Data-Driven Customer Segmentation and Personalized Information Provision in Public Transit
by Abhishek Arunasis Basu
B.Tech., Civil Engineering
Indian Institute of Technology Madras (2015)
Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of Master of Science in Transportation at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2018
© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

Author ............... 
Department of Civil and Environmental Engineering

Signature redacted 
May 17, 2018

Certified by .........
Jinhua Zhao
Edward H. and Joyce Linde Associate Professor of Transportation and City Planning

Signature redacted
Thesis Supervisor

Certified by ........
Haris N. Koutsopoulos
Professor of Civil and Environmental Engineering
Northeastern University
Thesis Supervisor

Certified by ..........
Rabi G. Mishalani
Visiting Professor, Department of Urban Studies and Planning

Signature redacted
Thesis Supervisor

Accepted by ........
Jesse Kroll
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee
Data-Driven Customer Segmentation and Personalized Information Provision in Public Transit

by

Abhishek Arunasis Basu

Submitted to the Department of Civil and Environmental Engineering on May 17, 2018, in partial fulfillment of the requirements for the degree of Master of Science in Transportation

Abstract

To ensure customer satisfaction, a transit agency must strive to understand and cater to its users’ needs. The goal of this research is to develop a framework that could help the transit agency to better understand its users and their behaviors. Segmentation of the market for transit users is the first step, since it allows for the understanding of heterogeneity in their characteristics and their varying requirements, at a granular level as opposed to an aggregate one. In this study, we create a framework, which uses smart card data, to identify customer segments.

The framework developed in this study includes a segmentation scheme that creates segments based on the spatial and temporal characteristics of the travel behavior of customers. Data from Hong Kong’s MTR system were used to demonstrate the practical application of the developed segmentation methodology. In doing so, a thorough analysis was conducted to interpret the specifics of the identified segments.

The segmentation scheme created in this study is capable of catering to meaningful applications that could serve both the agency and the users of the transit system. A few applications explored in the context of this study include the use of the customer segmentation framework for the provision of personalized information. It was demonstrated how targeted information could be provided to users who may likely be affected by a particular service disruption event. In addition, the segmentation framework was used to understand the impact of changes in the network, through a before-and-after analysis where the impact on customer travel patterns due to the provision of service on the newly opened South Island Line is adopted as a case study. Lastly, a predictive transit smart card attrition model was developed by using the features created for the purpose of segmentation. The framework for segmentation developed in this study was found to be useful for multiple applications. Furthermore, the framework is flexible and, therefore, could be generalized for use to address other applications and across other agencies.

Thesis Supervisor: Jinhua Zhao
Title: Edward H. and Joyce Linde Associate Professor of Transportation and City Planning

Thesis Supervisor: Haris N. Koutsopoulos
Title: Professor of Civil and Environmental Engineering
Northeastern University

Thesis Supervisor: Rabi G. Mishalani
Title: Visiting Professor, Department of Urban Studies and Planning
Acknowledgments

I wish to express my gratitude towards my advisors, Prof. Jinhua Zhao, Prof. Haris Koutsopoulos and Prof. Rabi Mishalani for their guidance, support and encouragement throughout this research. I am eternally grateful to you for your time, and I cannot express, in words, how much you’ve contributed to my growth.

I would also like to thank the staff at MTR, especially Pandora Leung, Felix Ng and Henry Lo for providing every possible support during this research, and ensuring our visits to Hong Kong were comfortable and fruitful.

I’d like to thank my friends on the MTR project, Kenji, Tianyou and Mike, for our interesting discussions. I’d like to thank Gabriel and Zhan for their generous support. I’d also like to thank all my colleagues in the Transit Lab, Andy, Eli, Leo, Yiwen, Jian; and my friends from IIT.

Thank you mom and dad, for your support. This thesis would not have been possible without you.
# Contents

1 Introduction 17
   1.1 Background and Motivation ........................................ 17
   1.2 Research objectives .............................................. 20
   1.3 Thesis organization .............................................. 21

2 Literature review 23
   2.1 Customer segmentation ............................................ 24
      2.1.1 Market segmentation in public transit .................... 24
   2.2 Information provision ............................................ 28
   2.3 Customer attrition ............................................... 31
   2.4 Lessons learnt ................................................ 34

3 Methodology 37
   3.1 Introduction .................................................. 37
      3.1.1 Characteristics of applications ......................... 38
      3.1.2 Description of overall approach ......................... 39
      3.1.3 Structure of segmentation ................................ 40
   3.2 Short-term segments ........................................... 41
      3.2.1 Features for short-term segmentation ................... 42
      3.2.2 Creation of short-term segments ......................... 43
      3.2.3 Temporal stability of short-term segments .............. 45
   3.3 Long-term segments ........................................... 47
      3.3.1 Features for long-term segmentation .................... 47
3.3.2 Creation of long-term segments ......................................... 52
3.3.3 Temporal stability of long-term segments .............................. 52
3.3.4 Segments based on similar spatial spread ............................. 52

4 Customer segmentation: results and analysis ............................ 57
4.1 MTR overview ................................................................. 57
4.2 Data used in the study ....................................................... 58
   4.2.1 Data used for creating short-term segments ......................... 58
   4.2.2 Data used for creating long-term segments ......................... 58
4.3 Short-term segments .......................................................... 59
   4.3.1 Correlation among features and PCA ................................. 59
   4.3.2 Determining the optimal number of clusters ....................... 61
   4.3.3 Temporal segments .................................................... 63
   4.3.4 Spatial segments ...................................................... 70
   4.3.5 Temporal stability of short-term segments ......................... 79
4.4 Long-term segments .......................................................... 80
   4.4.1 Correlation among features and PCA ................................. 80
   4.4.2 Determining the optimal number of clusters ....................... 83
   4.4.3 Four-segment configuration .......................................... 83
   4.4.4 Eight-segment configuration ........................................ 92
   4.4.5 Temporal stability of long-term segments ......................... 102
   4.4.6 Selection of configuration for further study ...................... 104
   4.4.7 Shifts among segments .............................................. 105
   4.4.8 Home station changes ............................................... 109
   4.4.9 Extending results to other years ................................... 109
   4.4.10 Spatial-spread based segments .................................... 111

5 Personalized information provision ........................................ 117
5.1 Overview ................................................................. 117
5.2 Framework ............................................................... 119
5.3 Targeted information provision during incidents ..................... 122
6 Before-and-after analysis: South Island Line

6.1 Overview of the South Island Line (SIL) ........................................... 135
6.2 Analysis Framework ................................................................. 136
6.3 Insights from short-term segments ............................................. 138
6.4 Insights from long-term segments ............................................. 139
6.5 Insights from home station inference ......................................... 143
6.6 Insights based on short-term features ........................................ 151
6.7 Discussion .................................................................................. 153

7 Smart card attrition

7.1 Attrition in the context of this study ........................................... 155
7.2 Problem formulation ................................................................. 157
7.3 Predicting attrition .................................................................... 160
  7.3.1 Month-specific attrition model ............................................. 161
  7.3.2 Unified attrition model ......................................................... 166
7.4 Potential pitfalls ......................................................................... 167

8 Conclusion

8.1 Summary .................................................................................... 171
8.2 Contributions ............................................................................. 172
8.3 Limitations and avenues for future research .............................. 173

A MTR Network

B Descriptive statistics and density plots of long-term features
List of Figures

3-1 Applications supported by the segmentation scheme .................. 38
3-2 Customers who are assigned each type of segment .................. 41
3-3 Spatial zones defined in this study ..................................... 44

4-1 Correlation among the temporal features ............................. 60
4-2 Cumulative variance explained versus the number of PCA components
 selected (temporal features) .............................................. 60
4-3 Correlation among the spatial features ............................... 61
4-4 Cumulative variance explained versus the number of PCA components
 selected (spatial features) .............................................. 62
4-5 Value of DB index versus number of clusters (temporal segments) .. 62
4-6 Value of DB index versus number of clusters (spatial segments) .. 63
4-7 Plot of TP values for each temporal segment .......................... 70
4-8 Plot of SP values of the five most important stations for each spatial
 segment ................................................................. 79
4-9 Correlation among the long-term features ............................ 82
4-10 Cumulative variance explained versus the number of PCA components
 selected (long-term features) ............................................ 83
4-11 Value of DB index versus number of clusters (long-term segments) .. 84
4-12 Distribution of feature values for the four-segment solution ........ 88
4-13 Number of users in each segment for four-segment configuration ... 89
4-14 Distribution of feature values for the eight-segment solution ........ 97
4-15 Number of users in each segment for eight segment solutions ....... 98
4-16 Average time of completion of first trip of day (and the deviation) for each segment ............................................................. 105
4-17 Number of customers in each long-term segment across months in the year 2015 ................................................................. 106
4-18 Aggregate number of shifts as a percentage of the sample (100,000 users) ................................................................. 107
4-19 Shifts among the long-term segments in the year 2015 .............. 114
4-20 Distribution of users based on the number of changes in the inferred home station ............................................................. 115
4-21 Distribution of users based on the number of unique home stations, in the year 2015 ................................................................. 115
4-22 Number of customers in each long-term segment across months in the year 2016 ................................................................. 116
4-23 Representative medoids for the spatial-spread segments ............. 116

