Passenger-to-Itinerary Assignment Model Based on Automated Data

A Dissertation Presented

by

Yiwen Zhu

to

The Department of Civil and Environmental Engineering

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the field of

Transportation

Northeastern University

Boston, Massachusetts

August 2017
ABSTRACT

Many subway systems around the world are experiencing large increases in demand, especially during peak periods. As a result, they operate at (or near) capacity, with crowding an increasingly important issue that operators must deal with. Systems can be very congested with passengers frequently unable to board the first train.

This dissertation addresses the important problem of evaluating the performance of subway systems operating near capacity, especially from the passenger’s point of view. It develops a key building block towards this goal, the Passenger-to-Itinerary Assignment Model (PIAM), to identify the boarding train(s) of each passenger using Automatic Fare Collection (AFC) and Automatic Vehicle Location (AVL) data.

PIAM is a model system, with a number of individual modules that interact but can also be used independently: the access/egress time model, left behind model, route choice model, and assignment model. The overall problem is challenging because of the large number of feasible itineraries a passenger may have. To deal with this, PIAM first estimates the left behind probabilities by station and time interval at the aggregate level and then assigns individual passengers to itineraries. The model is extended to incorporate trips involving route choice by estimating the route choice fractions and integrating them into the assignment model.

The research also proposes an alternative, computationally efficient method using principles from queuing theory to estimate time dependent aggregate crowding levels at stations and on platforms.

The methodology is validated using synthetic data and the performance compares favorably to a recent model (Hörcher et al., 2017). It is also applied using actual data from a congested, subway system during peak hours. A series of applications
are developed using PIAM output to assess the capacity utilization of the network, including train load estimation, passengers left behind, journey time components, crowding at stations, performance under special events, etc.
ACKNOWLEDGMENTS

I would love to show my deepest gratitude to my supervisors, Prof. Koutsopoulos and Prof. Wilson, for their support over the past five years. Without them, I wouldn’t have finished this work. Haris has been involved in my research since my first year at MIT and guided me through the whole research project from the very beginning. I always irritate him in many aspects but he is patient and trying to work me through things step by step. I learned a lot from him. He is very supportive, and I really really appreciate his time and effort on me. Nigel is my first academic and research supervisor. As the PhD supervisor for both, Haris (a long time ago) and me, he provided many critical advices to me and three of us together explored many interesting research ideas. Without their guidance, many of the work in this research would not be possible. I would also love to thank their support both, in research and in life. It has been an unforgeable five years with many things happened, and I couldn’t be more grateful for their encouragement and kindness.

I would like to thank and acknowledge other faculty members, Jinhua Zhao, John Attanucci, Fred Salvucci, Mikel Murga, Gabriel Sánchez-Martínez, Neema Nassir, Jay Gordon and Peter Furth for their support in my research and many other aspects. It has been a great time working in the lab with many excellent researchers and nice people. I am very grateful for having them.

It has been a great time to be with my dear lab mates, Peyman Noursalehi, Mike Ma, Danny Wang, Tianyou Liu, Pu Yue, Zhan Zhao, Cecilia Viggiano, Anson Stewart and the “younger generations” of transit lab. Thanks to my PhD fellows, Tianli Zhou, Linsen Chong, Li Jin, Haizheng Zhang, and Chao Zhang. Thank you for all the advices and help in the past five years. It has been a tough time (especially
in 2015 and this semester). And I really appreciate your support and help to lead me through. I wish all the best for you in the future.

My best friends, Shiyong Qiu, Shiying Zhou, Yang Liu, Zheng Yin, Xiaohua Wang, Shiyu Mei, Zexuan Xu, and Yi Su in Africa, I am very grateful to have you with me for so many years and helping me to overcome difficulties in life. You always light up my days. And my friends Can Zhang and Zhe Han from Tsinghua (the top university in China), as we are all poor PhDs in US, you encouraged me a lot both, in research and in daily life just by letting me know what you were doing. Thanks to Yue Chen and Peng Wang, my roommates at Ashdown, for all the support for the past 5 years. You are like my families. Thanks to my friends, Katie Pincus, Farah Machlab, Michel Babany. And I will miss Boston so much.

Finally, special thanks to my family, especially my mom, Linling Zheng and my husband, Chiwei Yan. I am very blessed to always have your support and you are the most precious treasure in life. I wish you all healthy and happy all the time.
# TABLE OF CONTENTS

## 1 Introduction

1.1 Problem Description ........................................ 3
1.2 Objectives .................................................... 4
1.3 Approach ...................................................... 5
1.4 Contributions ............................................... 9
1.5 Dissertation Structure ..................................... 9

## 2 Literature Review

2.1 Automated Data Sources .................................. 11
2.2 Passenger Assignment ...................................... 14
   2.2.1 Planning-Level Studies .............................. 15
   2.2.2 Operation-Level Studies ............................ 16

## 3 PIAM without Transfers

3.1 Introduction ............................................... 23
3.2 Methodology ............................................... 24
   3.2.1 Passenger Assignment Model .................... 25
   3.2.2 Train Load Estimation ............................ 30
   3.2.3 Left Behind Distribution ......................... 30
3.3 Model Validation and Application ...................... 32
   3.3.1 Validation with Synthetic Data ................. 32
   3.3.2 Validation with Actual Data .................... 39
   3.3.3 Performance Metrics .............................. 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.4</td>
<td>Visualization</td>
<td>43</td>
</tr>
<tr>
<td>3.4</td>
<td>Summary</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>General PIAM: Framework</td>
<td>47</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>47</td>
</tr>
<tr>
<td>4.2</td>
<td>Methodology</td>
<td>49</td>
</tr>
<tr>
<td>4.3</td>
<td>Summary</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>General PIAM: Left Behind Model</td>
<td>53</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>5.2</td>
<td>Methodology</td>
<td>54</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Problem Definition</td>
<td>55</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Problem Formulation</td>
<td>56</td>
</tr>
<tr>
<td>5.3</td>
<td>Model Validation</td>
<td>64</td>
</tr>
<tr>
<td>5.4</td>
<td>Application Using Actual Data</td>
<td>68</td>
</tr>
<tr>
<td>5.4.1</td>
<td>System Description</td>
<td>68</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Estimated Probabilities</td>
<td>70</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Comparison with Manual Observations</td>
<td>71</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>General PIAM: Assignment Model</td>
<td>75</td>
</tr>
<tr>
<td>6.1</td>
<td>Methodology</td>
<td>75</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Problem Definition</td>
<td>76</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Problem Formulation</td>
<td>78</td>
</tr>
<tr>
<td>6.2</td>
<td>Model Validation</td>
<td>83</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Synthetic Data Generation</td>
<td>83</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Assumption Validation</td>
<td>85</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Results</td>
<td>85</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Comparison with Hörcher et al. (2017)</td>
<td>88</td>
</tr>
<tr>
<td>6.3</td>
<td>Application with Actual Data</td>
<td>94</td>
</tr>
<tr>
<td>6.4</td>
<td>Extension to Trips with Route Choice</td>
<td>97</td>
</tr>
</tbody>
</table>
9.1 Summary .................................................. 134
9.1.1 PIAM for Trips without Transfers ....................... 134
9.1.2 General PIAM ............................................ 135
9.1.3 PIAM for Trips with Route Choice ...................... 135
9.1.4 Crowding Model ......................................... 136
9.1.5 Access/Egress Time Model .............................. 137
9.2 Contributions ............................................. 137
9.2.1 The PIAM Framework .................................. 138
9.2.2 Individual Modules .................................... 138
9.3 Future Work ............................................. 139
9.3.1 Methodological Enhancements .......................... 139
9.3.2 Applications ........................................... 140
References .................................................... 142
LIST OF FIGURES

1-1 Passenger-to-Itinerary Assignment Model (PIAM) ....................... 2
1-2 Time-Space diagram for a passenger with one transfer .................. 3
1-3 Access/egress time model ....................................................... 6
1-4 General PIAM framework ....................................................... 7
1-5 Relationship between models ................................................... 8

2-1 Density map between Vincennes and Nation for half-hour data (source: Aguilera et al., 2013) ............................................................... 14

3-1 Passenger time-space diagram ................................................... 25
3-2 Passenger-to-Itinerary Assignment Model (PIAM) ......................... 27
3-3 Case study network ................................................................. 32
3-4 Journey time distribution .......................................................... 33
3-5 Distribution of the number of feasible itineraries for passengers depart-
ing Station 2 ............................................................... 34
3-6 Distribution of the probability of boarding the actual train ............... 35
3-7 Distribution of the dominant probabilities ..................................... 36
3-8 Distribution of the number of times passengers are left behind at Sta-
tion 2 ........................................................................... 37
3-9 Passenger load estimation during the off-peak ................................ 38
3-10 Passenger load estimation during the peak .................................... 38
3-11 Passenger loads on trains vs headways ....................................... 39
3-12 Probability of boarding the first train .......................................... 40
3-13 Distribution of the number of times a passenger is left behind ........ 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-7</td>
<td>Distribution of the probabilities of assigning passengers to their actual itinerary</td>
<td>87</td>
</tr>
<tr>
<td>6-8</td>
<td>Transfer times as a function of time</td>
<td>87</td>
</tr>
<tr>
<td>6-9</td>
<td>Train load estimation</td>
<td>89</td>
</tr>
<tr>
<td>6-10</td>
<td>Distribution of transfer times</td>
<td>90</td>
</tr>
<tr>
<td>6-11</td>
<td>Distribution of the probabilities of boarding the actual itinerary</td>
<td>91</td>
</tr>
<tr>
<td>6-12</td>
<td>Distribution of the probability of boarding the actual train for different segments based on the Hörcher et al. model</td>
<td>91</td>
</tr>
<tr>
<td>6-13</td>
<td>Comparison of transfer time estimation</td>
<td>92</td>
</tr>
<tr>
<td>6-14</td>
<td>Train load estimation for the Blue Line (departing Station 1)</td>
<td>93</td>
</tr>
<tr>
<td>6-15</td>
<td>Expected number of times a passenger is left behind as a function of time</td>
<td>94</td>
</tr>
<tr>
<td>6-16</td>
<td>Distribution of the number of feasible itineraries</td>
<td>95</td>
</tr>
<tr>
<td>6-17</td>
<td>Estimated transfer time for each passenger</td>
<td>96</td>
</tr>
<tr>
<td>6-18</td>
<td>Distribution of passenger in-station times (Station 1)</td>
<td>97</td>
</tr>
<tr>
<td>6-19</td>
<td>PIAM structure for a passenger with route choice</td>
<td>98</td>
</tr>
<tr>
<td>6-20</td>
<td>PIAM modeling framework</td>
<td>99</td>
</tr>
<tr>
<td>6-21</td>
<td>Network with transfers and route choice</td>
<td>102</td>
</tr>
<tr>
<td>6-22</td>
<td>Probability of transferring at Station 2</td>
<td>103</td>
</tr>
<tr>
<td>6-23</td>
<td>Probability of assigning to the actual itinerary</td>
<td>103</td>
</tr>
<tr>
<td>6-24</td>
<td>Probability of transferring at Station 2</td>
<td>104</td>
</tr>
<tr>
<td>7-1</td>
<td>Simplified station layout</td>
<td>109</td>
</tr>
<tr>
<td>7-2</td>
<td>Maximum likelihood estimation results</td>
<td>114</td>
</tr>
<tr>
<td>8-1</td>
<td>Average queue length from $t_1$ to $t_2$</td>
<td>118</td>
</tr>
<tr>
<td>8-2</td>
<td>Average queue length from $t_1$ to $t_2$</td>
<td>119</td>
</tr>
<tr>
<td>8-3</td>
<td>Calculation of arrival rate at Station 2 example</td>
<td>121</td>
</tr>
<tr>
<td>8-4</td>
<td>Time-space diagram for passengers (1-2-3)</td>
<td>122</td>
</tr>
<tr>
<td>8-5</td>
<td>Passenger flow through a transfer station</td>
<td>127</td>
</tr>
<tr>
<td>8-6</td>
<td>Station crowding estimation</td>
<td>129</td>
</tr>
</tbody>
</table>
8-7 Headway distribution ...................................... 130
8-8 Number of times passengers are left behind at Station 2 and their wait
time ................................................................. 130
8-9 Crowding level at Station 2 .............................. 131
LIST OF TABLES

1.1 Summary of models ........................................ 5

5.1 Estimation result ......................................... 66

7.1 Comparison of the manual and estimation results ............ 115
Chapter 1

Introduction

Many subway systems around the world are experiencing large increases in demand, especially during peak periods. The Mass Transit Railway system (MTR) in Hong Kong has seen an 83.8% increase in patronage in the past 10 years (MTR Corporation, 2016). Transport for London (TfL) had an increase of about 70% in both rail and bus trips since 2000 (Mayor of London, 2015). The New York City subway system has experienced a 21.5% increase in weekday passengers in the past 10 years. As a result of this increased demand, many systems operate at (or near) capacity during the peak periods. Crowding has become an increasingly important issue that operators must deal with. During the peak, even with very short headways, systems can be very congested with passengers frequently unable to board the first train. These conditions can significantly affect system reliability and service quality, and can also raise safety concerns.

To better deal with the increasing demand, operators are interested in:

- Measuring the impact on passengers due to near capacity operations,
- Better understanding the performance of the system especially when it is stressed by large demand,
- Developing short-term strategies to improve capacity utilization,
- Providing information to passengers about crowding conditions.
This research aims to develop tools for transit operators to monitor operations of a rail transportation network using automated data. The Passenger-to-Itinerary Assignment Model (PIAM) serves as the most important building block for this research (see Figure 1-1).

Figure 1-1: Passenger-to-Itinerary Assignment Model (PIAM)

The model utilizes and integrates two main data sources–Automatic Fare Collection (AFC) and Automatic Vehicle Location (AVL) data. The emergence of means of collecting data automatically, such as AFC and AVL systems, is the enabling factor for addressing the problem. The methods developed require AVL and AFC data from subway systems with both entry and exit station transactions (tap-in and tap-out). The AFC data describes the passenger demand in the network by time of day and the AVL data provides detailed information about the movement of trains. The model aims to capture the interaction between the demand and supply in the transit network and provide more detailed information on individual trips (e.g. number of times a passenger is left behind, journey time components, crowding levels, etc.). It also enables a close examination of passenger movements in the system. The resulting out-
put supports system monitoring and performance measurement from the passenger’s point of view, as well as enhanced customer information. PIAM infers details of the journey passengers made on a particular day based on the AFC and AVL data, while the traditional schedule-based assignment models (e.g. Nuzzolo et al., 2001; Poon et al., 2004; Hamdouch and Lawphongpanich, 2008; Sumalee et al., 2009; Nuzzolo et al., 2012) are planning tools that target future conditions.

The remainder of the chapter is organized as follows. Section 1.1 describes the assignment problem. Section 1.2 discusses the objectives. Section 1.3 outlines the general approach used to address the problem. Section 1.4 summarizes the main contributions. Section 1.5 presents the structure of the dissertation.

1.1 Problem Description

Figure 1-2 shows the movements of a passenger who enters the system at $t^{in}$ and exits at $t^{out}$ with one transfer.

![Time-Space diagram for a passenger with one transfer](image)

Figure 1-2: Time-Space diagram for a passenger with one transfer

He/she can board Trains 1_1, 1_2, or 1_3 for the first segment, and Trains 2_1, 2_2,
or 2,3 for the second segment. The feasible itineraries are different combinations of the trains for the two segments. The Passenger-to-Itinerary Assignment Model (PIAM) aims at identifying passenger boarding events and hence inferring the specific train itineraries individuals used. The model estimates the probabilities of a passenger boarding each train, the expected number of times a passenger is left behind due to capacity constraints, journey time components, etc.

1.2 Objectives

The broad goal of this research is to understand how transit systems operate near capacity by examining the capacity utilization of the network.

More specifically, based on the assignment results, at the individual level, different journey time components (in-station access time, wait time, in-vehicle time, transfer time and in-station egress time) can be computed for each passenger. At the aggregate level, this information allows for accurate examination of the customer experience and the utilization of the network.

The model facilitates a number of applications:

- Providing a complete inference of a passenger’s movements at a high resolution,

- Developing service quality metrics from the passenger’s point of view, such as:
  - Probability distribution of the number of times passengers were left behind,
  - Travel time variability and overall service reliability,
  - Crowding on trains, at platforms, and in stations.

- Improving customer communication,
  - Enhanced travel planner and customized passenger information with crowding status at stations and on trains.
1.3 Approach

Table 1.1 summarizes the models developed in this dissertation.

<table>
<thead>
<tr>
<th>Tools</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access/Egress Time Model</td>
<td>AFC/AVL, Station layout</td>
<td>Access/egress time distribution</td>
</tr>
<tr>
<td>PIAM (Non-Transfer)</td>
<td>AFC/AVL, Station layout, Access/egress time distribution</td>
<td>Passenger movements, Left behind probabilities</td>
</tr>
<tr>
<td>Left Behind Model</td>
<td>AFC/AVL, Station layout, Access/egress time distribution</td>
<td>Left behind probabilities</td>
</tr>
<tr>
<td>General PIAM Route Choice Model</td>
<td>AFC/AVL, Station layout, Access/transfer/egress time distribution</td>
<td>Route choice fractions</td>
</tr>
<tr>
<td>Assignment Model</td>
<td>AFC/AVL, Station layout, Access/transfer/egress time distribution</td>
<td>Passenger movements, Left behind probabilities</td>
</tr>
<tr>
<td></td>
<td>Left behind probabilities</td>
<td>Journey time components, Train load</td>
</tr>
<tr>
<td></td>
<td>Route choice fractions</td>
<td>In-station crowding</td>
</tr>
<tr>
<td>Crowding Model</td>
<td>AFC/AVL, Route choice fractions</td>
<td>In-station crowding</td>
</tr>
<tr>
<td></td>
<td>Journey time components (Method I), left behind probabilities (Method II)</td>
<td></td>
</tr>
</tbody>
</table>

- Access/egress time model

The access/egress time distribution is a key input to the PIAM. A model is developed to estimate the mean and variance of the passenger access/egress time using AFC/AVL data and station layouts. The method consists of two modules, the walk speed model and the distance model. It can incorporate the impacts of station characteristics and passenger behavior in choosing their access and egress paths (see Figure 1-3).
The proposed model parameters can be estimated using maximum likelihood with data from trips with a single feasible itinerary. For such trips, the egress time is known. However, the sample is truncated (i.e. it includes only passengers with egress time less than the corresponding headway). The proposed formulation takes this important characteristic into account and corrects for biases due to the truncated nature of the sample.

- **PIAM for non-transfer trips**
  
  The PIAM is first developed for trips without transfers. This model serves as an important building block for the more general case of the problem involving transfers and route choice. The method estimates the probability of a passenger boarding each feasible train, and the left behind probabilities.

- **General PIAM (for trips with transfers and route choice)**
  
  The general problem is complex because of the large number of feasible itineraries a passenger may have, given their tap-in and tap-out times, especially for trips with transfers and route choice. Prior methods are either based on simplifying assumptions that bias the results and limit the applicability of the model or address special cases of the problem (e.g. no transfers). To address this problem, a two-step approach is proposed that uses as an input the access/egress time
distribution estimated by the access/egress time model (see Figure 1-4). This model consists of two modules: the left behind model and the assignment model. At the aggregate level, the left behind model groups passengers based on their estimated arrival time at the station platform and estimates the probability of being left behind by station and time interval using either maximum likelihood or Bayesian inference. At the disaggregate level, the assignment model assigns each passenger to their feasible itineraries based on the corresponding probability of being left behind, and the access, transfer and egress time distributions.

Figure 1-4: General PIAM framework

For trips involving route choice, a similar approach is used. Data about train departure/arrival times, left behind probabilities and access/transfer/egress time distribution are used to estimate route choice fractions by time interval using maximum likelihood. The estimated route choice fractions and left behind probabilities are used in the assignment step.

- **Crowding model**

  Calculation of crowding levels in stations using PIAM requires the application
of the model to all OD pairs in the network, and is computationally demanding. For that reason, this research proposes a crowding model to estimate the number of passengers in stations and on platforms using principles from queuing theory. These methods have computational advantages as they only require the application of PIAM to limited OD pairs. Two methods are proposed. Method I has as input the journey time decomposition results from PIAM (for trips without transfers). Method II has as input the left behind probabilities. This quantity can be estimated from PIAM results (for trips without transfers) or from the left behind model. The results can be used to infer station and platform crowding for any station (transfer and non-transfer).

Figure 1-5 summarizes the relationship among the different models developed in this dissertation. Using AFC/AVL data and station layouts, the access/egress time model estimates the access and egress time distributions as important inputs to PIAM. The crowding model uses as input the left behind probabilities estimated by the left behind model or the journey time components estimated by the passenger assignment model to estimate efficiently the crowding level in stations and on platforms. While the individual modules interact, they can also be used independently to calculate useful performance metrics.

![Figure 1-5: Relationship between models](image-url)
1.4 Contributions

The main contributions of this research are:

1. Develops a data-driven approach to the Passenger-to-Itinerary Assignment problem at the individual passenger level. The method is able to infer trip details, such as passenger movements, components of journey time, expected number of times left behind, etc.,

2. Generates various aggregate performance metrics for the operator to assess the capacity utilization of the network, such as the estimation of individual train loads, etc.,

3. Calculates left behind probabilities and route choice fractions using aggregate methods, such as maximum likelihood and Bayesian inference,

4. Proposes an efficient method to estimate time dependent crowding levels at stations and on platforms, both, with and without transfers,

5. Develops a method to estimate the walk speed and access/egress time distributions at a given station using AFC and AVL data.

