l \),

\( \hat{x}_l(r) = \) the value in the \( r^{th} \) entry of the observed boarding and alighting count vector \( \hat{x}_l \) on bus trip \( l \).

The observed trip-level APC counts determined by Equation 7.1 are unbiased. That is, the expectations of the observed trip-level APC boarding or alighting counts are assumed equal to the true trip-level boarding or alighting counts. This assumption is reasonable because the biases in measurement errors could be easily quantified and corrected for. As such, the remaining errors are random fluctuations around the true values. The observed trip-level APC counts determined by Equation 7.1 are rounded off to produce discrete values. Larger value of \( w \) results in larger variance of the observed trip-level APC counts.
In this chapter, the parameter $w$ for the distribution of the multiplicative factor $\beta$ is chosen from 0.1 to 0.4 in increments of 0.1 to simulate various levels of measurement errors in APC counts. Larger parameter $w$ represents larger variances in measurement errors in APC counts. In addition, the value of 0.15 is also considered since this value corresponds to generated observed APC boarding and alighting counts that are within 10% of the true APC boarding and alighting counts with a confidence level of 90%, which is the threshold used by many transit agencies for the acceptance of APC data (Boyle, 2008).

7.3 Effects of the Number of Bus Trips with APC Counts and OD Survey Sample Size on the Accuracy of OD Estimates

This section investigates the joint effects of the number of bus trips for which APC boarding and alighting counts are collected and OD survey sample size on the accuracy of OD estimates. Assuming that APC counts are free of measurement errors, APC counts on various numbers of bus trips and OD survey data with various sample sizes are considered for the estimation of the probability OD matrix. Given an assumed number of bus trips with APC counts, a volume OD flow matrix is generated from the true underlying probability OD flow matrix for each of the trips, and the corresponding trip-level boarding and alighting counts are derived. In addition, given the assumed OD survey sample size, the survey OD flow matrix is generated from the true underlying probability OD flow matrix. The period-level probability OD flow estimates are produced from the various methods based on the generated trip-level APC counts and the generated OD survey data, and the performance measures are calculated for each of the period-level
probability OD flow estimates. This process is replicated 100 times for a combination of number of bus trips with APC data and OD survey sample size, where with each replication different volume OD flow matrices and different OD survey data are generated. The entire process is then repeated for different combinations of number of bus trips with APC data and OD survey sample size.

The average $HD^2$ measure and the corresponding $RP$ measure as a function of the number of bus trips for which APC data are assumed to have been collected (across the 100 simulation replications for each number of bus trips) are presented in Figures 7.1 and 7.2 for the IPF, CM, HEM, and $OD_{gen}$ methods. The assumed OD survey sample size $tos$ in producing the results for each subplot of Figures 7.1 and 7.2 is presented in the title of each subplot. Zero OD survey sample size (i.e., $tos = 0$) means that the null base OD matrix is used for the IPF, CM and HEM methods. Since Figures 7.1 and 7.2 show similar patterns, the following discussion focuses on the results associated with the $RP$ measure shown in Figure 7.2.
Figure 7.1 $H D^2$ Performance Measure versus the Number of Bus Trips for which APC Data were Collected under Different OD Survey Sample Sizes (Given Measurement Error Free APC Counts)
Figure 7.2 $HD^2$-based $RP$ Measure versus the Number of Bus Trips for which APC Data were Collected under Different OD Survey Sample Sizes (Given Measurement Error Free APC Counts)

Obviously in Figure 7.2, increasing the number of bus trips with APC counts improves the accuracy of OD estimates. Nevertheless, the performance improvement of the CM and HEM estimates is larger than that of the IPF estimates. For example, when the null base OD matrix is used in the IPF, CM and HEM methods (i.e., $tos = 0$), the $RP$ measures of the IPF, CM and HEM estimates increases 0.0074, 0.1194 and 0.1916 respectively when the number of bus trips with APC counts increases from 30 to 500. In addition, the HEM algorithm does not work well when only a few bus trips have APC counts. However, the performance of the HEM algorithm surpasses that of the IPF and
CM methods at a fairly low number of bus trips with APC counts and becomes increasing superior as the number of bus trips with APC data continues to increase. These results are consistent with what has been seen in the empirical study in Chapter 6.

The $RP$ measures for the IPF, CM and HEM methods are larger than that of the $OD^{gen}$ method for a low number of bus trips for which APC data are assumed to have been collected. However, and not surprisingly, when the number of bus trips for which APC data were collected increases, the performance of the $OD^{gen}$ estimates improves much faster than that of the IPF, CM or HEM method. For example, the $OD^{gen}$ estimates outperform the IPF, CM and HEM estimates when the number of bus trips is larger than 5, assuming the null base OD matrix is used for the IPF, CM and HEM methods (i.e., $tos = 0$).

Nevertheless, the performance of the IPF and HEM estimates also improves (i.e., the $RP$ measure increases) as the number of bus trips for which APC data were collected increases. However, at a certain point (e.g., approximately 50 bus trips with APC data for $tos = 0$), the continued increase of the $RP$ measure for the IPF method is practically unnoticeable. By contrast, the $RP$ measure for the HEM algorithm continues to increase appreciably. Although the rate of the increase of the $RP$ measure of the HEM algorithm is less than that of the IPF method when only a few bus trips are assumed to have APC data, the HEM algorithm outperforms the IPF method when the number of bus trips with APC data reaches a certain value. This value is larger if the OD survey sample size is larger. In addition, the advantage of the HEM algorithm over the IPF method diminishes when the OD survey sample size increases. For example, the HEM algorithm outperforms the IPF
method when the number of bus trips with APC data reaches approximately 18 if the null base is used in the IPF and HEM methods, and the HEM algorithm outperforms the IPF method when the number of bus trips with APC data reaches approximately 40 if the OD survey sample size is 600. In both cases, the number of bus trips with APC counts that is needed such that the HEM algorithm outperforms the IPF method is small if buses are equipped with APC technologies.

The relationship between the number of bus trips with APC counts and the performance of the CM estimates depends on the OD survey sample size. Under the OD survey sample size of 0 or 100 (i.e., the two subplots in the first row of Figure 7.2), the performance of the CM estimates continues to improve (i.e., the $RP$ measure increases) when the number of bus trips for which APC data were collected increases. However, the performance of the CM estimates may deteriorate as the number of bus trips with APC counts increases under the OD survey sample size of 400 or 600 (i.e., the two subplots in the second row of Figure 7.2). For example, under the OD survey sample size of 400, the $RP$ measure of the CM algorithm decreases from 0.9259 to 0.9251 when the number of bus trips with APC counts increases from 200 to 500.

The general patterns of the IPF, CM and HEM methods in Figure 7.2 are understandable. The performance of the IPF estimates improves as the number of bus trips with APC counts increases. However, at certain point, the improvement is practically unnoticeable. The IPF method is based on the first moments of the trip-level boarding and alighting counts. When the number of bus trips reaches a certain value (approximately 50 in this example), the sample means of the trip-level APC boarding and
alighting counts are already very good approximation of the first moments of the trip-level boarding and alighting counts. Further increasing the number of bus trips with APC data would produce little improvement to the approximation of the first moments of the trip-level boarding and alighting count distribution and, therefore, the performance improvement of the IPF method is very small.

The performance of the HEM estimates continues to improve when the number of bus trips with APC counts increases. When the number of bus trips with APC counts is small, the HEM algorithm does not work well. The HEM estimates are based on the distribution of the trip-level boarding and alighting counts. More bus trips with APC data are needed to provide good approximation of the distribution of the trip-level boarding and alighting counts than good approximation of simply the first moments of the distribution. A small number of bus trips with APC data is not able to provide a good approximation of the distribution of the trip-level boarding and alighting counts. As a result, when the number of bus trips with APC data is small, the HEM algorithm does not perform well, and the IPF method is better than the HEM algorithm. Increasing the number of bus trips with APC data would improve the approximation of the distribution of the trip-level boarding and alighting counts. As a result, the performance of the HEM estimates continues to improve substantially and eventually is better than that of the IPF estimates.

The performance of the CM estimates may improve or deteriorate as the number of bus trips with APC counts increases, depending on the OD survey sample size. On the one hand, the CM estimates are based on the distribution of the trip-level boarding and
alighting counts. Increasing the number of bus trips with APC counts would improve the approximation of the distribution of the trip-level boarding and alighting counts and, therefore, the performance of the CM estimates improves. On the other hand, the CM algorithm estimates a mode of the joint posterior distribution of the probability OD matrix and the trip-level volume OD matrices on the bus trips for which the APC counts were collected. When the number of bus trips with APC counts increases, the dimension of the joint mode increases (as there would be much more trip-level volume OD matrices to be estimated). Given the limited information provided by the APC counts and the high dimension of joint mode to be estimated, the CM estimates may not be able to provide a good summary of the joint posterior distribution when the number of bus trips with APC counts is large. As a result, the results seen in Figure 7.2 in terms of the performance of the CM algorithm deteriorating with increased number of bus trips with APC data are likely to occur.

The effect of the number of bus trips with APC data has been discussed above. Recall that it was commented that more APC data are needed for the HEM algorithm to outperform the IPF method if the OD survey sample size is large. In addition, recall that the advantage of the HEM algorithm over the IPF method diminishes when the OD survey sample size increases. This result indicates that the performance of the IPF estimates improves faster than that of the HEM estimates as the OD survey sample size increases. The effect of OD survey sample size on the accuracy of OD estimates could be better seen in Figures 7.3 and 7.4, which present the average $HD^2$ measure and the corresponding $RP$ measure of various OD estimates as a function of the OD survey
sample size (across the 100 simulation replications for each OD survey sample size) under different numbers of bus trips with APC counts. The number of bus trips with APC counts $L$ for producing the results of each subplot in Figures 7.3 and 7.4 is presented in the title of each subplot. In Figures 7.3 and 7.4, zero OD survey sample size means that the null base OD matrix is used for the IPF, CM, and HEM methods. The $\text{OD}^{\text{gen}}$ estimates do not depend on the OD survey sample size and therefore the performance of the $\text{OD}^{\text{gen}}$ estimates is almost constant in Figures 7.3 and 7.4 (the small variations are due to the variations in the generated trip-level volume OD matrices in the simulation even though an average across 100 replications is taken). Since Figures 7.3 and 7.4 show similar patterns, the following discussion focuses on the results associated with the $RP$ measure shown in Figure 7.4.
200

Figure 7.3 $HD^2$ Performance Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data (Given Measurement Error Free APC Counts)
As expected, the performance of the IPF, CM and HEM estimates improves (i.e., the RP measure increases) as the OD survey sample size increases. For the IPF method, the rate of the performance improvement due to the increase of the OD survey sample size is not affected by the number of bus trips with APC counts. However, for the CM and HEM algorithms, the rate of the performance improvement is slower when the number of bus trips with APC counts is larger. As a result, when the number of bus trips with APC counts is large, the performance improvement due to the increase of OD
survey sample size is larger for the IPF method than for the CM and HEM methods. These results are consistent with what has been seen in the empirical study in Chapter 6.

The patterns described above are understandable. The IPF estimates are based on the first moments of the trip-level boarding and alighting counts. No matter how many bus trips have APC data, the IPF method treats the sample means of the trip-level boarding and alighting counts as the true first moments of the trip-level boarding and alighting counts. As a result, the IPF method uses constant weights for APC data and OD survey data in determining the final OD estimates, regardless of the number of bus trips with APC counts. Therefore, the rate of the increase of the $RP$ measure for the IPF method due to the increase of the OD survey sample size is not affected by the number of bus trips for which APC counts were collected, especially when enough bus trips with APC data are available to arrive at good estimates of the first moments of APC data.

By contrast, the CM and HEM estimates are based on the distribution of the trip-level boarding and alighting counts. When the number of bus trips with APC counts increases, better approximation to the distribution of the trip-level boarding and alighting counts could be obtained. In this case, the CM and HEM algorithms assign larger weight to APC data and smaller weight to OD survey data in determining the final OD estimates (the weights for APC data and survey OD data are determined automatically by the algorithms). As a result, the rate of the increase of the $RP$ measures for the CM and HEM algorithms due to the increase of the OD survey sample size is slower when the number of bus trips with APC counts is larger.
To examine the full range of the $HD^2$ and the $HD^2$-based $RP$ measures across the 100 replications, the ECDFs of the $HD^2$ and the $HD^2$-based $RP$ measures are presented in Figures 7.5 and 7.6 for different combinations of number of bus trips with APC counts and OD survey sample size. Similar conclusions can be drawn from Figures 7.5 and 7.6 as those discussed in relation to Figures 7.1 through 7.4. For example, when the null base OD matrix (i.e., $tos=0$) is used in the IPF, CM and HEM methods (i.e., row 1 of Figures 7.5 and 7.6), the ECDFs of the $RP$ measures are similar for the three methods when the number of bus trips with APC counts is one, indicating that the performance of the three methods is similar. The ECDF of the $RP$ measure of the HEM estimates moves from the left to the right of the ECDF of the $RP$ measure of the IPF estimates when the number of bus trips with APC counts increases, indicating that the HEM algorithm could outperform the IPF method when the number of bus trips with APC counts increases.

Examining the effect of $tos$ at a given number of bus trips with APC counts, consider the case where $L = 50$ (i.e., comparing the sub-plots down column 3 of Figures 7.5 and 7.6). Notice that the ECDF of the $RP$ measure of the IPF estimates is to the left of those of the CM and HEM estimates when $tos = 0$. However, as the OD survey sample size $tos$ increases, the ECDF of the RP measure of the IPF estimates is closer to those of the CM and HEM estimates, indicating that the performance improvement of the IPF estimates is larger than that of the CM and HEM estimates due to the increase of OD survey sample size.
Figure 7.5 ECDFs of the $HD^2$ Measures under Various Numbers of Bus Trips with APC Data and Various OD Survey Sample Sizes (Given Measurement Error Free APC Counts)

Figure 7.6 ECDFs of the $HD^2$-based $RP$ Measures under Various Numbers of Bus Trips with APC Data and Various OD Survey Sample Sizes (Given Measurement Error Free APC Counts)
7.4 Effects of OD Survey Sample Size and Measurement Errors in APC Counts on the Accuracy of OD Estimates

This section investigates the joint effects of OD survey sample size and measurement errors in APC counts on the accuracy of OD estimates. Assuming that APC boarding and alighting counts are collected on 200 bus trips, APC counts with various levels of measurement errors and OD survey data with various sample sizes are considered for the estimation of the probability OD matrix. The level of measurement errors in APC counts is quantified by the parameter $w$ defined in the measurement error model presented in Equation 7.1. Larger value of $w$ indicates larger variance of measurement errors in APC counts.

The simulation process is described as follows. First, the OD survey sample size $tos$ is set. This value is the number of passengers whose OD pairs are observed as generated based on the underlying true probability OD flow matrix. Second, the volume OD flow matrix is generated from the true underlying probability OD flow matrix and the trip-level total demand distribution for each of the 200 assumed bus trips, and the trip-level boarding and alighting counts are derived from the generated trip-level volume OD flow matrices. Third, given a value of the parameter $w$ for the distribution of the multiplicative factor $\beta$ in Equation 7.1, a multiplicative factor $\beta$ is generated from the assumed symmetric triangular distribution for the boarding and alighting counts at each bus stop on each of the 200 assumed bus trips. The observed boarding and alighting counts are generated according to the measurement error model in Equation 7.1 conditional on the generated multiplicative factor $\beta$ and the true boarding and alighting counts, respectively. Fourth, the observed APC boarding and alighting counts on each of
the 200 assumed bus trips are balanced such that the total boarding count is equal to the total alighting count and no negative passenger load exists between bus stops for each bus trip (see Section 6.3 of Chapter 6 for the balancing procedure). Finally, the IPF, CM and HEM methods are applied to the balanced trip-level APC counts on 200 assumed bus trips and the generated survey OD data to estimate the period-level probability OD matrix. The process described above is replicated 100 times, where for each replication different volume OD flow matrices, different observed APC boarding and alighting counts, and different OD survey data are generated. The entire process is then repeated for different combinations of OD survey sample size and variance of the measurement errors in APC counts.

The average $HD^2$ measure and the corresponding $RP$ measure as a function of the value of $w$ (across the 100 simulation replications for each $w$) are presented in Figures 7.7 and 7.8 for the IPF, CM, HEM, and $OD^{gen}$ methods. Larger values of $w$ represent larger variances of measurement errors in APC data. A zero value of $w$ represents measurement error free APC boarding and alighting counts. The assumed OD survey sample size $tos$ is presented in the title of each subplot in Figures 7.7 and 7.8. A zero value of the OD survey sample size (i.e., $tos = 0$) means the null base OD matrix is used for the IPF, CM, HEM methods. The $OD^{gen}$ estimates are not influenced by measurement errors in APC counts since it is only based on the volume OD flows assuming they were observed and, therefore, the performance of the $OD^{gen}$ estimates is almost constant in Figures 7.7 and 7.8 (the slight variation is due to variations in the generated volume OD flows even though the performance measures are averages over 100 replications). Since Figures 7.7
and 7.8 show similar patterns, the following discussion focuses on the RP measures in Figure 7.8.

Figure 7.7 $HD^2$ Performance Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes (Given 200 bus trips with APC Counts)
Not surprisingly, the performance of the IPF, CM and HEM estimates deteriorates when the value of the parameter $w$ increases (i.e., the variance of the measurement errors becomes larger), indicating that the presence of larger measurement errors in APC counts lead to worse probability OD flow estimates for the three methods. When the value of $w$ increases, the rate of the decrease of the $RP$ measures of the CM and HEM estimates is faster than that of the IPF estimates, indicating that the impact of measurement errors in APC counts on the CM and HEM estimates is greater than on the IPF estimates. For example, if the null base OD matrix is used for OD estimation (i.e., $tos = 0$), the $RP$
measure of the IPF estimates decrease from 0.73 to 0.72 and the \( RP \) measure of the HEM and CM estimates decrease from 0.93 to 0.82 and from 0.89 to 0.79, respectively, when the value of \( w \) increases from 0 to 0.4. These results are consistent with what has been seen in the empirical study.

Because the impact of measurement errors in APC counts is larger on the CM and HEM estimates than on the IPF estimates, the CM and HEM estimates could be worse than the IPF estimates when the value of \( w \) reach a certain point. In addition, as discussed in Section 7.3, given large quantities of APC data (which is the case in this exercise where APC counts are assumed to have been collected on 200 bus trips), the performance of the IPF estimates improves faster than that of the CM and HEM estimates when the OD survey sample size increases. As a result, the value of \( w \) at which the IPF method outperforms the CM and HEM algorithms is smaller when the OD survey sample size is larger. For example, if the OD survey sample size is 400, the HEM estimates are worse than the IPF estimates when the value of the \( w \) is about 0.31 or larger. If the OD survey sample size is 600, the HEM estimates are worse than the IPF estimates when the value of \( w \) is about 0.28 or larger.

Figure 7.8 shows that the impact of measurement errors in APC counts is smaller on the IPF estimates than on the CM and HEM estimates. This result is understandable. The IPF estimates are based on the first moments of the trip-level boarding and alighting counts. In the simulation, the observed trip-level APC counts are assumed to be unbiased. The sample means of the balanced trip-level APC counts still provide good approximations to the first moments of the trip-level boarding and alighting counts,
although the balancing procedure may introduce small biases in the approximations. Therefore, the impact of measurement errors in APC counts on the IPF method is small. The CM and HEM estimates are based on the distribution of the trip-level boarding and alighting counts. Because of the measurement errors in APC counts, the empirical distribution of the balanced trip-level APC counts is different from the theoretical distribution of the true trip-level boarding and alighting counts as assumed in the formulation. However, the CM and HEM algorithms assume that APC counts are free of measurement error free, and therefore treat the balanced trip-level APC data as the true trip-level boarding and alighting counts. As a result, the impact of measurement errors in APC counts on the CM and HEM estimates is large.

The ECDFs of the $HD^2$ and the $HD^2$-based $RP$ measures are presented in Figures 7.9 and 7.10 for different combinations of OD survey sample size and measurement errors in APC counts. Similar conclusions can be drawn from Figures 7.9 and 7.10 as those discussed in relation to Figures 7.7 and 7.8. For example, when the OD survey sample size is 600, the ECDF of the $RP$ measure of the IPF estimates moves from the left to the right of the ECDFs of the $RP$ measures of the CM and HEM estimates when the value of $w$ increases (i.e., variance of the measurement errors in APC counts become larger) (see the last row of Figure 7.10), indicating that the IPF method could outperform the CM and HEM algorithms when measurement errors in APC counts become large.
Figure 7.9 ECDFs of the $HD^2$ Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts (Given 200 bus trips with APC Counts)

Figure 7.10 ECDFs of the $HD^2$-based Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts (Given 200 bus trips with APC Counts)
7.5 Effects of the Number of Bus Trips with APC Counts and Measurement Errors in APC Counts on the Accuracy of OD Estimates

This section investigates numerically the joint effects of the number of bus trips with APC counts and measurement errors in APC counts on the accuracy of OD estimates. Assuming that the null base OD flow matrix is used, APC counts with various levels of measurement errors on various numbers of bus trips are considered for the estimation of the probability OD matrix. The level of measurement errors in APC counts is quantified by the parameter $w$ defined in the measurement error model presented in Equation 7.1. Larger values of $w$ indicate larger variance of measurement errors in APC counts.

The simulation process is described as follows. First, given the assumed number of bus trips with APC counts, volume OD flow matrices are generated from the true underlying probability OD flow matrix and the trip-level total demand distribution. The corresponding true trip-level boarding and alighting counts are derived from the generated trip-level volume OD flow matrices. Second, the effect of measurement errors is simulated for a value of the parameter $w$ in exactly the same manner as the third step of the process followed for the analysis in Section 7.4. Third, the observed APC boarding and alighting counts on each bus trip are balanced to achieve the same constraints of the comparable fourth step of the analysis of Section 7.4. Finally, the IPF, CM and HEM methods are applied to the balanced trip-level APC counts to estimate the period-level probability OD matrix. The process described above is replicated 100 times, where for each replication different volume OD flow matrices and different trip-level observed APC counts are generated. The entire process is then repeated for different combinations
of number of bus trips with APC counts and variance of the measurement errors in APC counts.

The average $HD^2$ measure and the corresponding $RP$ measure as a function of the number of bus trips for which APC data are assumed to have been collected (across the 100 simulation replications for each number of bus trips) are presented in Figures 7.11 and 7.12 for the IPF, CM, HEM, and OD$^{gen}$ methods. The assumed value of the parameter $w$ is presented in the title of each subplot of Figures 7.11 and 7.12. Recall that larger values of $w$ represent larger variance of measurement errors in APC counts and a zero value of $w$ represents measurement error free APC counts. Since Figures 7.11 and 7.12 show similar patterns, the following discussion focuses on the $RP$ measures in Figure 7.12.
Figure 7.11 $HD^2$ Performance Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts (Given a Null Base OD Flow Matrix)
As can be seen in Figure 7.12, the performance of the IPF, CM and HEM estimates improves as the number of bus trips increases. However, at a certain point, the performance improvement of the IPF estimates is practically unnoticeable. By contrast, the performance of the CM and HEM estimates continues to improve. These results are similar as what have been seen in Figure 7.2 of Section 7.3.