5-1 Screenshots of MTR’s smartphone application (Source: MTR) .......... 120
5-2 The overall personalized information provision system .................. 121
5-3 Distribution of long-term users who were affected due to the incident ................................................................. 129
5-4 Comparison of mean distance of users who return to incident location to users who do not ............................................................. 129
5-5 Dashboard to select users who would be provided targeted information in the event of an incident ................................................................. 132
5-6 Dashboard for customer metrics (segmentation features) ............. 133

6-1 MTR network map with the South Island Line marked .................. 136
6-2 Utilization of stations on the SIL for trips by long-term users .......... 137
6-3 November 2016 temporal short-term segments for long-term users who utilize SIL in March 2017 ................................................................. 140
6-4 November 2016 spatial short-term segments for long-term users who utilize SIL in March 2017 ................................................................. 141
6-5 Long-term segments for long-term users (November 2016 and March 2017) ................................................................. 144
6-6 Shifts (transitions) within long-term segments (November 2016 and March 2017) split based on low or high usage of SIL .................................. 145
6-7 Proportion of users versus number of trips, split by whether their inferred home station lies on SIL .................................................. 146
6-8 Number of customers versus inferred home station of users in November 2016 split by whether their inferred home station lies on SIL in March 2017 (Station codes in Table A.1) .................................................. 147
6-9 Most frequent inferred home stations (November 2016) .................. 147
6-10 Shifts (transitions) within long-term segments (November 2016 and March 2017) split based on whether inferred home station lies on SIL or not .................................................................................. 148
6-11 Average SP values for long-term users (November 2016 and March 2017) 149
6-12 Average TP values for long-term users (November 2016 and March 2017) 150
6-13 Difference in average SP values .................................................. 152
6-14 Difference in average TP values .................................................. 152

7-1 Two long-term user cards (2015) which show reduction in usage in 2016 158
7-2 Number of long-term user cards which attrition in each month .... 160
7-3 Decision tree solution for March 2016 data ................................. 161
7-4 Decision tree solution for April 2016 data ................................. 161
7-5 Decision tree solution for May 2016 data ................................. 162
7-6 Decision tree solution for June 2016 data ................................. 162
7-7 ROC curve for the logit model .................................................. 164
7-8 Cross-validation curve for penalized logistic regression model .... 166

A-1 MTR network map (*Source: MTR, Fall 2017*) ............................. 177

B-1 Density plots of long-term features used in this study ................. 185
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Temporal stability of temporal segments as measured against October 2016 data (base data: May 2016)</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Temporal stability of spatial segments as measured against October 2016 data (base data: May 2016)</td>
<td>81</td>
</tr>
<tr>
<td>4.3</td>
<td>Four-segment configuration: segment-wise average feature values</td>
<td>84</td>
</tr>
<tr>
<td>4.4</td>
<td>Eight-segment configuration: segment-wise average feature values</td>
<td>93</td>
</tr>
<tr>
<td>4.5</td>
<td>Temporal stability of four-segment solution (January to May and January to October)</td>
<td>103</td>
</tr>
<tr>
<td>4.6</td>
<td>Temporal stability of eight-segment solution (January to May and January to October)</td>
<td>103</td>
</tr>
<tr>
<td>4.7</td>
<td>Cross tabulation of membership among the four-segment and eight-segment solutions</td>
<td>104</td>
</tr>
<tr>
<td>4.8</td>
<td>Number of home station changes versus number of unique home stations</td>
<td>110</td>
</tr>
<tr>
<td>4.9</td>
<td>Temporal stability of the eight-segment solution in the year 2016</td>
<td>111</td>
</tr>
<tr>
<td>6.1</td>
<td>Spatial-spread based segment shifts for users whose home station is not on the SIL in the month of March 2017</td>
<td>149</td>
</tr>
<tr>
<td>6.2</td>
<td>Spatial-spread based segment shifts for users whose home station is on the SIL in the month of March 2017</td>
<td>150</td>
</tr>
<tr>
<td>7.1</td>
<td>Results for monthly models</td>
<td>163</td>
</tr>
<tr>
<td>7.2</td>
<td>Logit model estimation results</td>
<td>163</td>
</tr>
<tr>
<td>7.3</td>
<td>Results for unified model</td>
<td>164</td>
</tr>
<tr>
<td>7.4</td>
<td>Coefficient estimates for penalized logistic regression model</td>
<td>165</td>
</tr>
</tbody>
</table>
A.1 MTR station names and codes (*Source: MTR*) . . . . . . . . . . . . . . . . . . . 176

B.1 Descriptive statistics for the features selected for long-term segmentation180
Chapter 1

Introduction

1.1 Background and Motivation

Pervasive adoption of “smart” systems (sensing, monitoring, information and control) has ushered in an era that has witnessed an unprecedented rise in data collection. The public transportation industry is among many industries that have been able to tap into this data boom, with considerable share of data partly gathered through the use transit smart cards, which are well on their way towards becoming the ubiquitous fare payment mechanism (Pelletier et al., 2011) across transit agencies. Apart from serving as a seamless mechanism of fare payment, the rise in the use of smart cards in public transportation has led to the availability of disaggregate data on users’ travel patterns. Given the enormity of such data now available to transit agencies across the globe, and the fact that the amount of data available to agencies is expected to grow in the near future, it is important to question how these data might be utilized for the betterment of transit systems and their customers.

This ‘opportunistic’ data collected from transit smart card systems could help in transforming the practice of public transportation planning, service design and operations. While the data collected may be shallow in terms of not including demographic and socio-economic information about smart card users, when compared to traditional sources of data collected through self-report surveys which are expensive, labor intensive, and time-consuming to administer and as a result of very small sam-
ple sizes is cross-sectional in nature, they provide the benefit of being available for a large proportion of travelers and over a continuous and long time period. Therefore, as long as the user interacts with fare gates using smart card as their mode of payment, their activity within the transit network would be known to the agency. While self-reporting surveys—whether manually or electronically administered—will continue to have an valuable role in gathering important information about travelers, in the era of automated data, sources such as smart card systems will continue to increase in use as such systems are adopted by agencies and users.

The data collected from smart cards could provide valuable insights into the typical travel patterns of individual users and would allow for aggregation to unravel the dominant travel patterns. These dominant travel patterns, in turn, could be used to identify the customer segments that use the system. As we see in Chapter 2, the use of smart card data towards this goal is backed by past studies which have confirmed that these data could be utilized to identify differences in travel behavior among users of the transit system. More specifically, clustering users based on similar travel behavior to identify user segments could provide useful insights to the transit agency. Studies have also shown that smart card data could be potentially used to support other applications as well. For instance, as is demonstrated in this thesis, insights obtained from smart card data could be used to provide personalized information in the context of transit, study the impacts of changes in service, and investigate changes in usage over time.

This research develops to create a framework to aid transit agencies to understand the different customer groups that use its service. By analyzing the difference in spatial and temporal travel behavior among customers, and by then grouping or clustering together customers that exhibit similar travel behaviors, the transit agency can better understand the characteristics of typical users of the transit system. Through this understanding, the transit agency can gain insight into the specific needs of the customer groups, and can then strive to cater to the needs of these customer groups, thereby paving the way to better customer satisfaction. As elaborated in Chapter 2, research backs the potential of segmentation in identifying customer groups in order
to provide them with targeted services. Moreover, studies also show that such seg-
mentation of customers is feasible by utilizing transit smart card data.

Given the abundance of availability of transit smart card data, and given the fact
that the collection of such data comes from the use of smart cards as fare payment
media, it makes sense to utilize them to create customer segments. While smart
card data could be used to understand the spatial and temporal characteristics of the
travel pattern of users, and distinguish among them, it is necessary to recognize that
the quantity of data available for customers varies. Naturally, there are considerable
differences among infrequent users and frequent users of the service. The segmenta-
tion of customers, therefore, could benefit by additionally distinguishing customers
on the basis of varying levels of data availability pertaining to them. Doing so helps
in enhancing the insights gleaned from the various types of customers that utilize the
service.