1.5 Dissertation Structure

The remainder of the dissertation is organized as follows. Chapter 2 reviews previous research on passenger assignment models and related topics. Chapter 3 develops the methodology for the passenger assignment problem for trips without transfers. Chapter 4 proposes the overall framework for the general passenger assignment problem for trips with transfers, which consists of two modules: the left behind model and the assignment model. Chapters 5 develops the left behind model to infer the probabilities of being left behind at the aggregate level by station and time interval using data from trips without transfers. Chapter 6 develops the assignment model based on the left behind probabilities and extends the PIAM framework to trips involving route
choice. Chapter 7 discusses the methodology for estimation of the parameters of the access/egress time distribution as a key input to the PIAM. Chapter 8 proposes a crowding model that directly infers the number of passengers in a station and the number of passengers waiting on a platform based on the PIAM output. Chapter 9 summarizes the research findings and proposes areas for future research.
Chapter 2

Literature Review

This chapter reviews the literature related to the passenger assignment problem and related topics. The chapter is organized into two sections. Section 2.1 discusses the use of automated data to obtain a better understanding of transit system performance and passenger behavior. Section 2.2 focuses on the passenger assignment problem especially at the train/vehicle level.

2.1 Automated Data Sources

The availability of data from smart cards and train tracking systems is critical for the models developed in this research. Smart cards serve the function of fare collection for transit agencies, but they also constitute a significant data source that helps operators to better understand passenger behavior and system performance. Analysis of Automatic Fare Collection (AFC) data can support a number of important transit agency functions including: (i) understanding passenger travel behavior, (ii) monitoring transit demand and (iii) evaluating system performance (Bagchi and White, 2005; Agard et al., 2006; Zhao et al., 2007; Chan et al., 2007; Wilson et al., 2009; Pelletier et al., 2011; Ortega-Tong, 2013; Langlois et al., 2016; Noursalehi and Koutsopoulos, 2016).

Morency et al. (2007) used data mining approaches to identify variability of transit use to optimize vehicle allocation and improve operational efficiency. Bagchi and
White (2005) used smart card data to analyze consistency of users’ spatio-temporal travel patterns. Chapleau et al. (2008) used smart card data to monitor activity patterns at trip generators (such as schools) and provided high resolution analysis of both supply and consumption of transit service. Ortega-Tong (2013) classified public transport users into homogeneous clusters based on their travel characteristics, such as activity patterns, mode choice, temporal and spatial variability and socio-demographic characteristics. The travel profile of each group provides insights into user behavior to inform analysis of customer experience and transport planning studies. Langlois et al. (2016) investigated heterogeneity among transit users using four weeks of AFC data. By identifying user activity sequences, 11 clusters were generated with distinct activity structures. Socio-demographic information was combined with smart card transactions to explore the associations between travel patterns and socio-demographic attributes.

A number of studies have used AFC and AVL data to infer Origin-Destination (OD) flow at the system level. For example, Farzin (2008) developed a methodology to integrate AFC, AVL, and station location data to infer the destination zone for each trip. The results were validated using a household survey. Munizaga and Palma (2012) proposed a method to estimate the OD matrix from smart card and GPS data. Time and location of alighting events were inferred for over 80% of the AFC transactions.

Chu and Chapleau (2008) further enriched the inference process to identify transfers using spatio-temporal criteria in a bus network. Gordon et al. (2013) extended previous OD inference methods to include interchanges and analyzed full multi-modal travel itineraries in London. The AFC system for the London Underground records the tap-in/out times and stations for all rail transactions. For bus transactions, only boardings are recorded. Gordon et al. (2013) developed a rigorous process to infer the bus origin/destination and interchange(s) for each individual to build a full-journey OD matrix. Using this as a seed matrix, control totals from various sources (bus ridership, entry/exit counts at rail stations) were then used to scale up the seed OD matrix to represent the full population.
At a more detailed level, the problem of inferring details of passenger journeys, for example, journey time components, passenger locations, and crowding levels, has also drawn interest. Sun et al. (2012) developed a regression model based on AFC data and the distances between origin and destination stations to decompose the gate-to-gate journey time. The analysis was based on the assumption that three key parameters were constant for all users: (i) walk times between gates and platforms, (ii) train speed between stations and (iii) train dwell times at stations. Regression was used to estimate these quantities and their relationship with the total journey time. The estimated parameters were significant and were used to identify the approximate locations of individuals at a point in time and estimate passengers’ “spatio-temporal density” in Singapore’s metro system.

Cellular phone networks operating in underground rail systems also provide new possibilities to model passenger movements. Aguiléra et al. (2013) conducted experiments in Paris’ underground system to assess the potential of this emerging data source to infer travel times, train occupancies, and OD flows. They used the records of signaling events (GSM data) triggered by switched-on mobile phones when they changed locations. This data contained location information at the station level for each cell phone and were used to track passengers throughout their trips. Even though the data was sparse, the aggregation of large amounts of this data provided a way to infer passenger’s movements in the system. A density map was used to link train trajectories with mobile phone records (see Figure 2-1). In the map, each dot represents a cell-phone data record with the x-axis indicating the time at the origin station and y-axis indicating the time at the destination station. The dots fall in a rectangular area that corresponds to a train itinerary, hence associating the cell-phone signal with a train. The comparison of the estimation results with the AFC data analysis showed a high level of consistency, indicating that cell-phone data is a promising data source for operators and transport authorities to monitor system operations.
To better understand passenger’s experience in the underground system, Transport for London (TfL) conducted a four-week trial to collect WiFi connection data from 54 London Underground (LU) stations. The data was used to track the routes passengers used and infer passenger movements within stations. The data has great potential to facilitate advertising and providing crowding information to customers (Transport for London, 2017).

Loadweigh data from weighing systems installed in some systems have been used to infer passenger loads on rail cars. Frumin (2010) regressed loadweigh data against manual passenger counts and estimated the average passenger weight and vehicle tare weight (i.e. the weight when the vehicle is empty) to infer passenger loads on trains. Nielsen et al. (2013) estimated an “inverse model” with the number of passengers as an independent variable and used the weight as the explanatory variable. The model can be used to monitor the capacity utilization of the network (Nielsen et al., 2013).

2.2 Passenger Assignment

Previous studies on the problem of passenger assignment focused on two different needs. At the planning level, transit assignment models serve as a basis for evaluating
possible service adjustments or network changes. At the operational level, the models estimate the capacity utilization of the network based on archived data and provide service performance indicators. This dissertation focuses on the operational level analysis.

2.2.1 Planning-Level Studies

Two main approaches are used for the passenger assignment problem at the planning level: frequency-based and schedule-based. The frequency-based approach analyzes the transit system by line and computes the average passenger flow based on service frequency (Nguyen and Pallottino, 1988; Spiess and Florian, 1989; Cepeda et al., 2006; Schmöcker et al., 2008). The schedule-based approach explicitly considers individual vehicle trips and their scheduled departure/arrival times in assigning passengers to vehicles. This class of models can be used to predict the effects of schedule changes (Nuzzolo et al., 2001; Poon et al., 2004; Hamdouch and Lawphongpanich, 2008; Nuzzolo et al., 2012). Most of the work reviewed in this section uses the second approach.

Nguyen et al. (2001) developed a graph-theoretic framework for the transit assignment problem and considered departure times and route choice simultaneously similar to the traffic equilibrium problem. The disutility of paths was a function of travel time and penalty for late (early) departures (arrivals). Timetables and OD demand were the main inputs to the problem. Passenger flows were obtained using a convergent equilibrium algorithm.

Poon et al. (2004) proposed an optimization formulation for the equilibrium assignment problem which explicitly considered the vehicle capacity at each boarding station. Passengers were assumed to have full information about future network conditions and to select paths to minimize the total cost of the trip (a function of the in-vehicle time, wait time, walk time, and transfer penalty). Equilibrium was achieved by a simulation-based, iterative approach using the method of successive averages. The paths were generated using time-dependent shortest paths as in Tong and Richardson (1984). In each simulation run, the queuing delays were recorded based on the passenger arrival/departure times at stations and minimum cost paths
were calculated dynamically. Capacity constraints were explicitly considered for each vehicle by assuming first come first served (FCFS) queuing discipline at the boarding stations. The method incorporated route choice while other choice dimensions, such as departure times and entry, transfer and exit stations were not considered. The authors suggested that the model could be used as a tool for the evaluation of the performance of a transit system with pre-determined schedules and of the effects of service changes.

Hamdouch and Lawphongpanich (2008) proposed a user equilibrium transit assignment model by solving a dynamic program. Service, demand, and access/egress were presented in a “diachronic” graph. As in Ahuja et al. (1993) and Hamdouch et al. (2004), a time-expanded network was used to represent the system. The model assumed that passengers use travel strategies by specifying, at each station and time point, an ordered list of preferred transit lines. This information was incorporated in the travel costs of the paths. The cost consisted of in-vehicle time, fares, and other costs associated with the strategy, such as crowding, and opportunity costs associated with early departures and late arrivals. Passengers were allowed to cycle or drive to transit stations and transfer at nearby stations by walking. The user equilibrium was formulated as a variational inequality problem and solved by the method of successive averages. Sumalee et al. (2009) proposed a dynamic transit assignment model that incorporates the seating probability to differentiate standing and sitting passengers. The model investigated the effects of the discomfort level on passengers’ travel decisions.

2.2.2 Operation-Level Studies

In this section, models which evaluate (past) system performance based on archived data, also taking into account service disruptions, are reviewed.

Buneman (1984) used a reverse-time approach to assign passengers to trains assuming that they were served by the last train that arrived at their exit station. The paper also assumed that transfer passengers boarded the last feasible departing train. Using data from the Bay Area Rapid Transit (BART) system, the method produced
operational performance measures, such as delays, and served as a daily analysis tool for the operator. However, during peak periods, the last train arriving at the destination before a passenger taps-out could be the train after the one that the passenger actually boarded (if the egress time is longer than the preceding headway). This is quite possible at large stations during peak hours, when the trains are running with short headways.

Kusakabe et al. (2010) introduced choices between express trains, rapid trains, local trains, and multiple transfer stations in a study using data from the Japanese Railways. They solved the problem by developing a time-space network to represent the train trajectories, and by enumerating all possible paths for each individual. Based on AFC data, they assigned all passengers to the “shortest path” (the path with the minimum access time at the entry station, the minimum egress time at the destination station and the least number of transfers).

Paul (2010) developed a method to assign individual passengers to trains using train tracking and AFC data from the London Underground (LU). Instead of assuming a constant walk time as did Sun et al. (2012), the method relied on the distributions of access and egress times which were derived from two sources:

- London Underground surveys, in which the journey time components were measured by surveyors following random passengers and recording their walk times, wait times, in-vehicle times, etc.

- The subset of passengers (about 10%), from Oyster card transactions, who had only one possible train itinerary based on their tap-in and tap-out times. For this subset of passengers, their egress times were derived as the difference between their tap-out times and the train arrival times.

The ratio between the expected value of access and egress time distributions was assumed to be the same as the ratio of the average access time and egress time in the manual survey. Under this assumption, the egress time distribution was first estimated based on passengers with a single itinerary and then “scaled up” using the ratio based on the manual survey. After examining all the possible train itineraries, a
“same percentile” assumption was made to select the most likely train itinerary and corresponding route. That is, given the egress time calculated for each itinerary, the access time and interchange time were assumed to be at the same percentile of the cumulative distribution as the egress time is in the distribution at the exit station. If the time intervals related to access and transfer times in an itinerary were less than these percentiles, this itinerary would be eliminated since no wait times were included in it. However, during each step of the elimination, if all the itineraries were deleted, the method would keep all the itineraries and continue to the next step.

While the method advanced earlier approaches to this topic, it has a number of limitations:

- The distribution of the journey time components may be biased toward lower access/egress times because of the lack of representativeness of the sample. The passengers with a single itinerary are likely those with faster walk speeds which enable them to catch the first train arriving after they tap-in, and tap-out before the second train arrives at the destination station. Passengers with longer access and egress times generally have more feasible itineraries and will not be in the sample.

- Due to data limitations, the assumptions used, such as the “same percentile” assumption, could not be fully tested, especially when the access and egress times are influenced by external factors other than individual characteristics, such as station configuration, crowding level, etc. In the selection process, due to this assumption, all itineraries may be eliminated and the program reverts to the original itinerary set.

- The method is deterministic and neglects many other possible outcomes.

- The method requires data on access/egress times that may not be readily available.

Sun and Schonfeld (2015) developed a schedule-based assignment model using AFC data. Assuming minimum/maximum egress times, a number of passengers
have one feasible itinerary. Assuming that access time is equal to the minimum access time, the number of times those passengers were left behind is known. Non-transfer passengers were assigned based on the left behind probabilities estimated from passengers with one feasible itinerary. The authors further assume that transfer time is equal to the minimum transfer time. Under these assumptions, the number of times a passenger is left behind at the origin station, or transfer station, for each feasible itinerary, is known and the probability of using this itinerary was calculated based on corresponding left behind probabilities. The model relies heavily on the left behind probabilities which, as discussed above, are estimated based on many simplifying assumptions that are difficult to verify.

Zhu (2014) proposed a probabilistic methodology, the Passenger-to-Train Assignment Model (PTAM). The model used AFC/AVL data and the access/egress walk speed distribution (also estimated from AFC transactions using maximum likelihood). The output of the model is the probability that a given passenger took a specific train among the set of feasible itineraries. At the aggregate level, the train load was estimated based on the assignment results with good accuracy. This method is applicable to lines operating without binding capacity constraints without transfers.

Hörcher et al. (2017), in a recent paper focusing on estimating the crowding penalty in a discrete route choice framework, extended the PTAM framework proposed in Zhu (2014) to incorporate transfers and route choice. For non-transfer trips, the assignment was based on the egress time distribution estimated from trips with a single feasible itinerary. Based on the assignment results for those trips, a “delayed access time” distribution was estimated, defined as the time between tapping-in and boarding at the origin station. For transfer trips, the assignment was based on the delayed access time distribution at the origin station and egress time distribution at the destination station. The probability of choosing each itinerary was assumed to be proportional to the corresponding likelihood of delayed access time at the origin station and egress time at the destination station.

The model introduced several simplifications, that are not necessarily justified, to deal with transfer trips (Zhu et al., 2017c):
• It does not take into account the dynamics at the transfer station (left be\-hinds, cross platform transfers, etc). These can be very important. Consider for example, itineraries with the same egress time and similar likelihood of delayed access time. In this case, the likelihood of the transfer time and the left behind probability at the transfer station are critical to the correct assignment of passengers.

• The egress time distribution was derived using trips with a single feasible itinerary. As mentioned above, this can be biased and this dissertation proposes a method to correct for the bias and properly estimate the parameters of the distribution.

• In order to obtain the “delayed access time” distribution, a large sample of access times is required, therefore a long period of observation is needed. Furthermore, Zhu et al. (2017a) reported that the crowding level often changes very fast during the peak period, hence the delayed access time distribution observed over a long period of time only reflects the average crowding level. The peak of the peak has much higher left behind probabilities than the average. Hence, the approach proposed by Hörcher et al. (2017) may lead to biased results.

Zhao et al. (2017) proposed a method to integrate AFC data with timetables to infer the boarding train for non-transfer passengers and the route choice fractions in subway networks. Their approach is based on the assumption that the access, transfer, and egress times of the passengers are shorter than the train headways. Under this assumption, the last train in the passenger’s itinerary can be easily identified. Passengers were classified into three groups: (1) no-transfer-one-route, (2) one-transfer-one-route, and (3) multi-routes. Based on the above assumption, the train each passenger in group 1 boarded is known and hence the distribution of the left behind at the origin station. The left behind probabilities calculated for group 1 passengers at the origin station were then used to estimate left behind probabilities at transfer stations for group 2 passengers. For those passengers, Zhao et al. (2017) assumed that they boarded the last feasible train in the last segment of their
itinerary. Given the boarding train for the second segment, the number of times a passenger is left behind at the origin and transfer stations can be identified for each feasible itinerary based on the above assumptions. The left behind probabilities at the transfer stations were estimated by maximizing the likelihood of group 2 passengers passing through the exit gate at the observed times. The left behind probabilities were not used for itinerary assignment. Rather, they became input to a model to estimate route choice fractions through maximum likelihood.

The proposed approach does not take into account walk speed variability among passengers. The assumption that passenger access/transfer/egress time is shorter than the headway may not be valid, especially at large stations during peak hours when trains are running with short headways. In addition, under service disruptions and delays, headway variability may also lead to violation of this assumption. Furthermore, many stations in large systems are quite complex with long access/egress/transfer distances. The last train arriving at the destination could be the one after the passenger’s actual boarding train.

The approach we propose in this research focuses on the inference of the itinerary a passenger most likely used and deals with many of the limitations of previous research.

1. We propose a method to estimate access/egress time distribution using egress time observations from passengers with one feasible itinerary. The proposed formulation corrects for bias due to the truncated nature of the sample.

2. We propose a rigorous method to estimate left behind probabilities by station and time interval that is able to capture rapid changes of crowding levels at stations.

3. We explicitly consider the dynamics at origin, transfer, and destination stations, and use distributions of access/egress/transfer time and left behind probabilities to assign passengers to itineraries.

4. The model is probabilistic and applicable to trips with and without transfers and route choice under capacity constraints. The methods are validated with
synthetic data and actual data. The results show that the probability of assigning a passenger to the correct itinerary/route and the estimated journey time components are more accurate than the previous models.
Chapter 3

PIAM without Transfers

The chapter presents a methodology for assigning passengers to individual trains using: (i) fare transaction records from Automatic Fare Collection (AFC) systems and (ii) Automatic Vehicle Location (AVL) data from train tracking systems. The Passenger-to-Itinerary Assignment Model (PIAM) in this chapter focuses on non-transfer trips. However, it serves as an important building block for the general case with transfers and route choice.

The method estimates the probability of a passenger boarding each feasible train, and the probability distribution of the number of trains a passenger is unable to board due to capacity constraints. The methodology is applied using both synthetic and actual data.

The remainder of the chapter is organized as follows. Section 3.1 introduces the problem. Section 3.2 describes the approach. Section 3.3 validates the model using synthetic data generated based on actual AFC transactions and train movement data and by applying the model using actual data. Section 3.4 concludes the chapter.

3.1 Introduction

The general passenger assignment problem is complicated and challenging, especially for transfer trips, due to the large number of feasible itineraries. In this chapter, we develop a Passenger-to-Itinerary Assignment Model (PIAM) for trips without
transfers, which is a simpler problem to solve initially but serves as an important building block to address the general problem with transfers and route choice.

The proposed methodology is applicable to systems operating close to capacity (as opposed to the work by Paul, 2010, and Zhu, 2014 which dealt with off-peak conditions only), and hence, is able to capture situations where there are passengers left behind due to capacity constraints (Zhu et al., 2017c). Based on the set of feasible itineraries (the trains that a passenger could have boarded) and the distribution of access/egress times, the model estimates the probability of a passenger having boarded each train and the probability distribution of the number of times a passenger was unable to board a train due to capacity constraints.

3.2 Methodology

The Passenger-to-Itinerary Assignment Model (PIAM) assumes a closed Automatic Fare Collection (AFC) system where transactions are recorded at both entry and exit stations. Thus both the tap-in and tap-out times of passengers and the arrival/departure times of trains at stations are known.

Figure 3-1 illustrates the possible itineraries for a passenger who enters the system at time $t^{in}$ and exits at time $t^{out}$. Access time is defined as the time it takes to walk from the tap-in fare gate to the platform; wait time is the time a passenger waits on the platform until he/she boards a train; and egress time is the time to walk to the tap-out fare gate after alighting from the train. Depending on access and egress times, in general, the passenger could have boarded a number of trains. Under the conservative assumption that the minimum access and egress times could be zero, the passenger in Figure 3-1 has three feasible trains (1,2,3).
3.2.1 Passenger Assignment Model

Based on the tap-in and tap-out times, a feasible itinerary (train) is defined by the following conditions:

1. The train departs the origin station after the passenger arrives at the platform:

\[
t_{i}^{in} + \tau_{i}^{o} \leq DT_{j} \tag{3.1}
\]

2. The train arrives at the destination station before the passenger taps-out:

\[
AT_{j} \leq t_{i}^{out} - \tau_{i}^{e} \tag{3.2}
\]

where,
$i$: passenger index.

t_{in}^i, t_{out}^i$: tap-in/tap-out times of passenger $i$.

$\tau_{a}^i, \tau_{e}^i$: minimum access/egress times for passenger $i$ (set to zero).

$DT_j$: departure time of train $j$ from the origin station.

$AT_j$: arrival time of train $j$ at the destination station.

Setting the minimum access and egress times to zero is conservative since it results in the largest possible set of feasible itineraries. Given the set of feasible itineraries, the probability of boarding each itinerary can be derived based on the distribution of passenger access/egress times at stations.

The following notation is used throughout this dissertation:

t_{a}^i, t_{e}^i$: access/egress times of passenger $i$.

$JT_i$: journey time of passenger $i$, where $JT_i = t_{out}^i - t_{in}^i$.

$M_i$: the number of feasible itineraries for passenger $i$.

$DT_{i,j}$: the “relative” departure time at the origin station for the $j^{th}$ train in the feasible itinerary set after setting the tap-in time of passenger $i$ to zero.

$j \leq M_i$.

$AT_{i,j}$: the “relative” arrival time at the destination station for the $j^{th}$ train in the feasible itinerary set after setting the tap-in time of passenger $i$ to zero.

$j \leq M_i$.

$f_a(t), f_e(t)$: access/egress time probability distributions.

$P_n$: the probability of a passenger being left behind $n$ times

Figure 3-2 illustrates the problem of boarding a train considering all possibilities for passenger $i$. After passenger $i$ taps-in, he/she walks to the platform. The passenger has $M_i$ feasible trains. When train capacity is not binding, all passengers board the first train after they arrive at the origin station platform. In this case, if the arrival time at the platform for passenger $i$ is between trains $j - 1$ and $j$, the passenger will board train $j$ with certainty. During peak hours, capacity constraints may be binding. In this case, the probability of passengers being unable to board a train should be
considered. The branches with a solid line represent the case where a passenger boards the first train after his/her arrival at the platform; a dashed line represents the case of being left behind. For example, a passenger may have to board the second (or third) train after their arrival on the platform due to capacity constraints.