In addition, the HEM algorithm outperforms the IPF and CM methods when the number of bus trips with APC counts reaches a certain value, which is larger when the value of \( w \) is larger. For example, if the value of \( w \) is zero (i.e., the APC counts are
measurement error free), the HEM algorithm outperforms the IPF method when the number of bus trips with APC data reaches approximately 18. If the value of $w$ is 0.4, the HEM algorithm outperforms the IPF method when the number of bus trips with APC data reaches approximately 52.

The effects of the number of bus trips with APC counts and measurement errors in APC counts on the accuracy of OD estimates has been discussed in Sections 7.3 and 7.4, respectively. The results shown in Figure 7.12 demonstrate that in the presence of measurement errors in APC counts, the HEM algorithm still can outperform the IPF and CM methods if sufficient number of bus trips with APC counts are available. The number of bus trips needed such that the HEM algorithm can outperform the IPF and CM methods is larger when measurement errors in APC counts are larger. However, it is relatively small compared with what can be feasibly obtained from APC-equipped buses.

The ECDFs of the $HD^2$ and the $HD^2$-based $RP$ measures are presented in Figures 7.13 and 7.14. Similar conclusions can be drawn as those discussed in relation to Figures 7.11 and 7.12. For example, when APC counts are free of measurement errors (i.e., $w = 0$), the ECDFs of the $RP$ measures are similar for the IPF, CM and HEM methods when the number of bus trips with APC counts is one (see the top right subplot in Figure 7.14), indicating that the performances of the three methods are similar. The ECDF of the $RP$ measure of the HEM estimates moves from the left to the right of the ECDF of the $RP$ measure of the IPF estimates when the number of bus trips with APC counts increases (see the first row of subplots in Figure 7.14), indicating that the HEM algorithm could outperform the IPF method when the number of bus trips with APC counts increases.
Figure 7.13 ECDFs of the $HD^2$ Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts (Given a Null Base OD Flow Matrix)

Figure 7.14 ECDFs of the $HD^2$-based $RP$ Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts (Given a Null Base OD Flow Matrix)
Chapter 8  OD Estimation under Congestion on Buses

8.1 Introduction

In Chapters 3 through 7, the probability OD matrix is assumed to be stable across bus trips in the given homogeneous time-of-day period. This assumption may not be reasonable under congestion on buses due to bus bunching or high demand. In either case, high occupancies on some or all buses could result in passengers not being able to board the first upcoming bus. That is, they will be “left-behind”. For example, at bus stops downstream of the location where bus bunching occurs, the leading bus would have relatively higher boarding volumes than the following bus, resulting in high occupancy on the leading bus. In addition, high demand results in high occupancies on most or all buses operating on the route. The left behind passengers in either case need to wait for the next upcoming bus on the route or take a different route or mode. In both cases, the boarding probabilities, which are defined as the probabilities that a randomly selected passenger from the population boards at specified bus stops, are not stable across bus trips in a given time-of-day period. The instability of the boarding probabilities results in the instability of the probability OD matrix across bus trips in the given time-of-day period.

Under congestion on buses, it is more likely that the Alighting Probability Matrix (APM) is stable across bus trips in a given time-of-day period. The APM provides the probabilities that a passenger alights at downstream bus stops conditional on the bus stop
at which he or she boarded. Estimation of the APM could conceivably control for the effect of the instability of the boarding probabilities on the final estimates. This chapter presents approaches to estimate the APM from the trip-level APC boarding and alighting counts.

The chapter is organized as follows. In Section 8.2, the general assumptions are introduced. In Section 8.3, the posterior distributions and point estimates of the APM are presented. In Section 8.4, it is shown that the CM and HEM algorithms developed in Chapter 4 can be applied to estimate the APM. In Section 8.5, a numerical example is presented.

8.2 General Assumptions

As in Chapter 3, it is assumed that no passengers alight at the first (starting terminal) bus stop and no passengers board at the last (ending terminal) bus stop of a given bus route. It is also assumed that no passengers board and alight at the same bus stop (i.e., passenger flows along the diagonal of the OD matrix are zero). In addition, it is assumed that the APC counts and survey OD data are free of measurement errors. Moreover, it is assumed that OD survey data and trip-level APC counts are independent. The discussion about this independence assumption has been presented in Chapter 3.

The alighting probability vector is denoted by $\gamma(r,(r+1):Ns)$ representing the probabilities that a passenger who boarded at stop $r$ alights at downstream stops $(r + 1)$ through $Ns$, the total number of stops on the route. The assumed distributions on parameters, variables and data are presented in the following. The distributional
assumptions for the estimation of the APM are similar to those for the estimation of the probability OD matrix presented in Chapter 3.

(1) For a given bus trip \( l \), conditional on the boarding count \( b_l(r) \) at stop \( r \) and the alighting probability vector \( \gamma(r,(r+1):N_s) \), the OD flow vector \( T_l(r,(r+1):N_s) \) ( \( T_l(r,(r+1):N_s) \)) represents the volume boarding at stop \( r \) and alighting at downstream bus stops \( (r + 1) \) through \( N_s \) is assumed to be multinomially distributed:

\[
f(T_l(r,(r + 1):N_s) | b_l(r),\gamma(r,(r + 1):N_s)) = f(T_l(r,(r + 1):N_s) | \gamma(r,(r + 1):N_s)) \sim \text{Multinomial}(b_l(r),\gamma(r,(r + 1):N_s))
\]

\[
= \frac{b_l(r)!}{N_s!} \prod_{q=r+1}^{N_s} \gamma(r,q)^{T_l(r,q)} \prod_{q=r+1}^{N_s} T_l(r,q)! \quad r = 1, ..., N_s - 1
\]

(8.1)

where,

\[
\gamma(r,(r + 1):N_s) = (\gamma(r,r + 1), ..., \gamma(r,q), ..., \gamma(r,N_s))
\]

(8.2)

\[
T_l(r,(r + 1):N_s) = (T_l(r,r + 1), ..., T_l(r,q), ..., T_l(r,N_s))
\]

(8.3)

\( \gamma(r,q) \) = probability that a passenger alights at stop \( q \) conditional on that he or she boarded at stop \( r \), and

\( T_l(r,q) = \text{OD flow from stop } r \text{ to stop } q \text{ on bus trip } l \).

The dependence on the trip-level boarding count \( b_l(r) \) is suppressed in the second line of Equation 8.1 because the boarding count \( b_l(r) \) is assumed to be known and fixed for bus trip \( l \) in the model structure. As a result, subsequent references to the trip-level
OD matrix $T_l$ are made under the assumption that it is conditional on this known trip-level boarding count $b_l(r)$.

(2) Conditional on the trip-level OD matrix $T_l$, the alighting count $a_l(q)$ at stop $q$ on bus trip $l$ is equal to the sum of OD flows destined for stop $q$, which is given by:

$$a_l(q) = \sum_{r=1}^{q-1} T_l(r,q) \quad q = 2,\ldots,N_s$$  \hfill (8.4a)

Based on Equation 8.1, on bus trip $l$, the boarding count at stop $r$, denoted by $b_l(r)$, is considered fixed and is equal to the sum of OD flows originating from stop $r$:

$$b_l(r) = \sum_{q=r+1}^{N_s} T_l(r,q) \quad r = 1,\ldots,N_s-1$$  \hfill (8.4b)

(3) Conditional on the number of surveyed passengers $bos(r)$ who boarded at stop $r$ and the alighting probability vector $\gamma(r,(r+1):N_s)$, survey OD flow vector $z(r,(r+1):N_s)$ from stop $r$ to downstream bus stops $(r + 1)$ through $N_s$ is assumed to be multinomially distributed:

$$f(z(r,(r+1):N_s) | bos(r),\gamma(r,(r+1):N_s)) = f(z(r,(r+1):N_s) | \gamma(r,(r+1):N_s)) \sim \text{Multinomial}(bos(r),\gamma(r,(r+1):N_s))$$

$$= \frac{bos(r)!}{N_s!} \prod_{q=r+1}^{N_s} \gamma(r,q)^{z(r,q)} \quad r = 1,\ldots,N_s-1$$  \hfill (8.5)

where,

$$z(r,(r+1):N_s) = (z(r,r+1),\ldots,z(r,q),\ldots,z(r,N_s))$$  \hfill (8.6)

$$z(r,q) = \text{number of surveyed passengers travelling from stop } r \text{ to stop } q.$$
The dependence on the number of surveyed passengers $bos(r)$ who boarded at stop $r$ is suppressed in the second line of Equation 8.5 because $bos(r)$ is assumed to be known and fixed in the model structure. Subsequent references to the survey OD data are made under the assumption that it is conditional on the number of surveyed passengers at each bus stop $bos(r)$.

In addition, the prior distribution of the APM $\gamma$ is assumed to be Dirichlet, which is given by:

$$
\pi(\gamma(r,(r+1) : N_s)) \propto \prod_{q=r+1}^{N_s} \gamma(r,q)^{\mu(r,q)-1} \quad r = 1,...,N_s-1
$$

Any prior information about the APM, such as a model derived OD matrix, could be incorporated through the hyper-parameter $\mu$. When no prior information is available, a uniform prior distribution can be used. The uniform prior distribution is obtained by setting $\mu(r, q) = 1$ for all feasible OD pairs $(r, q)$.

8.3 Formulation: Posterior Distributions and Point Estimates

This section derives the joint posterior distribution of the APM and the trip-level OD matrices and the marginal posterior distribution of the APM based on the assumptions presented in Section 8.2. The next section will show that the CM and HEM algorithm developed in Chapter 4 can be used to estimate the APM based on the joint posterior distribution of the APM and the trip-level OD matrices and the marginal posterior distribution of the APM, respectively.
Conditional on the survey OD data $z$ and the trip-level APC data $x^c$, the joint posterior distribution of the APM $\gamma$ and the trip-level OD matrices $T^c$ can be shown to be given by:

$$f(\gamma, T^c | x^c, z) \propto \prod_{l=1}^{L} f(x_l | T_l) \times f(T_l | \gamma) \times f(z | \gamma) \times f(\gamma)$$

(8.8)

In Equation 8.8, $f(x_l | T_l)$ equals one if $T_l$ satisfies the APC counts $x_l$ as captured by Equations 8.4a and 8.4b and zero otherwise. The conditional distributions of the trip-level OD matrix $f(T_l | \gamma)$ and survey OD flow matrix $f(z | \gamma)$ have been given in Equations 8.1 and 8.5, respectively. The prior distribution of the APM $\gamma$ has been given in Equation 8.7. Therefore, Equation 8.8 can be expressed by:

$$f(\gamma, T^c | x^c, z) \propto \prod_{l=1}^{L} \left[ f(x_l | T_l) \times \prod_{r=1}^{N_l} \prod_{q=r+1}^{N_r} \frac{\gamma(r, q)^{T_l(r,q)}}{T_l(r,q)!} \right] \times \prod_{r=1}^{N_l} \prod_{q=r+1}^{N_r} \gamma(r, q)^{z(r,q)} \times \prod_{r=1}^{N_l} \prod_{q=r+1}^{N_r} \gamma(r, q)^{\mu(r,q)-1}$$

(8.9)

In Equation 8.9, Term A represents the likelihood of observing APC counts $x_l$ conditional on the true OD matrices $T_l$. Term A equals one if $x_l$ are determined by OD matrices $T_l$ as given by Equations 8.4a and 8.4b and zero otherwise. Term B represents the likelihood of observing an OD matrix $T_l$ conditional on the APM $\gamma$. Term C represents the likelihood of observing the survey OD data $z$ in the given time-of-day period conditional on the APM $\gamma$, and Term D represents the prior likelihood of the APM $\gamma$.  

223
The marginal posterior distribution of the APM $\gamma$ can be derived from the joint posterior distribution of Equation 8.9 by summing the joint posterior distribution over the feasible OD matrix $T_l$ for each bus trip $l$, where $T_l$ satisfies the APC counts $x_l$ as captured in 8.4a and 8.4b, which is given by:

$$f(\gamma | x^c, z) \propto \prod_{l=1}^{L} \left[ \sum_{T_l \in S_l} \prod_{r=1}^{N_s} \prod_{q=r+1}^{N_s} \frac{\gamma(r,q)^{T_l(r,q)}}{T_l(r,q)!} \right] \times \prod_{r=1}^{N_s} \prod_{q=r+1}^{N_s} \gamma(r,q)^{z(r,q)+\rho(r,q)-1}$$  \hspace{1cm} (8.10)

In Equation 8.10, $T_l \in S_l$ represents all OD matrices $T_l$ that satisfy the APC counts $x_l$ as given by Equations 8.4a and 8.4b. $T_l \in S_l$ reflects $f(x_l | T_l)$ in Equation 8.9.

### 8.4 Applying the CM and HEM Algorithms to Estimate the APM

In Chapter 4, the HEM algorithm approximately finds a mode of the probability OD matrix. The marginal posterior mode is used as a point estimate of the probability OD matrix. In addition, the CM algorithm was developed to approximately find a mode of the joint posterior distribution of the probability OD matrix and the trip-level OD matrices. The probability OD matrix of the resulting joint mode is also considered a point estimate of the probability OD matrix for comparison purposes.

The CM and HEM algorithms can also be used to estimate the APM since the probability OD matrix and the APM are related to each other. Specifically, if the probability OD matrix $\alpha$ is known, the boarding probability $\rho(r)$ that a passenger randomly selected from the population boards at stop $r$ is given by:

$$\rho(r) = \sum_{q=r+1}^{N_s} \alpha(r,q)$$  \hspace{1cm} (8.11)
The alighting probability $\gamma(r, q)$ that a passenger alights at stop $q$ conditional on that he or she boarded at stop $r$ is given by:

$$\gamma(r, q) = \frac{\alpha(r, q)}{\rho(r)} \quad (8.12)$$

In addition, if both the boarding probability $\rho(r)$ and the alighting probability $\gamma(r, q)$ are known, the probability OD flow $\alpha(r, q)$ is determined uniquely:

$$\alpha(r, q) = \rho(r) \times \gamma(r, q) \quad (8.13)$$

The following shows that the CM algorithm can be used to produce a point estimate of the APM by finding a mode of the joint posterior distribution of the APM and the trip-level OD matrices of Equation 8.9. The CM algorithm approximately finds a mode of the joint posterior distribution of the probability OD matrix and the trip-level OD matrices of Equation 3.15, which is reproduced here for convenience:

$$f(\alpha, T^c \mid \chi^c, z) \propto \prod_{l=1}^{L} \left[ f(x_l \mid T_l) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\alpha(r, q)^{T(r, q)}}{T_l(r, q)!} \right] \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q)^{\gamma(r, q) + \rho(r, q) - 1} \quad (8.14)$$

Based on Equation 8.13, Equation 8.14 could be written as a function of $\gamma(r, q)$ and $\rho(r)$, which is given by:
The exponent of $\rho(r)$ in Term B of Equation 8.15 is arrived at based on the fact that the boarding count at a stop is equal to the sum of passenger flows originating from that stop. In Equation 8.15, Term A is the joint posterior distribution of the APM and the trip-level OD matrix $f(\gamma, T^c | x^c, z)$ (see Equation 8.9). Term B can be seen as the marginal posterior distribution of the boarding probability vector $f(\rho | x^c, z)$. Therefore, Equation 8.15 can be represented by:

$$f(\alpha, T^c | x^c, z) \propto \prod_{l=1}^{L} \left[ \frac{f(x_l | T_l) \times \prod_{r=1}^{N_s} \gamma(r, q)^{T(r,q)} T(r,q)!}{\prod_{r=1}^{N_s} \gamma(r, q)^{T(r,q)} T(r,q)!} \right]$$

$$\times \prod_{r=1}^{N_s} \gamma(r, q)^{z(r,q)+\mu(r,q)-1}$$

**Term A**

$$\times \prod_{r=1}^{N_s} \rho(r)^{\sum_{s=1}^{N_s} h(r,s)+\mu(r,s)-\sum_{s=1}^{N_s} \mu(r,s)-(N_s-r)}$$

**Term B**

Equation 8.15

$$f(\alpha, T^c | x^c, z) = f(\gamma, T^c | x^c, z) \times f(\rho | x^c, z)$$

(8.16)

The CM estimates correspond to a mode of the joint posterior distribution of the probability OD matrix $\alpha$ and the trip-level OD matrices $T^c, f(\alpha, T^c | x^c, z)$. Based on Equation 8.16, if $(\alpha, T^c)$ are the CM estimates, then the combination of APM $\gamma^c$ derived from $\gamma$ based on Equation 8.12 and $T^c$ can be shown to be an estimate of a mode of
\( f(\gamma, T^c | x^c, z) \). And, the boarding probability vector \( \hat{\rho} \) derived from \( \hat{\alpha} \) based on Equation 8.11 can be shown to be an estimate of a mode of \( f(\rho | x^c, z) \).

To prove these claims, suppose they are not true. Then, in a sufficiently small region around \((\hat{\gamma}, \hat{T^c})\), there would be another APM estimate \( \gamma_1 \) and another trip-level OD matrices \( T_1^c \) such that:

\[
f(\gamma_1, T_1^c | x^c, z) > f(\hat{\gamma}, \hat{T^c} | x^c, z) \tag{8.17}
\]

Also, in a sufficiently small region around \( \hat{\rho} \), there would be another boarding probability vector estimate \( \rho_1 \) such that:

\[
f(\rho_1 | x^c, z) > f(\hat{\rho} | x^c, z) \tag{8.18}
\]

Based on Equations 8.17 and 8.18, the following inequality can be obtained:

\[
f(\gamma_1, T_1^c | x^c, z) \times f(\rho_1 | x^c, z) > f(\hat{\gamma}, \hat{T^c} | x^c, z) \times f(\hat{\rho} | x^c, z) \tag{8.19}
\]

Based on Equation 8.16, Equation 8.19 is reduced to:

\[
f(\alpha_1, T_1^c | x^c, z) > f(\hat{\alpha}, \hat{T^c} | x^c, z) \tag{8.20}
\]

where \( \alpha_1 \) is obtained from \( \gamma_1 \) and \( \rho_1 \) based on Equation 8.13. Equation 8.20 contradicts the preposition that \((\hat{\alpha}, \hat{T^c})\) are the CM estimates, a mode of \( f(\alpha, T^c | x^c, z) \). Therefore, the previous claims that the APM \( \hat{\gamma} \) (derived from the CM estimate \( \hat{\alpha} \) ) and \( \hat{T^c} \) are estimates of a mode of \( f(\gamma, T^c | x^c, z) \) and that the boarding probability vector \( \hat{\rho} \) (derived from the CM estimate \( \hat{\alpha} \) ) is an estimate of a mode of \( f(\rho | x^c, z) \) is true.

227
Based on the discussion above, a point estimate of the APM based on the joint posterior distribution of Equation 8.9 can be derived from the CM estimate of the probability OD matrix using Equation 8.12.

In a similar fashion, the HEM algorithm can be used to produce a point estimate of the APM by finding a mode of the marginal posterior distribution of the APM of Equation 8.10. The HEM algorithm approximately finds a mode of the marginal posterior distribution of the probability OD matrix of Equation 3.16, which is reproduced here for convenience:

\[
f(\alpha | x, z) \propto \prod_{l=1}^{L} \left[ \sum_{T_l \in S_l} \prod_{r=1}^{N_l-1} \prod_{q=r+1}^{N_l} \frac{\alpha(r, q) T_l(r, q)!}{T_l(r, q)!} \right] \times \prod_{r=1}^{N_l-1} \prod_{q=r+1}^{N_l} \alpha(r, q)^{z(r,q)+\mu(r,q)-1} \tag{8.21}\]

Based on Equation 8.13, Equation 8.21 could be written as a function of \(\gamma(r, q)\) and \(\rho(r)\), which is given by:

\[
f(\alpha | x, z) \propto \prod_{l=1}^{L} \left[ \sum_{T_l \in S_l} \prod_{r=1}^{N_l-1} \prod_{q=r+1}^{N_l} \frac{\gamma(r, q) \times \rho(r)^{z(r,q)+\mu(r,q)-1}}{T_l(r, q)!} \right] \times \prod_{r=1}^{N_l-1} \prod_{q=r+1}^{N_l} \gamma(r, q)^{z(r,q)+\mu(r,q)-1} = \underbrace{\prod_{l=1}^{L} \sum_{T_l \in S_l} \prod_{r=1}^{N_l-1} \prod_{q=r+1}^{N_l} \gamma(r, q)^{T_l(r,q)} T_l(r, q)!} \times \prod_{r=1}^{N_l-1} \prod_{q=r+1}^{N_l} \gamma(r, q)^{z(r,q)+\mu(r,q)-1} \underbrace{\prod_{r=1}^{N_l-1} \rho(r)^{\mu(r,q)-1} \prod_{r=1}^{N_l-1} \sum_{q=r+1}^{N_l} \mu(r,q)-(N_s-r)} \tag{8.22}\]

Term A

Term B
The exponent of $\rho(r)$ in Term B of Equation 8.22 is arrived at based on the fact that boarding count at a stop is equal to the sum of passenger flows originating from the stop. In Equation 8.22, Term A is the marginal posterior distribution of the APM $f(\gamma | x^c, z)$ (see Equation 8.10). Term B can be seen as the marginal posterior distribution of the boarding probability vector $f(\rho | x^c, z)$. Therefore, Equation 8.22 can be represented by:

$$f(\alpha | x^c, z) = f(\gamma | x^c, z) \times f(\rho | x^c, z)$$  \hspace{1cm} (8.23)

The HEM estimates correspond to a mode of the marginal posterior distribution of the probability OD matrix $\alpha$, $f(\alpha | x^c, z)$. Based on Equation 8.23, if $\hat{\alpha}$ is the HEM estimate, then the APM $\hat{\gamma}$ derived from $\hat{\alpha}$ based on Equation 8.13 can be shown to be an estimate of a mode of $f(\gamma | x^c, z)$. And the boarding probability vector $\hat{\rho}$ derived from $\hat{\alpha}$ based on Equation 8.12 can be shown to be an estimate of a mode of $f(\rho | x^c, z)$.

To prove these claims, suppose they are not true. Then, in a sufficiently small region around $\hat{\gamma}$, there would be another APM estimate $\gamma_1$ such that:

$$f(\gamma_1 | x^c, z) > f(\hat{\gamma} | x^c, z)$$  \hspace{1cm} (8.24)

Also, in a sufficiently small region around $\hat{\rho}$, there would be another boarding probability vector estimate $\rho_1$ such that:

$$f(\rho_1 | x^c, z) > f(\hat{\rho} | x^c, z)$$  \hspace{1cm} (8.25)

Based on Equations 8.24 and 8.25, the following inequality can be obtained:

$$f(\gamma_1 | x^c, z) \times f(\rho_1 | x^c, z) > f(\hat{\gamma} | x^c, z) \times f(\hat{\rho} | x^c, z)$$  \hspace{1cm} (8.26)

Based on Equation 8.23, Equation 8.26 is reduced to:
where $\alpha_1$ is obtained from $\gamma_1$ and $\rho_1$ based on Equation 8.13. Equation 8.27 contradicts the preposition that $\hat{\alpha}$ is the HEM estimate, a mode of $f(\rho | x^c, z)$. Therefore, the previous claims that the APM $\hat{\gamma}$ (derived from the HEM estimate $\hat{\alpha}$) is an estimate of a mode of $f(\gamma | x^c, z)$ and that the boarding probability vector $\hat{\rho}$ (derived from the HEM estimate $\hat{\alpha}$) is an estimate of a mode of $f(\rho | x^c, z)$ is true.