After the identification of customer segments, the transit agency can distinguish
among these customer segments to tailor specific services based on perceived needs of
these segments. Moreover, the shifts among these segments, especially in response to
changes such as fare changes and network expansions, could help the agency analyze
the impact of its decisions. Such impacts of changes need not be restricted to solely
service-related factors and could be extrinsic to the system, for instance changes in
characteristics of other modes. Such shifts could provide the agency insights into how
customer segments, grouped together on the basis of similar spatial and temporal
characteristics, behave in response to a change.

Real-time information for transit users is useful and provides multiple benefits. With
the availability of automated sources of data, transit agencies have been able to pro-
vide real-time information to users, often at no additional cost. However, the trend
across industries, shows the necessity to target this information to users who may
need that particular type of information. The knowledge of customer segments that
use the service could be used, therefore, to target this information towards users that
need it the most. Such targeting could be viewed as an effort towards personaliza-
tion and, therefore, customer segmentation could form the basis of such personalized
information provision to customers.

1.2 Research objectives

The primary goal of this research is to develop a framework that could help the transit agency to better understand its users. Segmentation of the market for transit users is the first step, since it allows for the understanding of heterogeneity in their travel characteristics and helps in inferring their varying needs, at a granular level as opposed to system-level aggregate analysis. The framework shall be created by using transit smart card data and should be capable of catering to meaningful applications that could serve both the transit agency and the users of the transit system.

The secondary goal of this research is to utilize the customer segmentation framework for the following,

- Provision of personalized information provision in the context of transit. Doing so includes the following:
  - Creating a mechanism which uses the customer segmentation framework as a building block for provision of targeted information.
  - Exploring its applicability for information provision during service disruptions.

- Developing a procedure for analyzing the heterogeneity in the impact of changes (intrinsic to the transit service or extrinsic) on different customer groups. Doing so includes the following:
  - Using components of the customer segmentation framework to show its applicability for analyzing corresponding ridership changes to obtain insights.
  - Demonstrate the process by using a relevant case study as an illustrative example of the use of framework for in-depth analysis.
Data from MTR’s smart card system, the Octopus card, were utilized in this study. It is important to point out that while data from a specific agency was used to illustrate the efficacy of the framework developed here, the framework developed is highly generalizable to other agencies that use smart card mode of payment, with requisite adjustments which we claim are relatively minor.

1.3 Thesis organization

The rest of this thesis is organized as follows. Chapter 2 explores past research conducted in the field of market segmentation, and more specifically in the domain of public transit. It also discusses literature in the topics of information provision and customer attrition. Chapter 3, covers the key methodological techniques which have been utilized in this study, to identify customer segments among transit users using Hong Kong’s MTR system. Following this, the data utilized in this study and the results of customer segmentation are presented in Chapter 4. The rest of the chapters, barring the last, each look at the applications that the segmentation scheme developed in this study could potentially serve. Chapter 5 looks at personalized information provision to transit users and discusses some novel ways in which targeted information could be provided to users to try and ensure that it is relevant to the user. Chapter 6 discusses how the customer segmentation could aid in the analysis of before-and-after studies and explores a specific case of network expansion as an illustrative example. Chapter 7 describes the creation of a predictive attrition model that could potentially aid the agency to predict smart cards which may exhibit reduction in usage in the near future. Finally, Chapter 8 summarizes the main findings in this study, and discusses some of the limitations and possible future research directions.
Chapter 2

Literature review

Smart cards offer numerous benefits, which could be classified into strategic, tactical and operational (Pelletier et al., 2011). Therefore, these benefits can range from aiding in day-to-day operations, to long-term network planning. Broadly, these smart card systems could be either tap-in only or tap-in and tap-out. For observing and analyzing the travel patterns, the latter system is arguably more convenient since it contains information about the user’s exact destination station and exit times, and therefore doesn’t require inference as pre-step for downstream analysis. MTR’s system is tap-in and tap-out, with most interchange stations, barring a few, connected internally and don’t require tapping out from one station to tap in at the next station during transfers in the journey. While this provides convenience to its users, it doesn’t provide information about the exact route chosen by passengers in cases where multiple routes exist between the origin and destination stations. It should be noted that, smart card data alone lacks information on trip purpose, and other personal attributed of the card user. Such information could be requested from the user on a voluntary basis, and could help to enrich the analysis (Utsunomiya et al., 2006). The rest of this chapter is arranged as follows: section 2.1 discusses customer segmentation and its application in public transit, section 2.2 discusses information provision, and finally section 2.3 discusses past research in customer attrition. Finally, section 2.4 summarizes the key take aways from the literature reviewed, as relevant for this study.
2.1 Customer segmentation

Smith, who introduced the concept of market segmentation, stated “Market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets, in response to differing preferences, attributable to desires of consumers for more precise satisfaction of their wants”. Such partitioning of a market can help in understanding the specialized needs of each homogeneous sub group, so that their needs are catered to specifically (Wedel and Kamakura, 2013). With this definition, we now explore previous attempts at understanding travel behavior of transit users and transit market segmentation in the subsequent subsection.

2.1.1 Market segmentation in public transit

Public transit agencies usually have a set of differentiated fares for certain age groups and socioeconomic classes. For instance, the MTR has special concessionary cards for students, senior citizens, etc. Similarly, by choosing to purchase a fare pass, customers self-select themselves into categories of users who perceive their transit needs to be in line with the benefits offered by that particular fare pass. While such specialized classes help in dividing transit users into broad categories, the users within these categories do not necessarily have a homogeneous spatial and temporal travel pattern. This is understandable since the goals of such classification may not be to identify users with similar mobility needs, although it is possible that the different card categories display some differences, for instance, Nishiuchi et al. (2013) analyzed variations in passenger’s travel patterns (spatial and temporal variability) using a month’s smart card data collected from Kochi City, Japan (bus and tram network), and their analysis revealed that different journey behavior was observed for different card categories. Similar rule based segmentation could also be useful for analysis, for instance, in looking at the variability of passenger’s perception of transit performance, Tyrinopoulos and Antoniou (2008) considered segmenting the population between male and female respondents and found that it provided greater insight in the analysis. A study by Morency et al. (2006), proposed measures that describe the
variability of travel behavior of transit users and used k-means to identify days that were similar in terms of boarding times. It was concluded that smart card data had the ability to reveal regularity aspects of travel behavior. Therefore, such segmentation could potentially be useful for other types of analysis.

Traditionally, transit users have been divided into two broad segments—captive users and choice users. Krizek and El-Geneidy (2007) emphasized that the broad division into captive and choice users was inadequate since it did not consider non-transit riders, and therefore the study used data from two surveys—one for current users, and one for non-users. Principal component factor analysis was utilized to understand how the questions in the survey were related to each other, and k-means cluster analysis was conducted using factor analysis as the basis (“reduced-form data”), and this revealed that there existed eight types of commuters who possessed varying preferences. It suggested that users of the system could be divided into regular and irregular users, and further, transit users could be divided into captive and choice riders, while transit non-users could be divided into auto captives and potential riders. Therefore, in this study, regularity of commuting habits was considered along with mode captivity to gain a deeper insight into the transit market. Jacques et al. (2013) were critical of the prevailing approach citing the ambiguity associated with the terms “captive” and “choice”, and that their conflicting interpretations could lead to negative policy implications. Moreover, they highlighted the fact that such classification focused on availability of choice rather than the practicality or enjoyability of certain modes over others. A large scale survey was conducted to uncover market segments that were applicable to four modes of transportation, which included active modes apart from private auto and public transit. A two-step cluster analysis, using k-means, was performed for each mode, given that there were both continuous and categorical variables, and 21 mode-based clusters were obtained, which were finally combined to form four market segments—true captivity, dedication, utilitarianism and convenience, which was presented as an alternative approach to transport market segmentation. Shiftan et al. (2008) used Utah Transit Authority’s (UTA) household survey and a methodology based on structured equation modeling (SEM) for identifying potential transit
markets based on traveler attitudes. Eight segments of customers were obtained, and
the factors used were sensitivity of time, need for fixed schedule and willingness to
use transit. van Lierop and El-Geneidy (2017) attempted to find market segments
across two Canadian transit providers, by aiming to consider the needs and desires
of customers. By utilizing five years of customer survey data, nine market segments
were obtained across different modes. Principal component factor analysis was uti-
liized to understand how the survey questions were interrelated, and was used as a
method to group questions that explained the variability in the data, and questions
which had insignificant factor loadings were dropped, thereby reducing the number
of variables considered for the cluster analysis. K-means was adopted for clustering
which revealed the different segments within each modal category. They found three
main categories as captive, choice and captive-by-choice users.
Kieu et al. (2015) considered the problem of transit customer segmentation by using
transit smart card data for public transport users in Queensland, Australia. They
identified segments using a priori market segmentation approach, and divided users
into four clusters based on temporal and spatial regularity and used the Density-
Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to mine
travel regularity of each user. The four segments of passengers defined in this study
were—regular OD passengers, habitual time passengers, transit commuters and irreg-
ular passengers. It was found that a majority of the smart card users were irregular
users—those who did not follow any travel pattern, since these users rarely travel
using public transit. Ma et al. (2013) used smart card data for data mining and un-
derstanding travel patterns of transit users in Beijing, China. The transit riders’ trip
chains were constructed followed by employing DBSCAN for detecting transit rider’s
historical travel patterns. Following this, clustering and rough set theory was used to
classify travel pattern regularity. Li et al. (2008) devised a hierarchical graph-based
similarity measurement framework, that ingested GPS log data input to model each
individual’s location history, which was then utilized for measuring similarity among
users. The results were validated through the survey volunteers to rate other users
in terms of similarity (based on their own understanding). Tao et al. (2014) used
smart card data in Brisbane, Australia, to examine the spatial-temporal dynamics of urban PT passenger travel behavior, by extracting bus service patterns, reconstructing travel trajectories and creating bus passenger flow comaps. These flow comaps were found to be effective in revealing the major travel pathways of passengers, and could be utilized to observe the spatial shift in their travel patterns. Ma et al. (2017) aimed to identify transit commuting patterns by using transit smart card data. With a month’s smart card data, the spatio-temporal regularity was measured. The results of the study were validated through a survey deployed on social media, and suggested that the proposed framework could reduce the necessity of longitudinal surveys and travel diaries.