![Passenger-to-Itinerary Assignment Model (PIAM)](image)

Figure 3-2: Passenger-to-Itinerary Assignment Model (PIAM)

Given the tap-in/out times, the boarding train for a specific passenger depends on their egress and access times and the train loads. Given the distribution of passenger’s access/egress times, the probability that passenger $i$ boarded train $j$, given that he/she tapped-out at $t_{i}^{out}$, can be derived using Bayes’ theorem (for $1 \leq j \leq M_i$):

$$P_i(\text{board train } j | t_{i}^{out}) = \frac{P_i(\text{board train } j, t_{i}^{out})}{P_i(t_{i}^{out})}$$

$$= \frac{P_i(\text{board train } j, t_{i}^{out})}{\sum_{j'=1}^{M_i} P_i(\text{board train } j', t_{i}^{out})} \quad (3.3)$$

The probability of boarding train $j$ and exiting at $t_{i}^{out}$, $P_i(\text{board train } j, t_{i}^{out})$, involves many possibilities, as shown in Figure 3-2. Since passenger $i$ may not be able to board the first train after his/her arrival at the platform, he/she may have arrived at the origin platform during any time interval before train $j$’s departure:

$$P_i(\text{board train } j, t_{i}^{out}) \quad \text{for } 1 \leq j \leq M_i$$

$$= \sum_{m=1}^{j} P_i(\text{board train } j, t_{i}^{out} | DT_{i,m-1} \leq t_{i}^{a} < DT_{i,m}) P(DT_{i,m-1} \leq t_{i}^{a} < DT_{i,m}) \quad (3.4)$$
where, $DT_{i,0} = 0$, and $P(DT_{i,m-1} \leq t^a_i < DT_{i,m})$ is the probability of arriving at the platform between trains $m - 1$ and $m$.

The probability of boarding train $j$ is assumed to be proportional to the likelihood of the corresponding egress time. For example, depending on the egress distance, an itinerary with 1 min egress time may be much more likely than one with 10 s. Therefore, the probability of boarding train $j$ and tapping-out at $t^{out}_i$, given that the passenger arrives at the platform between trains $m$ is:

\[ P_i(\text{board train } j, t^{out}_i | DT_{i,m-1} \leq t^a_i < DT_{i,m}) = \frac{P(t^e_i = JT_i - AT_{i,j})}{\sum_{j'=m}^{M_i} P(t^e_i = JT_i - AT_{i,j'})} \quad \text{for } m \leq j \leq M_i \quad (3.5) \]

Given a set of $M_i$ feasible itineraries, passenger $i$ may have one of $M_i$ possible egress times (see Figure 3-2). Hence, the conditional probability of the egress time $t^e_i = JT_i - AT_{i,j}$ is given by:

\[ P(t^e_i = JT_i - AT_{i,j}) = \frac{f_e(JT_i - AT_{i,j})}{\sum_{j'=1}^{M_i} f_e(JT_i - AT_{i,j'})} \quad \text{for } 1 \leq j \leq M_i \quad (3.6) \]

We use $P(t^e_i = JT_i - AT_{i,j})$ to denote the conditional probability of egress time given that the only possible values are $JT_i - AT_{i,j'}$ (for $j' = 1, 2, ... M_i$).

By substituting eq. (3.6) into (3.5):

\[ P_i(\text{board train } j, t^{out}_i | DT_{i,m-1} \leq t^a_i < DT_{i,m}) = \frac{P(t^e_i = JT_i - AT_{i,j})}{\sum_{j'=m}^{M_i} P(t^e_i = JT_i - AT_{i,j'})} \]
\[ = \frac{f_e(JT_i - AT_{i,j})}{\sum_{j'=m}^{M_i} f_e(JT_i - AT_{i,j'})} \quad \text{for } m \leq j \leq M_i \quad (3.7) \]

In general, the probability of passenger $i$ arriving at the platform between trains
$j - 1$ and $j$ can be calculated from the access time distribution. However, the probability is conditional on the fact that the last train in the feasible set left at $DT_{i,M_i}$, hence the access time has the last train’s “relative” departure time as an upper bound, $t_i^a < DT_{i,M_i}$.

$$P(DT_{i,j-1} \leq t_i^a < DT_{i,j}) = \frac{\int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt}{\int_0^{DT_{i,M_i}} f_a(t)dt} \quad \text{for } 1 \leq j \leq M_i \quad (3.8)$$

We use $P(DT_{i,j-1} \leq t_i^a < DT_{i,j})$ to denote the conditional probability that passenger $i$ arrived at the platform between trains $j - 1$ and $j$ given that his/her access times is shorter than $DT_{i,M_i}$ to simplify the notation. Combining eqs. (3.4), (3.7), and (3.8), the probability of boarding train $j$ and tapping-out at $t_i^\text{out}$ is:

$$P_i(\text{board train } j, t_i^\text{out}) = \sum_{m=1}^{M_i} P_i(\text{board train } j, t_i^\text{out} | DT_{i,m-1} \leq t_i^a < DT_{i,m})P(DT_{i,m-1} \leq t_i^a < DT_{i,m})$$

$$= \sum_{m=1}^{M_i} \sum_{j'=m}^{M_i} f_e(JT_i - AT_{i,j'}) \frac{\int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}{\int_0^{DT_{i,M_i}} f_a(t)dt} \quad \text{for } 1 \leq j \leq M_i \quad (3.9)$$

By substituting eq. (3.9) into (3.3), the probability that passenger $i$ boarded train $j$ is given by:

$$P_i(\text{board train } j | t_i^\text{out}) = \frac{P_i(\text{board train } j, t_i^\text{out})}{\sum_{j'=1}^{M_i} P_i(\text{board train } j', t_i^\text{out})}$$

$$= \frac{\sum_{m=1}^{M_i} \sum_{j'=m}^{M_i} f_e(JT_i - AT_{i,j'}) \frac{\int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}{\int_0^{DT_{i,M_i}} f_a(t)dt}}{\sum_{j'=1}^{M_i} \sum_{m'=1}^{M_i} \sum_{j''=m'}^{M_i} f_e(JT_i - AT_{i,j''}) \frac{\int_{DT_{i,m'-1}}^{DT_{i,m'}} f_a(t)dt}{\int_0^{DT_{i,M_i}} f_a(t)dt}} \quad (3.10)$$
3.2.2 Train Load Estimation

The train boarding probabilities derived in Section 3.2.1 can be used to estimate the expected load on individual trains. The expected load of train $m$ leaving station $k$ can be estimated recursively from the load leaving the previous station and the probabilities of alighting/boarding the corresponding train derived in the previous section:

$$E_m(k) = E_m(k - 1) - \sum_{\text{for all } i \text{ where } D_i = m} P_i(\text{board train } k|t_i^{out}) + \sum_{\text{for all } i \text{ where } O_i = m} P_i(\text{board train } k|t_i^{out}) \quad (3.11)$$

where,

- $k$: the station index.
- $E_m(k)$: the load of train $m$ departing station $k$ ($E_m(0) = 0$).
- $O_i$: the origin station for passenger $i$ ($1$ represents the terminal station).
- $D_i$: the destination station for passenger $i$ and $D_i > O_i$.

Since the focus of this chapter is on lines without transfers, transfer passengers are excluded.

3.2.3 Left Behind Distribution

The methodology presented in Section 3.2.1 provides the means to estimate the number of left behind passengers. The probability of a passenger being unable to board the first, second, etc. train after their arrival can be estimated based on the tree structure shown in Figure 3-2. For passenger $i$, the probability mass function of the
number of times, \( k (k = 1, 2, \ldots M_i - 1) \), a passenger is left behind is:

\[
P_i(\text{left behind } k \text{ times}) \quad \text{for } k = 1, 2, \ldots M_i - 1
\]

\[
= \sum_{j=1}^{M_i-k} P_i(DT_{i,j-1} \leq t_i^a < DT_{i,j}, \text{ board train } j + k|t_i^{out})
\]

\[
= \sum_{j=1}^{M_i-k} P_i(\text{board train } j + k, t_i^{out}|DT_{i,j-1} \leq t_i^a < DT_{i,j}) P_i(DT_{i,j-1} \leq t_i^a < DT_{i,j})
\]

\[
= \sum_{j=1}^{M_i-k} P_i(\text{board train } j + k, t_i^{out}|DT_{i,j-1} \leq t_i^a < DT_{i,j}) P_i(DT_{i,j-1} \leq t_i^a < DT_{i,j})
\]

\[
= \sum_{j=1}^{M_i-k} P_i(\text{board train } j + k, t_i^{out}|DT_{i,j-1} \leq t_i^a < DT_{i,j}) P_i(DT_{i,j-1} \leq t_i^a < DT_{i,j})
\]

\[
= \sum_{j=1}^{M_i-k} P_i(\text{board train } j, t_i^{out})
\]

By substituting eqs. (3.5), (3.8) and (3.9) into eq. (3.12) we get:

\[
P_i(\text{left behind } k \text{ times})
\]

\[
= \sum_{j=1}^{M_i-k} P_i(\text{board train } j + k, t_i^{out}|DT_{i,j-1} \leq t_i^a < DT_{i,j}) P_i(DT_{i,j-1} \leq t_i^a < DT_{i,j})
\]

\[
= \sum_{j=1}^{M_i-k} \frac{P(t_i^a = JT_i - AT_{i,j+k}) \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt}{\sum_{j'=1}^{M_i} P(t_i^a = JT_i - AT_{i,j'}) \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt}
\]

\[
= \frac{\sum_{j=1}^{M_i-k} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f_0(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}{\sum_{j'=1}^{M_i} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f_0(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}
\]

\[
= \frac{\sum_{j=1}^{M_i-k} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}{\sum_{j'=1}^{M_i} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}
\]

\[
= \frac{\sum_{j=1}^{M_i-k} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}{\sum_{j'=1}^{M_i} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}
\]

The expected number of times passenger \( i \) is left behind can be calculated:

\[
E_i(k) = \sum_{k=1}^{M_i-1} k P_i(\text{left behind } k \text{ times})
\]

\[
= \sum_{k=1}^{M_i-1} k \frac{\sum_{j=1}^{M_i-k} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}{\sum_{j'=1}^{M_i} \sum_{m=1}^{j''} \sum_{j'=1}^{M_i} f(JT_i - AT_{i,j'}) \int_{DT_{i,m-1}}^{DT_{i,m}} f_a(t)dt}
\]

The expected number of times passengers are left behind can be calculated for the entire population using the above results.
3.3 Model Validation and Application

Model validation is difficult because of the lack of ground truth data. The validation approach is two-fold: (i) synthetic data is used to validate all outputs of PIAM; and (ii) actual observations of passengers left behind at a station are compared with the left behind statistics obtained from the model.

3.3.1 Validation with Synthetic Data

The methodology is applied in a case study with data from a major, congested, subway system. Actual AFC transactions and train station arrival/departure times are used. Part of the line with 5 stations was used to eliminate transfers (see Figure 3-3).

![Figure 3-3: Case study network](image)

The access/egress time distribution is an important input to the model. Kim et al. (2006) developed a pedestrian mobility model and reported that the walk speed followed a log-normal distribution, with an average value of 1.26 m/s. Several of the studies of pedestrian movements also found that either a normal or log-normal probabilistic distribution can represent walk speeds accurately (Ottomanelli et al., 2012; Zhang et al., 2009; Daamen and Hoogendoorn, 2006).

Therefore, it was assumed that passenger walk speeds follow a log-normal distribution with mean 1.12 m/s and standard deviation 0.36 m/s (based on observations made by the agency). In Chapter 7, an access/egress time model is developed to estimate access/egress time distribution using AVL/AFC data.

Using the entry times from the AFC transactions, the train departure times, speeds drawn from the walk speed distribution, and passenger walk distances obtained from station plans and the gates passengers used to enter the system, the movement of the passengers through the system was simulated assuming that passengers board trains with available space on a first come, first served (FCFS) basis (based on their
arrival time at the platform). The output includes the train a passenger boarded, and his/her exit time. The time periods of interest are 14:00 to 15:00 for the off-peak period, and 18:00 to 19:00 for the peak period analysis.

Figure 3-4 shows the distribution of journey times based on the tap-in and tap-out times of the synthetic trips generated according to the above process, and compares them to the journey times actually experienced by the same passengers (i.e. based on the actual tap-in and tap-out times). As discussed above, the exit times from the AFC data are not used in the generation of synthetic data. The results show that the “synthetic” journey times are consistent with the actual times.

Figure 3-4: Journey time distribution

The distributions of the number of feasible itineraries are shown in Figure 3-5 for passengers departing Station 2, where passengers are more likely to be left behind. The numbers of feasible itineraries increase significantly in the peak (see Figure 3-5b).
Figure 3-5: Distribution of the number of feasible itineraries for passengers departing Station 2

During the off-peak, most passengers have only one feasible itinerary. During the peak at Station 2, the most congested station in this network, many passengers are left behind and, for that reason, there are larger numbers of feasible itineraries. The distribution of the number of feasible itineraries for the synthetic data is very similar to the corresponding distribution for the actual data.

Using the entry/exit times of the passengers and train arrival/departure times at stations, PIAM estimates the probabilities of each passenger boarding different trains. In the synthetic data, the train taken by each passenger is known, and is referred to as the “actual train”. The train with the highest probability of being boarded based on the output of the model, is defined as the “dominant train” and the corresponding probability as the “dominant probability”. The probability of boarding the actual train is also given by the model. Ideally, the dominant train is the actual train, and the corresponding probability is high.

The distribution of the estimated probabilities of boarding the actual train is an important metric since it relates to the robustness of the model. Figure 3-6 illustrates the estimated probabilities of boarding the actual train for passengers departing from Stations 1 and 2. For most off-peak passengers, the estimated probability of boarding the actual train is very high. During the peak period, the probability distribution of boarding the actual train shifts to the left due to the increased number of feasible
itineraries resulting from passengers being left behind. However, even in this case the performance of the model is very promising.

Figure 3-6: Distribution of the probability of boarding the actual train

The distributions of dominant probabilities are shown in Figure 3-7. During the off-peak (Figures 3-7a and 3-7b), the dominant probability for the majority of passengers is close to 1 (including passengers who have one feasible itinerary whose probability is exactly 1). During the peak (Figures 3-7c and 3-7d), the distributions shift to the left. This is due to the increased number of feasible itineraries (see Figure 3-5) resulting from both, the higher service frequency in the peak period and the higher probability of passengers being left behind. However, even in this case, almost all passengers have dominant probabilities greater than 0.50.
The fraction of passengers whose train with the highest probability (dominant train) is also the actual one differs slightly between the peak and off-peak periods. During the off-peak, the dominant train is the actual train for 99.27% of the passengers, while in the peak the figure is 92.44%.

The number of times passengers are unable to board a train due to crowding is an important service quality measure for many agencies. It is therefore, important to examine how accurately PIAM can estimate this measure. For each individual, the number is recorded based on his/her “actual experience” in the synthetic data. Using the model structure described in Figure 3-2, the probabilities a passenger is left behind once (arriving at the platform between trains $j - 1$ and $j$ and boarding train $j + 1$ is an event defined as being left behind once), or in general $n$ times
\( n \leq M_i - 1 \) are estimated. The expected number of times a passenger is left behind can be calculated from eq. (3.14).

The distributions of the estimated and actual number of times left behind are shown in Figure 3-8. The estimated distribution is very similar to the actual one. During the off-peak, no passengers are left behind (Figure 3-8a). During the peak period, due to capacity constraints, passengers are left behind up to 4 times. The distributions are also very similar to the actual during the peak period.

![Figure 3-8: Distribution of the number of times passengers are left behind at Station 2](image)

To evaluate the model’s ability to estimate loads on trains, Monte-Carlo simulation was used to assign passengers to trains, based on the estimated probabilities. Figures 3-9 (off-peak) and 3-10 (peak) show the 5th, 50th and 95th percentiles of the load estimation for each train (based on 2000 replications) compared to the “actual load”.

During the off-peak, the capacity constraint is not binding (Figure 3-9) and the estimated load is very close to the actual one. Furthermore, the estimated train load distributions have very small variance with the 5th and 95th percentiles close together. This is because, in most cases, the model identifies the actual train and the associated probability is very high.

The train load distributions for the peak period are shown in Figure 3-10. For Station 1, where the capacity constraint is not binding, PIAM estimates the passenger load on the trains very accurately, with the estimated 5th and 95th percentile distribu-
Figure 3-9: Passenger load estimation during the off-peak

tions close together. Most trains (starting with train 10) are full leaving Station 2 as shown in Figure 3-10b. PIAM captures these capacity constraints fairly accurately. The estimated load is close to the actual load with less than 10% difference in the median value across all trains in the period.

Figure 3-10: Passenger load estimation during the peak

Figure 3-11 shows the headways and the corresponding train loads. As expected, longer headways lead to higher train loads. Using the actual train movement data, instead of the timetable, allows for the evaluation of the impact of minor service disruptions. In the peak, other aspects of service quality caused by the capacity constraint, such as the increase in passengers’ wait time, number of times left behind, etc. are of interest. PIAM facilitates the development and use of a diverse set of service quality measures from the customer’s point of view.
3.3.2 Validation with Actual Data

In this section, the model is applied using the actual data (entry/exit times and train departure/arrival times) as input to PIAM. 27,571 trips between the same five stations during the period from 15:00 to 20:00 for a weekday were extracted. In this case, the actual train a passenger used is not known and cannot be used for validation. Instead, estimated probabilities of being left behind are compared with actual observations.

PIAM is able to estimate left behind probabilities at much higher resolutions than is possible with manually collected data. For example, Figure 3-12 shows the estimated probabilities of boarding the first train by time of day at Stations 1 and 2 in 5 min intervals. The peak period starts at 18:00, and this is when passengers start experiencing crowded trains and may be left behind. Most passengers can board the first train during the off-peak but only a small fraction of passengers during the peak
of the peak at Station 2 can do so. Station 2 is one of the most crowded stations in the system and many passengers are left behind several times there. Station 1, which is the terminal station for this line, is the least congested.

![Graph showing probability of boarding the first train at different times of day for Stations 1 and 2.]

Figure 3-12: Probability of boarding the first train

The agency conducted manual surveys at Station 2 to count the number of passengers left behind and the number of passengers boarding each train from 18:00 to 19:00 on the same day as the above analysis. The survey did not count the number of times a passenger is left behind (which is difficult due to the large passenger volume) and may not be able to capture passengers being left behind several times. By assuming a strict FCFS discipline, i.e. passengers left behind by the previous train board before new arrivals, the fraction of passengers boarding the first/second/third trains was roughly estimated.

Figure 3-13 compares the observed distribution of left behind passengers for the period 18:00-19:00 to the corresponding probabilities from the model. The results show that the model estimates are quite similar to the actual observations.
3.3.3 Performance Metrics

PIAM’s ability to allocate passengers to individual trains and estimate detailed left behind statistics at a high resolution allows the calculation of a number of unique performance metrics from both the system and passengers’ points of view. A number of such performance metrics are estimated for the actual transactions at the same stations as in the previous section. This discussion is used to illustrate the power of the model and the reasonableness of the results.

Figure 3-14 illustrates the probability of boarding the first train as a function of the station entry time. The solid line shows the average probability of boarding the first train in 30 s intervals. The dashed lines indicate train departures. As expected, passengers who arrive at the platform right after a train’s departure are more likely to board the next train, e.g. have a higher probability of boarding the first train. Passengers who arrive just before the next train’s departure are more likely to be left behind at the peak of the peak. Furthermore, Figure 3-14 also illustrates the impact of actual headways on the estimated probabilities.
The different components of a passenger’s journey time can also be estimated based on the PIAM output. One metric of interest is the time passengers spent at the origin station, including the access time and wait time on the platform. It is a function of the station configuration, in-station crowding, service frequency, and the number of times a passenger is left behind. Figure 3-15 shows the average in-station time for passengers departing from Stations 1 through 4.
During the off-peak, when passengers are not left behind, the average in-station time is determined by their access time and the headways. At around 16:30, when trains begin operating at a higher frequency—the start of the peak period—the in-station time initially decreases compared to the off-peak (since headways are shorter). However, during the peak of the peak, passengers are left behind, and their in-station time increases again mainly due to the extra wait time on the platform. For Station 2, the most congested station, the in-station time is even longer than in the off-peak. Even with higher service frequency, the capacity constraint results in longer wait times and increased passenger crowding on the platforms and the trains.

3.3.4 Visualization

Based on the PIAM output, the journey time components of passengers, and the crowding levels at stations and on platforms, can be inferred, and visualized, in detail. An animation has been developed to show the passenger/train movements, as well as the crowding during the time period analyzed. Figure 3-16 illustrates instances of the animation for the off-peak, and before and during the peak of the peak periods, respectively. The dots represent passengers boarding their most likely itinerary and the small rectangles the entry and exit gates. The platform is represented by the large rectangles. Crowding levels are shown in the bars which are color-coded based on the expected number of passengers. During the peak of the peak, Station 2 is very crowded as discussed previously and the train loads are much larger than during the off-peak.

This animation illustrates the ability of PIAM to capture individual passenger movements as well as to provide an estimate of crowding at different locations in the system.
Figure 3-16: Visualization of passenger movement through the system
3.4 Summary

The chapter presents a methodology to (retrospectively) assign passengers to individual trains using AFC and train tracking data. The method is probabilistic and, for the case of a single line (without transfers), able to capture the impact of capacity constraints. The method also serves as an important building block for the general case with transfers and route choice. The train a passenger boarded, the number of times a passenger is left behind, and the train loads can all be inferred.

The estimated train loads are influenced by the headway variability and serve as a good indicator of the capacity utilization and the level of service. During the peak periods, when the system operates at capacity, other level of service metrics, such as the number of times a passenger is left behind and station crowding, can be estimated.
Chapter 4

General PIAM: Framework

This chapter builds on the PIAM presented in Chapter 3 and proposes the overall framework for the general passenger assignment problem that is applicable to trips with and without transfers under capacity constraints. The general Passenger-to-Itinerary Assignment Model (PIAM) consists of two modules, the left behind model and the assignment model. The left behind model estimates the probability of being left behind by station and time interval and the assignment model assigns passengers to itineraries based on the left behind probabilities. The two modules will be discussed in more detail in Chapters 5 and 6.