Based on the discussion above, a point estimate of the APM based on the marginal posterior distribution of Equation 8.10 can be derived from the HEM estimate of the probability OD matrix using Equation 8.12.

### 8.5 Numerical Demonstration

#### 8.5.1 Overview

As discussed above, congestion on buses due to bus bunching and high demand results in instability of boarding probabilities, which in turn leads to instability of the probability OD matrix. The APM is more likely to remain stable under congestion on buses. This section evaluates the performance of various APM estimates in the presence and in the absence of congestion on buses. In addition, the performance of various probability OD flow estimates is also evaluated.

The numerical study presented in this section uses the structure of the CLS bus route of the OSU Campus Transit Lab. The distribution of the trip-level total demand and the underlying probability OD matrix are constructed from the manually collected true OD data (the same as those adopted in the numerical study of Section 5.4.3). The
capacity of buses running on the CLS bus route is about 65 passengers. Given this actually capacity and overall demand levels, the capacity in this study is artificially reduced to 20 to simulate congestion on buses.

Various probability OD flow and APM estimates are evaluated in two scenarios. The first scenario does not consider the bus capacity as a constraint. That is, all simulated passengers are able to board resulting in no congestion effect. Therefore, Scenario 1, where passengers can always board buses, is the same as the case considered in the numerical studies of the previous chapters. The second scenario imposes the bus capacity as a constraint where passengers cannot board a bus if there is no space available on the bus (i.e., the number of passengers on the bus is equal to the artificial capacity of 20). Naturally, Scenario 2 is prone to result in congested conditions with some passengers left behind to take subsequent buses.

In addition to the estimates by the CM and HEM algorithms, the probability OD flow and APM estimates produced by the IPF method and based on the generated “true” trip-level OD flows (OD\text{gen}) are also presented. The APM estimates for the IPF and OD\text{gen} methods are derived by dividing the cell values in the estimated probability OD matrix by the corresponding row total following Equation 8.12.

8.5.2 Performance Measures

The $HD^2$, $SSD$ and $Chi^2$ measures were defined in Chapter 5 to quantify the performance of various probability OD flow estimates. The same measures can also be used to quantify the performance of various APM estimates.
The $HD^2$ measure consists of the sum of the squared differences between the square root of the estimated alighting probabilities and the square root of the underlying true alighting probabilities:

$$HD^2 = \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \left( \sqrt{\hat{\gamma}(r,q)} - \sqrt{\gamma(r,q)} \right)^2$$

(8.28)

where, $\hat{\gamma}(r,q)$ is the alighting probability at stop $q$ conditional on boarding at stop $r$ “estimated” from one of the methods considered, and $\gamma(r,q)$ is the underlying, true alighting probability at stop $q$ conditional on boarding at stop $r$.

The $SSD$ is defined as the sum of the squared differences between the estimated alighting probabilities and the underlying true alighting probabilities:

$$SSD = \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} (\sqrt{\hat{\gamma}(r,q)} - \sqrt{\gamma(r,q)})^2$$

(8.29)

The $Chi^2$ statistic is defined as the sum of the ratios of the square difference between the estimated alighting probabilities and the underlying true alighting probabilities to the underlying true alighting probabilities:

$$Chi^2 = \sum_{r=1}^{N_r-1} \sum_{q=r+1}^{N_q} \frac{(\sqrt{\hat{\gamma}(r,q)} - \sqrt{\gamma(r,q)})^2}{\gamma(r,q)}$$

(8.30)

The definition of the $RP$ measure of the APM estimate is the same as that of the probability OD flow estimates presented in Equation 5.5 of Chapter 5.

8.5.3 Results and Discussion

In Scenario 1 (no congestion), based on the assumed probability OD matrix and the assumed distribution of the trip-level total demand, the volume OD flow matrix for
each of 200 bus trips is generated and the resulting trip-level APC counts are derived. Then, the period-level probability OD flow estimates and the APM estimate are produced by the various methods and the performance measures and their corresponding $RP$ measures are calculated for each of the period-level probability OD flow and APM estimates. The above process is repeated 100 times to produce the CDFs of the performance measures and their corresponding $RP$ measures for the various estimates. The null base OD flow matrix is used when producing the estimates from the IPF, CM, and HEM methods.

In Scenario 2 (congestion is present), based on the assumed probability OD matrix and the assumed distribution of the trip-level total demand, the volume OD flow matrix for each of 200 bus trips is generated and the resulting trip-level boarding and alighting counts are derived. Then for each of the 200 assumed bus trips, the observed trip-level APC counts are simulated recursively, beginning from the first boarding stop on the route and continuing downstream until the last alighting stop. At a given bus stop, if the number of generated passengers $m$ wishing to board (as given by the generated volume OD flows) is larger than the number of available spaces $n$ on the bus after passengers alight at that stop, $n$ passengers are selected randomly without replacement from the $m$ passengers (each with its corresponding destination stop) to board the bus, and the left-behind passengers are assumed to take buses on the same route for which boarding and alighting counts are not collected or buses running on other bus routes. The boarding count at the given bus stop, therefore, is set to be $n$ and the alighting counts at
downstream stops are determined accordingly considering the destination stops of the passengers who boarded.

Given the trip-level APC counts that are determined based on the above process, the period-level probability OD matrix and the APM estimates are produced by the various methods and the performance measures and their corresponding $RP$ measures are calculated for each of the period-level probability OD matrix and APM estimates. The above process is repeated 100 times to produce the CDFs of the performance measures and their corresponding $RP$ measures for the various estimates. The trials conducted under Scenarios 1 and 2 are independent, that is, Scenario 2 reflects a new set of generated bus trips to which the left-behind simulation process is applied. The null base OD flow matrix is used when producing the estimates from the IPF, CM, and HEM methods.

Figure 8.1 presents the simulated number of passengers who are not able to board buses by stop after applying the bus capacity constraint in one simulation trial (considering all bus trips for that trial). As can be seen, many passengers who are waiting for buses at bus stops 1 through 5 and 9 through 15 are not able to board buses due to congestion on buses.
Figures 8.2 through 8.4 present the ECDFs of the $HD^2$, $SSD$ and $Chi2$ measures and their corresponding $RP$ measures for estimates based on both the probability OD matrices and the APM, respectively. The results for the probability OD flow estimates and the APM estimates are presented in the left and right columns of Figures 8.2 through 8.4, respectively. The $IPF(\text{NC})$, $CM(\text{NC})$, $HEM(\text{NC})$, and $OD^{\text{gen}}(\text{NC})$ represent the CDFs of the performance or $RP$ measure of the IPF, CM, HEM, and $OD^{\text{gen}}$ estimates, respectively, in the absence of congestion on buses. The $IPF(\text{C})$, $CM(\text{C})$, $HEM(\text{C})$, and $OD^{\text{gen}}(\text{C})$ represent the ECDFs of the performance or $RP$ measure of the IPF, CM, HEM and $OD^{\text{gen}}$ estimates, respectively, in the presence of congestion on buses. Naturally, the $OD^{\text{gen}}(\text{C})$ estimates are based on the generated volume OD flows after simulating the left-
behind process. Since Figures 8.2 through 8.4 show similar patterns, the following
discussion focuses on the results associated with $HD^2$-based $RP$ measure shown Figure 8.2.

Figure 8.2 ECDFs of the $HD^2$ and $HD^2$-based $RP$ Measures of the Probability OD Matrix
and APM Estimates
Figure 8.3 ECDFs of the SSD and SSD-based RP Measures of the Probability OD Matrix and APM Estimates

Figure 8.4 ECDFs of the Chi2 and Chi2-based RP Measures of the Probability OD Matrix and APM Estimates
As can be seen in Figures 8.2, either in the presence or in the absence of congestion on buses, the OD\textsuperscript{gen} method produces the best estimates of the probability OD matrix and the APM. For methods that are based on boarding and alighting counts only, the HEM method produces better estimates of the probability OD matrix and the APM than the CM and IPF methods whether under congestion or not.

In addition, it is shown that congestion on buses would reduce the accuracy of the probability OD flow estimates for all methods. Specifically, the ECDF of the $RP$ measure of the probability OD flow estimates in the presence of congestion on buses is to the left of the ECDF in the absence of congestion on buses for all methods (the IPF, CM, HEM, and OD\textsuperscript{gen} methods).

Most importantly, the CM, HEM and OD\textsuperscript{gen} methods produce very similar APM estimates in the presence and in the absence of congestion on buses. The CDFs of the $RP$ measures under both scenarios are almost overlapped with each of the CM, HEM and OD\textsuperscript{gen} methods. Surprisingly, the performance of the APM estimate based on the IPF method is better in the presence of congestion than in the absence of congestion. It is not clear why such a result is seen. Further investigation is needed to provide a plausible explanation.

In summary, the numerical study demonstrates that the HEM method outperforms the CM and IPF methods whether the purpose is to estimate the probability OD matrix or the APM. Congestion on buses would reduce the accuracy of the probability OD flow estimates. However, similar estimates of the APM can be obtained in the presence and in the absence of congestion on buses based on the CM and HEM methods.
Whether to estimate the probability OD matrix or the APM depends on the understanding of the operation of the bus route of interest. On the one hand, for a bus route whether bus bunching and congestion on buses are rare, estimating the probability OD matrix is desirable since the probability OD matrix is more informative than the APM in describing passenger travel patterns. The APM could be easily derived from the probability OD matrix. However, the inverse is not true (that s, the probability OD matrix cannot be inferred from the APM unless the boarding distribution is known). On the other hand, for a bus route where bus bunching or congestion on buses takes place frequently, estimating the APM is desirable since the probability OD matrix is not stable across bus trips and, therefore, the estimated probability OD matrix would not be representative of the underlying passenger travel patterns on a typical bus trip.
Chapter 9   Summary and Future Research

9.1 Summary

This study presents a statistical formulation and a corresponding algorithm to estimate the probability OD flow matrix on a transit bus route in a homogeneous time-of-day period and investigates their performance. The main inputs of the proposed formulation are boarding and alighting volumes at bus trip level. These counts can be obtained from Automatic Passenger Counter (APC) technologies. OD survey data, if available, can also be incorporated in the proposed formulation.

Unlike previous methodologies that are based on boarding and alighting counts, the proposed methodology takes into account the distribution of the boarding and alighting data across bus trips when estimating the period-level probability OD flow matrix and not only the means of the distribution. As a result, it can take better advantage of the large quantities of boarding and alighting data that are now available from APC technologies implemented on many bus systems.

Although the proposed approach is theoretically attractive, developing an efficient algorithm to solve the estimation problem is challenging. The Expectation Maximization (EM) algorithm could conceivably be used to produce solutions for the proposed formulation by maximizing the marginal posterior likelihood of the probability OD matrix. However, the EM algorithm is computationally expensive. Therefore, a Heuristic Expectation Maximization (HEM) algorithm is developed to provide good
approximations to the solutions produced by the EM algorithm. The HEM algorithm is computationally feasible for realistically long bus routes.

Both the EM and HEM algorithms use the marginal posterior mode of the probability OD matrix as the point estimate of the probability OD matrix. In practice, the joint posterior mode has often been used as a point estimate when it is difficult to determine the marginal posterior mode. In this approach, a joint posterior mode of the probability OD matrix and the volume OD matrices on bus trips for which APC counts were collected is also determined, and the resulting probability OD matrix is used as an alternative point estimate of the probability OD matrix. The Conditional Maximization (CM) algorithm is developed for this purpose and is considered in the investigations throughout this dissertation.

The Iterative Proportional Fitting (IPF) method is also considered for comparison purposes in the investigations of this study. The IPF method is representative of methods in the literature where only the means in the APC data are considered. As discussed in Chapter 2, the IPF method produces the same or similar OD flow estimates as many other methods.

The proposed methodology is evaluated numerically and empirically. The numerical study is carried out on both an illustrative short bus route and an operational bus route that replicates the structure of the Campus Loop South (CLS) bus route operated by the Campus Area Bus Service (CABS) at the Ohio State University (OSU). The underlying true probability OD matrix on the illustrative bus route is assumed arbitrarily, whereas the underlying true probability OD matrix on the operational bus
route is constructed from large-scale manually collected true OD data on the CLS route. The IPF, CM and HEM methods are then evaluated empirically based on APC counts and manually collected true OD data on the CLS bus route. The empirical results confirm the findings in the numerical study.

In the numerical study, the quality of the various estimates is first evaluated on the simplistic, illustrative short bus route. On this route, it is possible to implement the EM algorithm that produces a solution of the proposed formulation. Simulation results demonstrate that the HEM algorithm arrives at good approximate solutions of the EM algorithm. In addition, it is shown that the HEM algorithm produces much better probability OD flow estimates than the CM and IPF methods. Specifically, two probability OD flow matrices, which have the same marginal values but different probability OD flows, are constructed to simulate the trip-level OD matrices and the resulting trip-level APC counts. The HEM algorithm produces probability OD flow estimates that are much closer to the corresponding underlying true values than the IPF and CM methods. In addition, the HEM algorithm can distinguish between the underlying probability OD matrices, whereas the IPF method cannot. The results are consistent with the fact that the IPF method is based on the first moments of the trip-level boarding and alighting counts, since the two assumed probability OD matrices have the same marginal values. The results also demonstrate the value of using the distribution of the trip-level boarding and alighting counts for OD estimation.

The quality of various estimates is then evaluated numerically using the structure of the CLS bus route. The numerical studies also show that the HEM algorithm produces
much better probability OD estimates than the IPF and CM methods. The EM algorithm is not considered in the evaluation on the CLS bus route because it is computationally prohibitive.

Additional dimensions of the probability OD flow estimation problem are also investigated and discussed numerically using the structure of the CLS bus route. Since both the CM and HEM algorithms require a starting value of the probability OD matrix to initialize the iterations, it is investigated whether the CM and HEM algorithms converge to the same or similar probability OD flow estimates when very different starting values are used. The numerical study shows that the overall performance of the CM and HEM algorithms does not depend on the starting values, and that the HEM algorithm outperforms the IPF and CM methods regardless of the starting values. Nevertheless, the numerical study indicates that the convergence of the CM and HEM algorithms depends on the OD pairs. For OD pairs with relatively high probability OD flows, the CM and HEM algorithms converge to values that are closer to the underlying true values than the IPF estimates. On the other hand, for OD pairs with very low probability OD flows, the CM and HEM algorithms may produce estimates that vary by starting values. However, the final estimates are still low for these OD pairs. The inability to converge for OD pairs with very low probability OD flows may not be an issue if the transit applications are more concerned with high probability OD flows than with very low probability OD flows.

In a recursive step of the HEM algorithm, a sequence is followed in order through the stops, beginning with the first alighting stop. A variant is proposed whereby the
sequence begins with the last boarding stop. In a numerical study using the structure of the CLS bus route, both sequences produce similar results.

The sensitivity of OD estimates to the underlying OD structure is also investigated numerically. The null base OD matrix is used for the IPF, CM and HEM methods in producing the probability OD flow estimates. Three probability OD matrices are constructed to represent three different OD structures for passengers travelling on the CLS bus route. The first probability OD matrix is constructed based on the manually collected true OD data. The second probability OD matrix is constructed such that it favors the IPF method using the null base. Specifically, the IPF method is applied to the marginal values of the first probability OD matrix using the null base to produce the second probability OD matrix. The third probability OD matrix is constructed such that it disfavors the IPF method using the null base more than the first probability OD matrix. Specifically, numerous probability OD matrices are generated and the one whose Chi-squared distance from the second probability OD matrix is approximately twice as large as that between the first and second probability OD matrices is chosen as the third probability OD matrix.

A numerical investigation shows that the quality of the results produced by the IPF method is very sensitive to the OD structure, and that the quality of the results produced by the CM algorithm is less sensitive to the OD structure than the IPF method. The HEM algorithm is robust with respect to the OD structure. Under the OD structure that favors the IPF method using the null base, the IPF method outperforms the CM and HEM algorithms, whereas under the other two OD structures, the HEM algorithm
outperforms the IPF and CM methods. In addition, the advantage of the IPF method over the HEM algorithm under the OD structure that favors the IPF method using the null base is much smaller than the advantage of the HEM algorithm over the IPF method under the other two OD structures. This investigation suggests that the HEM algorithm would tend to be preferred over the IPF and CM methods, since the underlying OD structure is unknown in practice.

The sensitivity of OD estimates to the distributional assumption of the trip-level OD flows is also investigated numerically. The CM and HEM algorithms are developed based on the assumption that the trip-level OD flows are multinomially distributed conditional on the probability OD flows and the trip-level total demand. In this investigation, the OD flows on a bus trip are generated independently from Negative Binomial distributions. The numerical investigation shows that the IPF, CM and HEM methods are robust with respect to the distributional assumption of the trip-level OD flows. It also shows that the HEM algorithm is still superior to the IPF and CM methods, even though the distributional assumption of the trip-level OD flows made in the formulation is violated.

The numerical investigations summarized above are carried out based on APC boarding and alighting counts on a large number of bus trips. Survey OD data are not considered in the investigations. In addition, APC counts are assumed to be free of measurement errors. The number of bus trips for which APC counts were collected, the availability of survey OD data and the existence of measurement errors in APC counts could all be important factors that affect the accuracy of OD estimates. The effects of
these three factors are investigated numerically using the structure of the CLS bus routes. The findings are summarized in the following.

When the number of bus trips with APC counts increases, the performance of the IPF method improves initially. However, the continued improvement in the IPF estimates is practically unnoticeable after a relatively small number of trips for which APC data are assumed to be collected. By contrast, the performance of the HEM estimates continues to improve noticeably as the number of bus trips with APC counts increases. Although the IPF estimates are slightly closer than the HEM estimates to the underlying true probability OD flows when only a very few bus trips have boarding and alighting (APC) counts, the HEM estimates become much better than the IPF estimates when APC data are assumed to be available from what can be considered a relatively small number of trips, compared to what can be obtained from APC-equipped buses. The better quality of the HEM estimates, compared to the quality of the IPF estimates, and the ability of the HEM algorithm to take greater advantage of increased APC data are seen to occur whether a null OD matrix or a matrix derived from a simulated OD survey is used as the base matrix in the algorithms and whether the APC counts are free of measurement errors or have large measurement errors. As far as the CM algorithm is concerned, the performance of the CM estimates may improve or deteriorate when the number of bus trips with APC counts increases, depending on the OD survey sample size. When the number of bus trips with APC counts increases, the performance of the CM algorithm continues to improve if the OD survey sample size is small, whereas the performance of the CM algorithm may deteriorate with an increased number of bus trips with APC data if
the OD survey sample size is large.

As expected, increasing the OD survey sample size improves the accuracy of OD estimates for the IPF, CM and HEM methods. When the number of bus trips with APC counts is small (e.g., less than 50), the rate of improvement due to the increase of the OD survey sample size is similar for the IPF, CM and HEM methods. However, when the number of bus trips with APC counts is large (e.g., larger than 100), the performance improvement of the IPF method is faster than that of the CM and HEM algorithms as the OD survey sample size increases. The results indicate that survey OD data are less important for the CM and HEM algorithms than for the IPF method in producing good probability OD flow estimates when large amounts of APC data are available.

Not surprisingly, the presence of measurement errors in APC counts deteriorates the performance of the IPF, CM and HEM methods. However, the impact of measurement errors in APC counts is greater on the CM and HEM estimates than on the IPF estimates. Even in the presence of measurement errors in APC counts, the HEM algorithm is still able to outperform the IPF and CM methods when APC counts from a sufficient number of bus trips are available.

The investigation on the effects of these three factors, the number of bus trips with APC data, the OD survey sample size and the presence of measurement errors in APC counts, on the accuracy of OD estimates reveals that the HEM algorithm can outperform the IPF and CM methods if the number of bus trips with APC counts is larger than a certain value. This value is larger when OD survey sample size is larger and when the magnitude of measurement errors in APC counts is larger. Nevertheless, the number
of bus trips with APC counts that is needed such that the HEM algorithm outperforms the
IPF and CM methods is considered small compared to what can be feasibly obtained by
APC systems.

The HEM algorithm is further evaluated using large amounts of empirical APC
data and large-scale manually collected true OD data on the CLS bus route. The
empirical APC counts are prone to realistic measurement errors. The probability OD flow
matrices estimated by the IPF, CM and HEM methods are evaluated by comparing the
results to the probability OD flow matrix estimated from the manually collected true OD
data. The results obtained in the empirical study are consistent with those found in the
numerical study. Specifically, the IPF, CM and HEM methods are applied to the
empirical APC counts using the null base to estimate the probability OD matrices for the
same period. The empirical study shows that the HEM algorithm does not work very well
when the number of bus trips with APC counts is small. However, when APC counts on
all available bus trips are considered for OD estimation, the HEM algorithm outperforms
the IPF and CM methods.

In addition, estimates are also produced using a base flow matrix constructed
from randomly sampled passengers of the manually collected true OD data. The sampled
data emulate data collected in an onboard OD survey. The assumed OD survey sample
size is considered relatively large compared with the size of the OD matrix for the CLS
bus route. In this case, the empirical study shows that the HEM estimates are worse than
the IPF or CM method when the number of bus trips with APC counts is small. However,
when APC counts on all available bus trips are considered for OD estimation, the HEM
and IPF method perform similarly, and both methods perform better than the CM method.

The results obtained in the empirical study also confirm the findings of the numerical study with respect to the effects of the number of bus trips for which APC counts were collected, the use of OD survey data and the presence of measurement errors in APC counts on the accuracy of OD estimates.

The probability OD matrix is assumed to be stable across bus trips in the given time-of-day period in most of the numerical and empirical investigations conducted in this study. However, under congestion on buses due to bus bunching and high demand, the probability OD matrix is not likely to be stable across bus trips. Under such conditions, it is suggested to estimate an Alighting Probability Matrix (APM), since the APM is more likely to be stable across bus trips in the given time-of-day period. It is demonstrated that the CM and HEM algorithms that were developed to estimate the probability OD matrix can be used to estimate the APM. A numerical study on the CLS bus route is carried out. The numerical study considers estimating the probability OD matrix and the APM in the presence and in the absence of congestion on buses. It is shown numerically that the HEM algorithm produces better probability OD matrix and APM estimates than the IPF and CM methods both in the presence and in the absence of congestion on buses. Congestion on buses reduces the accuracy of the probability OD flow estimates for all methods. However, similar estimates of the APM can be obtained by the CM and HEM algorithms in the presence and in the absence of congestion on buses.
9.2 Future Research

Despite the encouraging results seen, this study motivates future research along several dimensions. Based on the empirical and numerical results obtained in this study, the HEM algorithm is preferred over the CM and IPF methods for OD estimation when quantities of boarding and alighting counts that can be reasonably obtained from APC technologies are available. The number of bus trips with APC counts that is needed such that the HEM algorithm outperforms the CM and IPF methods depends, among other things, on the underlying OD structure, measurement errors in APC counts, and the number of bus stops on the given bus route. For example, if the underlying OD structure is close to the one that favors the IPF method using the null base, the HEM algorithm may not able to outperform the IPF method, even if APC counts have been collected on a large number of bus trips. In addition, large measurement errors in APC counts require more bus trips with APC data if the HEM algorithm is to outperform the IPF method. It is worthwhile to evaluate the performance of these methods under a wider set of conditions than those investigated in this dissertation.