Finding the ideal variables or features which could help describe the spatial and temporal characteristics of the user’s travel pattern is an essential step towards finding segments of users. Morency et al. (2007) discussed measures of spatial and temporal variability of transit use. The indicators of spatial variability included enumeration of bus stops used for boarding and the frequency of use of bus stops. The indicators of temporal variability included using data mining techniques to identify typical boarding patterns and the proportion of zero-boarding days. Ortega-Tong (2013) classified London’s public transport users using smart card data, and the features used to describe travel pattern were broadly grouped into those describing temporal variability, spatial variability, activity pattern variability, sociodemographic characteristics and public transport mode choice. Halvorsen (2015) used MTR’s smart card data to classify users and created features that described the spatial, temporal and frequency characteristics of users. Goulet-Langlois et al. (2016) did not aggregate variables, and instead created a longitudinal representation of user’s activity sequence. Ghaemi et al. (2017) tried to uncover temporal behavior of users in their monthly trips, and used smart card data from Gatineau, Canada for their analysis. They proposed a visual method for analysis of temporal user behavior, which involved reducing the high dimensional data into a lower space (semi-circle projection), and then utilized the agglomerative hierarchical clustering algorithm. The lower dimensional space provided the benefit of being interpretable.
A limitation of utilizing only one source of data—smart card data—is that the user’s activities outside the public transit system are largely unknown. In the present study, since only data pertaining to MTR’s subway system was available, user’s travel on other modes such as buses remained unknown. We postulate that if other sources of data were available at the time of this research, better insight into user’s travel patterns could be revealed. To support the above claim, consider research by Kusakabe and Asakura (2014), who proposed the fusion of person trip survey data with smart card data, which helped in estimating behavioral attribute of trips utilizing smart card data. Validation showed that the trip purpose of a majority of trips were correctly identified using the methodology proposed. The analysis also revealed the relationship between trip purpose and trip frequency, which helped to show how trip purposes affected the total number of trips. Moreover, the authors claimed that their methodology could be used for the estimation of actual origins and destinations, and that the proposed method could be used for continuous monitoring thereby allowing the agency to assume the cause of behavioral changes, especially due to changes in policy/fare revisions, etc.

2.2 Information provision

Apart from advantages in fleet management, the availability of real-time data has led to another major change in public transit, which is the development of real-time transit information systems. The investments in these systems are motivated by the expectation that it would improve passenger’s experience thereby leading to an increase in usage of public transit. Zhang et al. (2008) explored the impact of real-time information on the behavior and psychology of travelers by utilizing a panel survey data. It was found that real-time information was associated with a feeling of safety and an overall level of satisfaction. In a study by Brakewood et al. (2014), the impact of real-time information on bus riders in Tampa, Florida was explored. They found that among the advantages that real-time information provided—it led to significant decrease in level of anxiety and depression, and increase in level of productivity and
safety during daytime. Tang and Thakuriah (2011) found that there was a reciprocal relationship between attitudes and behavioral intentions, and therefore that attitudes may be altered after users experience real-time information provision systems. Initial mechanisms for providing real-time information to transit users consisted of bus stop displays. Although the potential of information provision systems may be limited, nevertheless, it could lead to a better transit experience. Dziekan and Kotthenhoff (2007), looked at the impact of at stop real-time information systems, and devised a framework to evaluate the effects of such information systems. Following this, two evaluation studies were discussed, with one showing that real-time information led to a drop in perceived wait times, and that passengers were found to adjust their walk speed based upon the information presented on the display. Grotenhuis et al. (2007) found that in case of integrated multi modal travel information in public transport, there was a distinction in the desired type of information at each stage of travel. In the pre-trip stage, the desired information pertained to the planning part of journey using public transit, wayside IMTI was desired when it helped the passenger to catch the correct vehicle and on-board travelers were concerned about their connections. Chow et al. (2014) looked at the impact of real-time passenger information signs in rail stations at the MBTA, and found that people reduced their overestimation of wait times by 50%, and that there were minor increases in ridership at stations with real-time information (after controlling for other variables).

By moving on to smartphones for providing this information, apart from the benefits associated with ease of access of information, the agency benefits from the cost savings associated with on-stop signage, however on the flip side this decision may alienate users who do not possess a smartphone, or other devices that are capable of providing adequate interface to receive this information. However, in the context of Hong Kong, most subway stations already have such signage, and around 74% of the population already have access to smartphones, a percentage that is expected to be close to 80% in the next four to six years. Therefore, it is reasonable to expect the agency to shift

---

and concentrate its efforts on ensuring that the smartphone application is effective and caters to their customer’s needs. Watkins et al. (2011) stated that one way to alleviate the reliability concerns associated with public transit (from the user’s perspective) was to provide real-time information to users. They developed a real-time information provision platform called OneBusAway in King’s country metro. This platform could be accessed via website, telephone, text-messaging and smartphone applications. A survey was conducted by the researchers, and the analysis revealed that the availability of real-time information decreased the perceived wait time by about 13%. Moreover, it was also found that real-time information also help reduce the actual wait time. Therefore, a major advantage of mobile real-time information is that it is available to passengers before they reach the stop. In Gooze et al. (2013), the authors investigated the impact of data accuracy in the real-time transit information provision system OneBusAway on rider experience. They conducted a survey, similar to their previous survey three years earlier, but which now included questions to understand the user’s accuracy expectations and their error reporting experience. The median error tolerance was found to be 4 to 6 minutes, however it was found that different population segments displayed varying expectations of prediction accuracy. It was found that there was a strong relationship between ridership frequency and accuracy expectations. In comparison to the previous survey, it was found that a greater number of users reported taking more public transit trips due to the real-time information available to them, and more users relied on the application as their main source of schedule and route information.

Tang and Thakuriah (2012) tried to quantify the ridership effect of the CTA Bus Tracker, which is a real-time bus information system, after controlling for other variables by using repeated observation data (longitudinal). They found that the Bus Tracker service did lead to an increase in bus ridership. Further analysis also showed that there were temporal variations in ridership gains among routes. Therefore, this was an attempt to provide empirical evidence of the ridership gains associated with the availability of real-time information. Carrel et al. (2013) surveyed users of San Francisco’s public transit users, and one of the results of their study was that issues
with real-time information system did not appear to force transit riders to leave, unless these issues occurred more frequently than other service problems and hypothesized that this could be due to the fact that the real-time information system was branded by the system manufacturer as opposed to the transit agency, or it may be due to users becoming accustomed to such errors. Brakewood et al. (2015) looked at the impact of real-time information on bus ridership in New York City. The study claimed that New York City had the advantage of simultaneous launch of real-time information on multiple interfaces, high levels of technology adoption and a marketing campaign to increase awareness about the real-time information system. Results showed that the longest routes experienced a significant increase in ridership. The topic of providing personalized transport information has been looked into before in Lathia et al. (2013), where the authors used fare card data in London, to first identify travel behavior heterogeneity among users, and then computed personalized trip time estimates. Personalization in other domains is widespread such as e-commerce (Linden et al., 2003) and news (Das et al., 2007), and therefore it is reasonable to expect that personalization would impact the way information is provided to public transit users in the future.