The remainder of this chapter is organized as follows. Section 4.1 introduces the problem. Section 4.2 develops the two-step framework. Section 4.3 concludes the chapter.

4.1 Introduction

The general assignment problem for transfer trips is complicated and challenging. During the peak periods, with very short headways, the number of feasible itineraries can be very large. If a trip has multiple transfers and each segment has multiple trains a passenger might have boarded, the number of feasible itineraries will be prohibitively large, considering all the combinations of trains along different segments. A straightforward extension of the approach for non-transfer trips presented in Chapter
3 will not result in accurate assignment of passengers to itineraries.

Figure 4-1 shows the movements of a passenger who enters the system at $t^{in}$ and exits at $t^{out}$ with one transfer. The figure also shows the feasible itineraries that correspond to the passenger’s tap-in and tap-out times. In defining the set of feasible itineraries, the minimum access, transfer, and egress times are all set to zero. This assumption results (conservatively) in the largest possible number of feasible itineraries.

The passenger in Figure 4-1 has two journey segments. He/she can board trains 1, 2, or 3 on the first segment, and trains 1, 2, or 3 on the second segment. Itineraries represent different combinations of trains for the two segments. Some itineraries are not feasible, either because the train for the first segment departs before the tap-in time, the train for the last segment arrives after the tap-out time, or the requirement for a minimum transfer time is not fulfilled at the transfer station.

Figure 4-1: Time-Space diagram for a passenger with one transfer

Given the set of feasible itineraries, the passenger may arrive at the platform in any interval between consecutive departures of feasible trains (based on their access/transfer time). Depending on the available capacity on the train, the passenger
may board the first train upon his/her arrival at the platform or be left behind.

Even with very optimistic assumptions about egress or access time (e.g. 30 s minimum), the number of feasible itineraries does not significantly decrease. As a result, the dimensionality and complexity of the problem increase the difficulty of solving the general problem. A straightforward extension of the approach for non-transfer trips presented in Chapter 3 will not result in accurate assignment of passengers to itineraries.

In this chapter, we propose a two-step general Passenger-to-Itinerary Assignment Model (PIAM) that is applicable to trips with and without transfers under capacity constraints. In the first step, the aggregate probabilities of being left behind by station and time interval are estimated using data from trips without transfers. AFC and AVL data and station layouts can be used to infer passenger’s walk speed. In the second step (individual level), each passenger (with and without transfers) is assigned to feasible itineraries. The probabilities of left behind estimated in the first step, the tap-in and tap-out times of a passenger, and the train arrival/departure times at stations as well as access/egress/transfer time distributions are used to infer the probabilities of boarding different feasible itineraries (consisting of trains on all segments of the trip) for each passenger. With the left behind probabilities known, the degrees of freedom for the individual assignment problem are reduced.

This approach has several advantages:

1. By incorporating the information on left behinds, the dimensionality and complexity of the assignment problem, especially for transfer trips, is reduced.

2. The left behind model, can be applied outside the assignment model as a preprocessing step (or used independently for the calculation of useful performance metrics).

### 4.2 Methodology

Figure 4-2 illustrates the overall approach:
• **Access/Egress time model**

The access/egress time distributions are important inputs to PIAM. We propose a method to estimate them based on AFC/AVL data and station layouts. The access/egress time model consists of two components: the walk speed model and the walk distance model. The speed model incorporates the impact of passenger and station characteristics on the walk speed. The distance model captures passenger behavior in terms of the paths they follow inside stations. The egress times of passengers with one feasible itinerary are used. The estimation is properly formulated as a maximum likelihood problem that takes into account the fact that the sample is truncated (since the egress times of the passengers in this sample have to be less than the corresponding headway). The transfer time distribution can also be estimated using the same method.
• **Left behind model**

While the behind probabilities can be calculated from the output of the assignment model (as in Chapter 3), the process can be time consuming to apply. Recognizing this difficulty, we propose a left behind model to estimate the probability of passengers being left behind for a given time period at the aggregate level using data from trips without transfers. The left behind model groups passengers based on their expected arrival time at the station platform and estimates the probability of being left behind \(n\) times \((n = 0, 1, 2, \ldots)\) by station and time interval using data from trips without transfers. An important assumption of PIAM is that the probability of being left behind is the same for transfer and non-transfer passengers at the same station and time period. With the left behind probabilities known, the degrees of freedom for the general assignment problem are reduced, especially for transfer trips.

• **Assignment model**

At the disaggregate level, the assignment model assigns each passenger to the feasible itineraries based on the corresponding probability of being left behind, and the access/egress/transfer time distributions. Probabilities of boarding different feasible itineraries (consisting of trains in all segments of the trip) for each individual passenger are estimated. The output of PIAM includes the train loads, crowding at stations, etc.

The left behind model, assignment model and access/egress time model will be discussed in more detail in Chapters 5, 6, and 7.

While the approach proposed in this chapter is applicable to trips without route choice, the framework can be extended based on the same principles to incorporate trips with route choice. The PIAM framework is extended to include trips with route choice in Chapter 6. Given the left behind probabilities, a route choice model is proposed to estimate route choice fractions similar to the left behind model using maximum likelihood. The degrees of freedom are progressively limited by the left
behind model and route choice model for the assignment of trips with transfers and route choice.

4.3 Summary

In this chapter, we focus on the development of the overall framework of the general PIAM for trips with and without transfers. The model uses as input the access/egress time distribution estimated by the access/egress time model and consists of two modules, the left behind model and the assignment model. The left behind model estimates the probability of being left behind by station and time interval, and the assignment model assigns passengers to itineraries based on the left behind probabilities. The left behind model, assignment model and the access/egress time model will be discussed in more detail in the following chapters. The approach is applicable to trips without route choice, however, the framework is extended to incorporate trips with route choice in Chapter 6.
Chapter 5

General PIAM: Left Behind Model

This chapter develops the left behind model, the first module of the general PIAM. The left behind model estimates the probabilities of being left behind by station and time interval using data from non-transfer trips.

The remainder of this chapter is organized as follows. Section 5.1 introduces the problem. Section 5.2 develops the left behind model and describes the estimation approaches. Section 5.3 validates the model using synthetic data generated from actual AFC transactions and train movement data. Section 5.4 applies the model using actual data during the peak period, and compares the results with manual observations. It also evaluates the system performance under sudden increases in demand (stress test). Section 5.5 concludes the chapter.

5.1 Introduction

In Chapter 4, we presented the overall framework of the general PIAM for trips with and without transfers. This chapter develops the first module, the left behind model, in more detail. The left behind model is important in its own right, as it provides useful performance metrics for peak hour operations. While left behind probabilities can be calculated from the output of PIAM as in Chapter 3, they are too complicated and time consuming to apply in the general case. Agencies sometimes hire surveyors to count the number of passengers left behind after a train departs, however this
process is expensive, inefficient and inaccurate. In this chapter, we propose an efficient method to directly infer left behind probabilities from AFC and AVL data, for systems in which entry and exit station transaction data are recorded.

This chapter examines the problem of estimating the probability distribution of the number of times passengers are left behind at a station in a given time period using data from non-transfer trips. Two approaches are used to estimate the left behind probability mass function (LBPMF) assuming that the number of times a passenger is left behind is a random variable with an unknown probability mass function: the first method is based on a maximum likelihood formulation of the problem and the second uses Bayesian inference to estimate the posterior distribution of the LBPMF.

### 5.2 Methodology

The problem can be viewed as identifying the underlying groups of passengers based on how many times they were unable to board a train. In the transportation literature, a number of problems have been studied using similar approaches. Barnhart et al. (2014) and Yan et al. (2016) for example, developed discrete choice models to identify the boarding flights of passengers. Kazali and Koutsopoulos (2013), used a mixture model to identify latent driver populations with different travel behavior using observations of travel times from an Automated Number Plate Recognition (ANPR) system.

In an application to Metro systems, Zhou et al. (2015) proposed a method to estimate path-selection proportions based on recorded entry and exit times from AFC data. Fu et al. (2014) estimated the posterior probabilities of choosing different routes by formulating the travel times as mixture distributions using smart card data (each corresponding to a different route). Lee and Sohn (2015) introduced the number of routes used as an unknown variable in a Bayesian framework and used reversible-jump Markov Chain Monte Carlo simulation for the estimation of the posterior distribution of different routes based on travel time observations.

A potential drawback of using only travel times to infer the latent groups of
passengers or their route choice is the inability to capture crowding effects, since during peak hours, longer journey times may be caused by severe congestion instead of passengers choosing an alternative route with a longer journey time.

5.2.1 Problem Definition

In this chapter, we use the same definition of feasible itinerary as in Chapter 3 for non-transfer trips assuming that the minimum access and egress times could be zero. Given the set of feasible itineraries, the passenger may arrive at the platform in any of the train departure intervals (based on his/her access time). Depending on the crowding on the train, the passenger may board the first train upon his/her arrival at the platform or be left behind if there is no available capacity.

The relationship of the variables involved in a passenger’s movement is depicted by the graphical model in Figure 5-1. A graphical model uses a directed acyclic graph to express the conditional dependence between random variables with arrows pointing to the dependent variables (Bishop, 2013). The shaded nodes represent the model parameters that are known (such as $\mu^a$, $\sigma^a$, $\mu^e$ and $\sigma^e$, the mean and standard deviation of the access/egress time distributions respectively). The access time ($t_a$) and egress time ($t_e$) are assumed random variables with a known distribution (e.g. the log-normal distribution in this study). The arrival time at the origin station platform ($A$) depends on the access time. $D$ is a random variable representing the number of times, $n$, a passenger is left behind. $n$ has a probability mass function with parameters $P_n$, for $n = 0, 1, 2, \ldots$. The arrival time at the platform and the number of times a passenger is left behind specify the boarding train of each passenger ($B$). The boarding train and the egress time define the journey time ($JT$) of the passenger, which is known from the AFC records. The objective of this chapter is to estimate the left behind probability mass function (LBPMF).
5.2.2 Problem Formulation

Figure 5-2 illustrates all possible instances for a passenger with $M_i$ feasible itineraries without transfers. The branches with a dashed line represent the left behind instances. For example, even if the passenger arrives before the first train, he/she can be left behind and have to board the second (or third) train due to capacity constraints.

Figure 5-2: Passenger-to-Itinerary Assignment Model (PIAM) without transfers

In this chapter, we assume that the access/egress time distributions are known. They can be estimated from manual surveys conducted at stations or using AFC and
AVL data as proposed in Chapter 7.

Assuming that during a very short time period, the probability distribution of being left behind at one station is constant, the parameters of the LBPMF for those passengers can be estimated by the likelihood function of the observations.

As shown in Figure 5-2, the probability of passenger $i$ arriving at the origin station platform between the departures of trains $j - 1$ and $j$ is:

$$P(DT_{i,j-1} \leq t^a_i < DT_{i,j}) = \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t) dt \text{ for } 1 \leq j \leq M_i,$$

(5.1)

with $DT_{i,0} = 0$. Given that the passenger arrived between trains $j - 1$ and $j$, the probability of boarding train $k$ is the probability of that passenger being left behind $k - j$ ($j \leq k \leq M_i$) times.

$$P(\text{board train } k|DT_{i,j-1} \leq t^a_i < DT_{i,j}) = P_{k-j}.$$

(5.2)

Given that the passenger boarded train $k$, the distribution of journey times/tap-out times can be derived based on the distribution of egress times.

$$P(t^e_i|\text{board train } k) = f_e(JT_i - AT_{i,k}).$$

(5.3)

The probability of a passenger arriving at the origin station platform between trains $j - 1$ and $j$, boarding train $k$, and tapping-out at $t^\text{out}_i$ can be derived using eqs. (5.1), (5.2) and (5.3).

$$P(t^\text{out}_i, \text{board train } k, DT_{i,j-1} \leq t^a_i < DT_{i,j})$$

$$= P(DT_{i,j-1} \leq t^a_i < DT_{i,j})P(\text{board train } k|DT_{i,j-1} \leq t^a_i < DT_{i,j})P(t^e_i|\text{board train } k)$$

$$= \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t) dt f_{k-j}(JT_i - AT_{i,k}).$$

(5.4)

The probability of passenger $i$ tapping-out at the observed time is the sum over all
possibilities in Figure 5-2.

\[ L_i(Z) = \sum_{j=1}^{M_i} \sum_{k=j}^{M_i} P(t^\text{out}_i, \text{board train } k, DT_{i,j-1} \leq t^a_i < DT_{i,j}) \]

\[ = \sum_{j=1}^{M_i} \sum_{k=j}^{M_i} \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt P_{k-j}f_e(JT_i - AT_{i,k}) \]

\[ = \sum_{j=1}^{M_i} \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt \sum_{k=j}^{M_i} P_{k-j}f_e(JT_i - AT_{i,k}), \quad (5.5) \]

where \( Z = [P_0, P_1, ..., P_{M-1}]^T \), is a vector of the parameters of the LBPMF. In this chapter, we assume that the maximum number of times a passenger can be left behind equals the maximum number of feasible itineraries in the group minus one, i.e. the length of \( Z \) is equal to \( M \).

For the whole population (group), assuming conditional independence between passengers, the probability of observing the journey times of all passengers in the group is given by:

\[ L(Z) = \prod_{i=1}^{N} L_i(Z) \]

\[ = \prod_{i=1}^{N} \prod_{j=1}^{M_i} \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt \sum_{k=j}^{M_i} P_{k-j}f_e(JT_i - AT_{i,k}). \quad (5.6) \]

**Maximum Likelihood Formulation**

The maximum likelihood estimation for the parameters of LBPMF uses eq. (5.6) to formulate the log-likelihood function:

\[ \mathcal{L}(Z) = \log [L(Z)] \]

\[ = \log \left[ \prod_{i=1}^{N} \prod_{j=1}^{M_i} \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt \sum_{k=j}^{M_i} P_{k-j}f_e(JT_i - AT_{i,k}) \right] \]

\[ = \sum_{i=1}^{N} \log \sum_{j=1}^{M_i} \int_{DT_{i,j-1}}^{DT_{i,j}} f_a(t)dt \sum_{k=j}^{M_i} P_{k-j}f_e(JT_i - AT_{i,k}). \quad (5.7) \]
Using MLE, the probability of a passenger being left behind $n$ times can be estimated as the solution of the following optimization problem:

\[
\max_{P_0, P_1, \ldots, P_{M-1}} \quad \sum_{i=1}^{N} \log \sum_{j=1}^{M_i} \int_{D T_{i,j}} f_a(t) dt \sum_{k=j}^{M_i} P_{k-j} f_e(JT_i - AT_{i,k}) \\
\text{s.t.} \quad \sum_{i=0}^{M-1} P_i = 1 \quad \text{(5.9)} \\
\quad \quad \quad \quad P_i \geq 0, \ \forall i = 0, 1, \ldots, M - 1 \quad \text{(5.10)}
\]

In the computational experiments in Sections 5.3 and 5.4, this MLE optimization problem is solved using the optimization package in SciPy (Jones et al., 2001).

**Bayesian Estimation**

Bayesian estimation treats the parameters of the distribution, $[P_0, P_1, P_2, \ldots, P_{M-1}]$, as a random vector while MLE provides only a point estimate of the parameters. Under the Bayesian framework, a prior distribution of $Z$ is assumed and after training, the posterior distribution is learned. Bayesian methods are flexible and can accommodate complex model structures with diverse parameters (Richardson and Green, 1997).

In Bayesian estimation, a prior distribution is usually chosen to have a closed form for the posterior distribution in the same family as the prior distribution. Such a prior distribution is called the conjugate prior. However, in this research, the journey times do not follow a standard distribution and there is no conjugate prior distribution. For that reason, we use the Dirichlet distribution as the prior. The Dirichlet is a multivariate distribution with the sum of its elements, $[P_0, P_1, P_2, \ldots, P_{M-1}]$ in this case, equal to 1. The degrees of freedom are thus $M - 1$. The Dirichlet distribution has a very flexible form in which the parameters to be estimated (the probabilities of being left behind $n$ times) can have various shapes for the density function. Assuming that $Z = [P_0, P_1, P_2, \ldots, P_{M-1}]^T$ follows a Dirichlet distribution with parameter vector
\[ \boldsymbol{Z} = [P_0, P_1, P_2, \ldots, P_{M-1}]^T \sim \text{Dir}(\boldsymbol{\alpha}), \]  
\[ P(\boldsymbol{Z}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{n=0}^{M-1} P_n^{\alpha_n-1}, \]  
\[ B(\boldsymbol{\alpha}) = \frac{\prod_{n=0}^{M-1} \Gamma(\alpha_n)}{\Gamma(\sum_{n=0}^{M-1} \alpha_n)}, \]  

where,

- \( \boldsymbol{\alpha} \): the parameter vector of the Dirichlet distribution, \( \boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{M-1}] \).
- \( P(\boldsymbol{Z}) \): the left behind probability density function.
- \( B(\boldsymbol{\alpha}) \): a normalization term for the probability density function of \( \boldsymbol{Z} \).
- \( \Gamma(\cdot) \): Gamma function, where \( \Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \).

By setting \( \alpha_n = 1, \forall n \in (0, 1, 2, \ldots, M - 1) \), the prior probability distribution of being left behind \( n \) times becomes identical for all \( n \) (a uniform distribution). This is used as the non-informative prior distribution for all the parameters:

\[ P(\boldsymbol{Z}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{n=0}^{M-1} P_n^{\alpha_n-1} = \frac{1}{B(\text{1})} \prod_{n=0}^{M-1} P_n^{1-1} = \text{constant}. \]  

The posterior distribution of \( \boldsymbol{Z} \) given observations \( \boldsymbol{X} \), which represent the journey times, can be derived using Bayes’ theorem:

\[ P(\boldsymbol{Z}|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Z})}{P(\boldsymbol{X})} \propto P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Z}). \]  

Given that \( P(\boldsymbol{Z}) = \text{constant} \), as in eq. (5.14), we have:

\[ P(\boldsymbol{Z}|\boldsymbol{X}) \propto P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Z}) \propto P(\boldsymbol{X}|\boldsymbol{Z}). \]  

where \( P(\boldsymbol{X}|\boldsymbol{Z}) \) is the likelihood of the journey time observations conditional on the LBPMF parameters. \( P(\boldsymbol{X}|\boldsymbol{Z}) \) is equivalent to \( L(\boldsymbol{Z}) \) in eq. (5.6). Hence, the posterior
distribution is given by:

\[
P(Z|X) \propto P(X|Z) = \prod_{i=1}^{N} \sum_{j=1}^{M_i} \int_{DT_{i,j}}^{DT_{i,j-1}} f_a(t) dt \sum_{k=j}^{M_i} P_{k-j} f_e \left( JT_i - AT_{i,k} \right). \quad (5.17)
\]

We use a constant \(Z_{\text{norm}}\) as the normalization term of the distribution to make the integral of eq. (5.17) 1:

\[
P(Z|X) = \frac{\tilde{P}(Z)}{Z_{\text{norm}}}, \quad (5.18)
\]

where,

\[
\tilde{P}(Z) = \prod_{i=1}^{N} \sum_{j=1}^{M_i} \int_{DT_{i,j}}^{DT_{i,j-1}} f_a(t) dt \sum_{k=j}^{M_i} P_{k-j} f_e \left( JT_i - AT_{i,k} \right). \quad (5.19)
\]

\(\tilde{P}(Z)\) is the kernel of the distribution \(P(Z|X)\) (i.e. obtained by omitting the terms that are not functions of any variables in the probability density function, such as the normalization term \(Z_{\text{norm}}\)).

The challenge of using eqs. (5.18) and (5.19) to estimate the probability distribution of the parameters of LBPMF is that \(Z_{\text{norm}}\) is unknown. Monte Carlo methods have been widely used in statistical inference to deal with this problem (Chen et al., 2012; Gamerman and Lopes, 2006; Gilks, 2005; Smith and Roberts, 1993; Geyer, 1992, 1991; Hastings, 1970). The Metropolis-Hastings (MH) algorithm is one of the best known methods in the MCMC class of methods. Similar to rejection sampling, the MH algorithm:

1. Generates a candidate sample from a proposal distribution \(Q\), as a random walk process. The proposal distribution is the density function used to generate the next sample given the previous one;

2. Accepts or rejects the new sample based on specific selection criteria (based on the proposal distribution and the target distribution).

The MH algorithm is more efficient than the basic rejection sampling methods
when the dimensionality of the problem is high. The rejection rate can be very low by appropriately selecting the proposal distribution, thus reducing the number of iterations required for convergence (Bishop, 2013).

With the MH algorithm, the evaluation of the acceptance of a sample does not require the normalization term of the target distribution. Knowing the kernel of the target distribution is sufficient. Thus the method circumvents the complication of calculating the normalization term.

For the generation of subsequent samples using the MH algorithm, at step $\tau$, a sample $Z^\star (Z^\star = [P_0^\star, P_1^\star, P_2^\star, ..., P_{M-1}^\star])$ is drawn from the proposal distribution $Q$. The choice of the proposal distribution affects the performance of the sampling. Large variance in the proposal distribution leads to a large step size, but the probability of accepting the new sample is low. Small variance leads to a smaller step size but it may take longer to explore the full sample space. A normal distribution centered at the current sample $Z^{(\tau)}$ is commonly used. In this research, we use the Dirichlet distribution with the mean of the distribution determined by the previous sample $Z^{(\tau)}$. The standard deviation can be determined by the values of $Z^{(\tau)}$ so that the distribution of $Z^\star | Z^{(\tau)}$ is centered at $Z^{(\tau)}$, where it also has the highest probability density.

$$Z^\star | Z^{(\tau)} \sim \text{Dir} (\alpha'(Z^{(\tau)})),$$  \hspace{1cm} (5.20) \\
$$Q (Z^\star | Z^{(\tau)}) = \frac{1}{B(\alpha')} \prod_{n=0}^{M-1} (P_n^\star)^{\alpha'_n - 1},$$  \hspace{1cm} (5.21) \\
$$\alpha'(Z^{(\tau)}) = \frac{Z^{(\tau)}}{\min (Z^{(\tau)})} \geq 1,$$  \hspace{1cm} (5.22)

where $B(\alpha')$ is a normalization term for the probability density function. Eq. (5.22) normalizes the value of $\alpha'$ to be larger than 1 so that the probability density function of the proposal distribution becomes convex.