The numerical and empirical studies presented in this dissertation are somewhat unique because a realistic period-level probability OD flow matrix can be constructed from large-scale manually collected true OD data. Nevertheless, the CLS bus route contains only 18 bus stops. Many bus routes have many more stops. Obtaining sufficient true OD data to produce a good depiction of passenger travel patterns for long bus routes would be difficult. Although difficult to implement, it would be worthwhile to conduct investigations similar to those conducted in this study on a long bus route based on a
realistic probability OD matrix. (Appendix F conducts similar investigation as in this study on a bus route with 78 bus stops. The investigation demonstrates that it is feasible to apply the HEM algorithm to bus routes that are longer than what was considered in this study. However, the underlying probability OD matrix is unknown and, therefore, is assumed arbitrarily in the numerical evaluation.)

The CM and HEM algorithms are developed based on the assumption that the APC data are free of measurement errors. In practice, trips that contain APC counts with large errors could be detected and the trips could be eliminated from use in OD estimation. APC counts with sufficiently small errors can first be balanced, as was done in the empirical study conducted, and the CM and HEM algorithms can be applied to the balanced APC counts. However, directly considering measurement error in the formulation and the estimation algorithm would also be desirable. In Chapter 3, incorporating measurement error in the formulation was seen to be straightforward. However, devising an algorithm to solve the estimation problem under such conditions for realistically long routes is likely to be challenging. Appendix B presents the MCMC algorithm that could provide the numerical marginal posterior distribution of the probability OD matrix for the formulation that considers measurement errors in APC counts. However, the MCMC algorithm is computationally expensive and therefore is not feasible to be applied on realistically long bus routes. Developing a computationally efficient algorithm for the formulation that directly considers measurement errors in APC counts would be a fruitful addition.

Two other valuable future research topics are outlined in the following
9.2.1 Uncertainty Measures

The HEM algorithm, as well as the CM and IPF methods, only provides a point estimate of the probability OD matrix. It is desirable to provide an uncertainty measure to indicate the accuracy of the estimate. The customary Bayesian measure of estimate accuracy is the posterior variance of the estimate (Berger, 1985). In Section 9.2.1.1, a procedure is introduced to approximate the posterior variance of the HEM estimate.

In addition, as discussed in Chapter 3, under the uniform Dirichlet prior distribution (the distribution used in this study), the marginal posterior likelihood of the probability OD matrix can be seen to be equivalent to the likelihood of observing the trip-level APC counts and the survey OD data. Therefore, the HEM estimate is also the Maximum Likelihood Estimate (MLE) in the classical frequentist formulation. In the classical frequentist formulation, a bootstrap approach is often used to approximate the distribution of an estimate when it is difficult to obtain the exact distribution of the estimate. The variance of the estimate can be obtained from the bootstrap distribution of the estimate. The bootstrap procedure in terms of the HEM estimate is described in Section 9.2.1.2.

9.2.1.1 Approximating the Posterior Variance of the HEM Estimate

If $\theta$ is a parameter with posterior distribution $\pi(\theta|y)$, where $y$ is the observed data, and $\delta$ is the estimate of $\theta$, the posterior variance of the estimate $\delta$ is given by (Berger, 1985):
\[ \text{var}(\delta \mid y) = \text{var}(\theta \mid y) + (E(\theta \mid y) - \delta)^2 \] (9.1)

where \( \text{var}(\delta \mid y) \) represents the posterior variance of the estimate \( \delta \), \( \text{var}(\theta \mid y) \) represents the posterior variance of the parameter \( \theta \) and \( E(\theta \mid y) \) represents the posterior mean of the parameter \( \theta \). Using Equation 9.1, if the posterior mean is used as the point estimate of \( \theta \), it can be shown that the posterior variance of the estimate \( \delta \) is smallest and equals the posterior variance of the parameter \( \theta \).

The HEM algorithm estimates the marginal posterior mode of the probability OD matrix and uses this estimate as the point estimate of the probability OD matrix. Based on Equation 9.1, the posterior variance of the HEM probability OD flow estimate \( \hat{\alpha}(i, j) \) from stop \( i \) to stop \( j \) is given by:

\[ \text{var}(\hat{\alpha}(i, j) \mid x^c, z) = \text{var}(\alpha(i, j) \mid x^c, z) + (E(\alpha(i, j) \mid x^c, z) - \hat{\alpha}(i, j))^2 \] (9.2)

Because the marginal posterior distribution of the probability OD matrix is not in closed form, it is difficult to calculate the posterior mean and variance of the probability OD flow. However, these two terms may be approximated as described in the following.

In this study it is assumed that trip-level OD flows are multinomially distributed and that the prior distribution of the probability OD matrix is Dirichlet. If the trip-level OD flows on all bus trips are known, the posterior distribution of the underlying period-level probability OD matrix is also Dirichlet, as presented in Equation 4.5 of Chapter 4. Conditional on the trip-level OD matrices \( T^c \) and survey OD data \( z \), the posterior mean of the probability OD flow \( \alpha(i, j) \) is given by (Gelman et al., 2004):

\[ E(\alpha(i, j) \mid T^c, z) = \frac{g(i, j)}{g_0} \] (9.3)
and the posterior variance of the probability OD flow $\alpha(i, j)$ is given by:

$$\text{var}(\alpha(i, j) | \mathcal{C}' , z) \approx \frac{g(i, j) \times (1 - \frac{g(i, j)}{g_0})}{g_0 \times (g_0 + 1)}$$

(9.4)

where,

$$g(i, j) = \sum_{l=1}^{L} T_l(i, j) + z(i, j) + \mu(i, j)$$

(9.5)

$$g_0 = \sum_{i=1}^{N_s - 1} \sum_{j=i+1}^{N_s} g(i, j)$$

(9.6)

The posterior mean and variance of the probability OD flow in Equation 9.2 may be approximated by Equations 9.3 and 9.4, respectively, by replacing the trip-level OD flows with those estimated in the HEM algorithm. Recall that the trip-level OD flows have been estimated in the E-step of the HEM algorithm for determining the final estimates of the period-level probability OD flows. With this substitution, the posterior variance of the HEM estimate may be approximated by:

$$\text{var}(\hat{\alpha}(i, j) | \mathcal{C}' , z) \approx \frac{g(i, j) \times (g_0 - g(i, j))}{g_0^2 \times (g_0 + 1)} + \left(\frac{g(i, j)}{g_0} - \hat{\alpha}(i, j)\right)^2$$

(9.7)

where $g(i, j)$ and $g_0$ are defined in Equations 9.5 and 9.6, respectively and the trip-level OD flow $T_l(i, j)$ in Equation 9.5 is replaced by the trip-level OD flow estimated in the HEM algorithm.

In the current OD estimation problem, the trip-level OD flows are not observed. Instead, the marginal values of the trip-level OD flows are observed in the form of
boarding and alighting counts. The approximation presented in Equation 9.7 assumes that a point estimate of the trip-level OD flow is the true trip-level OD flow. The resulting approximate posterior variance of the estimate could be larger or smaller than the true value.

It is worthwhile to investigate the accuracy of the approximation in Equation 9.7. Such an investigation may be difficult for long bus routes since it is difficult to obtain the true posterior variance of the estimate for long bus routes even when the analysis is conducted numerically. However, the accuracy of the approximation of the posterior variance of the HEM estimate can be evaluated using the illustrative short bus route as used in Chapter 5, since it is feasible to apply the MCMC algorithm developed in Appendix B to derive the numerical posterior distribution of the probability OD matrix. The numerical posterior mean and variance of the underlying probability OD flows can be derived from the numerical posterior distribution. The numerical posterior mean and variance of the underlying probability OD flow and the HEM estimate can then be used as inputs to Equation 9.2 to determine the posterior variance of the HEM estimate. The accuracy of the approximation can be evaluated by comparing the approximate value determined by Equation 9.7 with the posterior variance of the HEM estimate determined when using the procedure above.

It is necessary to point out that the uniform Dirichlet prior may be so strong that the resulting posterior variance of the probability OD flows would be very small for bus
routes serving a large number of bus stops. The hyper-parameters of the uniform Dirichlet prior are one for all feasible OD pairs. Therefore, the sum over the hyper-parameters of the uniform Dirichlet prior equals the number of feasible OD pairs, which is large when the number of bus stops is large. If the sum over the hyper-parameters of the prior Dirichlet is large, the parameter $g_0$ in Equation 9.6 would be large and, as a result, the posterior variance of the probability OD flow in Equation 9.4 would be very small. One possible remedy for this problem could be to aggregate adjacent bus stops so as to reduce the dimension of the resulting OD matrix. In this “stop aggregation” approach, boarding and alighting counts would be aggregated across bus stops according to the aggregation scheme defined for the stops. On the one hand, stop aggregation reduces the effect of the uniform Dirichlet distribution on the posterior variances of the probability OD flows. On the other hand, stop aggregation decreases the resolution of the data, which may reduce the accuracy of the final OD flow estimates. Therefore, it would be worthwhile to investigate the effect of stop aggregation on the accuracy of the final OD flow estimates. The effect of stop aggregation could be evaluated numerically as follows:

1. Simulate APC counts at stop level for each bus trip.

2. Estimate the probability OD matrix from APC counts at stop level and then aggregate the estimated probability OD matrix according to the aggregation scheme defined for the stops to produce the estimated probability OD matrix at the stop aggregation level.

3. Estimate the probability OD matrix from APC counts at the stop aggregation level.
4. Compare the probability OD matrices estimated in Steps (2) and (3) with the underlying true probability OD matrix (at the stop aggregation level). The difference between the performance of the probability OD flow estimates in Steps 2 and 3 would reflect the effect of stop aggregation on the accuracy of the final OD flow estimates.

9.2.1.2 Bootstrap Approximation of the Estimated Probability OD Flow Distribution

The bootstrap method is a computationally-intensive approach used to estimate the properties of an estimator and construct hypothesis tests by sampling from approximation distributions (Rice, 2006). In the bootstrap method, the empirical distribution of the observed data is used to approximate the population distribution. Under the assumption that the observations are independent samples from the same distribution, the bootstrap method samples randomly with replacement from the original dataset many times to construct the distribution of the estimator of interest. The variance of the estimator is then obtained from the resulting bootstrap distribution. Sampling with replacement in the bootstrap method guarantees that each sample is drawn from the same approximate distribution.

Assuming that the original dataset has APC counts on \( L \) bus trips and survey OD data of \( tos \) passengers, the distribution of the probability OD estimate may be obtained according to the bootstrap approach as follows:

For \( h=1, 2, \ldots, H \), apply the following three steps iteratively:

1. Obtain a set of simulated APC counts \( x_h^* \) on \( L \) bus trips by sampling \( L \) bus trips with replacement from the original APC counts.
2. Obtain a set of simulated OD survey data $z_h^*$ by sampling $t_{os}$ passengers with replacement from the original survey OD data.

3. Apply the HEM algorithm to the simulated sets of APC data $x_h^c$ and survey OD data $z_h^*$ to estimate the probability OD flows $\hat{\alpha}_h^*$. 

The estimated probability OD flows produced from the above procedure can be used to approximate the distribution of the probability OD estimates. The variance of the probability OD flow estimate can be approximated by:

$$\text{var}^* (\hat{\alpha}(i, j) | x^c, z) = \frac{\sum_{h=1}^{H} (\hat{\alpha}_h^*(i, j) - \overline{\alpha}^*(i, j))^2}{(H-1)}$$

(9.8)

where,

$$\overline{\alpha}^*(i, j) = \frac{\sum_{h=1}^{H} \hat{\alpha}_h^*(i, j)}{H}$$

(9.9)

This procedure would be a nonparametric version of the bootstrap method. In Steps 1 and 2 of the above procedure, the empirical distribution is used to approximate the population distribution. When the number of bus trips with APC counts is small or when the number of surveyed passengers is small, the approximation would not be good. In this case, a parametric version of the bootstrap method would be needed.

A parametric model describing the relationships among parameters, variables and data in the context of the probability OD flow estimation has been developed in Chapter 3. If the distributional assumptions of this model are valid, it could be used in the following approach:

1. Apply the HEM algorithm to the original sets of APC data $x^c$ and survey OD data $z$ to
estimate the probability OD flows $\hat{\alpha}$.

2. Assuming the estimated probability OD flows $\hat{\alpha}$ in Step 1 are the true underlying probability OD flows, for $h=1, 2, \ldots H$, apply the following three steps iteratively
   
   a. Simulate APC counts $x_h^{c^*}$ on $L$ bus trips based on the model structure presented in Figure 3.1. Specifically, sample $L$ bus trips with replacement from the original APC data and then use the trip-level total demands for the $L$ sampled bus trips and the assumed underlying probability OD matrix to simulate APC counts on $L$ bus trips.
   
   b. Simulate survey OD data $z_h^*$ for $tos$ passengers based on the model structure presented in Figure 3.1.
   
   c. Apply the HEM algorithm to the simulated sets of APC data $x_h^{c^*}$ and survey OD data $z_h^*$ to estimate the probability OD flows $\hat{\alpha}_h^*$.

The estimated probability OD flows produced from the above procedure can be used to approximate the distribution of the probability OD flow estimate. The variance of the probability OD flow estimate can be approximated by:

$$\text{var}^*(\hat{\alpha}(i,j) | x^c, z) = \sum_{h=1}^{H} (\hat{\alpha}_h^*(i,j) - \hat{\alpha}(i,j))^2 / (H - 1) \quad (9.10)$$

If the distributional assumptions are valid, the parametric version of the bootstrap method could be applied even when the number of bus trips with APC counts is small or when the number of surveyed passengers is small.

It is worthwhile to point out that the posterior variance of the estimate in Section 9.2.1.1 and the variance of the estimate in Section 9.2.1.2 are not comparable, since the
interpretations of the posterior variance of the estimate and the variance of the estimate are different. In Bayesian framework, the probability OD matrix is random, and the observed data are fixed. The posterior variance of the estimate reflects the subjective belief about the accuracy of the estimate after observing the data. The subjective belief is built on the prior information and the information obtained from the data. In frequentist framework, the probability OD matrix is deterministic and the observed data are random. The variance of the estimate reflects the variation of the data. Typically the posterior variance of the estimate is smaller than the variance of the estimate because of the effect of the prior belief in the posterior variance of the estimate.

9.2.2 Extension to a Two-Route with One Transfer Network

Although the HEM algorithm is developed for OD estimation at bus route level, they could be extended to estimate OD matrices for a two-route with one transfer network. OD estimation on a two-route with one transfer network has been discussed in Cui (2006) and Zhang (2008). In such a network, all passengers make at most one transfer, and the transfer occurs at the same bus stop. This section outlines the procedures of applying the HEM algorithm to estimate OD matrices for a two-route with one transfer network. It could be useful to investigate the performance of doing so.

For illustration, consider a simple hypothetical two-route with one transfer network as shown in Figure 9.1.
In the network illustrated in Figure 9.1, buses on Route 1 run from stops 11 to 15, and buses on Route 2 run from stops 21 to 25. Passengers originating from stop 11 or 12 on Route 1 may alight at stop 13 on Route 1, board at stop 23 on Route 2 and alight at stop 24 or 25 on Route 2. Stops 13 and 23 may be located at the same physical bus stop or at two different stops that are close to each other. In the following, Route 1 is referred to as transfer-producing route, and Route 2 is referred to as transfer-receiving route.

An OD matrix for a two-route with one transfer network can be decomposed into three sub-matrices (Cui, 2006), as illustrated in Table 9.1. Table 9.1 only considers one-direction travel on Routes 1 and 2 and transfer from Route 1 to Route 2. The sub-matrices located on the diagonal are referred to as non-transfer OD matrices, which summarize passenger flows between stop pairs on the same bus route. The sub-matrix located on the
off-diagonal is referred to as transfer OD matrix. The transfer OD matrix summarizes passenger flows from stops on Route 1 to stops on Route 2.

<table>
<thead>
<tr>
<th>Origins on Route 1</th>
<th>Destination on Route 1</th>
<th>Destination on Route 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-transfer OD matrix on Route 1</td>
<td>Transfer OD matrix from Route 1 to Route 2</td>
<td></td>
</tr>
<tr>
<td>Origins on Route 2</td>
<td>-</td>
<td>Non-transfer OD matrix on Route 2</td>
</tr>
</tbody>
</table>

Table 9.1 Structure of the Network OD Matrix for a Two-route with One Transfer Network (Cui, 2006)

Cui (2006) developed two methods to estimate OD matrices for a two-route with one transfer network. Besides APC boarding and alighting counts on both bus routes, the total transfer flow from one route to the other is assumed to be known in both methods. The methods and results presented in Cui (2006) are based on aggregated APC boarding and alighting counts, where the aggregation is performed by stop, over bus trips in the same time-of-day period. Zhang (2008) applied these two methods to empirical APC data on a pair of bus trips. A pair of bus trips consists of a bus trip on the transfer-producing route and the first bus trip on the transfer-receiving route departing after the transfer-producing bus trip arrived. Zhang (2008) developed an algorithm to identify the pair of bus trips from the time stamp information in APC data. The total transfer flow for one pair of bus trips is estimated by the smaller value of the alighting count at the transfer stop on the transfer-producing route and the boarding count at the transfer stop on the transfer-receiving route. Unlike Cui (2006), the transfer OD matrix in Zhang (2008) is represented in the form of probability. The probability transfer OD matrix describes the
probabilities that a randomly selected transfer passenger travels from specific stops on the transfer-producing route to specific stops on the transfer-receiving route.

Assuming APC counts and the total transfer flows on \( L \) pairs of bus trips are known, the procedure of applying of the HEM algorithm on a two-route with one transfer network is similar to that on a single bus route. For example, consider one pair of bus trips in the hypothesized network presented in Figure 9.1. A set of assumed boarding and alighting counts and the transfer total flows on the considered pair of bus trips is presented in Table 9.2.

<table>
<thead>
<tr>
<th>Stops on Route 1</th>
<th>Boarding</th>
<th>Alighting</th>
<th>Stops on Route 2</th>
<th>Boarding</th>
<th>Alighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>26</td>
<td>0</td>
<td>21</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>5</td>
<td>22</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>13 (transfer stop)</td>
<td>8</td>
<td>34</td>
<td>23 (transfer stop)</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>8</td>
<td>24</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>10</td>
<td>25</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>57</td>
<td>Total</td>
<td>62</td>
<td>62</td>
</tr>
<tr>
<td>Total transfer flows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
</tr>
</tbody>
</table>

Table 9.2 Boarding and Alighting Counts and Total Transfer Flows for One Pair of Bus Trips in the Hypothetical Network Presented in Figure 9.1

The boarding and alighting counts and the total transfer flows presented in Table 9.2 can be used to form the marginal values of the OD matrix for the given pair of bus trips, as illustrated in Table 9.3. In Table 9.3, dashed lines in cells are used to represent infeasible OD pairs. The sub-matrix located in the upper diagonal describes passenger flows that use OD pairs entirely on Route 1. The sub-matrix located in the lower diagonal describes passenger flows that use OD pairs entirely on Route 2. The sub-matrix located
in the upper off-diagonal describes passenger flows transferring from stops on Route 1 to stops on Route 2. The boarding and alighting counts are listed in the last column and last row of Table 9.3, respectively. For example, the boarding count at stop 11 on Route 1 is 26, which is presented in the last column of the row corresponding to stop 11. The boarding count at stop 23 on Route 2 is 26, among which 25 passengers transfer from Route 1. Only one passenger (i.e., 26-25 = 1) actually originates from stop 23. Therefore, the boarding count at stop 23 is set to be 1 in Table 9.3, which is presented in the last column of the row corresponding to stop 23. The alighting count at stop 13 on Route 1 is 34, from which 25 passengers transfer to Route 2. Only 9 passengers (i.e., 34-25 = 9) actually are destined for stop 13. Therefore, the alighting count at stop 13 is set to be 9 (i.e., 34-25) in Table 9.3, which is presented in the last row of the column corresponding to stop 13.
Table 9.3 Representation of Boarding and Alighting Counts and the Number of Transfer Passengers as the Marginal Values of an OD Matrix

<table>
<thead>
<tr>
<th>Ori\Des.</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>Route 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
</tbody>
</table>

The boarding and alighting counts and the total transfer flows for the other pairs of bus trips could be rearranged in the same way. After arranging the data according to this structure for all L bus trips, the HEM algorithm can be applied to the sets of marginal values of the OD matrices to estimate the probability OD flow matrix for the two-route with one transfer network.

In practice, it is difficult to obtain the total transfer flow for each pair of bus trips. In the absence of such information, the transfer OD matrix in the form of probability (i.e., probability transfer OD matrix) can also be estimated under the assumption that passenger travel behavior is the same for transfer and non-transfer passengers. The procedure described subsequently is similar to the Proportional Distribution (PD) method.
proposed by Cui (2006).

When using the approach proposed here, the HEM algorithm would first be applied to the trip-level APC counts to estimate the probability OD matrices for both the transfer-producing and transfer-receiving routes. Then, the probabilities in the column corresponding to the transfer stop in the estimated probability OD matrix of the transfer-producing route would be normalized by dividing the cell values by the column total. These normalized values would represent the probabilities that a transfer passenger originates from the specified bus stops and are referred to as origin probabilities. The probabilities in the row corresponding to the transfer stop in the estimated probability OD matrix of the transfer-receiving route would also be normalized by dividing the cell values by the row total. These normalized values would represent the probabilities that a transfer passenger is destined for specific bus stops and are referred to as destination probabilities. Given the origin and destination probabilities, the transfer probability OD matrix can be estimated by the IPF or HEM method using the null base OD matrix or a base OD matrix constructed from some other source of information. The difference between the transfer probability OD matrices estimated by the HEM and IPF methods is expected to be small, since only one set of marginal values of the transfer probability OD matrix is used.

For example, for the network in Figure 9.1, Table 9.4 shows the estimated route-level probability OD matrices for Routes 1 and 2. On Route 1, the probabilities that a transfer passenger originates from stops 11 and 12 are 0.25 (i.e., 0.1 / (0.1+0.3)) and 0.75 (i.e., 0.3 / (0.1+0.3)), respectively. On Route 2, the probabilities that a transfer passenger
is destined for stops 24 and 25 are 0.4 (i.e., 0.2 / (0.2+0.3)) and 0.6 (i.e., 0.3 / (0.2+0.3)), respectively. The origin and destination probabilities are the marginal values of the transfer probability OD matrix, which are presented in Table 9.5. With the marginal values, the transfer probability OD matrix can be estimated by the IPF or HEM method.

<table>
<thead>
<tr>
<th>Ori.\Des.</th>
<th>11</th>
<th>12</th>
<th>13 (transfer stop)</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>-</td>
<td>0.15</td>
<td>0.1</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>0.3</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.02</td>
<td>0.25</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.02</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ori.\Des.</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>-</td>
<td>0.02</td>
<td>0.03</td>
<td>0.1</td>
<td>0.03</td>
</tr>
<tr>
<td>22</td>
<td>-</td>
<td>-</td>
<td>0.04</td>
<td>0.1</td>
<td>0.03</td>
</tr>
<tr>
<td>23 (transfer stop)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>24</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.15</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 9.4 Estimated Route-level Probability OD Matrices

<table>
<thead>
<tr>
<th></th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 9.5 Estimate the Transfer Probability OD Matrix from the Origin and Destination Probabilities

In the procedure presented above, route-level probability OD matrices are estimated by the HEM method developed in this study. The IPF or HEM method could be applied to the marginal values of the transfer probability OD matrix derived from the
estimated route-level probability OD matrices to estimate the transfer probability OD matrix. In the future, it would be useful to quantify the impact of the accuracy of the route-level probability OD flow estimates on the transfer probability OD flow estimate.