2.3 Customer attrition

While most profit-driven private companies focus heavily on customer attrition/churn analysis, such discussion in the domain of public transit is relatively rare. Webb (2010) looked at customer loyalty in the public transportation context. This study used data from CTA’s 2008 customer experience survey, which consisted of a mix of choice and captive riders among survey respondents with varying accessibility, and covered aspects of customer satisfaction and loyalty in its questionnaire. Factor analysis and structured equation modeling (SEM) were utilized to understand the relationships between customer loyalty and its drivers. A major finding in this study, was that loyalty forms differently among different customer groups and that it was necessary to cater to these groups individually to improve their experience in order to increase
their loyalty.

We recognize that attracting and retaining customers are part of the positive cycle that leads to increase in transit ridership, with each component being equally important. While agencies focus more on attracting new customers (Webb, 2010), retaining customers should not be neglected. Through customer segmentation, the broad characteristics of these segments can be identified, and analyzing customer who are leaving the system could help in retaining customers and increasing loyalty, which could in turn have positive spillover effects, both in terms of increasing ridership further (word of mouth advertising) as well as curbing externalities (pollution, congestion, etc.) associated with the use of modes such as private auto. High rate of (avoidable) attrition might also tarnish the image of the service, effectively preventing other users from using it. Although a discussion of causal factors that lead to attrition may be beyond the scope of the present research, given the study’s reliance solely on smart card data, it may be useful to study whether a predictive model could be built such that it could identify users (or smart cards) who may leave the system (or drastically reduce their usage) in the near future.

Therefore, in this regard, lessons from other industries provide an interesting insight into creating predictive models for customer attrition or churn. Unlike the problem of identifying customer segments, the problem of churn prediction falls in the category of supervised machine learning since by using past dataset for training the model, customers could be labeled as churners and non-churners based on their present usage. Au (2013) highlighted the potential of methods such as tree-based methods and neural network based methods (multilayer perceptron) and Bin et al. (2007) adopted decision tree for prediction of churn. Hung et al. (2006) looked at the application of data mining techniques to telecom churn management. It was found that both a decision tree and neural network techniques could perform well and accurately predict churn. The data utilized for building the model included customer demographics, billing information, call detail records, contract/service status and service change log. Segmentation was incorporated as well into the churn prediction, but it was noted that since the churn rate was extremely small, the data size was insufficient (since
there were very few churners in some segments). Lemmens and Croux (2006) used bagging and boosting classification trees to predict churn for a telecom firm. It was noted that bagging and boosting provided substantially better performance than binary logit model. It was also noted that the model could be trained with a “balanced” sample and then the classifier could be corrected for the balanced sampling scheme. Xie et al. (2009) looked at customer churn prediction using improved balanced random forests (IBRF) in the banking industry. The proposed technique (IBRF) had the advantage of penalizing more heavily the misclassification of the minority class, while iteratively learning the best features. The experimental results showed its superiority over random forest algorithms.

A common issue related to churn is the imbalance in the dataset, since the number of customers who leave the system is often a small fraction of the total (churn is a rare event). This may cause common predictive techniques to have a suboptimal performance. One technique in literature that looks at tackling such imbalanced datasets is Synthetic Minority Over-sampling Technique (SMOTE) which shows that a combination of over-sampling the minority class and under-sampling the majority class can lead to better classifier performance in the ROC space, than simply under-sampling the majority class (Chawla et al., 2002). Burez and Van den Poel (2009) discussed handling class imbalance in customer churn prediction. They evaluated the efficacy of advanced sampling techniques such as CUBE, and concluded that it did not increase the predictive performance, and that this finding was in line with similar work done elsewhere. They also concluded that weighted random forests may work well and that it should be compared to logistic regression. It was also found that undersampling worked well, but undersampling such that there were as many churners as non-churners in the training set was not necessary. Moreover, AUC and lift were found to be good evaluation metrics. It also noted that boosting was a robust classifier but never outperformed any other technique.

Based on the above studies, we may infer that information on the user is required to truly predict customer attrition. Although a causal analysis of attrition may not be feasible by solely using transit smart cards without information on the actual user, it
may be possible to predict card attrition. The agency could then try to identify and target these users specifically and try to understand the circumstances underlying attrition, and whether it is avoidable or not. If it is the former, then the agency could strive to obtain better feedback on where it needs to improve and if possible, try to convince the customer that requisite action would be taken swiftly. While direct communication between the transit customer and transit rider has been far-fetched in the past (cost and labor considerations), with improved communication interfaces with the customer (as discussed earlier in the chapter) may make two-way communication a possibility. Webb (2010) also pointed towards the necessity of conducting surveys regularly and consistently in order to obtain time series data which could allow for comparison between the customer’s intent to reuse the service versus their actual usage. While it was recognized that asking customers for their smart card information was necessary for the above step (to match with survey results), it should be pointed out that findings from the previous section on information provision also hint towards the fact that such a step would be beneficial.

2.4 Lessons learnt

In their study, Lai and Chen (2011) explored the behavioral intentions of public transit passengers by utilizing passenger survey data from KMRT, Taiwan. This was deemed necessary since customer loyalty is necessary for long-term financial performance. It was found that passenger loyalty relied on passenger satisfaction, and that while better quality transit services were essential, increased passenger involvement was also significantly important. It was suggested that a more market-oriented approach than a traditional supply oriented approach was necessary, for instance by utilizing advertisements as a pull strategy, customer’s perception of the value offered by transit could be improved, which could lead to better satisfaction. In their study, Kieu et al. (2015) concluded that understanding each passenger type was vital towards transit strategic planning, and that before the transit agency considers attracting new customers, they should first focus on existing customers and encourage them to
adopt public transit as their primary mode. Moreover, it emphasized that transit agencies could ‘pay special interest’ to irregular users who were regular users before, and try to understand the underlying reasons for such shift in travel behavior. It was also stated that the passenger segmentation methodology has helped transit operators to ‘well-suited’ specific information and services.

From the literature surveyed, it was concluded that:

- Understanding the customer’s specific needs is essential, and segmentation is one way to distinguish among customers with varying needs.

- Better understanding could help in providing targeted services to these users which could improve customer loyalty and satisfaction.

- Providing information to these users is one demand side strategy which could be a low cost alternative to ensure a better customer satisfaction.

- By ensuring higher satisfaction, through the above, the agency could ensure that they do not leave the PT system.

With this understanding, we present a novel framework for customer segmentation in Chapter 3 and interpret these segments (Chapter 4). A discussion of various applications using the framework is covered in Chapters 5, 6 and 7.
Chapter 3

Methodology

3.1 Introduction

To ensure a better experience and to meet customer’s needs, a transit agency needs to strive to understand its customers. Traditional tools such as surveys have been used in the past to quantitatively and qualitatively capture a transit customer’s needs. The data-driven customer segmentation methodology being presented in this study provides a low-cost alternative, albeit with some limitations, to traditional tools for understanding the transit customer which often involve manual surveys. Since it relies solely on transit smart-card data for creation of segments, the cost of collecting this data is minimal. Another major advantage of utilizing smart-card data is that it captures the longitudinal aspect of a user’s travel pattern, and these insights could help in identifying shifts in behavior. Moreover, in this study, the process of discovering segments from automated data sources such as smart cards may itself be automated, whereby the segmentation mechanism can ingest new data to generate results which may be interpreted based on the specific features chosen for clustering. The segmentation being presented in this study aims to facilitate various applications which are described in the following subsection.
3.1.1 Characteristics of applications

The customer segmentation presented in this study aims to capture a user’s spatial and temporal travel pattern so that the agency could then provide targeted services based on their specific characteristics (and inferred needs). This study attempts to create a segmentation scheme that could be utilized across several applications. Some specific applications discussed in this study include personalized information provision, before-and-after analysis and predicting card attrition. Although these are some of the applications that would be demonstrated, we believe there are several other applications which could also benefit from the segmentation scheme presented in this study, and these applications are shown in Figure 3-1.

For the segmentation scheme to serve such diverse applications, the structure should be able to capture the details of each user’s specific spatial and temporal travel characteristics. In this study, we only utilized smart card data for user’s trips made using MTR’s service, and no other data source apart from this was used. While a lot of past studies have either looked at extremely disaggregate level or have resorted to scalar aggregation of features, we believe that the balance lies between the two. This
type of structure could benefit multiple applications which could have varying levels of data requirements, and hence the applications driven by this framework could serve users of both types: those who have travel infrequently on the service, and frequent users of the service.