An advantage of using the Dirichlet distribution in eq. (5.20) as the proposal distribution is that the likelihood of extreme values, i.e. some elements are equal to zero, is relatively high. Figures 5-3 illustrates the density functions where the
The sampling algorithm is described below:

1. At iteration $\tau$, sample $Z^*$ from $Q(Z^*|Z^{(\tau)})$.

2. Calculate the acceptance ratio:

$$A(Z^*|Z^{(\tau)}) = \min \left\{ 1, \frac{\hat{P}(Z^*)}{P(Z^{(\tau)})} \frac{Q(Z^{(\tau)}|Z^*)}{Q(Z^{(\tau)}|Z^{(\tau)})} \right\}.$$ (5.23)

3. Sample $U_0$ from the uniform distribution $U[0,1]$.

4. Accept or reject the sample $Z^*$:

$$Z^{(\tau+1)} = \begin{cases} Z^* & \text{if } U_0 \leq A(Z^*,Z^{(\tau)}), \quad \text{(accept)} \\ Z^{(\tau)} & \text{otherwise,} \quad \text{(reject).} \end{cases}$$ (5.24)
5. $\tau = \tau + 1$, go back to step 1.

Samples following the posterior distribution are generated after an initial warm up period (e.g. the first 1000 iterations). By monitoring the sample distribution, we can select the appropriate number of remaining draws. The posterior distribution of the parameters (the probabilities of being left behind $n$ times) can be approximated by the sample distribution.

### 5.3 Model Validation

The method is applied using AFC and AVL data from the same congested subway system used in Chapter 3 between 18:00-18:20 on 2012/08/30 on a section of one of the system lines (depicted in Figure 5-4 below). One heavily used OD pair from Station 2 to Station 3 was analyzed. The details of demand and train movements data we use here will be described in the beginning of the Section 5.4 when we present analysis involving another date. Because there is no true information about the actual left behind distribution, we use tap-in times from actual AFC transactions and train arrival/departure times to generate synthetic data to verify the model using the same process as in Chapter 3. Passenger walk speeds are assumed to follow a log-normal distribution with mean 1.12 m/s and standard deviation 0.36 m/s.

![Figure 5-4: Case study network](image)

Two different train capacities were used to test the performance of the estimation methods at two different levels of crowding representing off-peak and peak conditions. The distribution of journey times in the synthetic data is shown in Figure 5-5. As expected, the journey times for passengers during the peak are longer than in the off-peak because of passengers being left behind.
The distribution of the number of feasible itineraries is shown in Figure 5-6. During the off-peak, most passengers have two feasible itineraries. During the peak, with longer journey times, many passengers have more feasible itineraries. The maximum number of feasible itineraries is 6 in both cases, therefore the number of parameters to be estimated is 6 ($P_0, P_1, P_2, P_3, P_4, P_5$).

Tables 5.1a and 5.1b summarize the estimation results from the two methods for the off-peak and peak periods respectively, with the peak period being more capacity constrained. Both estimations are very similar to the actual values from the synthetic data in both periods.
Table 5.1: Estimation result

(a) Off-peak

<table>
<thead>
<tr>
<th></th>
<th>$P_0$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $P$</td>
<td>0.94</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{ML}$</td>
<td>0.95</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Posterior Mean</td>
<td>0.99</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

(b) Peak

<table>
<thead>
<tr>
<th></th>
<th>$P_0$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $P$</td>
<td>0.36</td>
<td>0.32</td>
<td>0.23</td>
<td>0.08</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>$P_{ML}$</td>
<td>0.38</td>
<td>0.34</td>
<td>0.24</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Posterior Mean</td>
<td>0.40</td>
<td>0.34</td>
<td>0.23</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The values of the parameters of the MCMC algorithm at each iteration (including the warm up period) are shown in Figure 5-7, for the off-peak (5-7a) and peak (5-7b) periods. The different colors represent the first five dimensions of the vector $Z$ (i.e. $P_0, P_1, ..., P_4$). The warm up period for the MH algorithm in this case ends before the 1000th iteration and samples afterwards are used. During the off-peak, the value of $P_0$, the probability of boarding the first train, is high, while during the peak period, it is around 0.35 to 0.45, indicating that most passengers are left behind.

(a) Off-peak  
(b) Peak

Figure 5-7: Sampling process of the MH algorithm
The marginal posterior distributions of the left behind probabilities can be seen in Figure 5-8. During the peak, the variance of the probabilities is much larger than for the off-peak as there is more uncertainty, and the confidence intervals are larger. During the off-peak, almost all the passengers board the first train.

<table>
<thead>
<tr>
<th></th>
<th>Off-peak</th>
<th>Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td><img src="image1.png" alt="Off-peak P0" /></td>
<td><img src="image2.png" alt="Peak P0" /></td>
</tr>
<tr>
<td>P1</td>
<td><img src="image3.png" alt="Off-peak P1" /></td>
<td><img src="image4.png" alt="Peak P1" /></td>
</tr>
<tr>
<td>P2</td>
<td><img src="image5.png" alt="Off-peak P2" /></td>
<td><img src="image6.png" alt="Peak P2" /></td>
</tr>
<tr>
<td>P3</td>
<td><img src="image7.png" alt="Off-peak P3" /></td>
<td><img src="image8.png" alt="Peak P3" /></td>
</tr>
</tbody>
</table>

Figure 5-8: Marginal posterior distributions of left behind probabilities
5.4 Application Using Actual Data

The model is applied using actual data collected between 18:00 to 19:00 on 2012/08/30 and 2014/10/03 on the same section of the system as in Section 5.3 (Figure 5-4). Two heavily used OD pairs were analyzed, Station 1 to Station 3 and Station 2 to Station 3.

5.4.1 System Description

The system operates at near capacity and 2012/08/30 represents a typical day. Due to a special event, Station 1 attracted a large volume of passengers on 2014/10/03 that increased the demand significantly and stressed the system. During the same hour, the total number of trips originating at Station 1 increased from 24017 on 2012/08/30 to 27029 on 2014/10/03, a 13% increase. As the system was already operating near capacity, the 13% increase in demand resulted in severe crowding.

The passenger flows between the first five stations in the eastbound direction (leaving Station 1) are shown in Figure 5-9. Figure 5-9a shows the passenger flow from Station 1 to Stations 2 through 5, while Figure 5-9b shows the passenger flow from Station 2 to Stations 3 through 5. The peak period starts at 17:00 and the peak of the peak at both stations occurs between 18:10 and 18:30.

![OD flow within first five stations](image)
The service frequency during the off-peak on 2014/10/03 was higher than on 2012/08/30 while during the peak it was almost the same. There was a service disruption at around 18:30 at Station 2 (see Figure 5-10) on 2014/10/03.

![Figure 5-10: Headways at Station 2](image)

The journey time distribution for the two OD pairs (1-3 and 2-3) is shown in Figure 5-11 for both days. For the first OD pair (1-3) on 2012/08/30, the journey times are similar across the whole time period, while an increase can be observed during the peak on 2014/10/03. For the second OD pair (2-3), journey times increase during the peak on both days, and are much longer on 2014/10/03.

![Figure 5-11: Journey time distribution](image)
5.4.2 Estimated Probabilities

The left behind probabilities for passengers boarding at Station 2 are estimated for three periods (18:00 to 18:20, 18:20 to 18:40 and 18:40 to 19:00 on 2012/08/30). The distributions of the probabilities of passengers being left behind are shown in Figure 5-12. The red dashed lines indicate the MLE results. At Station 2, the left behind probabilities during the 18:20-18:40 period are higher than in the 18:00-18:20 and 18:40-19:00 periods with only 15% of the passengers able to board the first train. The estimation results from the MCMC and MLE methods are quite consistent.

<table>
<thead>
<tr>
<th></th>
<th>18:00-18:20</th>
<th>18:20-18:40</th>
<th>18:40-19:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
<td><img src="image3" alt="Diagram" /></td>
</tr>
<tr>
<td>P1</td>
<td><img src="image4" alt="Diagram" /></td>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
</tr>
<tr>
<td>P2</td>
<td><img src="image7" alt="Diagram" /></td>
<td><img src="image8" alt="Diagram" /></td>
<td><img src="image9" alt="Diagram" /></td>
</tr>
<tr>
<td>P3</td>
<td><img src="image10" alt="Diagram" /></td>
<td><img src="image11" alt="Diagram" /></td>
<td><img src="image12" alt="Diagram" /></td>
</tr>
</tbody>
</table>

Figure 5-12: Marginal posterior distributions of left behind probabilities
5.4.3 Comparison with Manual Observations

The agency conducted manual surveys on the numbers of passengers left behind from 18:00 to 19:00 at Station 2 on 2012/08/30. As shown in Figure 5-13, the survey results are similar to the estimation results. The estimated probability of boarding the first train is 44% for both the MLE and 44% Bayesian estimations compared to 47% from the survey. The survey assumes a strict FCFS discipline to estimate the number of times a passenger is left behind, which may underestimate the actual number.

![Figure 5-13: Left behind probabilities at Station 2](image)

As discussed earlier, on 2014/10/03, there was a special event that increased the ridership from Station 1 significantly. The results from applying the MLE method are used to examine how the system performed when the demand at Station 1 increased by 13%. Figure 5-14 shows the estimated left behind probabilities on 2014/10/03 compared to 2012/08/30.

![Figure 5-14: Left behind probabilities at Station 1](image)
On 2012/08/30, Station 1, the terminal station, does not experience heavy crowding, and most people can board the first train. However, on 2014/10/03, the situation is much worse with less than 60% of passengers able to board the first train. In the second period, 51% of passengers were left behind once and 10% twice. The 13% increase in demand at Station 1 results in a significant increase in the numbers of left behind passengers and overcrowding on the platforms. This also affects Station 2 as shown in Figure 5-15.

![Figure 5-15: Left behind probabilities at Station 2](image)

Figure 5-16 shows the expected number of times a passenger was left behind at Station 2 by time of day. Starting at 18:00, on both days, the expected number of
problem was much more severe and involved two periods (one at around 18:25 and the other at 18:55). The high resolution of estimating the left behind statistics makes it a good indicator of station/on-train crowding.

The results show similar accuracy for both approaches. The Bayesian approach can provide distributional information, so that a confidence interval or the standard deviation of the parameter estimation could be obtained. However, it is also computationally more intensive.

5.5 Summary

The chapter developed the first module of the general PIAM and presented a methodology to estimate the probability of passengers being left behind for a given time period. Maximum likelihood estimation (MLE) and Bayesian inference implemented with Markov Chain Monte Carlo sampling were used. The model examined the performance of a congested system during peak periods and under different conditions. The estimated probabilities of passengers being left behind using both methods are similar to manual survey results and provide crowding information at a much more detailed level. A stress test is used to illustrate the richness of the model and evaluate the system performance under increased demand.
Chapter 6

General PIAM: Assignment Model

This chapter develops the second module of the general Passenger-to-Itinerary Assignment Model (PIAM), the assignment model, that is applicable to trips with and without transfers based on the left behind estimation presented in Chapter 5. The tap-in and tap-out times of a passenger, the train arrival/departure times at stations and the left behind probabilities by station and time interval are used to infer the probabilities of taking different itineraries. The framework is extended to incorporate trips with route choice and estimate route choice fractions.

The remainder of the chapter is organized as follows. Section 6.1 describes the approach. Section 6.2 validates the model using synthetic data generated from actual AFC transactions and train movement data. Section 6.3 applies the model using actual data in a congested network. Section 6.4 extends PIAM to include trips with route choice. The model is validated using synthetic data and applied with actual data. Section 6.5 concludes the chapter.

6.1 Methodology

The general PIAM has as input the tap-in/out times from the AFC data and the train arrival/departure times at stations from the AVL data. The left behind model developed in the previous chapter estimates the probability a passenger is left behind $n$ times ($n = 0, 1, 2, ...$) by station and time interval. The assignment model then
assigns each passenger to itineraries based on the left behind probabilities, and the access/egress/transfer time distributions.

### 6.1.1 Problem Definition

Figure 6-1 shows the movements of a passenger who enters the system at $t^{in}$ and exits at $t^{out}$ with one transfer. The figure also shows the feasible itineraries that correspond to the passenger’s tap-in and tap-out times. The passenger has two journey segments. He/she can board trains 1, 1, 2 or 1, 3 on the first segment, and trains 2, 1, 2 or 2, 3 on the second segment. Itineraries represent different combinations of trains for the two segments.

Considering a passenger $i$ with $S_i$ journey segments, we define an itinerary $I$ as $I = \{I[1], I[2], ..., I[S_i]\}$, where $I[1]$ represents the $I[1]^{th}$ train that departs the origin station after the passenger’s tap-in time. For $s \geq 2$, $I[s]$ represents the $I[s]^{th}$ train after the arrival of train $I[s - 1]$ at the transfer station.

Figure 6-1: Time-Space diagram for a passenger with one transfer

In general, for trips with or without transfers, we consider $I$ to be feasible if it satisfies the following conditions:
1. For the first segment, the train departs the origin station after the passenger’s arrival at the platform:

\[ t_i^{in} + \tau_i^a \leq DT_{i,I[1]} \]  

(6.1)

2. For the last segment, the train arrives at the destination station before the passenger’s tap-out time:

\[ AT_{i,I[S_i]}^{s} \leq t_i^{out} - \tau_i^e \]  

(6.2)

3. If itinerary \( j \) consists of more than one segment (i.e. with transfers), for any pair of consecutive segments, segment \( s \) and \( s + 1 \) (\( 1 \leq s \leq S_i - 1 \)), the train for segment \( s + 1 \) departs after the passenger alights from the previous train and arrives at the transfer platform:

\[ DT_{i,I[s]}^{s+1} \geq AT_{i,I[s]}^{s} + \tau_i^{s,t} \]  

(6.3)

where,

- \( S_i \): number of segments for passenger \( i \).
- \( DT_{i,I[s]}^{s} \): in itinerary \( I \), the departure time of train \( I[s] \) on the \( s^{th} \) segment.
- \( AT_{i,I[s]}^{s} \): in itinerary \( I \), the arrival time of train \( I[s] \) on the \( s^{th} \) segment.
- \( \tau_i^{s,t} \): minimum transfer time for passenger \( i \) (set conservatively to zero) at the departure station of segment \( s \).

Given the set of feasible itineraries, the passenger may arrive at the platform in any interval between consecutive departures of feasible trains (based on their access/transfer times). Depending on the available capacity of the train, the passenger may board the first train upon his/her arrival at the platform or be left behind.
6.1.2 Problem Formulation

Figure 6-2 illustrates all possible instances for a passenger with one transfer. The branches with a dashed line represent the left behind instances. For example, at the origin station, even if the passenger arrives before train 1_1, he/she can be left behind and have to board the second (or third) train. Similarly, at the transfer station, the passengers can board the first train upon arrival at the transfer station platform or be left behind. The probability of assigning passenger $i$ to each itinerary (e.g. trains 1_2 and 2_3) is the sum over all possibilities in Figure 6-2.

![PIAM for a passenger with one transfer](image)

Figure 6-2: PIAM for a passenger with one transfer

The following notation is used throughout this dissertation:

$I_i$: the feasible itinerary set for passenger $i$.

$DT_{i,j}^1$: the “relative” departure time of the $j^{th}$ train at the origin station in the feasible itinerary set (after setting the tap-in time of passenger $i$ to zero).

$DT_{i,j}^{s,1}$: the “relative” departure time of the $j^{th}$ train in segment $s$ ($2 \leq s \leq S_i$)
after the arrival time of train $I[s - 1]$ (setting the arrival time of train $I[s - 1]$ to zero).

$AT_{i,j}^1$: the “relative” arrival time of the $j^{th}$ train in the feasible itinerary set (after setting the tap-in time of passenger $i$ to zero) in the first segment.

$AT_{s,i,j}^1$: the “relative” arrival time of the $j^{th}$ train in segment $s$ ($2 \leq s \leq S_i$) after the arrival time of train $I[s - 1]$ (setting the tap-in time of passenger $i$ to zero).

$t_{i,s}^{t,t}$: passenger $i$’s transfer time at the departure station of segment $s$.

$f_{s,t}(t)$: transfer time distribution at the departure station of segment $s$.

$P_{n}^{s}(t)$: the probability of being left behind $n$ times for a passenger who is expected to arrive at the platform of the departure station of segment $s$ ($\forall s = 1, 2...S_i$) at time $t$. The left behind model proposed in Chapter 5 estimates these probabilities using data from trips without transfers.

An important assumption of the general PIAM is that the probability of being left behind is the same for transfer and non-transfer passengers at the same station and time period. These probabilities are used for the assignment of both transfer and non-transfer trips. The assumption will be validated in Section 6.2.

As shown in Figure 6-2, the probability of passenger $i$ arriving at the origin station platform between trains $j - 1$ and $j$ ($j \geq 1$) is:

$$P(DT_{i,j-1}^1 \leq t_i^n < DT_{i,j}^1) = \int_{DT_{i,j-1}^1}^{DT_{i,j}^1} f_{s,t}(t)dt$$  \hspace{1cm} (6.4)$$

with $DT_{i,0}^1 = 0$.

Given that the passenger arrived between trains $j - 1$ and $j$ ($1 \leq j \leq I[1]$), the probability of boarding train $I[1]$ is the probability that the passenger is left behind $I[1] - j$ times, given that his/her expected arrival time on the platform is $t_i^{in} + \mathbb{E}[t_i^a]$, where $\mathbb{E}[t_i^a]$ is the expected access time.

$$P(\text{board train } I[1] | DT_{i,j-1}^1 \leq t_i^n < DT_{i,j}^1) = P_{I[1]-j}^1(t_i^{in} + \mathbb{E}[t_i^a])$$  \hspace{1cm} (6.5)$$
The probability of a passenger arriving at the origin station platform between trains $j - 1$ and $j$, and boarding train $I[1]$ can be derived using eqs. (6.4) and (6.5).

\[
P(\text{board train } I[1], DT_{i,j-1}^{1} \leq t_{i}^{a} < DT_{i,j}^{1})
\]

\[
= P(DT_{i,j-1}^{1} \leq t_{i}^{a} < DT_{i,j}^{1})P(\text{board train } I[1]|DT_{i,j-1}^{1} \leq t_{i}^{a} < DT_{i,j}^{1})
\]

\[
= \int_{DT_{i,j,j-1}^{1}}^{DT_{i,j}^{1}} f_{a}(t)dt P_{I[1]|_{j-1}^{1}}^{1}(t_{i}^{i} + \mathbb{E}[t_{i}^{a}]) \quad (6.6)
\]

To board train $I[1]$, passenger $i$ may arrive in any time interval before train $I[1]$’s departure. The probability of passenger $i$ boarding train $I[1]$ is given by:

\[
P(\text{board train } I[1])
\]

\[
= \sum_{j=1}^{I[1]} P(\text{board train } I[1], DT_{i,j-1}^{1} \leq t_{i}^{a} < DT_{i,j}^{1})
\]

\[
= \sum_{j=1}^{I[1]} P(DT_{i,j-1}^{1} \leq t_{i}^{a} < DT_{i,j}^{1})P(\text{board train } I[1]|DT_{i,j-1}^{1} \leq t_{i}^{a} < DT_{i,j}^{1})
\]

\[
= \sum_{j=1}^{I[1]} \int_{DT_{i,j,j-1}^{1}}^{DT_{i,j}^{1}} f_{a}(t)dt P_{I[1]|_{j-1}^{1}}^{1}(t_{i}^{i} + \mathbb{E}[t_{i}^{a}]) \quad (6.7)
\]

Similarly, given that passenger $i$ boarded train $I[s]$ in segment $s$, to board train $I[s+1]$ in the next segment $s + 1$, he/she may arrive at the transfer platform at any time before train $I[s+1]$’s departure. The conditional probability of boarding train $I[s+1]$ given that the passenger boarded train $I[s]$ in the previous segment is given by:

\[
P(\text{board train } I[s+1]|\text{board train } I[s])
\]

\[
= \sum_{j=1}^{I[s+1]} P(\text{board train } I[s+1], DT_{i,j-1}^{s+1} \leq t_{i}^{s+1} < DT_{i,j}^{s+1}|\text{board train } I[s]) \quad (6.8)
\]

with $DT_{i,j,0}^{s+1} = 0$.

Assuming that the probability a passenger is left behind $n$ times is the same for transfer and non-transfer passengers at the corresponding station, $P_{n}^{s}(t)$, estimated by the left behind model using data from non-transfer trips, applies for transfer
passengers as well. The probability of boarding train \(I_{s+1}\) in segment \(s+1\), given that the passenger arrived at the transfer platform between trains \(j-1\) and \(j\) and boarded train \(I_s\) in segment \(s\), is the probability of being left behind \(I_{s+1} - j\) times at the transfer station at time \(t_i^{in} + AT^{s,I}_{i,I_s} + E[t_i^{s+1,t}]\), which is the expected arrival time at the transfer platform of passenger \(i\) in segment \(s+1\) given boarding train \(I_s\).

\[
P(\text{board train } I_{s+1} | DT^{s+1,I}_{i,j-1} \leq t_i^{s+1,t} < DT^{s+1,I}_{i,j}, \text{ board train } I_s) = P_{I_{s+1}-j}(t_i^{in} + AT^{s,I}_{i,I_s} + E[t_i^{s+1,t}])
\]  

(6.9)

By substituting eq. (6.9) into eq. (6.8), the probability of boarding train \(I_{s+1}\) given that the passenger boarded train \(I_s\) in the previous segment is given by:

\[
P(\text{board train } I_{s+1} | \text{board train } I_s) = \sum_{j=1}^{I_s+1} P(\text{board train } I_{s+1} | DT^{s+1,I}_{i,j-1} \leq t_i^{s+1,t} < DT^{s+1,I}_{i,j}, \text{ board train } I_s)
\]

(6.10)

The probability of a passenger using itinerary \(I\) can be derived using eqs. (6.7) and
Given that the passenger used itinerary $I$, the probability of passenger $i$ tapping-out at the observed time can be derived from the distribution of egress time and the arrival time of train $I[S_i]$ at the arrival station of the last segment of itinerary $I$.

\[
P(t_{i_{\text{out}}} | I) = f_e(JT_i - AT_{i,I[S_i]})
\]
(6.12)

The probability of using itinerary $I$ and exiting at $t_{i_{\text{out}}}$ can be derived using eqs. (6.11) and (6.12).

\[
P(I, t_{i_{\text{out}}}) = P(t_{i_{\text{out}}} | I) P(I)
\]

For trips with no transfers $S_i = 1$, and eq. (6.13) reduces to:

\[
P(I, t_{i_{\text{out}}}) = f_e(JT_i - AT_{i,I[I]}^1) \sum_{j=1}^{I[I]} \int_{DT_{i,j}^1} f_a(t) dt P_{I[I]}^1(t_{i_{\text{in}}} + E[t^a_i])
\]
(6.14)
can be derived using Bayes’ theorem.

\[
P(I | t_{i_{out}}) = \frac{P(I, t_{i_{out}})}{P(t_{i_{out}})} = \frac{P(I, t_{i_{out}})}{\sum_{t' \in T_i} P(I', t_{i_{out}})}
\]  

(6.15)

where \( P(I, t_{i_{out}}) \) and \( P(I', t_{i_{out}}) \) are calculated as in eq. (6.13).