Although further investigations are needed, it appears that the formulation and the HEM algorithm proposed in this study provide a promising approach to take better advantage of the large quantities of boarding and alighting data that are now being made available from APC-equipped systems. Transit agencies have been implementing APC technologies for a variety of applications, mostly centered on load and utilization measurements. The results of this study suggest that passenger OD flow estimation could be a promising additional use of available APC data.
Reference


269


Appendix A  Advantage of the Proposed Approach

A.1  Introduction

To highlight the difference between the proposed approach and existing OD estimation methods, the proposed approach is compared with four existing methods: the Iterative Proportional Fitting (IPF) method (Ben-Akiva, 1987; Ben-Akiva et al., 1985), the Poisson Maximum Likelihood Estimator (PMLE) (Spiess, 1987), Vardi’s formulation (Vardi, 1996), and the Normal Maximum Likelihood Estimator (NMLE) (Cascetta and Nguyen, 1988). All methods are applied to the trip-level boarding and alighting counts to estimate the probability OD matrix. The boarding and alighting counts are assumed to be free of measurement errors.

Table A.1 presents the assumptions of the five methods under comparison. The IPF method does not rely on distributional assumption. In addition, notice that the PMLE method and Vardi’s formulation have the same distributional assumption about the trip-level OD flows. However, the PMLE method additionally assumes that the trip-level boarding and alighting counts across bus stops and bus trips are independently.
<table>
<thead>
<tr>
<th>Model</th>
<th>Distributional Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPF</td>
<td>No distributional assumptions</td>
</tr>
</tbody>
</table>
| PMLE          | Trip-level OD flows are independently Poisson distributed conditional on the expected OD flows  
|               | Trip-level boarding and alighting counts are independently across bus stops and bus trips |
| Vardi’s formulation | Trip-level OD flows are independently Poisson distributed conditional on the expected OD flows |
| NMLE          | Trip-level OD flows are multivariate Normal distributed conditional on the expected OD flows and the variance-covariance matrix |
| Proposed      | Trip-level OD flows are Multinomial distributed conditional on the probability OD flows and the trip-level total demand |

Table A.1 Assumptions of Five Methods under Comparison

In the following, it will be shown that the IPF, PMLE and NMLE methods use strictly the first moments of the trip-level boarding and alighting data to estimate the probability OD matrix, and that Vardi’s formulation and the proposed approach use the distribution of the trip-level boarding and alighting data to estimate the probability OD matrix.

The IPF, PMLE and NMLE methods and Vardi’s formulation were originally used to estimate the expected OD flows. By definition, the probability OD flows can be obtained by dividing the expected trip-level OD flows by the expected trip-level total demand, which is given by:

\[ \alpha(r,q) = \frac{E(T(r,q))}{E(tot)} \]  

(A.1)

The subscript that denotes bus trips is omitted in Equation A.1 since the trip-level OD flow and the trip-level total demand are referred to a general bus trip.
A.2 Method Comparison

**IPF method**

The IPF method could be seen as an algorithm for an optimization problem as follows (Ben-Akiva, 1987):

\[
\min_{\alpha \geq 0} \sum_{i=1}^{N_s-1} \sum_{j=i+1}^{N_s} \alpha(i, j) \times \log \left( \frac{\alpha(i, j)}{t_0(i, j)} \right)
\]

s.t.

\[
\sum_{i=1}^{N_s-1} \alpha(i, j) = \frac{\sum_{l=1}^{L} a_l(j)}{\sum_{l=1}^{L} \text{tot}_l} \quad j = 2, \ldots, N_s
\]

\[
\sum_{j=i+1}^{N_s} \alpha(i, j) = \frac{\sum_{l=1}^{L} b_l(i)}{\sum_{l=1}^{L} \text{tot}_l} \quad i = 1, \ldots, N_s - 1
\]

where \(t_0(i,j)\) is passenger flow from stop \(i\) to stop \(j\) in the base OD matrix. The other notations have been defined in Table 3.1 of Chapter 3.

Information in boarding and alighting counts that is used by the IPF method is reflected in the constraints of the optimization problem above. As can be seen in Equations A.2b-A.2c, given boarding and alighting counts on \(L\) bus trips, the IPF method aggregates boarding and alighting counts by stop across bus trips first, and then uses the ratio of the aggregated boarding count to the aggregated total demand and the ratio of the aggregated alighting count to the aggregated total demand for each bus stop as the constraints the marginal values of the true probability OD matrix should satisfy. Therefore, the IPF probability OD estimates are based on the aggregated boarding and alighting counts or the sample means (i.e., the first moments) of the trip-level boarding and alighting data.
**PMLE method**

Based on the assumption that the trip-level OD flows are independently Poisson distributed, it can be shown that the trip-level boarding and alighting counts across bus stops are multivariate Poisson distributed (Karlis and Meligkotsidou, 2005). Since the multivariate Poisson distribution is analytically intractable, the PMLE method additionally assumes that the trip-level boarding and alighting counts across bus stops and bus trips are independent. Based on this assumption, conditional on the probability OD flow matrix $\alpha$ and the expected trip-level total demand $E(tot)$, the distribution of boarding and alighting counts on bus trip $l$ can be shown to be given by:

$$f(x_i | \alpha, E(tot)) \propto \prod_{r=1}^{N_r} e^{-E(tot)\alpha(r, \cdot)} \times (E(tot) \times \alpha(r, \cdot))^{b_r}$$

$$\times \prod_{q=2}^{N_s} e^{-E(tot)\alpha(\cdot, q)} \times (E(tot) \times \alpha(\cdot, q))^{a_q}$$

(A.3)

where $\alpha(r, \cdot) = \sum_{q=r+1}^{N_s} \alpha(r, q)$ and $\alpha(\cdot, q) = \sum_{r=1}^{q-1} \alpha(r, q)$. Term A in Equation A.3 represents the likelihood of observing boarding count $b_r$ at bus stop $r$ on bus trip $l$ conditional on the probability OD matrix $\alpha$ and the expected trip-level total demand $E(tot)$, and Term B in Equation A.3 represents the likelihood of observing alighting count $a_q$ at bus stop $q$ on bus trip $l$ conditional on the probability OD matrix $\alpha$ and the expected trip-level total demand $E(tot)$. Equation A.3 reflects the assumption that the trip-level boarding and
boarding and alighting counts are independent across bus stops. Based on the fact that \( \sum_{r=1}^{N_s-1} \alpha(r, \cdot) = 1 \),

\[
\sum_{q=2}^{N_s} \alpha(\cdot, q) = 1, \quad \text{and} \quad \sum_{r=1}^{N_s-1} \sum_{q=2}^{N_s} a_i(q) = \text{tot}_i,
\]

Equation A.3 could be written by:

\[
f(x_i | \alpha, E(\text{tot})) \propto \prod_{r=1}^{N_s-1} \alpha(r, \cdot)^{b_i(r)} \times \prod_{q=2}^{N_s} \alpha(\cdot, q)^{a_i(q)}
\times e^{-2 \times E(\text{tot})} \times E(\text{tot})^{2 \times \text{tot}_i}
\]

(A.4)

In addition, based on the assumption that boarding and alighting counts across bus trips are independent, the likelihood of observing boarding and alighting counts on all bus trips \( L \) is given by:

\[
f(x^c | \alpha, E(\text{tot})) \propto \prod_{r=1}^{N_s-1} \sum_{i=1}^{L} b_i(r) \times \prod_{q=2}^{N_s} \sum_{i=1}^{L} a_i(q)
\times e^{-2 \times L \times E(\text{tot})} \times E(\text{tot})^{2 \times \sum_{i=1}^{L} \text{tot}_i}
\]

(A.5)

It can be seen in Equation A.5 that the PMLE method aggregates boarding and alighting counts by stop across bus trips in the likelihood of observing boarding and alighting counts on all bus trips. As a result, the PMLE method only uses information in the first moments of the trip-level boarding and alighting data to estimate the probability OD matrix.

**NMLE method**

For the NMLE method, for convenience, the probability and trip-level OD matrices are rearranged in the form of vectors. The relationship between the trip-level OD vector and the boarding and alighting counts on bus trip \( l \) is represented by \( x_l = \tau \times T_l \).

Based on the assumption that the trip-level OD flows are multivariate Normal distributed,
it can be shown that the trip-level boarding and alighting counts are also multivariate Normal distributed:

\[ f(x_l | \alpha, E(tot)) \sim MVN(E(tot) \times \tau \times \alpha, \Sigma) \]  

(A.6)

where \( \Sigma \) is the variance-covariance matrix of the trip-level boarding and alighting counts, which quantifies the variation of the trip-level boarding and alighting counts across bus trips. Theoretically, \( \Sigma = \tau \times \Phi \times \tau' \) (\( \Phi \) is the variance-covariance matrix of the trip-level OD flows). However, the variance-covariance matrix of the trip-level OD flows \( \Phi \) is usually unknown. Therefore, \( \Sigma \) is often estimated exogenously and then is assumed to be known in the NMLE method.

Based on Equation A.6, the sample means of the trip-level boarding and alighting counts are also multivariate Normal distributed:

\[ f(\bar{x} | \alpha, E(tot)) \sim MVN(E(tot) \times \tau \times \alpha, \Sigma_L) \]  

(A.7)

where \( \bar{x} \) is the sample means of the trip-level boarding and alighting counts across bus trips. The log likelihood of the trip-level boarding and alighting counts for Equation A.7 is proportional to:

\[ \log(f(\bar{x} | \alpha, E(tot))) \propto -L \times (\bar{x} - \tau \times E(tot) \times \alpha)' \times \Sigma^{-1} \times (\bar{x} - \tau \times E(tot) \times \alpha) \]  

(A.8)

The variance-covariance matrix of the trip-level boarding and alighting counts \( \Sigma \) is estimated exogenously and assumed to be known in Equation A.8. Although the accuracy of the estimate of the variance-covariance matrix \( \Sigma \) would affect the final OD estimates, the NMLE method relies on the sample means of the trip-level boarding and alighting counts to estimate the probability OD flows as reflected in Equation A.8.
In Vardi’s formulation, conditional on the probability OD flow matrix $\alpha$ and the expected trip-level total demand $E(tot)$, the distribution of boarding and alighting counts on bus trip $l$ can be shown to be given by:

$$f(x_l | \alpha, E(tot)) = \sum_{T_l \in S_l} \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{e^{-E(tot)\times \alpha(r,q)}}{T_l(r,q)!} \times (E(tot) \times \alpha(r,q))^T_l(r,q) \times \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q)^{T_l(r,q)}}{T_l(r,q)!}$$

$$= e^{-E(tot)} \times E(tot)^{tot} \times \sum_{T_l \in S_l} \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q)^{T_l(r,q)}}{T_l(r,q)!}$$

In Equation A.9, the term inside the summation in the right hand side of the first equality sign represents the likelihood of observing an OD matrix $T_l$ on bus trip $l$ conditional on the probability OD flow matrix $\alpha$ and the expected trip-level total demand $E(tot)$, where the OD matrix $T_l$ satisfies the constraints of the boarding and alighting counts $x_l$ as reflected in the following Equations:

$$\sum_{r=1}^{q-1} T_l(r,q) = a_l(q) \quad q = 2, ..., N_s \quad (A.10a)$$

$$\sum_{q=r+1}^{N_q} T_l(r,q) = b_l(r) \quad r = 1, ..., N_s - 1 \quad (A.10b)$$

In addition, $T_l \in S_l$ in Equation A.9 represents all OD matrices $T_l$ satisfy the constraints of the true boarding and alighting count $x_l$ as reflected by Equations A.10a and A.10b. The likelihood of observing boarding and alighting counts $x_l$ on bus trip $l$ is obtained by summing the likelihoods of all feasible OD matrices $T_l$ that satisfy the constraints of the boarding and alighting counts $x_l$. Based on the assumption that boarding and alighting counts across bus trips are independent, the likelihood of observing boarding and
alighting counts on all bus trips is obtained by the product of the likelihoods of observing the trip-level boarding and alighting counts across all bus trips:

$$f(x^c | \alpha, E(tot)) = e^{-L_x E(tot)} \times E(tot)^{L_x} \times \prod_{l=1}^{L} \sum_{T_l \in S_l} \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q) T_l(r,q)}{T_l(r,q)!}$$  \hspace{1cm} (A.11)

As seen in Equation A.11, Vardi’s formulation is based on the distribution of the trip-level boarding and alighting data and not just the first moments of the trip-level boarding and alighting data.

Proposed approach

In the proposed approach, conditional on the probability OD flow matrix \( \alpha \), the distribution of boarding and alighting counts on bus trip \( l \) can be shown to be given by:

$$f(x_l | \alpha) \propto \sum_{T_l \in S_l} \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_q} \frac{\alpha(r,q) T_l(r,q)}{T_l(r,q)!}$$  \hspace{1cm} (A.12)

In Equation A.12, the term inside the summation represents the likelihood of observing an OD matrix \( T_l \) on bus trip \( l \) conditional on the probability OD flow matrix \( \alpha \), where the OD matrix \( T_l \) satisfies the boarding and alighting counts \( x_l \) as reflected in Equations A.10a and A.10b. The likelihood of observing boarding and alighting counts \( x_l \) on bus trip \( l \) is obtained by summing the likelihoods of all possible OD matrices \( T_l \) that satisfy boarding and alighting counts \( x_l \). Based on the assumption that boarding and alighting counts across bus trips are independent, the likelihood of observing boarding and alighting counts on all bus trips is obtained by the product of the likelihoods of observing the trip-level boarding and alighting counts across all bus trips:
As seen in Equation A.13, the proposed approach is based on the distribution of the trip-level boarding and alighting data and not just the first moments of the trip-level boarding and alighting data.

Theoretically, distribution is more informative than the first moments. Conceivably, all moments of the trip-level boarding and alighting data could be utilized if OD estimation is based on the full distribution of the trip-level boarding and alighting data. To illustrate the advantage of using the full distribution of the trip-level boarding and alighting data over using the first moments of the trip-level boarding and alighting data for OD estimation, the relationship between the probability OD flows and the first and second moments of the trip-level boarding and alighting counts are presented in the following.

A.3 Relationship between the Probability OD Flows and the First and Second Moments of the Trip-level Boarding and Alighting Counts

A.3.1 First Moments of the Trip-level Boarding and Alighting Counts

The trip-level boarding and alighting counts are linear functions of the trip-level OD flows (see Equations A.10a and A.10b). Therefore, the expectations of the trip-level boarding and the alighting counts are also linear functions of the expectations of the trip-level OD flows. Based on the relationship between the mean of the trip-level OD flow and the probability OD flow as capture by Equation A.1, the relationship between the

\[ f(x^c \mid \alpha) \propto \prod_{l=1}^{L} \left( \sum_{\tau_l} \prod_{r=1}^{N_r} \prod_{q=r+1}^{N_r} \frac{\alpha(r,q)^T_l(r,q)!}{T_l(r,q)!} \right) \]  

(A.13)
probability OD flows and the first moments of the trip-level boarding and alighting counts is given by:

\[
\sum_{q=r+1}^{N_s} \alpha(r, q) = \frac{E(b(r))}{E(tot)} \quad r = 1, \ldots, N_s - 1
\]  \hspace{1cm} (A.14a)

\[
\sum_{r=1}^{q-1} \alpha(r, q) = \frac{E(a(q))}{E(tot)} \quad q = 2, \ldots, N_s
\]  \hspace{1cm} (A.14b)

Equations A.14a and A.14b do not depend on any model assumptions regarding the distribution of the trip-level OD flows and are reflected in all OD estimation methods. Totally, there are \(2 \times N_s - 2\) equations in Equations A.14a and A.14b. However, only \(2 \times N_s - 3\) equations are independent since the total boarding count equals the total alighting count. In addition, the number of feasible OD pairs is \((N_s - 1) \times N_s/2\). The number of feasible OD pairs would be larger than the number of independent equations as long as the number of bus stops is larger than three. As a result, multiple probability OD flow matrices can satisfy Equations A.14a and A.14b, which makes it necessary for most traditional methods to bring additional information, such as survey OD data, in OD estimation.

In the following, it will be shown that the second moments of the trip-level boarding and alighting counts are also informative for Vardi’s formulation and the proposed approach. Doing so illustrates how more generally the full distribution of the trip-level boarding and alighting counts could provide additional information for OD flow estimation.
A.3.2 Second Moments of the trip-level Boarding and Alighting Counts

For Vardi’s formulation, the covariance between the boarding count \( b(r) \) at stop \( r \) and the alighting count \( a(q) \) at stop \( q \) equals the expected OD flow from stop \( r \) to stop \( q \) (Vardi, 1996). The expected trip-level OD flow as a function of the probability OD flow has been given in Equation A.1. Therefore, the covariance between the boarding count at stop \( r \) and the alighting count at stop \( q \) is given by:

\[
\text{cov}(b(r), a(q)) = E(T(r, q)) = E(tot) \times \alpha(r,q)
\]  

(A.15)

Based on Equation A.15, the probability OD flow \( \alpha(r,q) \) can be written as a function of the covariance \( \text{cov}(b(r), a(q)) \) between the boarding count at stop \( r \) and the alighting count at stop \( q \) and the expected trip-level total demand \( E(tot) \):

\[
\alpha(r,q) = \frac{\text{cov}(b(r), a(q))}{E(tot)}
\]  

(A.16)

For the proposed approach, the trip-level OD flows are assumed to be multinomially distributed conditional on the trip-level total demand \( tot \) and the probability OD matrix \( \alpha \). The probability OD matrix \( \alpha \) is assumed to be stable across bus trips and the trip-level total demand \( tot \) varies across bus trips. According to the law of total covariance (Rice 2006), the covariance between the boarding count \( b(r) \) at stop \( r \) and the alighting count \( a(q) \) at stop \( q \) is given by:

\[
\text{cov}(b(r), a(q)) = E(\text{cov}(b(r), a(q) | tot)) + \text{cov}(E(b(r) | tot), E(a(q) | tot))
\]  

(A.17)

Each component in the right hand side of Equation A.17 is developed in the following.

Conditional on the trip-level total demand \( tot \), the covariance of the boarding count \( b(r) \) at stop \( r \) and the alighting count \( a(q) \) at stop \( q \) is given by:
\[
\text{cov}(b(r), a(q) \mid \text{tot}) = \text{cov}\left( \sum_{s=r+1}^{N_r} T(r,s) , \sum_{n=1}^{q-1} T(n,q) \mid \text{tot} \right) \\
= \sum_{s=r+1}^{N_r} \sum_{n=1}^{q-1} \text{cov}(T(r,s), T(n,q) \mid \text{tot}) 
\]

(A.18)

Based on the Multinomial distribution assumption, conditional on the trip-level total demand \( \text{tot} \), the covariance of the trip-level OD flow \( T(r,s) \) from stop \( r \) to stop \( s \) and the trip-level OD flow \( T(n,q) \) from stop \( n \) to stop \( q \) is given by:

\[
\text{cov}(T(r,s), T(n,q) \mid \text{tot}) = -\text{tot} \times \alpha(r,s) \times \alpha(n,q) 
\]

(A.19)

And the variance of the trip-level OD flow \( T(r,q) \) from stop \( r \) to stop \( q \) is given by:

\[
\text{var}(T(r,q) \mid \text{tot}) = \text{cov}(T(r,q), T(r,q) \mid \text{tot}) = \text{tot} \times \alpha(r,q) \times (1 - \alpha(r,q)) 
\]

(A.20)

Substituting Equations A.19 and A.20 into Equation A.18 leads to:

\[
\text{cov}(b(r), a(q) \mid \text{tot}) = \sum_{s=r+1}^{N_r} \sum_{n=1}^{q-1} \text{cov}(T(r,s), T(n,q) \mid \text{tot}) \\
= -\text{tot} \times \sum_{s=r+1}^{N_r} \sum_{n=1}^{q-1} \alpha(r,s) \times \alpha(n,q) + \text{tot} \times \alpha^2(r,q) \\
+ \text{var}(T(r,q) \mid \text{tot}) \\
= -\text{tot} \times \sum_{s=r+1}^{N_r} \sum_{n=1}^{q-1} \alpha(r,s) \times \alpha(n,q) + \text{tot} \times \alpha^2(r,q) \\
+ \text{tot} \times \alpha(r,q) \times (1 - \alpha(r,q)) \\
= -\text{tot} \times \alpha(r,) \times \alpha(, q) + \text{tot} \times \alpha^2(r,q) \\
+ \text{tot} \times \alpha(r,q) \times (1 - \alpha(r,q)) \\
= -\text{tot} \times \alpha(r,) \times \alpha(, q) + \text{tot} \times \alpha(r,q)
\]

(A.21)

Based on Equation A.21, the first component in the right hand side of Equation A.17 is given by:

\[
E(\text{cov}(b(r), a(q) \mid \text{tot})) = -E(\text{tot}) \times \alpha(r,\cdot) \times \alpha(\cdot, q) + E(\text{tot}) \times \alpha(r,q)
\]

(A.22)
The second component in the right hand side of Equation A.17 is developed in the following. The trip-level boarding counts combine the trip-level OD flows according to the stops the trip-level OD flows originate from (see Equation A.10a). Based on the assumption that trip-level OD flows are multinomially distributed conditional on the probability OD matrix $\alpha$ and the trip-level total demand $tot$, the boarding counts are also multinomially distributed since the Multinomial distribution is preserved when the counting variables are combined (Rice 2006). Specifically, for a given bus trip $l$, conditional on the trip-level total demand $tot_l$ and the probability OD matrix $\alpha$, the boarding counts across bus stops are multinomially distributed:

$$f(b_l(1),...,b_l(N_s-1) | tot_l,\alpha) \sim multinomial(tot_l,(\alpha(1,\cdot),...,\alpha(N_s-1,\cdot)))$$  \hspace{1cm} (A.23)

The probability OD matrix $\alpha$ is assumed to be stable across bus trips $l$ and the trip-level total demand $tot$ varies across bus trips. Therefore, for a general bus trip, the expected boarding count $b(r)$ at stop $r$ conditional on the trip-level total demand $tot$ is given by:

$$E(b(r) | tot) = tot \times \alpha(r,\cdot)$$  \hspace{1cm} (A.24)

Similarly, the expected alighting count $a(q)$ at stop $q$ conditional on the trip-level total demand $tot$ is given by:

$$E(a(q) | tot) = tot \times \alpha(\cdot,q)$$  \hspace{1cm} (A.25)

Based on Equations A.24 and A.25, the second component in the right hand side of Equation A.17 is given by:

$$\operatorname{cov}(E(b(r) | tot),E(a(q) | tot)) = \operatorname{cov}(tot \times \alpha(r,\cdot),tot \times \alpha(\cdot,q))$$
$$= \operatorname{var}(tot) \times \alpha(r,\cdot) \times \alpha(\cdot,q)$$  \hspace{1cm} (A.26)
Substituting Equations A.22, A.26 into Equation A.17 leads to Equation A.27 as follows:

\[
\text{cov}(b(r), a(q)) = E(\text{cov}(b(r), a(q) | \text{tot})) + \text{cov}(E(b(r) | \text{tot}), E(a(q) | \text{tot}))
\]
\[
= -E(\text{tot}) \times \alpha(r,\cdot) \times \alpha(\cdot,q) + E(\text{tot}) \times \alpha(r,q)
\]
\[
+ \text{var}(\text{tot}) \times \alpha(r,\cdot) \times \alpha(\cdot,q)
\]
\[
= E(\text{tot}) \times \alpha(r,q) + (\text{var}(\text{tot}) - E(\text{tot})) \times \alpha(r,\cdot) \times \alpha(\cdot,q)
\]

(A.27)

Equation A.27 could be simplified by introducing variances of boarding count \(b(r)\) at stop \(r\) and alighting count \(a(q)\) at stop \(q\).