### 3.1.2 Description of overall approach

We develop a customer segmentation scheme that could provide the transit agency an in-depth knowledge of its customers. While creating this, the aim is to:

- Identify customer segments that are representative of typical users who use MTR's service, based on similar spatial and temporal characteristics, which in turn are captured through features defined in this study.

- The customer segments obtained should be interpretable with the help of features used to define these, and should be distinguishable based on the features used to create these.

- The features used for creating these segments should be interpretable themselves and should offer value to be used on their own, or in conjunction with the segments, in meaningful applications.

- The customer segments should be able to facilitate multiple downstream applications as defined in the scope of this research.

- The segments obtained should be stable over a period of time for comparisons to be drawn across time periods. Hence, while individual customers might shift across segments, the segments themselves should not disappear.

After the framework is developed, it would be used for:

- Understanding the customer better by utilizing the insights gained from the segments they belong to.
• Used as a building block for the provision of targeted information with the explicit aim to increase the relevance of the information and the implicit aim to improve customer’s experience.

• Understanding, quantitatively, how incidents might impact different customer groups.

• Observing shifts in customer segments, especially in the event of an extrinsic or intrinsic change.

• Trying to predict smart card attrition.

With the above specification, we now describe the structure of segmentation adopted in this study in the following subsection.

3.1.3 Structure of segmentation

Based on the characteristics of the segmentation considered, a two-tiered segmentation scheme was created in this study. The idea was to capture a given user’s general travel attributes such as the stations and times of their trips in the first tier, and to summarize detailed characteristics of their trips—such as the distance they travel on average, and the number of unique locations they visit, etc.—in the second tier. Therefore the overall segmentation structure could be divided into two broad parts: short-term segments (tier 1) and long-term segments (tier 2). All users who travel in a certain month may be assigned a short-term segment, although to avoid issues arising due to window effects, only users who have traveled at least once in the previous month are assigned a short-term segment. Long-term segments are assigned only to (‘long-term’) users who have traveled for a long period of time (over a year) and are frequent users of the service (use MTR’s service on at least half of the weekdays in a month). In this study, we focus only on weekday travel pattern, and claim that a similar methodology would be applicable for the analysis of weekend travel patterns. The eligibility of allocation of segments to users based on their service usage has been shown in Figure 3-2.
The rationale for the two-tiered structure is that while the agency might possess limited travel data on infrequent users, nevertheless these users could be characterized based on the the locations and the times when they travel (general characteristics). However, when it comes to looking at their detailed characteristics, there might not be enough data on these users to conclusively determine these characteristics (stability problems). Therefore, by splitting the segmentation structure into two tiers, we ensure that each user, regardless of their usage, is assigned a customer segment which is sufficient for at least a few applications. To expand on this further, consider the provision of automated information updates to users in the event of a disruption in service. For this, the general characteristics of the user’s travel pattern may be sufficient to determine if they might be potential beneficiaries of this information. Whereas, for determining if a customer is going to leave the system, more information, in the form of detailed characteristics of their trips might be necessary.

Note: Henceforth, the terms Group and Segment would be used interchangeably to refer to clusters obtained in this study. The cluster name assigned to a certain cluster may also be used to refer to the same.

### 3.2 Short-term segments

As mentioned previously, the purpose of short-term segmentation is to study the general characteristics of travel behavior of a larger fraction of the population that uses
the MTR system. Hence, the features utilized for segmentation and the segmentation structure (tier 1) are modified to suit this goal. Short-term segments classify customers that utilize similar stations and travel during similar times of the day, and therefore can be used for applications that need these general details of travel patterns. For instance, targeting information to specific users who use similar stations at certain times could benefit from the short-term segmentation. Moreover, such a segmentation would allow for the observation of micro-changes in customer's spatio-temporal travel patterns from one month to the other.

3.2.1 Features for short-term segmentation

To measure the general details of the customer’s travel pattern, two features are defined: Temporal Probability (TP) and Spatial Probability (SP). These features are created for each customer, defined over the period of a month, and they help in identifying the average probability of a user’s trip ending in a particular hour and involving a particular spatial zone, respectively. These are defined as follows,

\[ TP(h)_{\text{customer}} = \frac{\text{Number of days when a customer's trip ends in hour } h}{\text{Total number of days the customer travels}} \]  

(3.1)

\[ SP(z)_{\text{customer}} = \frac{\text{Number of days when a customer's trip either begins or terminates in zone } z}{\text{Total number of days the customer travels}} \]  

(3.2)

where \( h \) is the hour when the trip ends and \( z \) is zone of travel. In the definition of TP, \( h \) or hour could take a value between 5 and 25, where 5 is 5 AM and 25 is 1 AM (following day). The specific choice of range of \( h \) would be clear in Chapter 4. For computing TP, we use the mathematical floor value of the tap out time for hour \( h \). For computing SP, the stations in the network are divided into 33 zones in the following way: stations that belong to the same line in a contiguous portion of the line are assigned the same zone, and interchange stations are assigned a separate zone. In some cases, based on knowledge of similar stations, these are assigned to the same zone. The final zonal assignment can be observed in Figure 3-3. Therefore, in the definition of SP, the value of \( z \) (zone) can take a value between 1 and 33. This conversion to zones helps in meaningfully compressing the spatial variables from 90 to 33 before clustering, thereby helping to achieve a concise representation allowing
for quicker computation during clustering. While looking at the results of clustering, we would analyze and plot the SP values for individual stations (rather than zones) for ease of interpretation. The SP value for a station is defined as,

\[
SP(z)_{\text{customer } i} = \frac{\text{Number of days when a customer's trip either begins or terminates at station } s}{\text{Total number of days the customer travels}}
\]  

(3.3)

Therefore, \(TP(h)\) for a particular customer represents the probability of trip ending in hour \(h\). The value of \(TP(h)\) for a particular value of \(h\) may lie between zero and one. Higher the value of \(TP(h)\) for a given user, more certain it is that the user may use MTR’s system such that a tap out is observed in hour \(h\), and therefore it is reasonable to assume that the user travels on the system between hour \(h - 1\) to hour \(h + 1\). Similarly, the value of \(SP(z)\) (or \(SP(s)\)) also lies between zero and one for a particular value of \(z\) (or \(s\)), and a high value indicates that on a day when the user is using MTR, the likelihood of the user accessing zone \(z\) (or station \(s\)) is high. To illustrate an interpretation of SP and TP values, consider as an example a hypothetical customer \(i\). If for customer \(i\), \(SP\text{(Tsuen Wan)} = 0.9\) and \(TP(14) = 0.8\), then this implies that if the user is using the system on a given day, then the likelihood of observing the user in the system during 1 PM to 3 PM is 0.8, based on past behavior of the user, and the likelihood that the user accesses Tsuen Wan station is 0.9. During the clustering process, separate segments are created based on \(TP\) and \(SP\) values to yield temporal segments and spatial segments respectively. While customers that belong to the same temporal segment travel during similar periods of the day, customers that belong to the same spatial segments travel to/from similar set of stations (zones) on the network. Each customer is therefore assigned a spatial segment and a temporal segment.

### 3.2.2 Creation of short-term segments

Since the labels (segments) were unknown, the task of classifying users is unsupervised learning, and in this study the k-means clustering technique has been adopted for the creation of clusters of users with similar characteristics. While supervised algorithms like k-Nearest Neighbors require known labels for the sample, unsupervised
algorithms like k-means create clusters in data based on its internal structure and thus involves unlabeled data (Duda et al., 2012). A short summary of the k-means algorithm (Lloyd, 1957) is as follows,

- The initialization step involves the selection of initial centroids, one for each cluster.
- The second step assigns each sample to its nearest centroid.
- Step three is the cluster centroid update step, whereby new centroids are created by obtaining the mean value of all the samples assigned to a given cluster.
- The second and third steps are repeated until the cluster centers no longer move significantly. Therefore, the k-means algorithm tries to obtain the optimal clustering by minimizing the sum of distances between samples and the centroid.

Figure 3-3: Spatial zones defined in this study
Therefore, the k-means algorithm tries to obtain the optimal clustering by minimizing the sum of distances between samples and the centroid of each cluster. The distance used in this study is the Euclidean distance. The k-means solution could converge to a suboptimal local minimum and this is influenced by the initial centroid selection. Therefore, as a safeguard, during the initialization, centroids were chosen based on k-means++ (Arthur and Vassilvitskii, 2007). A total of 150 iterations were run for each value of k (the number of clusters) to avoid convergence to a suboptimal solution. The value of k was varied and the Davies-Bouldin (DB) index was used as a criterion for selecting the ideal number of clusters (or optimal value of k), with lower the DB index value indicating better separation of the clusters and closeness within the inside the clusters (Davies and Bouldin, 1979).