### 6.2 Model Validation

Since data about the actual passenger itineraries is not available, in order to validate the proposed methodology, synthetic data was generated using the tap-in times and the train movement data from a major subway system during the peak period on 2012/09/07, similar to the synthetic data generation in Chapters 3 and 5. Three heavily used OD pairs, 1-2, 2-3, and 1-3 were analyzed (see Figure 6-3). Passengers traveling from Station 1 to 3 transfer at Station 2.

![Figure 6-3: Network section for testing](image)

**6.2.1 Synthetic Data Generation**

As in Chapters 3 and 5, it is assumed that passenger walk speeds follow a log-normal distribution with mean 1.12 m/s and standard deviation 0.36 m/s (from actual ob-
servations of passengers at various stations in the system). The parameters of the distribution can also be estimated using automated data as in Chapter 7. The movement of the passengers through the system was simulated assuming that passengers board trains with available space on a first come, first served (FCFS) basis (according to their arrival times at the platform).

Figure 6-4 shows the distribution of journey times based on the tap-in and tap-out times of the synthetic trips, and compares them to the journey times actually experienced by the same passengers (i.e. based on the actual tap-in and tap-out times in the AFC data). The results show that the synthetic journey times are consistent with the actual journey times for the three OD pairs (1-2, 1-3, and 2-3).

The number of feasible itineraries for non-transfer trips (1-2 and 2-3) ranges from 1 to 8 with passengers having to wait for more than one train due to capacity constraints. The number of feasible itineraries for transfer trips (1-3) ranges from 1 to over 40. Over 40% of the passengers have more than 10 feasible itineraries because of the numerous combinations of trains they could have used. The distribution of the number of feasible itineraries in the actual data is similar.
6.2.2 Assumption Validation

An important assumption of the proposed model is that transfer and non-transfer passengers experience similar left behind probabilities. Figure 6-5 compares the actual number of times transfer passengers (blue dots) and non-transfer passengers (green lines) are left behind at the same station (synthetic data). The x-axis shows the actual arrival time at the platform and y-axis the number of times the passenger is left behind. The vertical lines indicate the train departures. The left behind probabilities for transfer passengers are the same as for non-transfer passengers who arrive at the platform at the same time (transfer passengers alight in a batch from the connecting train). Within each headway, the probability a passenger being left behind increases as the passenger’s arrival time at the platform gets closer to the train arrival.

![Figure 6-5: Number of times a passenger is left behind at the transfer station](image)

6.2.3 Results

The probability a passenger being left behind at Stations 1 and 2 is estimated using the maximum likelihood method proposed in Chapter 5 by station and time interval,
using data for non-transfer trips. Figure 6-6 shows the estimation of the expected number of times a passenger is left behind at Station 2 compared to the synthetic data. The vertical lines indicate train departures. The estimation is consistent with the synthetic data and captures the rapid change over time of the crowding levels at the various stations.

![Graph showing expected number of times a passenger is left behind at Station 2](image)

**Figure 6-6:** Expected number of times a passenger is left behind at Station 2

In the synthetic data, the actual itinerary taken by each passenger is known. The distribution of the probabilities of taking the actual itinerary estimated by PIAM is shown in Figure 6-7 for non-transfer trips (6-7a and 6-7b), and transfer trips (6-7c). Even with the very large number of feasible itineraries for transfer trips, the probability of assigning the passenger to the actual (correct) itinerary is high.
Figure 6-7: Distribution of the probabilities of assigning passengers to their actual itinerary

Figure 6-8 compares the estimated and actual transfer times for transfer passengers at Station 2. Transfer time includes walk time to the transfer platform after alighting and wait time for the connecting train. The estimated transfer times are very close to the actual. The transfer time increases as the crowding level increases from 18:00 to 19:00, and more passengers are left behind. Based on the assignment results, the expected load on individual trains can be estimated as the sum of the probabilities of taking the corresponding itineraries as in Chapter 3. Figure 6-9 shows the train
load estimation for the two lines. The solid lines are the total loads and the dashed lines are the loads corresponding to non-transfer trips only. Most trains departing Stations 1 and 2 are fully loaded and PIAM captures the capacity constraints fairly accurately. The estimated loads are very close to the actual loads.

### 6.2.4 Comparison with Hörcher et al. (2017)

In this section, we compare the PIAM results to the model proposed by Hörcher et al. (2017), using synthetic data from the same network between 15:00 and 20:00 on 2012/09/07. Hörcher et al. (2017) extended the Passenger-to-Train Assignment Model (PTAM) framework proposed in Zhu (2014) to incorporate transfers and route choice using a number of simplifying assumptions. Chapter 2 discusses the implications of these assumptions: (i) the dynamics at the transfer station (left be hind s, transfer distance, walk speed, etc.) are not considered. For example, if two itineraries have the same egress time and similar likelihood of the corresponding delayed access time, the distribution of transfer time and left behind probabilities at the transfer station would be important for accurate assignment; (ii) the empirical egress time distribution was derived using trips with a single feasible itinerary. This results in overestimation of passengers’ walk speed, especially in the case with short headways; (iii) the “delayed access time” distribution is estimated based on a large sample of access times. With station crowding levels changing quickly during the peak periods, an accurate estimation of delayed access time distribution requires long periods of observation. It may also reflect average conditions and underestimate the crowding levels at the peak of the peak (Zhu et al., 2017b).

The Hörcher et al. model can be summarized in the following steps for the analyzed network with 3 OD pairs:

1. Estimate empirical egress time distribution at Station 2, $f^2_e(t)$, based on trips from Station 1 to 2 with a single feasible itinerary,

2. Estimate empirical egress time distribution at Station 3, $f^3_e(t)$, based on trips from Station 2 to 3 with a single feasible itinerary,
Figure 6-9: Train load estimation

(a) Red Line (departing Station 2)

(b) Blue Line (departing Station 1)
3. Assign trips from Station 2 to 3 with multiple feasible itineraries based on $f^3_e(t)$,
4. Assign trips from Station 1 to 2 with multiple feasible itineraries based on $f^2_e(t)$,
5. Estimate empirical in-station time (including access time and wait time) distribution at Station 1, $f^1_u(t)$, based on the assignment results for trips from Station 1 to 2,
6. Assign transfer trips from Station 1 to 3 based on $f^1_u(t)$ and $f^3_e(t)$, assuming the probability of choosing each itinerary is proportional to the corresponding likelihood of in-station time and egress time.

Figure 6-10 shows the empirical distribution of egress times at Station 3 as described in Step 2, compared with the true distribution for the whole population. The empirical egress times distribution estimated using the truncated sample is biased and underestimates the probability of having long egress times.

![Figure 6-10: Distribution of transfer times](image)

The distribution of the estimated probabilities of boarding the actual itinerary is shown in Figure 6-11 for trips with more than one feasible itinerary. Both PIAM and the Hörcher et al. model perform well for non-transfer trips. For transfer trips, the probability of boarding the actual itinerary according to the Hörcher et al. model is almost uniformly distributed between 0 and 1 (Figure 6-11b). PIAM is able to
estimate the probability of the actual itinerary more accurately, with performance similar to the one for non-transfer trips.

Figure 6-11: Distribution of the probabilities of boarding the actual itinerary

Figure 6-12 shows the distribution of the probability of boarding the actual train for the first and last segments respectively for transfer trips based on the Hörcher et al. model. For transfer trips, the Hörcher et al. model tends to assign passengers to an earlier train in the first segment (see Figure 6-12). Using only the “delayed access time” and egress time distributions without considering the dynamics at the transfer station is insufficient to correctly assign transfer passengers.

Figure 6-12: Distribution of the probability of boarding the actual train for different segments based on the Hörcher et al. model
Figure 6-13 compares the estimated transfer times with the actual times for transfer passengers using the two methods. The Hörcher et al. model overestimates the transfer time, while PIAM’s estimation is close to the actual. The estimated distribution of transfer times is further used by the Hörcher et al. model for the assignment of trips with multiple routes, resulting in biased results.

![Figure 6-13: Comparison of transfer time estimation](image)

To evaluate the model’s accuracy and robustness in estimating train loads, Monte-Carlo simulation was used to assign passengers to itineraries based on the corresponding probabilities as estimated by both methods. Figure 6-14 shows the 5\textsuperscript{th}, 50\textsuperscript{th}, and 95\textsuperscript{th} percentiles of estimated the train load based on 2000 replications for the Blue Line (departing Station 1). The Root Mean Square Error (RMSE) of the train load estimation is 11.3 pax/train based on the Hörcher et al. model and 6.8 pax/train based on PIAM. For the Red Line, the RMSE is 5.3 pax/train according to the Hörcher et al. model and 3.9 pax/train based on PIAM. The train load distribution estimated by the Hörcher et al. model also has a larger variance compared with the PIAM results. The reason is that the distribution of probabilities of boarding the actual itinerary for transfer trips estimated by the Hörcher et al. model is almost uniform (see Figure 6-11b).
Figure 6-14: Train load estimation for the Blue Line (departing Station 1)

(a) Hörcher et al. (2017)

(b) PIAM
6.3 Application with Actual Data

In this section, the model is applied using actual data from the same system (entry/exit times and train arrival/departure times). 13,741 trips during the period from 15:00 to 20:00 on 2012/09/07 were used for the analysis.

The left behind model as discussed in Chapter 5 is used to estimate the distribution of the number of times a passenger is left behind at Stations 1 and 2 by time interval. Figure 6-15 shows the estimation results along with the train departure times (vertical lines). Station 2 is very crowded as expected, since it is one of the busiest stations in the network with large transfer volumes. The train frequency on the Red Line doubles starting at 17:00 and the crowding level starts increasing at 18:00.

![Figure 6-15](image)

(a) Station 1  
(b) Station 2

Figure 6-15: Expected number of times a passenger is left behind as a function of time

The distributions of the number of feasible itineraries are shown in Figure 6-16 for the three OD pairs. The number of feasible itineraries for transfer trips is very large as expected.

PIAM estimates the different components of a passenger’s journey time (access time, in-vehicle time, egress time, etc.). Based on this output, a metric of interest, specifically for transfer trips, is the time passengers spend at transfer stations. Figure 6-17 shows the distribution of transfer times for passengers traveling from Station 1 to Station 3. With the train frequency doubling at 17:00, the transfer time initially
Figure 6-16: Distribution of the number of feasible itineraries
decreases. However, due to the higher crowding levels and left behinds at the transfer station, the transfer time increases again at the peak of the peak. Compared with synthetic data, the peak of the peak comes earlier, possibly due to the large volume of demand from other stations/lines, while only three OD pairs are considered in the generation of synthetic data.

Figure 6-17: Estimated transfer time for each passenger

Figure 6-18 shows the estimation of total in-station times at the origin station for passengers traveling from Station 1 to 3 and Station 1 to 2 respectively. Including access and wait times, the in-station time is a function of the station characteristics, service frequency, and the number of times a passenger is left behind. The in-station time at the origin station, Station 1, is the same for transfers (1-3) and non-transfers (1-2). Some passengers have very long in-station times. Those are probably passengers who conduct other activities in the station, such as meeting with friends, etc. Note that, those outliers are not included in the estimation of left behind probabilities.
6.4 Extension to Trips with Route Choice

The framework for assigning passengers to itineraries presented in the previous sections can be extended to the case of trips with route choice.

A number of papers in the literature have proposed methods to estimate route choice fractions using AFC data with journey time observations. They usually model the travel time distribution as a mixture of different distributions (each corresponding to a single route) and use Bayesian inference or maximum likelihood for the route choice fraction estimation (Fu et al., 2014; Lee and Sohn, 2015; Sun et al., 2015). The applicability of those models is limited due to the effect of capacity constraints. During peak periods, a “shorter” route may have passengers who are left behind. However, the above models cannot distinguish if a passenger with a long journey time chose a longer route or was left behind multiple times on a shorter route.

Zhao et al. (2017) considered the congestion effects and used the left behind probabilities at origin and transfer stations estimated at an earlier step of their model (discussed in Section 1) to estimate route choice fractions using maximum likelihood. They assumed that the likelihood of passengers’ tap-out times is equal to the sum of the likelihood of choosing each itinerary on each route, calculated as the product of the corresponding left behind probabilities at transfer/origin stations assuming that
passengers’ access/transfer/egress times are shorter than the headway.

Hörcher et al. (2017) estimated the route choice fractions for each individual passenger based on the estimated distributions of delayed access times, egress times and transfer times. By enumerating feasible itineraries for each route, they assumed that the probability of a passenger choosing each route is proportional to the sum of the likelihood of corresponding feasible itineraries. As previously pointed out, the distributions of journey time components are biased.

In this section, we extend the framework we presented in previous sections to include itineraries with route choice. Figure 6-19 illustrates the possible instances for a passenger with two possible routes (each with one transfer). $P(r_1)$ and $P(r_2)$ denote the probabilities of choosing routes $r_1$ and $r_2$ respectively. For this passenger, the number of feasible itineraries can be very large.

Figure 6-19: PIAM structure for a passenger with route choice

We therefore propose a three-step approach extending the framework proposed in Chapter 4 (see Figure 4-2). Figure 6-20 illustrates the approach of incorporating route choice in PIAM:

- The left behind model estimates the probability of being left behind using data
from trips without route choice or transfers;

- The route choice model estimates the route choice fractions by time interval given the left behind probabilities;

- The assignment model assigns passengers to itineraries based on the left behind probabilities and route choice fractions.

![PIAM modeling framework](image)

**Figure 6-20: PIAM modeling framework**

### 6.4.1 Estimation of Route Choice Fractions

The estimation of route choice fractions follows the left behind model (uses the left behind probabilities as input) and is based on the same principles used in Chapter 5. For a trip with multiple routes, the conditional probability of using each itinerary and tapping-out at the observed tap-out time, given the corresponding route choice, is given by eq. (6.13), and is a function of the access/transfer/egress time distributions and the left behind probabilities. Let $I_{i,r}$ denote the feasible itinerary set for passenger $i$ given route $r$, and $R_i$, the set of possible routes for passenger $i$. The conditional
probability of using itinerary $I \in \mathcal{I}_{i,r}$ and tapping-out at $t_{i}^{out}$ given route $r \in \mathcal{R}_{i}$ is:

$$P(I, t_{i}^{out} | r)$$

$$= f_{e}(JT_{i} - AT_{i, I[S_{i}]}) \sum_{j=1}^{I_{i}[1]} \int_{DT_{i,j}^{1}}^{DT_{i,j}^{1}} f_{a}(t) dt P_{i}[1]-j(t_{i}^{in} + E[t_{i}^{a}])$$

$$\times \prod_{s=1}^{S_{i}-1} \left[ \sum_{j=1}^{I_{i}[s+1]} \int_{DT_{i,j}^{s+1}}^{DT_{i,j}^{s+1}} f_{s+1,i}(t) dt P_{i}[s+1]-j(t_{i}^{in} + AT_{i,j}^{s+I} + E[t_{i}^{s+1+1}]) \right] \tag{6.16}$$

Let $P(r) \ \forall r \in \mathcal{R}_{i}$ be the probability of choosing route $r$. The probability of passenger $i$ tapping-out at the observed time is the sum over all possibilities:

$$P(t_{i}^{out}) = \sum_{r \in \mathcal{R}_{i}} \sum_{I \in \mathcal{I}_{i,r}} P(I, t_{i}^{out} | r)P(r) \tag{6.17}$$

where $P(I, t_{i}^{out} | r)$ is given by eq. (6.16).

For the whole population, assuming conditional independence among passengers, the likelihood of observing the journey times of all passengers in the group is given by:

$$L = \prod_{i=1}^{N} P(t_{i}^{out}) = \prod_{i=1}^{N} \sum_{r \in \mathcal{R}_{i}} \sum_{I \in \mathcal{I}_{i,r}} P(I, t_{i}^{out} | r)P(r) \tag{6.18}$$

and the corresponding likelihood function is:

$$\mathcal{L} = \sum_{i=1}^{N} \log P(t_{i}^{out}) = \sum_{i=1}^{N} \log \sum_{r \in \mathcal{R}_{i}} \sum_{I \in \mathcal{I}_{i,r}} P(I, t_{i}^{out} | r)P(r) \tag{6.19}$$

The probability of choosing different routes can be estimated by solving the fol-
lowing problem with respect to $P(r)$.

$$\max_{P(r) \forall r \in \mathcal{R}} \quad \mathcal{L} = \sum_{i=1}^{N} \log \sum_{r \in \mathcal{R}_i} \sum_{t \in \mathcal{T}_{i,r}} P(I, t_{i}^{\text{out}} | r) P(r)$$

(6.20)

s.t. \quad \sum_{r \in \mathcal{R}} P(r) = 1 \quad (6.21)

$$P(r) \geq 0, \quad \forall r \in \mathcal{R} \quad (6.22)$$

where $\mathcal{R}$ is the set of all possible routes for the analyzed OD pair.

### 6.4.2 Assignment Model

With known left behind probabilities and route shares by time interval, the probability of using different itineraries for a given passenger, can be calculated as in eq. (6.15).

$$P(I, r | t_{i}^{\text{out}}) = \frac{P(I, r, t_{i}^{\text{out}})}{P(t_{i}^{\text{out}})} \quad \forall I \in \mathcal{T}_{i,r}, \quad \forall r \in \mathcal{R}_i \quad (6.23)$$

$$= \frac{P(I, t_{i}^{\text{out}} | r) P(r)}{P(t_{i}^{\text{out}})}$$

where $P(I, t_{i}^{\text{out}} | r)$ and $P(t_{i}^{\text{out}})$ are given by eqs. (6.16) and (6.17) respectively.

### 6.4.3 Application with Synthetic Data

The model is applied to four OD pairs, 1-2, 2-3, 1-3, and 4-2 (Figure 6-21). From Station 1 to 3, passengers can transfer from the Blue Line to the Red Line at Station 2 or Station 4. At Station 2, due to crowding during the peak period, passengers may be left behind multiple times. Some transferring passengers who are familiar with the system may choose to transfer at Station 4, which is the terminal station for the Red Line and does not have left behind issues.

In order to validate the proposed method, synthetic data for the four OD pairs, 1-2, 2-3, 1-3, and 4-2, was generated using the tap-in times and the train movement data. Passengers were assigned to a path (1-2-3 or 1-4-3) according to pre-defined
fractions. At transfer stations, (Stations 2 and 4), transfer passengers and tap-in passengers are loaded onto the trains based on first come, first served (FCFS) rules according to their arrival time at the platform.

PIAM is applied using the generated data. The left behind probabilities at Stations 1, 2, and 4 are calculated using the method discussed in Chapter 5 by station and time interval. No passengers are left behind at Station 4. The route choice fractions were estimated using maximum likelihood as described in Section 6.4.1. Given the estimated route fractions and left behind probabilities, the assignment model is used to estimate the probability of each feasible itinerary for each passenger.

Figure 6-22 shows the estimated probability, $P(r_1)$, of choosing Station 2 (i.e. route 1-2-3) as the transfer station compared with the “true” in the synthetic data in 30 min intervals and for different “true” probabilities. The estimated probability of choosing Station 2 as the transfer station is very consistent with the actual (synthetic) data.

The distribution of the estimated probabilities of assigning individual passengers to their actual itineraries is shown in Figure 6-23 for trips with route choice (1-3). The probability of assigning passengers to their actual itinerary is high despite the fact that the number of feasible itineraries is very high (60 in some cases).
6.4.4 Application with Actual Data

Figure 6-24 shows the estimated probability of choosing Station 2 as the transfer station in 60 min intervals using actual data. As expected, during the peak period, due to the possibility of being left behind and overcrowding at Station 2, some passengers chose to transfer at Station 4 instead.
6.5 Summary

This chapter developed the second module of the general Passenger-to-Itinerary Assignment Model (PIAM), the assignment model, that is applicable to trips with or without transfers. The assignment model uses as input the probability a passenger is left behind by station and time interval and access/transfer/egress time distributions. For the assignment of trips with route choice, the model estimates the route choice fractions and assigns trips to itineraries based on the estimated route choice fractions and left behind probabilities. Synthetic data is generated to validate the model. The results, in terms of the probability of assigning passengers to their actual route and itinerary, are accurate. The model output includes a complete inference of a passenger’s movements through the network at a high resolution. The model output can also be used to estimate aggregate performance measures, such as train loads, average waiting time, transfer time, etc.
Chapter 7

Access/Egress Time Model

The access/egress time distributions are key inputs to the PIAM. This chapter proposes a maximum likelihood method to estimate the parameters of the walk speed/distance distributions using AFC and AVL data, circumventing the need for manual observations. The access/egress time distributions can thus be estimated.