Based on the law of total variance (Rice 2006), the variance of the boarding count \(b(r)\) at stop \(r\) is given by:

\[
\text{var}(b(r)) = \text{var}(E(b(r) | \text{tot}) + E(\text{var}(b(r) | \text{tot}))
\]

(A.28)

The expected boarding count \(b(r)\) at stop \(r\) conditional on the trip-level total demand \(\text{tot}\) has been presented in Equation A.24. Therefore, the first term in the right hand side of Equation A.28 is given by:

\[
\text{var}(E(b(r) | \text{tot}) = \text{var}(\text{tot} \times \alpha(r,\cdot)) = \alpha(r,\cdot)^2 \times \text{var}(\text{tot})
\]

(A.29)

Since trip-level boarding counts are multinomially distributed conditional on the trip-level total demand \(\text{tot}\) (see Equation A.23), the variance of boarding count \(b(r)\) at stop \(r\) conditional on the trip-level total demand \(\text{tot}\) is given by:

\[
\text{var}(b(r) | \text{tot}) = \text{tot} \times \alpha(r,\cdot) \times (1 - \alpha(r,\cdot))
\]

(A.30)

Based on Equation A.30, the second term in the right hand side of Equation A.28 is given by:

\[
E(\text{var}(b(r) | \text{tot})) = E(\text{tot} \times \alpha(r,\cdot) \times (1 - \alpha(r,\cdot)))
\]
\[
= E(\text{tot}) \times \alpha(r,\cdot) \times (1 - \alpha(r,\cdot))
\]

(A.31)
Substituting Equations A.29, A.31 into Equation A.28 leads to Equation A.32 as follows:

\[
\begin{align*}
\text{var}(b(r)) &= \alpha(r,)^2 \times \text{var}(\text{tot}) + E(\text{tot}) \times \alpha(r,) \times (1 - \alpha(r,)) \\
&= E(\text{tot}) \times \alpha(r,) + (\text{var}(\text{tot}) - E(\text{tot})) \times \alpha^2(r,)
\end{align*}
\]  
(A.32)

The first term in the right hand side of Equation A.32 is actually the expected boarding count \(b(r)\) at stop \(r\) (see Equation A.24). Therefore, Equation A.32 is reduced to:

\[
\text{var}(b(r)) = E(b(r)) + (\text{var}(\text{tot}) - E(\text{tot})) \times \alpha^2(r,)
\]  
(A.33)

Similarly, it can be shown that the variance of the alighting count \(a(q)\) at stop \(q\) is given by:

\[
\begin{align*}
\text{var}(a(q)) &= \text{var}(E(a(q) \mid \text{tot}) + E(\text{var}(a(q) \mid \text{tot})) \\
&= \text{var}(\text{tot}) \times \alpha^2(\cdot, q) + E(\text{tot}) \times \alpha(\cdot, q) \times (1 - \alpha(\cdot, q)) \\
&= E(\text{tot}) \times \alpha(\cdot, q) + (\text{var}(\text{tot}) - E(\text{tot})) \times \alpha^2(\cdot, \cdot, q) \\
&= E(a(q)) + (\text{var}(\text{tot}) - E(\text{tot})) \times \alpha^2(\cdot, \cdot, q)
\end{align*}
\]  
(A.34)

Based on Equations A.33 and A.34, the following Equation can be obtained:

\[
(\text{var}(\text{tot}) - E(\text{tot})) \times \alpha(r,) \times \alpha(\cdot, q) = \sqrt{(\text{var}(b(r)) - E(b(r)))} \\
\times \sqrt{(\text{var}(a(q)) - E(a(q)))}
\]  
(A.35)

Substituting Equation A.35 into Equation A.27 leads to Equation A.36:

\[
\alpha(r, q) = \frac{\text{cov}(b(r), a(q)) - \sqrt{(\text{var}(b(r)) - E(b(r))) \times (\text{var}(a(q)) - E(a(q)))}}{E(\text{tot})}
\]  
(A.36)

As can be seen in Equations A.36, the variance and covariance of the trip-level boarding and alighting counts could also inform the estimation of the probability OD matrix in the proposed approach.

For Vardi’s formulation, Equations A.10a, A.10b and A.16 that describe the relationships between the probability OD flows and the first and second moments of the
trip-level boarding and alighting counts are used simultaneously to estimate the probability OD flows in the method of moments based EM algorithm developed by Vardi (1996). Notice that the terms in the right hand sides of Equations A.10a, A.10b and A.16 could be estimated directly from the trip-level boarding and alighting counts. For the proposed approach, the information about the probability OD flows contained in the second moments of the trip-level boarding and alighting counts, as reflected in Equation A.36, along with the other higher moments is captured in the estimation algorithms described in Chapters 4. Equations A.36 are not used explicitly in the estimation algorithms developed subsequently. It is presented to demonstrate that there is theoretically valuable information in the distribution of the trip-level boarding and alighting data going beyond the first moments, in this case as represented by the second moments.
Appendix B  Markov Chain Monte Carlo Algorithms

B.1 Introduction

Chapter 3 introduced two formulations for the estimation of the probability OD matrix in a homogeneous time-of-day period. Both formulations consider passenger demand variation across bus trips. However, one formulation considers measurement errors in APC counts while the other formulation assumes APC counts are measurement error free. This appendix introduces Markov Chain Monte Carlo (MCMC) algorithms to simulate the marginal posterior distributions of the probability OD flows for these two formulations.

MCMC algorithm is a general method to draw samples from a target distribution. The MCMC algorithm draws samples from distributions sequentially. The samples drawn in the current step depends on the values of the previous samples. The series of drawn samples form a Markov Chain. After a large number of iterations, the samples will converge to the target distribution (Gelman et al., 2004).

One popular Markov chain algorithm used in multidimensional problems is Gibbs sampler, also called alternating conditional sampling (Gelman et al., 2004). The Gibbs sampler divides the parameter vector into several components, and draws samples for a joint posterior distribution from the posterior distribution of one component conditional on the values of the other components. This method is useful when it is difficult to
sample from the marginal posterior distribution directly, while it is easy to sample from the conditional posterior distributions.

When it is difficult to draw samples from the conditional posterior distributions in the Gibbs sampler, the Metropolis-Hastings step can be incorporated into the overall Gibbs sampler framework to produce the marginal posterior distribution (Gelman et al., 2004). The Metropolis-Hastings step uses the acceptance/rejection rule to converge to the target conditional posterior distribution. This step draws a sample from a proposed distribution and then accepts it with an acceptance probability.

Two MCMC algorithms are developed in this appendix. One MCMC algorithm is for the formulation that directly incorporates measurement errors in APC counts. This algorithm is introduced in Section B.2. The other is for the formulation that assumes APC counts are free of measurement errors. This algorithm is introduced in Section B.3. Some comments about the MCMC algorithms are made in Section B.4. A numerical study is carried out on an illustrative short bus route in Section B.5.

B.2 MCMC1 Algorithm for the Formulation that Directly Incorporates Measurement Errors in APC Counts

The measurement error model was not specified in the formulation that directly incorporates measurement errors in APC counts in Chapter 3. For demonstration purposes, this section assumes that the observed trip-level APC counts are independently Poisson distributed, conditional on the true trip-level boarding and alighting counts. This assumption implies large measurement errors in APC counts. For example, assuming the true boarding or alighting count is 10, the observed APC count could range from 5 and 15
with the probability of 0.9. If a measurement error model has been estimated from empirical APC data, the estimated measurement error model may be considered in the MCMC1 algorithm using similar procedure as presented in the following.

The marginal posterior distribution of the probability OD matrix has been presented in Equation 3.13 of Chapter 3. Drawing samples directly from the marginal posterior distribution is difficult. Therefore, the Gibbs sampler is used to draw samples for the joint posterior distribution of the probability OD matrix, the trip-level volume OD matrices and the true trip-level APC counts from several conditional posterior distributions. The necessary conditional posterior distributions are presented in the next section. Eventually, the samples of the probability OD flows from the Gibbs sampler can be used to approximate the marginal posterior distributions of the probability OD flows.

### B.2.1 Conditional Posterior Distributions for MCMC1 Algorithm

In the formulation that considers explicitly measurement errors in APC counts, the unknown parameter and variables are the probability OD matrix $\alpha$, the trip-level OD matrices $T^r$, and the true trip-level APC counts $x^c$. If the trip-level OD matrices are known, the true trip-level boarding and alighting counts $x^c$ are determined uniquely. This relationship has been presented in Equations 3.8a and 3.8b of Chapter 3, which are reproduced here for convenience:

\[
\sum_{q=r+1}^{N} T_l(r,q) = b_l(r) \quad r = 1,\ldots,N_r-1; l = 1,\ldots,L
\]  
\[
\sum_{r=1}^{q-1} T_l(r,q) = a_l(q) \quad q = 2,\ldots,N_q; l = 1,\ldots,L
\]

As discussed above, the Gibbs sampler divides the parameter vector into several
components, and draws samples for the joint posterior distribution from the posterior distribution of one component conditional on the values of other components. The necessary conditional posterior distributions are derived from the joint posterior distribution in the following. More specifically, the conditional posterior distribution of the probability OD matrix $\alpha$ and the conditional posterior distribution of OD flow for each feasible OD pair on each bus trip are derived in the following.

The joint posterior distribution of the probability OD matrix $\alpha$, the trip-level OD matrices $T^c$ and the true trip-level boarding and alighting counts $x^c$ has been presented in Equation 3.12 of Chapter 3, which is reproduced here for convenience:

$$f(\alpha, T^c, x^c | \hat{x}^c, z) \propto \prod_{l=1}^{L} f(\hat{x}_l | x_l) \times f(x_l | T_l) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\text{tot}_l^{(r,q)} \times \alpha(r,q)^{T(r,q)}}{T_l^{(r,q)}!} \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r,q)^{x^c(r,q)-\mu(r,q)-1}$$

(B.2)

In Equation B.2, $f(x_l | T_l)$ equals one if $x_l$ is determined by $T_l$ as captured by Equations B.1a and B.1b and zero otherwise.

The conditional posterior distribution of the probability OD matrix $\alpha$ can be derived from the joint posterior distribution of Equation B.2. Conditional on the survey OD data $z$, the observed trip-level APC data $\hat{x}^c$, the trip-level OD matrices $T^c$ and the true trip-level boarding and alighting counts $x^c$, the distribution of the probability OD matrix $\alpha$ depends on only the trip-level OD matrix $T^c$ and the survey OD data $z$, because the probability OD matrix $\alpha$ affects $\hat{x}^c$ and $x^c$ through $T^c$. Therefore, the conditional posterior distribution of the probability OD matrix $\alpha$ is given by:
\[ f(\alpha \mid \hat{x}^e, z, T^e, x^e) = f(\alpha \mid z, T^e) \]
\[
\propto \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q)^{z(r,q)+\mu(r,q)-1} \]
\[
= \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \alpha(r, q)^{\sum_{t=1}^{L} T_t(r,q) + z(r,q) + \mu(r,q)-1} \]

(B.3)

Equation B.3 is the density function of the \textit{Dirichlet}(G) distribution, where \( G \) is the collection of \( g(r,q) \) and \( g(r,q) \) is given by:

\[ g(r,q) = \sum_{l=1}^{L} T_t(r,q) + z(r,q) + \mu(r,q) \]  

(B.4)

The conditional posterior distribution of OD flow for each feasible OD pair on each bus trip can also be derived from the joint posterior distribution of Equation B.2. Equation B.2 reflects the assumptions that the trip-level OD matrices \( T_t \) are independent and that the trip-level OD matrices \( T_t \) and survey OD data \( z \) are independent for all bus trips \( l \) conditional on the probability OD matrix \( \alpha \). Therefore, conditional on the probability OD flow matrix \( \alpha \) and the observed APC count \( \hat{x}_l \) on bus trip \( l \), the distribution of the trip-level OD matrix \( T_t \) and the true trip-level boarding and alighting counts \( x_l \) can be obtained from Equation B.2 by removing terms that are not related to \( T_t \) and \( x_l \) from Equation B.2, which is given by:

\[ f(T_t, x_l \mid \hat{x}_l, \alpha) \propto f(\hat{x}_l \mid x_l) \times f(x_l \mid T_t) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \frac{\text{tot}_l \times \alpha(r,q)^{T_t(r,q)}}{T_t(r,q)!} \]  

(B.5)
Based on Equation B.5 and the assumption that the observed trip-level APC counts are independently Poisson distributed conditional on the trip-level true boarding and alighting counts, Equation B.5 can be shown to be given by:

\[
f(T_l, x_l | \hat{x}_l, \alpha) \propto \prod_{r=1}^{N_r-1} e^{-\hat{h}(r)} \times b_j(r) \hat{b}(r) \times \prod_{q=2}^{N_q} e^{-q_i(q)} \times a_i(q) \hat{a}(q) \\
\times f(x_l | T_l) \times \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \left( \frac{\text{tot}_l b(r, q) T_l(r, q)!}{T_l(r, q)!} \right)
\]

(B.6)

In Equations B.6, if the trip-level OD matrix \( T_l \) is known, the true trip-level boarding and alighting counts \( x_l \) is determined uniquely by Equations B.1a and B.1b. Otherwise, the likelihood is zero (i.e., \( f(x_l | T_l) = 0 \)). Therefore, the conditional posterior distribution in Equations B.6 can be expressed as follows:

\[
f(T_l, x_l | \hat{x}_l, \alpha) = f(x_l | T_l) \times f(T_l | \hat{x}_l, \alpha)
\]

(B.7)

and \( f(T_l | \hat{x}_l, \alpha) \) is given by:

\[
f(T_l | \hat{x}_l, \alpha) \propto \prod_{r=1}^{N_r-1} \left[ e^{-\sum_{q=r+1}^{N_q} T_l(r, q)} \times \left( \sum_{q=r+1}^{N_q} T_l(r, q) \right)^{\hat{b}(r)} \times \prod_{q=2}^{N_q} e^{-\sum_{r=1}^{q-1} T_l(r, q)} \times \left( \sum_{r=1}^{q-1} T_l(r, q) \right)^{\hat{a}(q)} \right] \\
\times \text{tot}_l \prod_{r=1}^{N_r-1} \prod_{q=r+1}^{N_q} \left( \frac{\alpha(r, q) T_l(r, q)!}{T_l(r, q)!} \right)
\]

(B.8)

The conditional posterior distribution of OD flow for one feasible OD pair on bus trip \( l \) can be derived from Equation B.8. Conditional on the observed APC counts \( \hat{x}_l \) on bus trip \( l \), the probability OD matrix \( \alpha \), and OD flows \( T_l(s-, t-) \) for feasible OD pairs other than \( (s, t) \) on bus trip \( l \), the distribution of OD flow for feasible OD pair \( (s, t) \) on bus trip \( l \) can be shown to be given by:

293
\[ f(T_i(s,t) \mid T_i(s-,t-), \hat{\lambda}_i, \alpha) \propto e^{\sum_{q=1}^{N_s} T_i(s,q)} \times \left( \sum_{q=1}^{N_s} T_i(s,q) \right)^{\hat{\lambda}_i(s)} \times \left( \sum_{r=1}^{N_t} T_i(r,t) \right)^{\hat{\lambda}_i(t)} \times \frac{\text{tot}_i \times \alpha(s,t)T_i(s,t)}{T_i(s,t)!} \]  

(B.9)

Notice that only OD flows for feasible OD pairs originating from stop \( s \) or destined for stop \( t \) show up in Equation B.9. The other known OD flows are reflected in the proportionality constant of the distribution.

### B.2.2 Implementation of the MCMC1 Algorithm

The MCMC1 algorithm repeatedly uses the conditional posterior distribution of the probability OD matrix in Equation B.3 and the conditional posterior distribution of OD flow for one feasible OD pair on one bus trip in Equation B.9. The implementation of MCMC1 algorithm can be described as follows:

1. Start with some (likely crude) estimates of trip-level OD matrices.

2. for \( h=1,2,\ldots \) apply the following two steps iteratively:

   - **Step 1:** Draw sampled values of the probability OD flows from the Dirichlet \((G)\) distribution presented in Equation B.3. The definition of the parameter vector \( G \) is presented in Equation B.4.

   - **Step 2:** Conditional on the sampled probability OD flows, for each bus trip, simulate a new OD matrix by sequentially sampling OD flow for each feasible OD pair based on Equation B.9, in which \( T_i(s-,t-) \) is set at their most recent sampled values.
Step 1 samples the probability OD matrix from the Dirichlet($G$) distribution. A fast method to sample from Dirichlet($G$) is to draw $d(r,q)$ for all feasible OD pairs from independent Gamma distribution with a common scale parameter and a shape parameter $g(r,q)$, and let (Gelman et al., 2004):

$$\alpha(r,q) = \frac{d(r,q)}{\sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} d(i,j)}$$  \hspace{1cm} (B.10)

Directly sampling in Step 2 based on the non-standard conditional distribution in Equation B.9 is difficult. Therefore, the Metropolis-Hastings step is useful here. The Metropolis-Hastings step generates a candidate value from a proposed distribution and accepts it with probability:

$$\min(1, \frac{f(T^*_l(s,t) \mid \cdots) \times q(T^*_l(s,t))}{f(T_l(s,t) \mid \cdots) \times q(T^*_l(s,t))})$$  \hspace{1cm} (B.11)

where

$f(\cdot) =$ conditional posterior distribution in Equation B.9,

$q(\cdot) =$ proposed probability density function,

$T^*_l(s,t) =$ candidate value of OD flow for OD pair $(s,t)$ generated from the proposed function,

$T_l(s,t) =$ current value of OD flow for OD pair $(s,t)$.

The proposed distribution used in this algorithm is Poisson($\text{tot}_l \times \alpha(s,t)$), in which $\text{tot}_l$ is the sum of the most recently values of OD flows on bus trip $l$ and $\alpha$ is the most recently sampled value of the probability OD matrix.
However, the MCMC1 algorithm cannot be used directly for the formulation that assumes APC counts are free of measurement errors. Step 2 of the MCMC1 algorithm samples OD flow for each feasible OD pair conditional on the most recently sampled values of the other feasible OD pairs on the same bus trip. When APC counts are measurement error free, knowing OD flows for a proportion of feasible OD pairs would uniquely determine OD flows for the other feasible OD pairs because of the constraints of the true boarding and alighting counts in Equations B.1a and B.1b. Therefore, it is not necessary to sample all feasible OD pairs. More critically, the MCMC1 algorithm will be trapped at the starting values and therefore cannot converge to the target distribution if the MCMC1 algorithm is applied to boarding and alighting counts without measurement errors.

For example, consider OD matrix for a bus trip on a hypothetical short bus route with four bus stops as presented in Table B.1. Conditional on the true boarding and alighting counts, OD flows $T(1,2)$ and $T(3,4)$ are determined uniquely by the true boarding and alighting counts because passengers alighting at stop 2 must have boarded at stop 1 and passengers boarding at stop 3 must alight at stop 4. In addition, as long as OD flow for one of the other four feasible OD pairs is known, OD flows for the rest three feasible OD pairs can be shown to be determined uniquely by the constraints of the true boarding and alighting counts.

When the MCMC1 algorithm is applied to the true boarding and alighting counts on the bus route of interest, Step 2 of the MCMC1 algorithm samples OD flow for one feasible OD pair conditional on the most recently sampled values of the other five
feasible OD pairs on the same bus trip and the true boarding and alighting counts. However, the value of the current feasible OD pair is unique if the OD flows for the other five feasible OD pairs and the true boarding and alighting counts are known. Therefore, any candidate value from the proposed distribution in the Metropolis-Hastings step will be rejected unless the candidate value is the same as the current value. As a result, the MCMC1 algorithm will be trapped at the starting values and cannot converge to the target distribution.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Boarding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>T(1,2)</td>
<td>T(1,3)</td>
<td>T(1,4)</td>
<td>b(1)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>T(2,3)</td>
<td>T(2,4)</td>
<td>b(2)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>T(3,4)</td>
<td>b(3)</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>a(2)</td>
<td>a(3)</td>
<td>a(4)</td>
<td></td>
</tr>
</tbody>
</table>

Table B.1 Trip-level OD Matrix for a Hypothetical Bus Route

B.3 MCMC2 Algorithm for the Formulation that Assumes that APC Counts are Free of Measurement Errors

This section develops the MCMC2 algorithm for the formulation that assumes that APC counts are free of measurement errors. Unlike the MCMC1 algorithm, the MCMC2 algorithm draws samples for a proportion of feasible OD pairs for each bus trip in Step 2, and the values for the rest feasible OD pairs are determined by the constraints of APC counts.

The MCMC2 algorithm is similar to the MCMC algorithm developed by Tebaldi and West (1998). However, the MCMC algorithm developed by Tebaldi and West is based on the assumption that OD flows in one observation period are independently
Poisson distributed, while the MCMC2 algorithm is based on the assumption that OD flows on one bus trip are multinomially distributed. In addition, the MCMC algorithm developed by Tebaldi and West is applied to link flows in one observation period, while the MCMC2 algorithm is applied to APC counts on multiple bus trips.

B.3.1 Foundation for Step 2 of the MCMC2 Algorithm

In the MCMC2 algorithm, for convenience, the probability OD matrix $\alpha$, the trip-level OD matrices $T^c$ and the survey OD flow matrix $z$ are rearranged in the form of vectors. The relationship between the trip-level OD vector and the trip-level APC count vector is represented by $x_i = \tau \times T_i$.

In the equation $x_i = \tau \times T_i$, the true trip-level APC counts $x_i$ on bus trip $l$ consists of $(N_s - 1)$ boarding counts and $(N_s - 1)$ alighting counts. Actually, only $r = 2 \times (N_s - 1) - 1$ counts are linearly independent since the total boarding count equals the total alighting count. Assuming $\tau^+$ is of full rank $r$ derived from the assignment mapping matrix $\tau$ by removing one row from $\tau$, $x_i = \tau \times T_i$ is equivalent to the following equation:

$$x_i^+ = \tau^+ \times T_i$$

(B.12)

where $x_i^+$ is obtained by removing one entry from the true APC count vector $x_i$ on bus trip $l$. The removed entry corresponds to the row that is removed from $\tau$ to produce $\tau^+$. Since the number of feasible OD pairs $N$ is usually larger than the number of linearly independent equations $r$, multiple OD matrices $T_i$ can satisfy Equation B.12.

According to the linear algebra theory (Lipson and Lipschutz, 2001), if OD flows of $(N - r)$ OD pairs are known, OD flows of the other $r$ OD pairs can be determined
uniquely by Equation B.12. In the following, the former \((N - r)\) OD pairs are referred to as free OD pairs and the latter \(r\) OD pairs are referred to as pivot OD pairs. The free OD pairs cannot be chosen arbitrarily. Special function will be introduced later to choose the free OD pairs.