The features may have correlations which if left unaccounted for could lead to extra weight being levied on some variables. To avoid this, prior to clustering, Principal Component Analysis (PCA) can be used to identify a linear combination of the features that best summarizes the variance. Clustering based on components obtained through PCA therefore, helps to select uncorrelated vectors that represent the features, and also helps in reducing the dimensionality which speeds up computation. The first principal component obtained captures maximum variability in the original data, and the proportion of variance captured by each subsequent component decreases. In this study, components which when combined together explain 90% of the total variance (or up to 0.9 cumulative proportion of variance is explained) are selected, and the original data transformed in the component space is then used for clustering.

### 3.2.3 Temporal stability of short-term segments

The clusters created in this study were verified for temporal stability in light of issues identified in similar studies in the past (Ortega-Tong, 2013). This step is indispensable if these clusters are used to classify users in subsequent months. In
In this study, the segments are created with the aim that results are comparable across months, and therefore the segments obtained need to be generalizable to other time periods, or in other words the segments should be stable across time. Therefore, the cluster temporal stability is an important issue which cannot be overlooked as this would determine if the cluster results obtained in one month could be generalized to data pertaining to other months. In the present study, a methodology similar to Goulet-Langlois et al. (2016) was adopted to examine cluster stability. For measuring temporal stability, a random sample of data from the base period (period 1, training period) was selected and cluster analysis was carried out. Following this, a random sample of data was selected from another period of time (period 2, period with which stability is compared) and cluster analysis was carried out, and let these clusters assigned to each point in the sample be denoted as $C^*$. Now, the data from period 2 was classified with respect to clusters identified in period 1, and for this data sample selected from period 2 was transformed and projected onto the principal components obtained in period 1, and then clusters were assigned to each data point based on shortest Euclidean distance from the nearest centroid (period 1). Let us call the clusters assigned to each point as $C$. Now, each user was assigned one value of $C^*$ and one value of $C$. The temporal stability of the clusters was identified as the proportion of users assigned the same segment under both cluster assignments as described above. Visualized in a tabular format, if the rows and columns represents assignments under $C^*$ and $C$, the higher the sum of the diagonal elements as a proportion of total, greater is the stability of the clusters. For brevity, we call this value as the Correspondence Score and is defined as follows (for a square matrix of $C^*$ and $C$, with elements $a_{ij}$),

$$Correspondence\ Score = \frac{\sum_{i=1}^{n} a_{ii}}{\sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij}} \quad (3.4)$$

A high Correspondence Score implies that the clusters are stable across time periods, and this is highly desirable.
3.3 Long-term segments

The second tier of segmentation consists of looking at the detailed characteristics of travel patterns of frequent users who are long-term customers—defined as users who utilize the system on at least half of the weekdays in each month, and throughout the year. In this section, the features utilized for long-term segmentation have been explained in detail. A separate set of features were created for this purpose, which were distinct from the features defined previously for the creation of short-term segments. These features provide deeper insight into a customer's travel pattern and help to further understand the spatial, temporal and frequency aspects of their travel pattern. For users with less data, many of the features defined here may be less stable, and therefore these segments were created only for long-term users.

3.3.1 Features for long-term segmentation

The features are chosen in a way to capture the spatial, temporal and frequency aspects of the customer's travel pattern. However, the spatial and temporal features here are aggregated. For creation of features, we measure these travel pattern attributes either aggregated to the level of a month or aggregated to the level of a day. For attributes that are measured over the period of a day, we consider the mean value of the feature and standard deviation of the feature over a month, and include the mean and standard deviation in the final set of features used for clustering. For example, if we record the number of trips made by a particular user each day, then the features we consider for clustering are the mean number of trips per day (average trips per day over the month), and the standard deviation of trips per day (over a month). Therefore, the final set of features only consists of attributes defined (aggregated) over the period of a month.

Attributes measured over a month

These attributes are computed for each customer measured over the period of a month. A total of ten features are measured over a month, and these features primarily
capture the frequency and the spatial characteristics of the customer’s travel pattern.
A detailed description of these features follows,

**Frequency**

1. *Active*: Number of weekdays traveled in a month as a proportion of the total
   number of weekdays in that month. This feature can be used to infer a user’s
   dependence on MTR’s service.

2. *Range*: The difference between the first weekday and the last weekday traveled
   as a fraction of total number of days in that month. This feature gives an
   idea of the temporal spread of a customer’s travel over a month. Controlling for
   *Active*, when the range is low, it would imply that a user’s trips are concentrated
   temporally in the month whereas when the range is high, it implies the user’s
   trips are spread temporally, where the unit of time is days.

**Spatial characteristics**

1. *RI*:

   \[
   (RI)_{station \, x, \, customer \, i} = \frac{(\text{Number of entries})_{station \, x, \, customer \, i} + (\text{Number of exits})_{station \, x, \, customer \, i}}{\text{(Total number of entries and exits at all stations)_{customer \, i}}} \quad (3.5)
   \]

   This effectively represents the importance of each station in a customer’s travel
   pattern and therefore describes the distribution of the customer’s travel along
   different stations on the system. *RI* is computed for the period of a month,
   for each customer and for each station. To reduce the final number of features
   selected, only the first and second order statistics, along with the maximum
   value and the cardinality of the set of stations accessed by each customer, are
   used. Therefore, these four features summarize the *RI* for a customer and are
   utilized as features for clustering.

   (a) *µ*<sub>*RI*</sub>: The average value of relative importance for stations.

   (b) *σ*<sub>*RI*</sub>: The deviation in the relative importance for stations.

   (c) max *RI*: The fraction of trips that start or end for the station that a
   particular user visits the most number of times, in a month.
(d) \(n(RI_d)\): The total number of unique stations visited over the period of a month.

2. \(RI_d\): Relative Importance for origin-destination (OD) pairs is defined as follows,

\[
(RI_d)_{OD \ z, \ customer \ i} = \frac{\text{(Number of trips)}_{OD \ z, \ customer \ i}}{\text{(Total number of trips)}_{customer \ i}}
\]  

This effectively represents the prevalence of travel on a particular OD pair in a customer’s travel pattern. Similar to the previous case, to reduce the final number of features selected, only the first and second order statistics, along with the maximum value and the cardinality of the set of unique ODs for each customer are used. Along with \(RI_s\), this provides further information on the spatial variability of the customer’s trips.

(a) \(\mu_{RI_d}\): The average value of relative importance for OD pairs.

(b) \(\sigma_{RI_d}\): The deviation in the relative importance for OD pairs.

(c) \(\max RI_d\): The fraction of trips on the OD pair on which the user travels most number of times.

(d) \(n(\max RI_d)\): The number of distinct OD pairs on which the user travels over the period of a month.

**Attributes measured over a day**

These features are computed for each customer over the period of a day and then their mean and standard deviation over a month are utilized as final features for clustering purposes. These features are used to measure frequency, spatial and temporal characteristics of the customer.

**Frequency**

1. \(Trips\): The number of trips a customer makes per day when they are ‘active’ on the system. This feature provides knowledge about the intensity or frequency of using MTR’s service. Final features chosen are \(\mu_{Trips}\) and \(\sigma_{Trips}\). Note that,
the average trips per day \( \mu_{Trips} \) would also be referred to as ‘trip intensity’ elsewhere in this study.

**Spatial characteristics**

1. **Unique:** The number of distinct stations a given customer visits on any ‘active’ day. This helps in understanding (along with trip intensity) how many unique stations the customer usually visits each day (as opposed to over a month as captured by \( n(RI_s) \)). Final features chosen are \( \mu_{Unique} \) and \( \sigma_{Unique} \).

2. **Homebased:** The number of trips in which the customer either starts or ends their trip at the inferred home station (defined subsequently). This provides information on the centrality of the customer’s trips around their inferred home station. Final features chosen are \( \mu_{Homebased} \) and \( \sigma_{Homebased} \).

3. **Symmetric:** The number of symmetric trips (OD pairs) in a day. So, if a customer travels from A to B and subsequently between B to A on the same day, then these two trips are symmetric in nature. The feature Symmetric is the ratio of all trips which are symmetric to all trips made during that day. Along with Homebased, this feature provides an idea of the customer’s flexibility in mode choice and station choice. It could be inferred that the customer is probably using another mode beside MTR if the value of Symmetric is low. Final features chosen are \( \mu_{Symmetric} \) and \( \sigma_{Symmetric} \).

4. **Distance:** This denotes total distance a customer usually travels on any given day. This helps in inferring how spatially separated the customer’s activity centers are (along with Unique). Final features chosen are \( \mu_{Distance} \) and \( \sigma_{Distance} \). Note that, we utilize the Euclidean distance here.