Section 7.1 introduces the problem. Section 7.2 develops the walk speed model and illustrates the underlying assumptions. Section 7.3 proposes a walk distance estimation method based on station configuration. Section 7.4 describes the maximum likelihood estimation method using AFC and AVL data. Section 7.5 applies the methodology using both synthetic and actual data, and validates the results. Section 7.6 draws conclusions.

7.1 Introduction

The access and egress time distributions are important inputs to the PIAM. Some agencies, e.g. MTR in Hong Kong and TfL in London, occasionally collect access/egress time and walk speed data manually. In this chapter, we discuss a method to estimate the walk speed/distance distributions of passengers in stations using AVL and AFC data without the need for manual data collection. The advantage of estimating the speed/distance distributions instead of the access/egress time distribution directly is that the former is more general and can accommodate characteristics of the
station, such as configuration, walk distances, etc., and characteristics of passengers. As such, it can be used throughout the network.

The access time $t^a_i$ and egress time $t^e_i$ for passenger $i$ can be calculated as:

$$t^a_i = \frac{s^a_i}{v^a_i}$$  \hspace{1cm} (7.1)

$$t^e_i = \frac{s^e_i}{v^e_i}$$  \hspace{1cm} (7.2)

where $s^a_i$ and $s^e_i$ are the access/egress distances respectively and $v^a_i$ and $v^e_i$ the access/egress speeds respectively for passenger $i$. In the following sections, we will discuss models for estimating the speed and distance used in eqs. (7.1) and (7.2). Based on the speed/distance distributions, the access and egress time distributions can be derived.

### 7.2 Walk Speed Model

Passenger’s walk speed at different stations is a function of:

- Individual characteristics, such as age and gender. Familiarity with the system and the stations is also a factor. Trip purpose may also be important: commuters, for example, will generally be faster than tourists.

- Station configuration, such as ramps, stairs, and escalators.

- The degree of crowding. Especially during peak hours, crowding increases the interactions among passengers and reduces their walk speeds.

The walk speed is modeled as a random variable following a specific distribution (for example normal or log-normal). The mean (and variance) of this distribution can be expressed as a function of many factors.

$$\mu_i = \alpha_0 + \alpha X_i$$  \hspace{1cm} (7.3)

Where,
\( \mu_i \): the mean of the distribution for passenger \( i \).

\( \alpha \): vector of parameters.

\( X \): set of explanatory variables.

Examples of explanatory variables include:

\[
X_{1,i} = \begin{cases} 
1 & \text{for senior card holders} \\
0 & \text{o.w.}
\end{cases}
\]

\[
X_{2,i} = \begin{cases} 
1 & \text{for frequent travelers (e.g., more than 8 trips per week)} \\
0 & \text{o.w.}
\end{cases}
\]

\( X_{3,i} \): measure of station complexity.

\[
X_{4,i} = \begin{cases} 
1 & \text{for trips during peak hours} \\
0 & \text{o.w.}
\end{cases}
\]

The access and egress speeds can be modeled as a joint distribution.

\[
(v_i^e, v_i^a) \sim f_{V_i^e, V_i^a}(\mu_i^e, \sigma_i^e, \mu_i^a, \sigma_i^a, \rho)
\]

(7.4)

Where,

\( f_{V_i^e, V_i^a} \): the joint distribution of access/egress speeds for passenger \( i \).

\( \rho \): the correlation between access and egress speeds.

\( \mu_i^e, \sigma_i^e, \mu_i^a, \sigma_i^a \): parameters of the distribution.

### 7.3 Walk Distance Model

The tap-in/out gates for each individual are known from the AFC data. However, even though the transactions gates are recorded, the exact locations on the platforms each passenger alighted at, or boarded, and the path he/she chose are not observed. To deal with this issue, the access and egress distances can be modeled as random
variables with distributions:

\[ s^a_i \sim f_{S^a_i}(s^a_i) \quad (7.5) \]
\[ s^e_i \sim f_{S^e_i}(s^e_i) \quad (7.6) \]

Where,

\( s^a_i, s^e_i \): access and egress distances of passenger \( i \) respectively.

\( f_{S^a_i}(s^a_i), f_{S^e_i}(s^e_i) \): probability distributions of access and egress distances respectively for passenger \( i \).

Different models at varying degrees of sophistication can be used to capture the distance distribution. Here we propose a model where we assume that passengers are divided into two groups. Passengers in the first group optimize their platform location to minimize the distances to/from the gates. Passengers in the second group randomly choose their platform locations to board (and alight) the train. In this case, some passengers may have long walk distances.

To capture this behavior, we introduce a parameter representing the proportion (or the probability) of passengers who optimize their locations on the platform. Then, the walk distance distribution for a random passenger can be expressed as:

\[ f_{S^a_i}(s^a_i) = pf_{S^a_o}(s^a_i) + (1-p)f_{S^a_r}(s^a_i) \quad (7.7) \]
\[ f_{S^e_i}(s^e_i) = pf_{S^e_o}(s^e_i) + (1-p)f_{S^e_r}(s^e_i) \quad (7.8) \]

Where,

\( f_{S^a_o}(s^a_i), f_{S^a_r}(s^a_i) \): access/egress distance distributions respectively for passenger \( i \) given that he/she optimizes their platform locations.

\( f_{S^a_r}(s^a_i), f_{S^e_r}(s^e_i) \): access/egress distance distributions respectively for passenger \( i \) given that he/she is located randomly on the platform.

\( p \): proportion of passengers who optimize their locations.

The parameter \( p \) can be estimated as a parameter of the access/egress time model.
7.3.1 Distance for “Optimizing” Passengers

For this group, the lower bound on the walk distance can be obtained by assuming that all passengers alight at the section of platform that is closest to their tap-out gates. These passengers position themselves on the boarding platform accordingly. An optimizing passenger can move to the train door through which he/she can alight at the section of the destination platform with the shortest egress distance while waiting at the origin station. Each platform is divided into sections and each gate group is assigned to the closest section of the platform.

Figure 7-1 shows a simplified station layout with the orange area representing the concourse (fare gate) level and the blue area representing the platform level.

The fare gates are organized into groups A, B, C and D, based on their locations. The platform is divided into four sections based on the closest fare gate group, with boundaries drawn at the mid-point of the distance between adjacent escalator entries. Access points to the platform can also inform the location of the boundaries.

An optimizing passenger who tapped-in through fare gate group A and tapped-out though fare gate group B, would select path “A-A-B’-B’”. The passenger walked from fare gate group A to section A, moved from part A to part B on the
origin platform during his/her waiting time, alighted at section B on the destination platform, and exited through fare gate group B. At both entry and exit stations, the walk distances are minimized. Note that the distance on the origin platform from section A to section B is not counted in the access distance because the passenger has arrived at the platform and the access time that determines his/her boarding train is not affected by this extra walking. However, if there was not enough time for this passenger to move along the origin platform (such as when the passenger arrived at the platform and the train is about to leave), this passenger could also follow the path “A-A-A’-B’”. In this case, the egress distance will be longer than for the previous path.

We assume that under most circumstances, an optimizing passenger can always minimize the walk distances at both entry and exit stations. The walk distances for this type of passengers are estimated from the fare gate groups and the station layouts.

### 7.3.2 Distance for Random Passengers

Passenger in this group randomly choose their locations to board (and alight) the train. Under this assumption, passengers may alight from one end of the platform, walk to the other end of the platform and tap-out at their fare gate. In Figure 7-1, such passengers may take a longer path such as “A-A-A’-B’”, or “A-C-C’-B’”. We should keep in mind that, even passengers who are not familiar with the stations, can still take the optimal path by chance. However, the average walk distance for this group should be longer than the “optimizing” group.

### 7.4 Estimation

Measurements of the walk speed of passengers can be obtained through direct observations by having surveyors follow passengers when they enter the system. MTR for example, has conducted such surveys for key stations. However, such data typically represents small samples. Manual data collection is also expensive and time
consuming, explaining the small sample sizes.

The alternative proposed in this chapter, is to use AFC and AVL data for estimation of the walk speed parameters. More specifically, from this data, passengers with one feasible itinerary can be identified (see for example in Chapter 3). For this subset of trips, the actual egress time for passenger $i$ can be calculated as the difference between the train arrival time and the passenger’s tap-out time:

$$t_i^e = t_i^{out} - AT_{i,1}$$

Where,

- $t_i^e$: egress time of passenger $i$,
- $t_i^{out}$: tap-out time,
- $AT_{i,1}$: the arrival time at the destination of the only feasible train for passenger $i$.

A naive method to estimate the desired distribution for the egress time would be to use these observed egress times. However, the sample of such observations is not representative of the population, and using it directly would result in biased parameters (overestimating the walk speed). The reason for this is that in order for a passenger to have only one feasible itinerary, their egress time has to be less than the headway between the train they arrived on and the next train. Otherwise, they would have more than one feasible itinerary. Therefore, these observed egress times have an upper bound which is equal to the corresponding headway. This implies that the observations are truncated and do not represent a random sample from the underlying distribution of interest. This is an important characteristic of the data, especially in cases where headways are short.

The (truncated) sample of passengers with one feasible itinerary has an egress time distribution that follows the distribution of the general population conditional on $t_i^e \leq h_i$, where $h_i$ is the headway between the train passenger $i$ took and the next train.

The probability density function of the speed distribution in the truncated sample
will be:

\[ f_{V_{e} | V_{min,i}}(v_{e} | v_{min,i}) = \frac{f_{V_{e}}(v_{e})}{1 - F_{V_{e}}(v_{min,i})} \]  \hspace{1cm} (7.10)

Where,

\( i \): passenger index.

\( f_{V_{e} | V_{min,i}}(v_{e} | v_{min,i}) \): conditional probability density function of walk speed for passenger \( i \).

\( f_{V_{e}}(v_{e}) \): probability density function of walk speed for passenger \( i \).

\( F_{V_{e}}(v_{e}) \): cumulative distribution of walk speed for passenger \( i \).

\( v_{min,i} \): minimum speed for passenger \( i \), \( v_{min,i} = \frac{s_{i}}{b_{i}} \).

Following the discussion in Section 7.3, we assume that the egress distance follows a distribution with probability density function \( f_{S_{e}}(s_{e}) \). Based on eq. (7.10), the probability that passenger \( i \), in the truncated sample, has egress time less than \( t_{e} \) is given by:

\[ P(t \leq t_{e}) = \int_{0}^{t_{e}} \int_{s_{e}}^{\infty} \frac{f_{V_{e}}(v)}{1 - F_{V_{e}}(v_{min,i})} f_{S_{e}}(s_{e}) dv ds_{e} \]  \hspace{1cm} (7.11)

The probability density function of the egress time distribution, \( f_{T_{e}}(t) \) for the truncated sample is given by:

\[ f_{T_{e}}(t) = \int_{0}^{t_{e}} \int_{s_{e}}^{\infty} \frac{f_{V_{e}}(v)}{t_{e}^{2}} \frac{s_{e}^{2}}{1 - F_{V_{e}}(v_{min,i})} f_{S_{e}}(s_{e}) dv ds_{e} \]  \hspace{1cm} (7.12)

For egress time observation, \( t_{e} \), the log-likelihood function is given by:

\[ L^{*} = \sum_{i} \log f_{T_{e}}(t_{e}) \]  \hspace{1cm} (7.13)

By maximizing the log-likelihood function as in eq. (7.13) with respect to the parameters of the walk speed and distance distributions, the mean and variance of the walk speed and any parameters associated with the distance distribution (e.g.
percentage of passengers with optimizing behavior) can be estimated.

Given the walk speed distribution, based on the walk distance distribution at origin/destination/transfer stations, the access/egress/transfer time distributions can be easily derived.

### 7.5 Validation

#### 7.5.1 Validation with Synthetic Data

Synthetic data is used to validate the walk speed estimation method, assuming a log-normal speed distribution, consistent with the literature. Kim et al. (2006) presented a detailed mobility model for pedestrians based on data from a 13-month tracking experiment. They reported that the walk speed followed a log-normal distribution with an average of 1.26 m/s. Several other studies on pedestrian movements also found that either a normal or log-normal distribution is a good probability distribution to represent walk speed. Especially, when the walk speed is asymmetric, a log-normal distribution is better (Ottomanelli et al., 2012; Zhang et al., 2009; Daamen and Hoogendoorn, 2006).

For the generation of the synthetic data, the walk distance is assumed known and constant for all passengers using the same gates. Different parameters of the walk speed distribution are also assumed in order to test the robustness of the methodology. The subset of passengers with a single feasible itinerary is used as the (truncated) sample in the maximum likelihood estimation. With a fixed walk distance, the likelihood function, eq. (7.13), reduces to:

\[
L^* = \sum_i \log \left( \frac{f_{V_i^e}(v_i^e)}{1 - F_{V_i^e}(v_{min,i})} \right) \tag{7.14}
\]

As discussed above, a naive method to estimate the desired distribution for the egress time would be to use these observed egress times without recognizing that the sample is truncated, i.e. using \( f_{V_i^e}(v_i^e) \) in the likelihood function directly. Figure 7-2 compares the estimated and the actual distributions for different walk speed mean and
standard deviation values. The green dash line represents the true distribution, the red line the distribution estimated using eq. (7.14), and the blue line the distribution estimated using the naive method without recognizing that the sample is truncated.

![Graph showing distribution comparison](image)

(a) Mean: 1.2; Standard deviation: 0.74  
(b) Mean: 1.5; Standard deviation: 1.23

Figure 7-2: Maximum likelihood estimation results

The proposed method replicates the actual distribution well, while the naive method, as expected, overestimates the walk speed due to the biases introduced.

### 7.5.2 Validation Using Actual Data

The proposed method for estimating the speed distribution was also validated using actual data on passenger speeds. The agency occasionally collects data on walk speed using surveyors who follow individual passengers at key stations.

The parameters of the walk speed distribution are estimated using transactions for the busiest OD pair in the network of interest with 1,603 trips having one feasible itinerary. The egress distances were measured from the station plan and the locations of the tap-out gate of each AFC transaction. It is assumed that the egress distance is uniformly distributed between the shortest and longest paths from the platform to the exit gate.

Table 7.1 compares the parameters estimated with the proposed method with the parameters based on manual observations. The estimation results are consistent with the survey results, despite the relatively simple walk distance model used. While the manual survey is time-consuming and expensive, the proposed method results in a
good estimate of the parameters of the walk speed distribution using AFC and AVL data.

<table>
<thead>
<tr>
<th></th>
<th>Mean (m/s)</th>
<th>Std (m/s)</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey (Male)</td>
<td>1.12</td>
<td>0.31</td>
<td>102</td>
</tr>
<tr>
<td>Survey (Female)</td>
<td>1.02</td>
<td>0.19</td>
<td>114</td>
</tr>
<tr>
<td>Estimated</td>
<td>1.11</td>
<td>0.50</td>
<td>1,603</td>
</tr>
</tbody>
</table>

Table 7.1: Comparison of the manual and estimation results

7.6 Summary

The access/egress time distributions are important inputs to the PIAM. This chapter proposed a method to estimate them using only AFC and AVL data. The proposed approach consists of two components: a walk speed model and a walk distance model. The method can incorporate the impact of station characteristics and passenger behavior in choosing their access and egress paths. The model parameters can be estimated using maximum likelihood and data from trips with only one feasible itinerary. For these trips, the egress time is known precisely. However, such a sample is truncated (i.e. consists of passengers with egress time less than the corresponding headway). The proposed formulation recognizes this and corrects for biases due to the truncated nature of the sample. An example with synthetic/actual data was used to illustrate the feasibility of the approach. The estimation results compare well with actual observations.
Chapter 8

Crowding Inference

In this chapter, we present two methods to infer the number of passengers in stations and the number of passengers waiting on platforms based on the PIAM results. The aim of these methods is to provide a computationally efficient way to calculate crowding levels.

Section 8.1 introduces the crowding problem and discusses related literature. Section 8.2 develops the model and describes the two estimation approaches. Section 8.3 validates the model using simulation. Section 8.4 applies the model using actual data. Section 8.5 concludes the chapter.

8.1 Introduction

Calculation of crowding levels in stations using the PIAM requires the application of the model to all OD pairs in the network, and is computationally intensive. Measurement through video cameras is currently not practical (Lam et al., 1999; Jiang et al., 2009).

In this chapter, two methods to infer the number of passengers in stations and the number of passengers waiting on platforms based on the PIAM results are proposed. The methods have computational advantages as they only require the application of the model to limited OD pairs with the subject station as the origin.
8.2 Methodology

The main idea behind these methods is that the average in-station time and platform wait time can be interpreted as the average wait time, \( W_q \), in a queuing system. Figure 8-1 illustrates the relationship between the average wait time and the average queue length in a queuing system.

We are interested in calculating the average queue length during the period from \( t_1 \) to \( t_2 \), \( L_q \):

\[
L_q = \frac{\int_{t_1}^{t_2} L_q(t)\,dt}{t_2 - t_1} \quad (8.1)
\]

where \( L_q(t) \) is the queue length at \( t \).

Assuming that the arrival and departure rates of passengers during the period from \( t_1 \) to \( t_2 \) are constant (i.e. the curves between \( A_1 \) and \( A_2 \) and between \( B_1 \) and \( B_2 \) are straight lines), the average queue length during this time period is equal to the queue length at \( t_3 = \frac{t_1 + t_2}{2} \) (the length of \( AB \)).
In Figure 8-2, $t_1'$ and $t_2'$ denote the entry times for passengers who exit the system at $t_1$ and $t_2$ respectively. Assuming that the arrival rate from $t_1'$ to $t_2'$ is also constant (i.e. the curve between $C_1$ and $C_2$ is a straight line), the wait time for passengers who enter the system at $t_1'$ and exit at $t_2$, $W_q$, is equal to the average wait time for passengers who enter the system from $t_1'$ to $t_2'$ (exit from $t_1$ to $t_2$).

Similar to Little’s formula (Larson and Odoni, 1981), from the $\triangle ABC$:

\[ L_q = AB = \tan \angle ACB \times BC = \lambda W_q \]  \hspace{1cm} (8.2)

where,

$L_q$: the average queue length from $t_1$ to $t_2$.

$W_q$: the average wait time for passengers who exit the system from $t_1$ to $t_2$.

$\lambda$: the average arrival rate from $t_3'$ to $t_3$. Assuming that the arrival rate from $t_1$
to \( t_3' \) is close to the arrival rate from \( t_3' \) to \( t_3 \), the arrival rate from \( t_3' \) to \( t_3 \) can be approximated by the arrival rate from \( t_1 \) to \( t_2 \).

The two methods differ with respect to the calculation of the average wait time. The first method has as input the journey time decomposition results from PIAM as in Chapter 3 (for non-transfer passengers for a limited number of OD pairs). As the individual boarding time is estimated, the average wait time and in-station time for those passengers who exit the queuing system, i.e. board a train, in the time period of interest can be calculated.

The second method uses as input the expected number of times a passenger is left behind to estimate the wait time. This quantity can be estimated from the PIAM results as in Chapter 3. It can also be estimated using the maximum likelihood and Bayesian inference methods developed in Chapter 5 and reported in Zhu et al. (2017a).

However, the expected number of times a passenger is left behind is usually estimated as a function of passenger’s arrival time on the platform. In order to estimate the average wait time aggregated according to the boarding time, i.e. the exit time from the queuing system, the average wait time aggregated according to the arrival time on the platform should be estimated first for all the time intervals.

Let us assume that the average wait time is estimated over small intervals relative to the interval of interest \( t_2 - t_1 \) (which is actually the case for the PIAM output). Then the cumulative departure function \( D(t) \) can be approximated by shifting \( A(t) \) by \( W_q(t) \), the average wait time for passengers arriving during a short interval centered around \( t \). With function \( D(t) \) approximated, the wait time \( W_q \) for passengers exiting the system between \( t_1 \) and \( t_2 \) can be approximated.

The arrival rate, \( \lambda \), is a common input for both methods and can be calculated from available data, as discussed in the following section.

### 8.2.1 Arrival Rate Estimation

The arrival rate, \( \lambda \), consists of the arrival rate for passengers who tap-in at the station, \( \lambda^{NT} \), and the transfer rate, \( \lambda^T \). \( \lambda^{NT} \) can be estimated from AFC data directly by
counting the number of passengers tapping-in in the corresponding time interval. The
transfer rate, $\lambda^T$, can be estimated from the exit OD matrix, route shares, and the
tavel times between OD pairs. The exit OD matrix is calculated by aggregating AFC
transactions based on their exit times at the destination. In Figure 8-3, for example,
Station 2 is a transfer station where we need to estimate the arrival rate of transfer
passengers. Let $N_h$ denote the exit OD demand, i.e. the demand from Station 1 to
Station 3, in time period $h$ at the destination, and $p$ the route share for the route
transferring at the subject station, i.e. Station 2. The number of transfer passengers
at Station 2 who tapped-out at Station 3 in time period $h$ is $N_hp$.

![Figure 8-3: Calculation of arrival rate at Station 2 example](image)

For a passenger tapping-out at $t_{i\text{out}}$ at Station 3 and transferring at Station 2,
he/she transferred at Station 2 at $t_{i\text{out}} - \Delta ST$, where $\Delta ST$ is the time from the
arrival time at the subject station, i.e. Station 2, to the tap-out time. Considering
Figure 8-4 for example, assuming that $15 \leq \Delta ST \leq 30$, passengers transferring at
Station 2 in time period 18:00-18:15 exited Station 3 in time periods 18:15-18:30
and 18:30-18:45. Therefore, the exit OD flows at Station 3 in the two time periods
contribute to the transfers at Station 2 in time period 18:00-18:15. The fraction of
passengers (transferring at Station 2 in time period 18:00-18:15) can be approximated
by $\frac{30-\Delta ST}{15}$ in time period 18:15-18:30, and $\frac{\Delta ST-15}{15}$ in time period 18:30-18:45.
In general, in time period \( h \) (in 15 min intervals), the total number of transfers contributing to this OD pair at the subject station, \( \lambda^T_{h,OD} \), is given by:

\[
\lambda^T_{h,OD} = N_{h+\lfloor \Delta ST / 15 \rfloor} + \frac{15 - \Delta ST \mod 15}{15} + N_{h+\lfloor \Delta ST / 15 \rfloor + 1} \frac{\Delta ST \mod 15}{15}
\]  \hspace{1cm} (8.3)

The time \( \Delta ST \) can be calculated from AFC data as follows:

\[
\Delta ST = t^t + t^w + \Delta TT
\]  \hspace{1cm} (8.4)

where \( t^t \) is the transfer time at the subject station, \( t^w \) the wait time, \( \Delta TT \) the time from boarding at the subject station to the tap-out time.