For example, for the trip-level OD matrix in Table B.1, the number of feasible OD pairs \(N\) is 6 and the number of linearly independent counts \(r\) is 5 (i.e., \(2 \times (N - 1) - 1 = 2 \times (4 - 1) - 1\)). If OD flow of one (i.e., \(N - r = 6 - 5\)) OD pair is known, OD flows of the other five OD pairs are determined uniquely by the constraints of APC counts. In Table B.1, the free OD pair could be either OD pair (1,3), (1,4), (2,3) or (2,4).

Based on Equation B.12, OD flows of the pivot OD pairs could be written as a function of the OD flows of the free OD pairs as follows (Lipson and Lipschutz, 2001):

\[
T_l^1 = T_l^{h,1} + \beta \times T_l^2
\]  

(B.13)

where:

\(T_l^2\) = OD flow vector of \((N - r)\) free OD pairs on bus trip \(l\),

\(T_l^1\) = OD flow vector of \(r\) pivot OD pairs on bus trip \(l\),

\(T_l^{h,1}\) = OD flow vector of \(r\) pivot OD pairs in a solution to Equation B.12,

\(\beta\) = matrix with the size \(r \times (N - r)\).

The free OD pairs and matrix \(\beta\) can be determined using Gauss Jordan elimination with partial pivoting as implemented in the “rref” function in Matlab (Lipson and Lipschutz, 2001). The input of the “rref” function is the full rank assignment mapping matrix \(\tau^+\) and the outputs are \(R (r \times N)\) and \(jb\). \(R\) is the reduced row echelon
form of $\tau^*$ and $jb$ specifies the columns of $R$ that correspond to the pivot OD pairs. Let $jb^-$ represents the free OD pairs. The matrix $\beta$ equals $-R(:, jb^-)$, where $R(:, jb^-)$ extracts the columns corresponding to the free OD pairs $jb^-$ from matrix $R$.

Equation B.13 requires a solution to Equation B.12. This solution should be integer numbers since OD flows are discrete values. An integer solution to Equation B.12 can be obtained by a recursive procedure. The recursive procedure determines OD flows recursively, beginning at the first alighting stop on the route and continuing downstream until the last alighting stop. The simple First In, First Out (FIFO) rule would lead to a simple integer solution to Equation B.12. That is, at a given bus stop, it is assumed that passengers who boarded the bus first would alight at the given bus stop first.

Table B.2 is used to demonstrate the procedure to determine an integer OD flow matrix solution to Equation B.12 from boarding and counts on a bus trip. In Table B.2, dashed line represents infeasible OD pairs. The recursive procedure to determine an integer OD matrix from the APC boarding and alighting counts is as follows:

Stop 2: the alighting count at Stop 2 is 2 and there is only one feasible OD pair (1,2). Therefore, OD flow from Stop 1 to Stop 2 $T(1,2) = 2$.

Stop 3: the alighting count at Stop 3 is 6. At Stop 3, the number of onboard passengers who boarded at Stop 1 is 4 (the total boarding at Stop 1 is 6 and 2 passengers have alighted at Stop 2) and the number of onboard passengers who boarded at Stop 2 is 4. Based on the FIFO rule, passengers who boarded at Stop 1 should alight first and therefore OD flow from Stop 1 to Stop 3 $T(1,3) = 4$. After that, set OD flow
from Stop 2 to Stop 3 \( T(2,3) = 2 \) such that OD flows destined for Stop 3 equal the total alighting count at Stop 3.

Stop 4: the alighting count at Stop 4 is 10. Since this is the last bus stop of the given bus route, all onboard passengers must alight at Stop 4, which results in OD flow from Stop 1 to Stop 4 \( T(1,4) = 0 \), OD flow from Stop 2 to Stop 4 \( T(2,4) = 2 \) and OD flow from Stop 3 to Stop 4 \( T(3,4) = 8 \).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Boarding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Alighting</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B.2 Determine an OD Matrix from Boarding and Alighting Counts on a Bus Trip

In the next subsection, the trip-level OD vector is rearranged such that the first \( r \) elements are values of the pivot OD pairs \( T_i^1 \) and the rest \( (N - r) \) elements are values of the free OD pairs \( T_i^2 \). Therefore, \( T_i = [T_i^1; T_i^2] \). The probability OD vector and the survey OD flow vector are also rearranged in the same manner. In addition, the column of the assignment mapping matrix \( \tau^+ \) is reordered such that the first \( r \) columns correspond to the pivot OD pairs, and the rest \( (N - r) \) columns correspond to the free OD pairs.

### B.3.2 Conditional Posterior Distributions for the MCMC2 Algorithm

In the formulation that assumes that APC counts are free of measurement errors, the unknown parameter and variables are the probability OD vector \( \alpha \) and the trip-level OD vectors \( T^c \). The necessary conditional distributions for the MCMC2 algorithm are
derived from the joint posterior distribution in the following. More specifically, the conditional posterior distribution of the probability OD vector $\alpha$ and the conditional posterior distribution of OD flow for each free OD pair on each bus trip are derived.

The joint posterior distribution of the probability OD matrix $\alpha$ and the trip-level OD matrices $T^c$ has been presented in Equation 3.15 of Chapter 3, which is reproduced here in the form of vectors for convenience:

$$f(\alpha, T^c | x^c, z) \propto \prod_{l=1}^{L} f(x_l | T_l) \times \prod_{t=1}^{N} \frac{\alpha(t)^{T_l(t)}}{T_l(t)!} \times \prod_{t=1}^{N} \alpha(t)^{z(t)+\mu(t)-1}$$  \hspace{1cm} (B.14)

The conditional posterior distribution of the probability OD vector $\alpha$ can be derived from the joint posterior distribution of Equation B.14. Conditional on the survey OD data $z$, the trip-level OD vectors $T^c$ and the trip-level APC counts $x^c$, the distribution of the probability OD vector $\alpha$ depends on only the trip-level OD vectors $T^c$ and the survey data $z$, because probability OD vector $\alpha$ affects $x^c$ through $T^c$. Therefore, the conditional posterior distribution of the probability OD vector $\alpha$ is given by:

$$f(\alpha | x^c, z, T^c) = f(\alpha | z, T^c) \propto \prod_{t=1}^{N} \alpha(t)^{\sum_{l=1}^{L} T_l(t)+z(t)+\mu(t)-1}$$  \hspace{1cm} (B.15)

Equation B.16 is the density of the Dirichlet($G$) distribution, where $G$ is the collection of $g(t)$ and $g(t)$ is given by:

$$g(t) = \sum_{l=1}^{L} T_l(t) + z(t) + \mu(t)$$  \hspace{1cm} (B.16)

The conditional posterior distribution of OD flow for each free OD pair on each bus trip can also be derived from the joint posterior distribution of Equation B.14.
Equation B.14 reflects the assumptions that the trip-level OD matrices $T_l$ are independent and that the trip-level OD matrices $T_l$ and survey OD data $z$ are independent for all bus trips $l$ conditional on the probability OD matrix $\alpha$. Therefore, conditional on the probability OD flow matrix $\alpha$ and the true APC counts $x_i$ on bus trip $l$, the distribution of the trip-level OD matrix $T_l$ can be obtained from Equation B.15 by removing terms that are not related to $T_l$ from Equation B.15, which is given by:

$$f(T_l | x_i, \alpha) \propto f(x_i | T_l) \times \prod_{i=1}^{N} \frac{\alpha(t)^T(i)}{T_l(t)!}$$

(B.17)

In Equation B.17, if OD flows for the free OD pairs $T_l^2$ are known, OD flows for the pivot OD pairs $T_l^1$ are determined uniquely by Equation B.13. Therefore, the conditional distribution in Equations B.17 can be expressed as follows:

$$f(T_l^1, T_l^2 | x_i, \alpha) = f(T_l^1 | T_l^2, x_i, \alpha) \times f(T_l^2 | x_i, \alpha)$$

(B.18)

where $f(T_l^1 | T_l^2, x_i, \alpha) = 1$ if Equation B.13 is satisfied. Otherwise, $f(T_l^1 | T_l^2, x_i, \alpha) = 0$, and $f(T_l^2 | x_i, \alpha)$ is given by:

$$f(T_l^2 | x_i, \alpha) \propto \prod_{i=1}^{r} \frac{\alpha(i)^T(i)}{T_l^1(i)!} \times \prod_{i=1}^{N-r} \frac{\alpha(i)^T(i)}{T_l^2(i)!}$$

(B.19)

where $T_l^1$ is determined by $T_l^1 = T_l^{b,1} + \beta \times T_l^2$ (i.e., Equation B.13).

The conditional posterior distribution of OD flow for one free OD pair on bus trip $l$ can be derived from Equation B.19. Conditional on the APC counts $x_i$ on bus trip $l$, the probability OD matrix $\alpha$, and OD flows $T_l^2(s-)$ for the free OD pairs other than $s$ on bus
trip \( l \), the distribution of OD flow for free OD pair \( s \) on bus trip \( l \) can be shown to be given by:

\[
f(T^2_i(s) | T^2_i(s-), x_i, \alpha_s) \propto \prod_{t=1}^r \frac{\alpha(t)^{T^2_i(t)}}{T^2_i(t)!} \times \frac{\alpha(s)^{T^2_i(s)}}{T^2_i(s)!}
\]

(B.20)

In Equation B.20, OD flows \( T^2_i(s-) \) for free OD pairs other than \( s \) are known and therefore are reflected in the proportionality constant of the distribution.

### B.3.3 Implementation of the MCMC2 Algorithm

The MCMC2 algorithm repeatedly uses the conditional posterior distribution of the probability OD vector in Equation B.15 and the conditional posterior distribution of the OD flow for each free OD pair on each bus trip in Equation B.20. The implementation of the MCMC2 algorithm can be described as follows:

1. Start with some (likely crude) estimates of the trip-level OD vectors.
2. for \( h=1,2,\ldots \) apply the following two steps iteratively:
   - Step 1: Draw sampled values of the probability OD flows from the \textit{Dirichlet} \((G)\) distribution presented in Equation B.15. The definition of the parameters \( G \) has been presented in Equation B.16.
   - Step 2: Conditional on the sampled probability OD flows, for each bus trip, simulate a new OD matrix by sequentially sampling OD flow for each free OD pair based on Equation B.20, in which \( T^2_i(s-) \) is set at their most recently sampled values; when drawing sample for an free OD pair,
the pivot OD flows $T^1_i$ is explicitly reevaluated using Equation B.13

with the most recently sampled values of free OD pairs $T^2_i$.

The procedure to draw samples from the Dirichlet distribution has been described in Section B.2.1.2. Directly sampling in Step 2 based on the non-standard conditional distribution in Equation B.20 is difficult. Therefore, sampling in Step 2 of the MCMC2 algorithm also uses the Metropolis-Hastings step. The Metropolis-Hastings step generates a candidate value from a proposed distribution and accepts it with probability:

$$
\min(1, \frac{f(T^*_{i}(s,t) | \cdot) \times q(T_{i}(s,t))}{f(T_{i}(s,t) | \cdot) \times q(T^*_{i}(s,t))})
$$

(B.21)

where

$f(\cdot) =$ conditional posterior distribution in Equation B.22,

$q(\cdot) =$ proposed probability density function,

$T^*_{i}(s,t) =$ candidate value of OD flow for free OD pair $(s,t)$ generated from the proposed function,

$T_{i}(s,t) =$ current value of OD flow for free OD pair $(s,t)$.

The proposed distribution used in this algorithm is $\text{Poisson}(\text{tot}_l \times \alpha(s,t))$, in which $\text{tot}_l$ is the sum of the most recently values of OD flows on bus trip $l$ and $\alpha$ is the most recently sampled value of the probability OD matrix.

**B.4 Technical Comments on the MCMC Algorithms**

The convergence of the MCMC algorithm has been proved in the literature (Gelman et al., 2004). After the iterations of the MCMC algorithm have run long enough,
the samples of the probability OD flows can be used to approximate the marginal posterior distributions of the probability OD flows.

Since the early iterations are influenced by the starting values, generally certain amount of iterations in the beginning of the series of iterations is disregarded, which is referred to as burn-in in Markov chain simulation.

In addition, samples in one iteration of the MCMC algorithm depend on the values of the previous iteration. As a result, the samples are correlated. To reduce the correlation in the samples, the sequence of samples can be thinned by keeping the samples of every $k^{th}$ iteration and disregarding the rest.

It is worth to point out that both the MCMC1 and MCMC2 algorithms are computationally expensive. The MCMC1 algorithm needs to draw samples for each feasible OD pairs on each bus trip in each iteration of the algorithm. The MCMC2 algorithm needs to draw samples for each free OD pairs on each bus trip in each iteration of the algorithm. The simulation time increases tremendously when the number of bus stops and the number of bus trips increase. In addition, the algorithm may converge very slowly to the target distribution since the trip-level OD flows are highly correlated in the posterior distribution.

B.5 Numerical Evaluation on a Simplistic, Illustrative Bus Route

B.5.1 Experiment Overview

The two MCMC algorithms developed in Sections B.2 and B.3 are evaluated numerically on an illustrative short bus route. There are four bus stops on the illustrative short bus route beginning at stop 1 and terminating at stop 4. The assumed underlying
true probability OD matrix is presented in Table B.3. The trip-level total demand is assumed to be Negative Binomial $NB(r,p)$ distributed. The number of successes $r$ is set to be 30 and the probability of success $p$ is set to be 0.5. The resulting mean and variance of the trip-level total demand is 30 and 60, respectively.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.07</td>
<td>0.09</td>
<td>0.30</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.35</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B.3 Assumed Probability OD Flow Matrix on an Illustrative Bus Route

The trip-level volume OD flow matrices and the trip-level true APC counts on 50 bus trips are simulated based on the model structure presented in Figure 3.1 of Chapter 3 and the assumed underlying true probability OD flow matrix and the assumed trip-level total demand distribution. The observed trip-level APC counts are generated independently conditional on the simulated true trip-level boarding and alighting counts using the Poisson distribution. Specifically, for each bus stop on each bus trip, the observed boarding or alighting count is generated randomly using the Poisson distribution conditional on the simulated true boarding or alighting count.

The MCMC1 algorithm is applied to the observed trip-level APC counts and the MCMC2 algorithm is applied to the true trip-level boarding and alighting counts to estimate the probability OD flows. The performance difference between the MCMC2 estimates and the MCMC1 estimates reflects the effect of measurement errors in APC counts on the final MCMC estimates. In addition, for comparison purposes, OD flow
estimates produced by the IPF method are also considered in the demonstration. The IPF method is applied to the true trip-level boarding and alighting counts to estimate the probability OD flow matrix.

**B.5.2 Results**

Figures B.1 presents the numerical posterior distributions of the probability OD flows determined by the MCMC algorithms and the point estimates of the probability OD flows determined by the IPF method. The assumed underlying true probability OD flows are also indicated by a vertical dashed line in Figures B.1. The null base OD flow matrix is assumed for the IPF method and the MCMC algorithms in producing the results shown in Figure B.1.

![Figure B.1 Numerical Posterior Distributions of the Probability OD Flows](image)

308
Each MCMC algorithm is run for a total 6,000 samples. The first 1,000 samples are discarded and the rest samples are thinned by keeping every 50\textsuperscript{th} samples. The posterior distributions of the probability OD flows are summarized based on 100 samples. The convergence of each MCMC algorithm is confirmed by checking the trace plot. The trace plot presents the sampled probability OD values for each feasible OD pair against the iteration number. The trace plots do not show obvious trends.

In Figure B.1, The MCMC1 and MCMC2 algorithms and the IPF method perform similarly for OD pairs 1-2 and 3-4. The underlying probability OD flows are close to the modes of the numerical posterior distributions of the probability OD flows determined by the MCMC1 and MCMC2 algorithms. The IPF estimates are also close to the underlying probability OD flows. OD pairs 1-2 and 3-4 are special because the trip-level OD flows in these two OD pairs are uniquely determined by the trip-level true APC counts (passengers alighting at stop 2 must have boarded at stop 1 and passengers boarding at stop 3 must alight at stop 4).

For the other four OD pairs, the MCMC2 algorithm produces the best probability OD flow estimates. The numerical posterior distributions of the probability OD flows determined by the MCMC2 algorithm are almost centered on the underlying probability OD flows. The underlying probability OD flows are also in the range of the numerical posterior distributions of the probability OD flows determined by the MCMC1 algorithm. However, the numerical posterior distributions determined by the MCMC1 algorithm have larger variances than those determined by the MCMC2 algorithm. This result is expected since measurement errors in APC counts introduce additional variations in the
final estimates (recall that the MCMC1 algorithm is based on the observed APC counts and the MCMC2 algorithm is based on the true APC counts). In addition, it can be seen that the IPF estimates deviate from the underlying probability OD flows.

The overall performance of different probability OD flow estimates can be summarized by the $HD^2$, $SSD$ and $Chi2$ performance measures and their corresponding $RP$ measures as presented in Tables B.4 (See Section 5.3 for the definition of the performance and $RP$ measures). In Tables B.4, the performance measures for the MCMC estimates are calculated based on the means of the numerical posterior distributions of the probability OD flows.

<table>
<thead>
<tr>
<th>Data</th>
<th>IPF</th>
<th>MCMC2</th>
<th>MCMC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$HD^2$</td>
<td>0.056</td>
<td>0.001</td>
<td>0.010</td>
</tr>
<tr>
<td>$HD^2$-based $RP$</td>
<td>0.51</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>$SSD$</td>
<td>0.044</td>
<td>0.0008</td>
<td>0.008</td>
</tr>
<tr>
<td>$SSD$-based $RP$</td>
<td>0.45</td>
<td>0.99</td>
<td>0.90</td>
</tr>
<tr>
<td>$Chi2$</td>
<td>0.271</td>
<td>0.005</td>
<td>0.044</td>
</tr>
<tr>
<td>$Chi2$-based $RP$</td>
<td>0.50</td>
<td>0.99</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table B.4 Performance and $RP$ Measures of Various Probability OD Flow Estimates

As can be seen in Tables B.4, the MCMC2 estimates are better than the MCMC1 estimates since the MCMC2 estimates are based on the true trip-level APC counts while the MCMC1 estimates are based on the observed trip-level APC counts with measurement errors. In addition, based on APC counts that are free of measurement errors, the MCMC2 estimates are much better than the IPF estimates. Moreover, even if APC counts have large measurement errors, the MCMC1 estimates based on the
observed APC counts are also much better than the IPF estimates based on the true APC counts.
Appendix C  Use of the IPF Method in the CM Algorithm

This appendix provides the argument supporting the use of the IPF method to obtain a trip-level OD matrix \( T_l \) that approximately maximize the distribution of the trip-level OD matrix \( T_l \) in Equation 4.3 of Chapter 4. Equation 4.3 represents the distribution of the trip-level OD matrix \( T_l \) conditional on the probability OD flow matrix \( \alpha \) and APC count data \( x_l \), which is reproduced here for convenience:

\[
f(T_l | \alpha, x_l) \propto \prod_{r=1}^{N_s-1} \prod_{q=r+1}^{N_s} \frac{\alpha(r,q)^{T_l(r,q)}}{T_l(r,q)!} \times f(x_l | T_l)
\]  

(C.1)

In Equation C.1, \( f(x_l | T_l) \) equals one if \( T_l \) satisfies the constraints of APC counts \( x_l \) on bus trip \( l \) and zero otherwise.

Maximizing the likelihood in Equation C.1 is equivalent to maximising the natural logarithm of the likelihood. What follows is the maximization of the latter:

\[
l = \sum_{r=1}^{N_s-1} \sum_{q=r+1}^{N_s} (T_l \times \log(\alpha(r,q)) - \log(T_l(r,q)!))
\]  

(C.2)

with the constraints that the OD matrix \( T_l \) satisfies the given APC counts:

\[
\sum_{q=r+1}^{N_s} T_l(r,q) = b_l(r) \quad r = 1, \ldots, N_s - 1
\]  

(C.3a)

\[
\sum_{r=3}^{q-1} T_l(r,q) = a_l(q) \quad q = 2, \ldots, N_s
\]  

(C.3b)

Equations C.3a and C.3b reflects \( f(x_l | T_l) \) in Equation C.1.
The Lagrange function for the maximization problem above is introduced as follows:

\[
L = \sum_{r=1}^{N_s-1} \sum_{q=r+1}^{N_s} (T_i \times \log(\alpha(r,q)) - \log(T_i(r,q))) \\
+ \sum_{r=1}^{N_s-1} C_r \times \left( \sum_{q=r+1}^{N_s} T_i(r,q) - b_i(r) \right) + \sum_{q=2}^{N_s} D_q \times \left( \sum_{r=1}^{q-1} T_i(r,q) - a_i(q) \right) 
\]  
\[\text{where } C_r \text{ and } D_q \text{ are the Lagrange multipliers. If } T'_i \text{ is the OD matrix that maximizes Equation C.2 with the constraints of Equations C.3a and C.3b, then there exist } C_r \text{ and } D_q \text{ such that the partial derivatives of the Lagrange function in Equation C.4 are zero (i.e., the first order necessary condition for the maximization problem) (Lipson and Lipschutz, 2001).}
\]

Equation C.4 could be simplified using Stirling’s approximation for the logarithm of factorial terms in Equation C.4. Stirling’s approximation is given by (Paris and Kaminsky, 2001):

\[
\log(T_i(r,q)!) \approx T_i \log(T_i(r,q)) - T_i(r,q) 
\]  
\[\text{(C.5)}\]

Substituting Equation C.5 into Equation C.4 yields:

\[
L = \sum_{r=1}^{N_s-1} \sum_{q=r+1}^{N_s} (T_i \times \log(\alpha(r,q)) - (T_i \log(T_i(r,q)) - T_i(r,q))) \\
+ \sum_{r=1}^{N_s-1} C_r \times \left( \sum_{q=r+1}^{N_s} T_i(r,q) - b_i(r) \right) + \sum_{q=2}^{N_s} D_q \times \left( \sum_{r=1}^{q-1} T_i(r,q) - a_i(q) \right) 
\]  
\[\text{(C.6)}\]

Based on Equation C.6, the first order necessary condition of the maximization problem above is given by:

\[
\frac{\partial L}{\partial T_i(r,q)} = \log(\alpha(r,q)) - \log(T_i(r,q)) + C_r + D_q = 0 \quad r = 1, \ldots, N_s-1, q = r+1, \ldots, N_s 
\]  
\[\text{(C.7a)}\]
\[
\frac{\partial L}{\partial C_r} = \sum_{q=1}^{N_s} T_i(r,q) - b_i(r) = 0 \quad r = 1,\ldots,N_s - 1 \quad (C.7b)
\]

\[
\frac{\partial L}{\partial D_q} = \sum_{r=1}^{q-1} T_i(r,q) - a_i(q) = 0 \quad r = 1,\ldots,N_s - 1 \quad (C.7c)
\]

Equation C.7a can be rearranged as follows:

\[
T_i(r,q) = e^{C_r} \times \alpha(r,q) \times e^{D_q} \quad (C.8)
\]

Let \( A_r = e^{C_r} \) and \( B_q = e^{D_q} \), Equation C.8 is given by:

\[
T_i(r,q) = A_r \times \alpha(r,q) \times B_q \quad (C.9)
\]

Equation C.9 represents the relationship between the probability OD matrix \( \alpha \) and the trip-level OD matrix \( T_i \) that maximizes Equation C.2 with the constraints of Equations C.3a and C.3b.