The total journey time \( JT \) for passengers tapping-in at the subject station and tapping-out at the same destination is known from the AFC data (e.g. in Figure 8-3, the total travel time for passengers tapping-in at Station 2 and tapping-out at Station 3). For those passengers, \( JT \) is given by:

\[
JT = t^a + t^w + \Delta TT
\]  \hspace{1cm} (8.5)
where $t^a$ is the access time, $t^w$ the wait time, and $\Delta TT$ the time from boarding to tapping-out at the destination, as defined above.

Assuming that average wait time for transfer and non-transfer passengers are the same, combining eqs. (8.4) and (8.5), the shifting time can be calculated as:

$$\Delta ST = JT - t^a + t^t$$  (8.6)

$t^a$ and $t^t$ can be approximated by the expected access and transfer times (using the walk speed distribution and the corresponding walk distances). Therefore, the shifting time for each path of the OD pair can be calculated using AFC data. By shifting the demand of the path by the corresponding shifting time, $\Delta ST$, the transfer rate at the subject station can be estimated based on eq. (8.3).

### 8.2.2 Method I: Based on Wait Time Estimation

PIAM as in Chapter 3 provides as output the in-station time, platform wait time, and expected boarding time for individual non-transfer trips. Hence, the average platform wait time, $W^T_{q}^{NT}$, and the in-station time, $W^{NT}_{q}$, for non-transfer passengers who (are expected to) board a train in the time period of interest can be estimated directly from the PIAM results. The total in-station time for transfer trips is the sum of the platform wait time and transfer time. Under the assumption that the average platform wait time for transfer trips is the same as for non-transfer trips, $W^T_q = W^{NT}_q$, the total in-station time for transfer trips is given by:

$$W^{T}_{qs} = \mathbb{E}(\tau_t) + W^T_q = \mathbb{E}(\tau_t) + W^{NT}_{q}$$  (8.7)

where, $\mathbb{E}(\tau_t)$ is the expected transfer time.

The average in-station time for non-transfer passengers, $W^{NT}_{qs}$, is estimated directly by PIAM as in Chapter 3.

Assuming that $W^T_q = W^{NT}_q$, the total number of passengers in the station, $L_{qs}$,
and on the platform, $L_q$, can be calculated using Little’s formula:

$$L_{qs} = \lambda^T W_{qs} + \lambda^{NT} W_{qs}^{NT}$$

$$= \lambda^T (\mathbb{E}(\tau_i) + W_q^{NT}) + \lambda^{NT} W_{qs}^{NT}$$

(8.8)

and

$$L_q = \lambda^T W_q^T + \lambda^{NT} W_q^{NT} = (\lambda^T + \lambda^{NT}) W_q^{NT}$$

(8.9)

Eqs. (8.8) and (8.9) allow the calculation of the station/platform crowding based on the PIAM output.

### 8.2.3 Method II: Based on Left Behind Estimation

Method II requires less information (compared to Method I), namely the left behind distribution at the subject station (estimated either from direct observations, from the PIAM output as in Chapter 3, or from the left behind model as in Chapter 5). The left behind distribution is usually estimated as a function of a passenger’s arrival time on the platform. Therefore, the average platform wait time and the in-station time estimated using the left behind distribution can be calculated (aggregated by the arrival time on the platform) using the method discussed below.

The platform wait time can be inferred based on the average number of times a passenger is left behind and the distribution of train headways.

$$W_q = W_0 + W_n$$

(8.10)

where, $W_0$ is the wait time until the arrival of the first train after the passenger arrives at the platform and $W_n$ is the “extra” time a passenger would wait if left behind $n$ times. $n$ is a random variable with a probability density function (estimated for example, as discussed in Chapter 5).

$W_0$ can be estimated from the mean, $\mathbb{E}(h)$, and variance, $Var(h)$, of the headway distribution in the period of interest, assuming that a passenger’s arrival is a random
incident (Larson and Odoni, 1981).

\[
W_0 = \frac{\mathbb{E}(h)}{2} + \frac{Var(h)}{2\mathbb{E}(h)}
\]  

(8.11)

where \(h\) is the departure headway (a random variable).

Assuming \(h\) and \(n\) are independent, \(W_n = \mathbb{E}(n)\mathbb{E}(h)\), where \(\mathbb{E}(n)\) is the expected number of times a passenger is left behind. The total wait time is then given by:

\[
W_q = W_0 + W_n = \frac{\mathbb{E}(h)}{2} + \frac{Var(h)}{2\mathbb{E}(h)} + \mathbb{E}(n)\mathbb{E}(h)
\]  

(8.12)

The average number of times, \(\mathbb{E}(n)\), a passenger is left behind during a period is known (for example from the left behind model as in Chapter 5). Transfer and non-transfer passengers are assumed to have the same average wait time on the platform.

The in-station time for non-transfer passengers is the sum of access time and wait time. Based on the walk speed distribution, the average access time \(\mathbb{E}(\tau_a)\) can be calculated as a function of the access distance. The total in-station time for non-transfer passengers is:

\[
W_{qs}^{NT} = \mathbb{E}(\tau_a) + W_q = \mathbb{E}(\tau_a) + \frac{\mathbb{E}(h)}{2} + \frac{Var(h)}{2\mathbb{E}(h)} + \mathbb{E}(n)\mathbb{E}(h)
\]  

(8.13)

Similarly, the in-station time for transfer passengers can be calculated as:

\[
W_{qs}^T = \mathbb{E}(\tau_t) + W_q = \mathbb{E}(\tau_t) + \frac{\mathbb{E}(h)}{2} + \frac{Var(h)}{2\mathbb{E}(h)} + \mathbb{E}(n)\mathbb{E}(h)
\]  

(8.14)

As mentioned above, the average wait time/in-station time is aggregated according to the arrival time on the platform. In order to estimate the average wait time/in-station time aggregated according to the boarding time, i.e. the exit time from the queuing system, the cumulative departure function \(D(t)\) is approximated by shifting \(A(t)\) by the corresponding estimated average wait time/in-station time during the time interval centered around \(t\), as discussed in the beginning of Section 8.2.
Applying Little’s formula, the total number of passengers in the station and on the platform is approximated by:

\[
L_{qs} = \lambda^T W_{qs}^T + \lambda^{NT} W_{qs}^{NT}
\]

\[
= \lambda^T (E(\tau_t) + W_q) + \lambda^{NT} (E(\tau_a) + W_q)
\]

\[
= \lambda^T E(\tau_t) + \lambda^{NT} E(\tau_a) + \lambda \left( \frac{E(h)}{2} + \frac{Var(h)}{2E(h)} + E(n)E(h) \right)
\]

(8.15)

and

\[
L_q = \lambda W_q = \lambda \left[ \frac{E(h)}{2} + \frac{Var(h)}{2E(h)} + E(n)E(h) \right]
\]

(8.16)

### 8.2.4 Discussion

Crowding in stations and on platforms can be calculated using a more aggregate approach with eqs. (8.8), (8.9), (8.15) and (8.16). The first method, based on the estimation of wait time from PIAM as in Chapter 3, can be used to infer crowding levels at a relatively high level of granularity and is more accurate than the second method since the wait time is estimated at the individual passenger level. This method is computationally more demanding than the second method but capable of capturing the changes of the crowding level within a short period of time (e.g. 5 min). The second method is easier to apply, as it uses as input the headway distribution and the expected number of times a passenger is left behind, which can be obtained from direct observations, from the PIAM output as in Chapter 3, or from the left behind model as in Chapter 5.

### 8.3 Validation

Because there is no data about station crowding, an in-station simulation model is used to validate the two methods. The simulation has as input the OD demand and train tracking data.

The model is station based and simulates passenger and train movements (e.g. departure and arrival) at a station. Three types of passengers are included: through...
passengers, tap-in passengers, and transfer passengers (see Figure 8-5). Through passengers pass through the subject station without alighting the train. They utilize the train capacity but are not considered in the station crowding.

![Passenger flow through a transfer station](image)

**Figure 8-5:** Passenger flow through a transfer station

To facilitate the analysis, the simulation method includes the terminal station hence, the number of through passengers at the transfer station is easy to calculate. Passengers arrive at the transfer station according to a Poisson process based on the corresponding arrival rates which are calculated using the method proposed in Section 8.2.1 using AFC data. The walk speed for transfer/tap-in passengers follows a log-normal distribution, with mean 1.12 m/s and standard deviation 0.36 m/s. Their arrival time at the platform is calculated based on their walk speeds and walk distances measured from station layouts. Based on their arrival time at the platform, transfer/tap-in passengers are loaded onto the trains with available capacity on a FCFS basis.

Trains arrive and depart the station based on the actual train movement data with fixed capacity (per train). If the capacity is reached, transfer/tap-in passengers will be left behind. The output of the model includes numbers of passengers in the station and on the platform.
8.3.1 Results

Figure 8-6a shows the crowding estimation results using method I aggregated in 5 min intervals, and Figure 8-6b using method II in 15 min intervals. The blue lines show the number of passengers in the station and the red lines the number of passengers on the platform. The estimation results are consistent with the observations from the simulation model.

8.4 Application

The two methods are applied using actual data to estimate crowding levels at Station 2, one of the busiest transfer stations in the network of interest. The distribution of the number of times a passenger is left behind and in-station/platform wait time are known from the PIAM results. The arrival rate is estimated using the OD matrix, route shares, and the expected travel times between OD pairs as described in Section 8.2.1.

Figure 8-7 shows the distribution of train departure headways at Station 2 by 15 min intervals. The red central mark represents the median value, and the whiskers extend to max and min values. The train frequency is doubled between 17:00 and 20:00. The headway distribution and the distribution of the number of times a passenger is left behind are used to estimate the wait time in eqs. (8.15) and (8.16). Figure 8-8 shows the distribution of the number of times a passenger is left behind and the platform wait time estimated by PIAM, which are important inputs to the crowding model. At the peak of the peak (between 18:45 and 19:00), passengers are left behind multiple times and their wait time increases. Figure 8-9 shows the estimation results. $Station_w$ and $Platform_w$ correspond to the results using method I in 5 min intervals. $Station_h$ and $Platform_h$ correspond to the results using method II aggregated in 15 min intervals. The number of passengers in the station and on the platform are represented by the solid and dashed lines respectively. At Station 2, the peak of the peak occurs between 18:30 and 18:45, which is consistent with actual observations. With larger demand and longer wait times, the in-station crowding
(a) Method I: based on wait time estimation

(b) Method II: based on left behind estimation

Figure 8-6: Station crowding estimation
Figure 8-7: Headway distribution

Figure 8-8: Number of times passengers are left behind at Station 2 and their wait time
level increases dramatically during the peak of the peak.

Figure 8-9: Crowding level at Station 2

8.5 Summary

This chapter proposed two methods to estimate station/platform crowding levels based on the outputs of PIAM. Incorporating the total arrival rate of passengers at a station and the PIAM outputs in terms of the distribution of the number of times a passenger is left behind or their wait time, crowding levels can be inferred. The methods can be used for crowding analysis at any station in the system, with or without transfers. The estimation results are similar to the results from the simulation model.
Chapter 9

Conclusion

This dissertation addresses the important problem of evaluating the performance of subway systems operating near capacity, especially from the passenger’s point of view. Near capacity operations result in crowded conditions both in stations and on trains, deteriorating level of service significantly. Therefore, operators are concerned about developing effective strategies to deal with this problem and also providing crowding information to passengers.

The methodology presented in this dissertation aims at developing a Passenger-to-Itinerary Assignment Model (PIAM) to identify the boarding train(s) of each passenger using Automatic Fare Collection (AFC) and Automatic Vehicle Location (AVL) data. PIAM is considered a key building block in assisting operators to deal with near capacity issues.

The model is applicable to general networks that require tap-in and tap-out transactions. It provides valuable information at both disaggregate and aggregate levels. A series of applications based on results from the PIAM are developed to examine the capacity utilization of the network, including:

- Service quality metrics from the passenger’s point of view, such as the probability of being left behind (denied boarding) at key stations, time spent at stations, etc.,

- Individual train loads and crowding hot-spots (on rail lines and at stations),
• Customer communications with better informed route suggestions, crowding levels, etc.

Such applications are useful to address crowding problems many systems, operating near capacity, face.

Section 9.1 summarizes the research findings and discusses the limitations of the methodology. Section 9.2 outlines the specific contributions of this research. Section 9.3 suggests future research directions.

9.1 Summary

The main elements of the methodology developed in this dissertation are:

9.1.1 PIAM for Trips without Transfers

The PIAM for trips without transfers uses the distributions of egress/access times to estimate the probability of boarding different trains based on the feasible itineraries (the set of trains that a passenger could possibly have boarded). The left behind probabilities for each passenger can also be estimated. The method is applied in a case study with data from a major, congested, subway system during peak hours. Based on actual AFC and train tracking data, synthetic data was generated to validate the model. The results, both, in terms of the trains passengers are assigned to and train loads, are close to the “true” observations from the synthetic data. The probability of a passenger being left behind (due to capacity constraints) in the actual system is also estimated by time of day and compared with survey data collected by the agency at the same station. The left behind probabilities can be accurately estimated from the PIAM results. The train load for individual trains can also be estimated. The results show that the individual train load is very sensitive to the headway.
9.1.2 General PIAM

A two-step general Passenger-to-Itinerary Assignment Model (PIAM) is proposed that is applicable to trips with or without transfers under capacity constraints. The model consists of two modules, the left behind model and the assignment model. At the aggregate level, the left behind model groups passengers based on their expected arrival time at the station platform and estimates the probability of being left behind by station and time interval. At the disaggregate level, the assignment model assigns each passenger to their feasible itineraries based on the corresponding probability of being left behind, and the access/egress/transfer time distributions. For trips with and without transfers, the model provides a complete inference of a passenger’s movements at a high resolution. Aggregate information, such as train load estimation, journey time decomposition, etc. provides useful performance metrics to assess the capacity utilization and evaluate the impact of near capacity operations on passengers.

The model compares favorably with an earlier model proposed by Hörcher et al. (2017) which was based on the PIAM without transfers. The results show that the probability of assigning a passenger to the correct itinerary and the estimated journey time components are more accurate than the Hörcher et al. model.

9.1.3 PIAM for Trips with Route Choice

The PIAM framework was extended to infer the route choice fractions based on AFC and AVL data. While previous literature proposed models to estimate route choice fractions by formulating the travel time distribution as a mixture of different distributions (each corresponding to a single route) and using mixture models to estimate the route choice fractions (Fu et al., 2014; Lee and Sohn, 2015; Sun and Schonfeld, 2015), the applicability of those models is limited during the peak hours due to the effects of congestion.

In this research, we propose a data-driven approach that uses information about train departure/arrival times, left behind probabilities, and access/transfer/egress time distributions to estimate route choice fractions. The method was validated with
synthetic data and the estimated route choice fractions were similar to the “true” observations. The model was also applied with actual data to infer the proportion of passengers choosing different transfer stations by time of day.

Subsequently, the estimated route choice fractions and left behind probabilities are used in the assignment of trips with transfers and route choice. The model was applied with synthetic data. The results show that the probability of assigning passengers to the correct route and itinerary is high.

9.1.4 Crowding Model

Two methods are developed to estimate time dependent crowding levels at stations and on platforms, both, with and without transfers that are easier to apply compared to the full PIAM. The first method has as input the journey time decomposition results from the PIAM (Chapter 3) and uses the passenger’s average wait time and access time to estimate crowding. The second method uses as input the expected number of times a passenger is left behind. This quantity is estimated from the PIAM results (Chapter 3). It can also be estimated using methods such as maximum likelihood and Bayesian inference proposed in Chapter 5.

The first method can be used to infer the crowding level at a relatively high level of granularity and is more accurate since the wait time is estimated at the individual passenger level. This method is computationally more demanding but capable of capturing the changes of the crowding level within a short period of time (e.g. 5 min). The second method is easier to apply, as it uses as input the expected number of times a passenger is left behind and the headway distribution. An in-station simulation model is used to validate the two methods. The estimation results are consistent with the simulation for both methods.

The crowding models can be used as a convenient tool for the operator to monitor system operations and identify crowding hot-spots and potential safety issues at stations.
9.1.5 Access/Egress Time Model

The access/egress time distributions are important inputs to PIAM. The dissertation proposes a method to estimate them based on AFC and AVL data without manual data collection. The egress time of the subset of passengers with only a single feasible itinerary is used for the estimation.

A general walk speed model is proposed to explicitly consider different factors that affect passengers’ walk speeds at stations (including individual characteristics, station configuration, degree of crowding, etc.).

Since there is heterogeneity among passengers in terms of their familiarity with the system, the walk distances at stations may vary from person to person. To capture this behavior, a distance model is introduced that captures the passengers’ access and egress path choices, for example, a fraction of the passengers may “optimize” their walk paths.

To estimate the walk speed distribution, the egress times of passengers with a single feasible itinerary are used. The estimation is properly formulated as a maximum likelihood problem that takes into account the fact that the sample is truncated (since the egress times of the passengers in this sample have to be less than the corresponding headway). The estimation results are similar to actual observations (both from synthetic data and manual observations of the same station).

9.2 Contributions

The passenger assignment problem is becoming increasingly important. In light of near capacity operations, the work in this dissertation provides useful performance metrics to assess capacity utilization and evaluate the impact on passengers. It can be used to monitor performance and also provide information to passengers. The main contribution of the research presented in this dissertation to address the above problems are:
9.2.1 The PIAM Framework

- *The proposed framework rigorously addresses the passenger assignment problem in general networks.*

The Passenger-to-Itinerary Assignment Model (PIAM) is probabilistic and infers detailed passenger movements. The two-step approach effectively reduces the dimensionality and complexity of the assignment problem, especially for transfer trips and trips with route choice.

9.2.2 Individual Modules

While PIAM solves the assignment problem satisfactorily, several modules of PIAM (such as the left behind model, route choice model, etc.) can also be used on their own to address important problems:

- *A method to estimate the left behind probabilities by station and time period.*

While aggregate metrics, such as left behind probabilities, can be calculated from the output of the assignment model, the process can be time consuming to apply. Recognizing this difficulty, we propose a left behind model to estimate the probability of passengers being left behind for a given time period at the aggregate level using automated data.

- *A method to estimate the proportion of passengers choosing different routes for a given OD pair.*

Using information about left behind probabilities, access/transfer/egress time distributions, and incorporating the train tracking data, the route choice fractions for OD pairs with alternative paths can be inferred by time of day. Compared with models that use only travel times to infer route shares, our approach is able to capture crowding effects and therefore is more accurate.

- *Efficient methods to calculate crowding levels.*

Station crowding can be calculated from the PIAM results by applying it to all OD pairs. However, this approach can be computationally expensive. The
crowding model proposed in this research has significant computational advantages as it only requires the application of PIAM to limited OD pairs. Based on the estimation of wait time and left behind probabilities using data from non-transfer trips, the results can be used to infer station crowding for both transfer and non-transfer stations. The crowding model is easier and much efficient to apply and serves as a convenient tool for the operator to monitor system operations.

- *A method to estimate the walk speed/distance distributions of passengers in stations.*

We propose a walk speed model and a distance model to estimate access and egress time distributions. The method uses AVL and AFC data to estimate the various parameters, recognizing the truncated nature of the data. The method provides good estimates of the speed distribution parameters for the whole population. The results show that the method corrects for biases due to the truncated nature of the sample.

### 9.3 Future Work

#### 9.3.1 Methodological Enhancements

The route choice model is validated using synthetic data in Chapter 6 for an OD pair with two paths. Further analysis of the route choice estimation is required with more comprehensive case studies to test the model’s ability to capture the route choice fractions, especially in complex networks with OD pairs with several paths. Aggregation of origins (destinations), represented as hyper nodes, can facilitate the estimation of the route choice fractions, especially for OD pairs with limited numbers of trips.

The definition of the set of feasible itineraries can be improved based on further study of passenger behavior at stations. In this dissertation, the conservative assumption of zero access/egress time was used to retain all feasible itineraries. However,
especially for large stations, the minimum access/egress times can be much larger than zero. Hence, they can be used to eliminate itineraries from the feasible set and further improve the assignment accuracy.

For the application presented in this dissertation, simplified walk speed/distance models are used. Using automated data, more detailed analysis can be conducted to estimate and assess the effects of different factors, such as the individual characteristics, and station characteristics. The underlying walk distances can also be modeled more accurately based on assumptions about passenger behavior inside stations.

New sources of data, such as train load data at the car level from pressure sensors, can provide additional information for both validation and model enhancement. However, this data has serious limitations. The data can be noisy and hence the inference of the total car weight may be erroneous; in addition, the inference of the number of passengers from the total weight depends on assumptions regarding the distribution of individual passenger weight as well as carry-on luggage. Therefore, the translation of the sensor readings to passenger counts will require careful calibration of key parameters. A study by Frumin (2010), with data from TfL’s London Overground system, provides insight into the issues involved.

### 9.3.2 Applications

A number of other applications can be developed based on the PIAM’s output, for example:

- Information dashboard to convey information on crowding levels and left behind probabilities to help operators monitor station crowding.

- Journey time decomposition (access, egress, wait and in-vehicle times), which can be used by operators to assess service reliability from the customer’s point of view.

- Advanced customer information, such as expected crowding and route choice recommendations integrated with journey planners.
• Application to open systems, i.e., systems without fare transactions at the exit gate, could be developed based on similar structure using access/transfer time distributions and train tracking data.

• Finally, real time application where passengers are provided with information about crowding conditions is a related area. The problem is complex. The concepts developed in this dissertation can provide useful insights on how the real time problem can be approached.
REFERENCES


MTR Corporation (2016). *MTR Patronage Figures*.