Equation C.9 is also the IPF estimator, which uses the probability OD matrix \( \alpha \) as the based OD matrix (Ben-Akiva et al., 1985). In the IPF method, \( A_r \) and \( B_q \) are factors for row \( r \) and column \( q \). The values of \( A_r \) and \( B_q \) are determined by an iterative procedure such that the resulting OD matrix satisfies the given APC counts (i.e., Equations C.7b and C.7c).

Based on the discussed above, the IPF method can be used to produce a \( T_i \) that approximately maximizes Equation C.1 (that is also Equation 4.3). The current probability OD matrix \( \alpha \) is treated as the base OD matrix, and the trip-level APC counts \( x_i \) are treated as the marginal values to which the reconstructed OD matrix in the IPF method would converge.
Appendix D  Convergence of the EM Algorithm in the Context of the Proposed Probability OD Estimation Problem

The Expectation Maximization (EM) algorithm is an iterative method for finding the mode of the marginal posterior distribution from the joint posterior distribution (Meng and vanDyk, 1997). Take the current OD estimation problem as an example, the EM algorithm finds the mode of the marginal posterior distribution of the probability OD matrix from the joint posterior distribution of the probability OD matrix and the trip-level OD matrices. The basic procedure of applying the EM algorithm to the current OD estimation problem is as follows:

1) Start from an initial (typically crude) guess of the probability OD matrix.

2) E(xpectation)-step: determine the expected log joint posterior distribution of the probability OD matrix and the trip-level OD matrices with respect to the trip-level OD matrices. The expectation is performed based on the distribution of the trip-level OD matrices conditional on the probability OD matrix determined in the previous iteration and the observed data.

3) M(aximization)-step: find a new value of the probability OD matrix that maximizes the expected log joint posterior distribution determined in the E-step.

The above procedure is repeated until convergence. The convergence of the EM algorithm has been proved in the literature (Meng and vanDyk, 1997). The convergence of the EM algorithm in the context of the proposed probability OD estimation problem is
presented in this appendix. Each iteration of the EM algorithm increases the marginal posterior density of the probability OD matrix, and eventually the EM algorithm converges to a local mode of the marginal posterior density of the probability OD matrix. To start with, the marginal posterior distribution of the probability OD matrix can be expressed by:

\[
f(\alpha | x^c, z) = \frac{f(T^c, \alpha | x^c, z)}{f(T^c | \alpha, x^c, z)}
\]

(D.1)

Taking the logarithm on both sides of Equation D.1 leads to:

\[
\log(f(\alpha | x^c, z)) = \log(f(T^c, \alpha | x^c, z)) - \log(f(T^c | \alpha, x^c, z))
\]

(D.2)

As discussed above, any iteration of the EM algorithm involves two steps: the E-step and the M-step. Equation D.2 is the key for both steps. The following considers the \(h^{th}\) iteration of the EM algorithm. The E-step in the \(h^{th}\) iteration is described first, followed by the M-step.

E-step

Taking expectations to both sides of Equation D.2 with respect to the trip-level OD matrices \(T^c\) (The expectation is performed based on the distribution of the trip-level OD matrices \(T^c\) conditional on the probability OD flows \(\alpha^{h-1}\) determined in the \((h - 1)^{th}\) iteration of the EM algorithm, the trip-level APC counts \(x^c\) and survey OD data \(z f(T^c | \alpha^{h-1}, x^c, z)\), where the trip-level OD matrices \(T^c\) are treated as random variables, leads to:

\[
\log(f(\alpha | x^c, z)) = E(\log(f(T^c, \alpha | x^c, z)) | \alpha^{h-1}, x^c, z)
- E(\log(f(T^c | \alpha, x^c, z)) | \alpha^{h-1}, x^c, z)
\]

(D.3)
The left hand side of Equation D.3 is the same as that of Equation D.2 because the left hand side of Equation D.2 does not depend on the trip-level OD matrices $T^c$. The expectation $E(\cdot|\alpha^{h-1}, x^c, z)$ reflects the averaging of the function represented by the dot over the trip-level OD matrices $T^c$. The expectation is performed based on the distribution $f(T^c|\alpha^{h-1}, x^c, z)$.

The fact that the second term on the right hand side of Equation D.3 is maximized at $\alpha=\alpha^{h-1}$ is important in presenting the M-step subsequently. This result can be seen from the following:

\[
E(\log(f(T^c|\alpha^{h-1}, x^c, z))|\alpha^{h-1}, x^c, z) - E(\log(f(T^c|\alpha, x^c, z))|\alpha^{h-1}, x^c, z)
= \sum_{T^c} \log\left( \frac{f(T^c|\alpha^{h-1}, x^c, z)}{f(T^c|\alpha, x^c, z)} \right) \times f(T^c|\alpha^{h-1}, x^c, z)
= -\sum_{T^c} \log\left( \frac{f(T^c|\alpha, x^c, z)}{f(T^c|\alpha^{h-1}, x^c, z)} \right) \times f(T^c|\alpha^{h-1}, x^c, z)
\geq -\sum_{T^c} \left( \frac{f(T^c|\alpha, x^c, z)}{f(T^c|\alpha^{h-1}, x^c, z)} - 1 \right) \times f(T^c|\alpha^{h-1}, x^c, z)
= -(\sum_{T^c} f(T^c|\alpha, x^c, z) - \sum_{T^c} f(T^c|\alpha^{h-1}, x^c, z))
= -(1 - 1)
= 0
\]  

(D.4)

The inequality in Equation D.4 holds since that:

log($x$) $\leq x - 1$ $\forall x > 0$ and log($x$) $= x - 1$ iff $x = 1$  

(D.5)

Equation D.5 holds since $f(x) = \log(x) - (x - 1)$ is a concave function and the maximum value of $f(x)$ is zero when $x = 1.$
Now consider a new $\alpha^h$ that maximizes the first term on the right hand side of Equation D.3, $E(\log(f(T^c, \alpha | x^c, z)) | \alpha^{h-1}, x^c, z)$. Since the second term in the right side of Equation D.3, $E(\log(f(T^c | \alpha^h, x^c, z)) | \alpha^{h-1}, x^c, z)$ will not be larger than $E(\log(f(T^c | \alpha^{h-1}, x^c, z)) | \alpha^{h-1}, x^c, z)$ given Equation D.4, the following inequality is obtained:

$$E(\log(f(T^c, \alpha^h | x^c, z)) | \alpha^{h-1}, x^c, z) - E(\log(f(T^c | \alpha^h, x^c, z)) | \alpha^{h-1}, x^c, z) > E(\log(f(T^c, \alpha^{h-1} | x^c, z)) | \alpha^{h-1}, x^c, z) - E(\log(f(T^c | \alpha^{h-1}, x^c, z)) | \alpha^{h-1}, x^c, z)$$

(D.6)

Based on Equation D.2, the term on the left hand side of Equation D.6 equals the marginal posterior log-likelihood of the probability OD matrix $\alpha^h$ determined in the $h^{th}$ iteration $\log(f(\alpha^h | x^c, z))$ and the term on the right hand side of Equation D.6 equals the marginal posterior log-likelihood of the probability OD matrix $\alpha^{h-1}$ determined in the $(h-1)^{th}$ iteration $\log(f(\alpha^{h-1} | x^c, z))$. Therefore, Equation D.6 is reduced to:

$$\log(f(\alpha^h | x^c, z)) > \log(f(\alpha^{h-1} | x^c, z))$$

(D.7)

As demonstrated in Equation D.6, the E-step only needs to evaluate the first term on the right hand side of Equation D.2, $E(\log(f(T, \alpha | x^c, z)) | \alpha^{h-1}, x^c, z)$, and then the M-step finds a new value of the probability OD matrix $\alpha^h$ to maximize $E(\log(f(T, \alpha | x^c, z)) | \alpha^{h-1}, x^c, z)$. Eventually, the marginal posterior density of the probability OD matrix $f(\alpha | x^c, z)$ increases at each iteration of the EM algorithm (see Equation D.7) and then the algorithm converges to a local mode of the marginal posterior density.
Appendix E  SSD and Chi2 Results from the Analysis of Chapter 7

Chapter 7 summarized the performance of various probability OD flow estimates using the $HD^2$ measure and the $HD^2$-based $RP$ measure. This appendix presents the results in Chapter 7 using SSD and Chi2 measures and their corresponding $RP$ measures. The results based on the $HD^2$, SSD and Chi2 measures are consistent. The correspondence of the plots in this Appendix and in Chapter 7 is provided in Table D.1.

<table>
<thead>
<tr>
<th>Chapter 7 ($HD^2$)</th>
<th>Appendix D (SSD)</th>
<th>Appendix D (Chi2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 7.1</td>
<td>Figure D.1</td>
<td>Figure D.7</td>
</tr>
<tr>
<td>Figure 7.2</td>
<td>Figure D.2</td>
<td>Figure D.8</td>
</tr>
<tr>
<td>Figure 7.3</td>
<td>Figure D.3</td>
<td>Figure D.9</td>
</tr>
<tr>
<td>Figure 7.4</td>
<td>Figure D.4</td>
<td>Figure D.10</td>
</tr>
<tr>
<td>Figure 7.5</td>
<td>Figure D.5</td>
<td>Figure D.11</td>
</tr>
<tr>
<td>Figure 7.6</td>
<td>Figure D.6</td>
<td>Figure D.12</td>
</tr>
<tr>
<td>Figure 7.7</td>
<td>Figure D.13</td>
<td>Figure D.17</td>
</tr>
<tr>
<td>Figure 7.8</td>
<td>Figure D.14</td>
<td>Figure D.18</td>
</tr>
<tr>
<td>Figure 7.9</td>
<td>Figure D.15</td>
<td>Figure D.19</td>
</tr>
<tr>
<td>Figure 7.10</td>
<td>Figure D.16</td>
<td>Figure D.20</td>
</tr>
<tr>
<td>Figure 7.11</td>
<td>Figure D.21</td>
<td>Figure D.25</td>
</tr>
<tr>
<td>Figure 7.12</td>
<td>Figure D.22</td>
<td>Figure D.26</td>
</tr>
<tr>
<td>Figure 7.13</td>
<td>Figure D.23</td>
<td>Figure D.27</td>
</tr>
<tr>
<td>Figure 7.14</td>
<td>Figure D.24</td>
<td>Figure D.28</td>
</tr>
</tbody>
</table>

Table E.1 Correspondence of the Plots in Appendix D and in Chapter 7
E.1 Effects of the Number of Bus Trips with APC Counts and OD Survey Sample Size on the Accuracy of OD Estimates

SSD

Figure E.1 SSD Performance Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes
Figure E.2 SSD-based RP Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes
Figure E.3 SSD Performance Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data
Figure E.4 SSD-based RP Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data
Figure E.5 ECDFs of the SSD Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes

Figure E.6 ECDFs of the SSD-based RP Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes
Figure E.7 \textit{Chi2} Performance Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes
Figure E.8 Chi2-based RP Measure versus the Number of Bus Trips for which APC Data are Collected under Different OD Survey Sample Sizes
Figure E.9 Chi2 Performance Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data
Figure E.10 *Chi2*-based *RP* Measure versus the OD Survey Sample Size under Different Numbers of Bus Trips with APC Data
Figure E.11 ECDFs of the Chi2 Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes

Figure E.12 ECDFs of the Chi2-based RP Measures under Various Numbers of Bus Trips for which APC Data were Collected and Various OD Survey Sample Sizes

329
E.2 Effects of OD Survey Sample Size and Measurement Errors in APC Counts on the Accuracy of OD Estimates

**SSD**

Figure E.13 SSD Performance Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes
Figure E.14 SSD-based RP Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes
Figure E.15 ECDFs of the SSD Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts

Figure E.16 ECDFs of the SSD-based Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts
Figure E.17 \textit{Chi2} Performance Measure versus the Level of Measurement Errors in APC Counts under Different OD Survey Sample Sizes
Figure E.18 *Chi2*-based *RP* Measure versus the Level of Measurement Errors in APC Counts Assuming under Different OD Survey Sample Sizes
Figure E.19 ECDFs of the *Chi2* Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts

Figure E.20 ECDFs of the *Chi2*-based Measure under Various OD Survey Sample Sizes and Various Levels of Measurement Errors in APC Counts
E.3 Effects of the Number of Bus Trips with APC Counts and Measurement Errors in APC Counts on the Accuracy of OD Estimates

$SSD$

Figure E.21 $SSD$ Performance Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts
Figure E.22 SSD-based RP Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts
Figure E.23 ECDFs of the SSD Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts

Figure E.24 ECDFs of the SSD-based RP Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts
Figure E.25 Chi2 Performance Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts
Figure E.26 *Chi2*-based *RP* Measure versus the Number of Bus Trips with APC Counts under Different Levels of Measure Errors in APC Counts
Figure E.27 ECDFs of the Chi2 Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts

Figure E.28 ECDFs of the Chi2-based RP Measure under Various Numbers of Bus Trips with APC Counts and Various Levels of Measurement Errors in APC Counts
Appendix F  Empirical and Numerical Studies on COTA Route 7

F.1  Introduction

The Campus Loop South (CLS) bus route considered in this study has only 18 bus stops. This appendix evaluates various probability OD flow estimates on a much longer bus route, Route 7 Southbound (SB) in the Central Ohio Transit Authority (COTA) network. Route 7 SB is 17.6 km long and serves 78 bus stops and 3003 feasible OD pairs. Buses on Route 7 provided services from University City to Whittier Street & Seymour Avenue as shown schematically in Figure F.1. Along route 7, the Ohio State University and the downtown area are two major trip generation and attraction centers.

![Route Map of Bus Route 7]

Figure F.1 Schematic Route Map of Bus Route 7
Boarding and alighting counts were collected by the APC systems on 37 bus trips on Route 7 SB between 8am and 10am from January 1, 2007 to July 11, 2007. The empirical APC data are used for OD estimation in the empirical study presented in Section F.2. Because large scale manually collected true OD data are not available, a probability OD matrix is constructed arbitrarily for demonstration purposes in the numerical study presented in Section F.3.

It is worth to point out that the purpose of the empirical and numerical studies on Route 7 SB is not to demonstrate that the HEM algorithm is superior to the IPF and CM methods. Instead, the purpose is to demonstrate that it is feasible to apply the CM and HEM algorithms on long bus routes. The numerical study also demonstrates that the HEM algorithm takes better advantage of the large quantities of APC data than the IPF method on long bus routes.

F.2 Empirical Study

The IPF, CM and HEM methods are applied to APC boarding and alighting counts on 37 bus trips. The null base OD flow matrix is used for the IPF, CM and HEM methods in producing the probability OD flow estimates. Because a good estimate of the underlying probability OD matrix is not available, the probability OD flow estimates produced by the CM and HEM methods are compared with those produced by the IPF method in Figure F.2.
Figure F.2 shows that the probability OD flow estimates produced by the three methods are different. However, it is hard to tell which method performs better, since the underlying true probability OD matrix is unknown. Recall that the empirical and numerical studies on the CLS bus route indicate that the IPF estimates are better than the CM and HEM estimates when the number of bus trips with APC counts is small. The findings on the CLS bus route seem to indicate that the IPF estimates are very likely to be better than the CM and HEM estimates on Route 7 SB since only a few bus trips had APC data and the number of feasible OD pairs on Route 7 SB is much larger than that on the CLS bus route.
The hypothesis that the IPF estimates are better than the CM and HEM estimates in the current experiment seems to be supported by the comparison of the ECDFs of the Passenger Distance Travelled (PDT) derived from the probability OD matrices estimated by the IPF, CM and HEM methods. As discussed in Chapter 2 and demonstrated in Chapters 5 and 6, the IPF method using the null base tends to overestimate both the short and long passenger trips. Figure F.3 presents the ECDFs of the PDT derived from the IPF, CM and HEM probability OD flow estimates. As can be seen in Figure F.3, the proportion of short and long passenger trips (e.g., passenger trips shorter than 2.9 km and longer than 12.5 km) derived from the CM or HEM probability OD flow estimates is larger than that derived from the IPF probability OD flow estimates. This result indicates that the CM and HEM estimates are worse than the IPF estimates.
Obtaining sufficient true OD data to produce a good depiction of passenger travel patterns for a long bus route like Route 7 SB would be difficult. Therefore, the underlying true probability OD flow matrix for Route 7 SB used in this numerical study is constructed arbitrarily. In addition, The Negative Binomial distribution with parameters $r = 6.75$ and $p = 0.17$ is used to generate the trip-level total demand in the simulation-based analysis. The parameters of the Negative Binomial distribution are estimated from the trip-level APC counts collected on 37 bus trips.
Trip-level OD matrices and the resulting boarding and alighting counts on 500 bus trips are generated based on the model structure in Figure 3.1 of Chapter 3 and the assumed underlying probability OD matrix and the assumed trip-level total demand distribution. The period-level probability OD estimates are produced by the various methods and the performance measures are calculated for each of the period-level probability OD estimates. The null base OD flow matrix is used when producing the estimates from the IPF, CM, and HEM methods. Based on boarding and alighting counts on 500 bus trips, the computational times for producing the period-level probability OD flow estimates by the CM and HEM algorithms are about one minute and half minute on a 2.80 GHz Intel CPU, respectively, demonstrating that it is feasible to apply the CM and HEM algorithms on long bus routes.

Figure F.4 compares the probability OD flow estimates produced by the IPF, CM and HEM methods to the underlying true values in one simulation trial. The horizontal and vertical values of one data point in Figure F.4 represent the assumed underlying true and estimated probability OD flows (by the IPF, CM or HEM method) for one feasible OD pair. By visual inspection, the HEM estimates are better than the IPF and CM estimates. The results are consistent with what has been seen in Figure 5.6 for the CLS bus route.
The procedure of generating boarding and alighting counts on 500 bus trips is repeated 100 times and the probability OD matrix is estimated by various methods in each replication. The resulting probability OD estimates are used to produce the ECDFs of the probability OD flow estimates. The ECDFs of the probability OD flow estimates for 16 OD pairs are presented in Figure F.5 for illustration purposes. The subplots in Figure F.5 are sorted in descending order by the magnitude of the underlying true probability OD flows from the left to the right and from the top to the bottom of the figure. As can be seen in Figure F.5, the ECDFs of the probability OD flows estimated by the HEM algorithm are closer to the underlying true values than those estimated by the
IPF and CM methods for most OD pairs. The results are consistent with what has been seen in Figure 5.7 for the CLS bus route.

Figure F.5 ECDFs of the Probability OD Flow Estimates for Individual OD Pairs on Route 7 SB

The overall performance of various estimates could be summarized by the performance and relative performance measures, which is convenient given the large number of feasible OD pairs. Figures F.6 presents the ECDFs of the $HD^2$, $SSD$ and $Chi^2$ measures and their corresponding $RP$ measures for various estimates. As can be seen, the HEM algorithm performs the best among the methods that are based on boarding and alighting counts (i.e., the IPF, CM and HEM methods). The CM algorithm is better than
the IPF method, but it is worse than the HEM algorithm. The results are consistent with what has been seen in Figure 5.8 for the CLS bus route.

Figure F.6 ECDFs of the Performance and $RP$ Measures of OD Estimation methods on Route 7 SB

To demonstrate that the HEM algorithm takes better advantage of the large quantities of APC data than the IPF method, the effect of the number of bus trips with APC counts on the accuracy of OD estimates is investigated on Route 7 SB. The APC counts are assumed to be free of measurement errors (i.e., $w = 0$ in the measurement error 
model in Equation 7.1 of Chapter 7) or contain acceptable measurement errors (i.e., \( w = 0.15 \) in the measurement error model in Equation 7.1 of Chapter 7).

The simulation process is described as follows. Firstly, given the assumed number of bus trips with APC counts, volume OD flow matrices are generated from the true underlying probability OD flow matrix and the corresponding true trip-level boarding and alighting counts are derived from the generated trip-level volume OD flow matrices. Secondly, given an assumed value of the parameter \( w \) for the distribution of the multiplicative factor \( \beta \) in Equation 7.1 of Chapter 7, a multiplicative factor \( \beta \) is generated based on the assumed symmetric triangular distribution for boarding or alighting count at each bus stop on each of the assumed bus trips. The observed boarding or alighting count is generated according to the measurement error model in Equation 7.1 of Chapter 7 conditional on the generated multiplicative factor \( \beta \) and the true boarding or alighting count. Thirdly, the observed boarding and alighting counts on each bus trip are balanced such that the total boarding count is equal to the total alighting count and no negative passenger load exists between bus stops for each bus trip (See Section 6.3 for the balance procedure). Finally, the IPF, CM and HEM methods are applied to the balanced trip-level boarding and alighting counts to estimate the period-level probability OD matrix. The process described above is replicated 100 times, where with each replication different volume OD flow matrices and different observed trip-level boarding and alighting counts are generated.

Figures F.7 and F.8 present the relationship between the average \( HD^2 \) performance measures and the corresponding \( HD^2 \)-based \( RP \) measures of various
estimates and the number of bus trips with APC counts, respectively. The results in the plot of the left column assume that the APC counts are free of measurement errors (i.e., $w=0$) and the results in the plot of the right column assumes that APC counts contain acceptable measurement errors (i.e., $w=0.15$).

Figure F.7 $HD^2$ Performance Measure versus the Number of Bus Trips with APC Counts
Similar with what have been seen in Figures 7.11 and 7.12 for the CLS bus route, the performance of the IPF estimates improves as the number of bus trips with APC counts increases. However, the improvement is practically unnoticeable at certain point. By contrast, the performance of the CM and HEM estimates continues to improve. At certain point, the CM and HEM estimates outperform the IPF estimates. In addition, the HEM estimates are consistently better than the CM estimates.

The ECDFs of the $HD^2$ measures and the $HD^2$-based RP measures are presented in Figures F.9 and F.10, respectively. Similar conclusions as drawn in Figures F.6 and F.7 can be obtained in Figures F.9 and F.10.
Figure F.9 ECDFs of the \( HD^2 \) Measure under Various Numbers of Bus Trips with APC Counts

Figure F.10 ECDFs of the \( HD^2 \)-based \( RP \) Measure under Various Numbers of Bus Trips with APC Counts

354
The performance of various estimates is also evaluated based on the SSD and the \textit{Chi2} measures. The results are presented in Figures F.11-F.18. Similar conclusions as those associated with the $HD^3$ measure can be obtained in Figures F.11-F.18.

Figure F.11 \textit{SSD} Performance Measure versus the Number of Bus Trips with APC Counts
Figure F.12 SSD-based RP Performance Measure versus the Number of Bus Trips with APC Counts
Figure F.13 ECDFs of the SSD Measure under Various Numbers of Bus Trips with APC Counts

Figure F.14 ECDFs of the SSD-based RP Measure under Various Numbers of Bus Trips with APC Counts
Figure F.15 Chi2 Performance Measure versus the Number of Bus Trips with APC Counts
Figure F.16 _Chi2-based RP_ Performance Measure versus the Number of Bus Trips with APC Counts
Figure F.17 ECDFs of the \( \text{Chi2} \) Measure under Various Numbers of Bus Trips with APC Counts

Figure F.18 ECDFs of the \( \text{Chi2-based RP} \) Measure under Various Numbers of Bus Trips with APC Counts