5. **Displacement:** This denotes the distance between the starting (origin station of first trip) and ending (destination station of last trip) stations for a customer during a given day. Displacement helps in inferring if the customer utilizes another mode for return trips. Final features chosen are \( \mu_{Displacement} \) and \( \sigma_{Displacement} \).
Temporal characteristics

1. *First*: The time of completion of the first journey of the day. The final features which are utilized are $\mu_{First}$ and $\sigma_{First}$.

2. *Last*: The time of completion of the last journey of the day. The final features which are utilized are $\mu_{Last}$ and $\sigma_{Last}$.

Together these features provide information on when the customer intends to reach their activity locations. The variability can help in inferring the possible time flexibility they may possess in their itinerary. These can coincide for customers with relatively low trip intensity (since they would make only one trip per day usually).

Inferred home station

Since the actual home location of the user was unknown in this study, we devised a set of rules to infer a user’s home station based on their trips. For each user the following rule-based method was used to infer their home station: each user’s trips on each day in a given month was analyzed, and then a table was constructed such that it contained two columns—stations and a counter (initialized to zero) for that corresponding station. Hence, for each user, and for each day of the month:

- If the user makes only a single trip on a given day, then the counter for the origin station of the trip is incremented by one.

- If the user visits makes more than one trip on a given day, then:
  
  - The counter for the origin station of the first trip of the day is incremented by one.
  
  - The counter for the destination station of the last trip of the day is incremented by one.

- The station with the highest corresponding value of the counter is the inferred home station for that user. In case two or more stations have the highest value
of counter, then the station where the user starts more of their trips is chosen as the home station.

Therefore by using the above methodology, each user in the data set is assigned an inferred home station for each month. This is utilized primarily for computing the feature Homebased, as defined earlier.

### 3.3.2 Creation of long-term segments

As in the case of short-term segments, a similar methodology was utilized for the creation of long-term segments. Since Euclidean distances were used in k-means in this study, it necessitates the scaling of features so that their ranges are not significantly distinct (order of magnitude). Ignoring the previous step could lead to distortion since the distances along some axes defined on a larger scale may dominate other features defined on a smaller scale. And therefore, in the case of long-term segmentation, the features are scaled prior to the PCA step, and then clustering is carried out on the PCA reduced data. The rest of the methodology remains the same, as defined earlier for short-term segments, with one small exception. Here, the number of clusters $k$ is varied between 2 and 9, and is therefore restricted to single digit values, since the primary purpose is to obtain a broad and simple representation of long-term users that is temporally stable. After clustering, the distribution of the features for each segment was interpreted to understand the corresponding segments.

### 3.3.3 Temporal stability of long-term segments

The methodology adopted for evaluating temporal stability remains the same as seen in the case of short-term segments. Temporal stability was evaluated to ensure that the segments identified were generalizable across time.

### 3.3.4 Segments based on similar spatial spread

To recap, relative importance for stations $RI_s$ is defined over a month and a value exists for each user and for each station in the network. This value could vary between
0 and 0.5, the former value is assumed when the user never visits a particular station, and the latter value is assumed when all the user’s trips involve that particular station, either as the origin or as the destination of the trip. Observing the vector of Relative Importance (for all stations) for a particular user provides useful insight about the spatial aspect of a user’s travel over a month. The distribution of $RI_s$ which is the relative importance for stations may be used to infer a particular user’s spatial spread in the absolute sense. While, by utilizing the features such as $\mu_{RI_s}$, $\sigma_{RI_s}$, $\max RI_s$ and $n(RI_s)$, a user’s absolute spatial spread is captured in the long-term segmentation, an observation of the relative sense in which users make these trips could make the spatial spread easily comparable and observable among users. This relative spatial spread in turn denotes how these users choose to travel among the stations on the MTR’s network and the relative frequency among these (rank ordered set of stations for each user and the frequency of trips made utilizing each of these). For computing the spatial spread in the relative sense, we disregard the trips to particular stations, and only retain the information on the frequency of these trips to each unique station. Hence, while the specific stations and the frequency of visits may be unique to each user, the pattern in which users visit these stations is easily comparable among users.

**Methodology**

The overall strategy is influenced by techniques used in Jiang et al. (2013), and has been altered to suit the specific purpose in this study. Prior to the creation of segments, we first create a relative frequency distribution of each user’s trips and the following strategy is utilized:

- Extract the set of stations a user has visited in a given month and the frequency of visits to each of these. This would correspond to the vector of $RI_s$ for the user.
- The $RI_s$ vector is sorted in the descending order of the trips and the top 20 stations² (and corresponding $RI_s$ values) are retained for each customer.

²Since from the data it was found that more than 95% long-term customers visited less than or equal to twenty unique stations in a month.
• Since some customers might have visited fewer than 20 stations, the following formula is utilized to smoothen the vector:

\[ P(x) = \frac{P(x) + \delta}{1 + \delta|\mathcal{D}|} \]  

(3.7)

Here \(|\mathcal{D}|\) represents the number of stations chosen, and the value of \(\delta\) is chosen as 0.001 since it is a conservative lower bound for the relative frequency of station visits.

• The vectors are now scaled so that the sum is 1 for each customer (a requirement for a legitimate probability mass function (PMF)).

After the above process concludes, we have obtained a PMF that represents a frequency distribution of a rank ordered set of stations which a particular customer visits. We now move on to clustering the PMFs and to identify segments based on similar spatial spread. For this, the following procedure is utilized:

• Compute the dissimilarity matrix, where we use Kullback - Leibler divergence.

• After obtaining the dissimilarity matrix, the Partition Around Medoids (PAM) technique is utilized to obtain the segments (and the representative medoids).

• The Davies - Bouldin index (DB Index) criterion is used to select the optimal number of clusters \((k)\).

In the above, Kullback - Leibler divergence is an "information-based measure of disparity among probability distributions" (Joyce, 2011). Partition Around Medoids (or PAM) involves the search for \(k\) representative objects among the objects of the data set and these representative objects are termed medoids (Kaufman et al., 1990). After obtaining the segments, the profile of the representative medoids may be compared to understand the relative differences in spatial spread among the segments. This is useful, especially to observe the changes in the spatial spread for a particular user due to a given event. For instance, a significant change in the network may impact
the spatial spread of customers (esp. those who depend on it). Clustering and then observing shifts among these clusters may reveal shifts in spatial spread.
Chapter 4

Customer segmentation: results and analysis

4.1 MTR overview

The two tiers of segments, long-term and short-term, were created using a smart card dataset sample made available by the MTR. The smart card system in Hong Kong is the Octopus card, and could be used to access multiple modes of transport, apart from MTR’s service. For use in the MTR, the smart card types, apart from staff cards, are adult, child, student, senior citizen and disabled. Apart from adult card type, the other card types are concessionary cards. Although the Octopus card may contain a customer’s transactions on modes apart from the MTR, however for this study, only data pertaining to the user’s transaction on the MTR system was available. Therefore, data pertaining to transactions on the Hong Kong MTR’s 9 heavy rail lines which are the Island Line (ISL), Kwun Tong Line (KTL), Tsuen Wan Line (TWL), Tseung Kwan O Line (TKL), Tung Chung Line (TCL), Airport Express (AEL), East Rail Line (ERL), Ma On Shan Line (MOS) and West Rail Line (WRL), were made available for this study. Further details of the MTR network are provided in Appendix A.

As such, the fields in the smart card database (shared) are intuitive in that they record the entry station, the exit station, the times associated with entry and exit,
the card type, card value (amount of money available on the card), transaction value and discount as applicable. Apart from the above, a field called business date existed, which was defined as beginning at 5:00 AM each day and ending at 4:59 AM of the following day. For the purpose of the research, the transaction times were used along with the business date as the limit of one transit day. Hence, trips between 12:00 AM and 4:59 AM would correspond to the previous business day.

4.2 Data used in the study

4.2.1 Data used for creating short-term segments

For the creation of short-term segments, we utilize a month’s smart card data: specifically May 2016. For this purpose, a sample is chosen from the set of all eligible customers in that month, who are customers that have traveled at least once in the previous month (i.e., April 2016). The reason for the above condition, as mentioned in the previous chapter, is to prevent ‘window effects’—a user new to the system may start traveling towards the end of the month and therefore an erroneous cluster may be assigned to this user. The sample of customers selected for short-term segmentation includes the following card types: Adult, Children, Students, Senior citizen and Disabled, which is the set of general card types in use in the MTR system. A representative random sample of 100,000 users was selected for creating the clusters using short-term features as described in Chapter 3. Sampling helped since the computation time was lower as compared to choosing the entire population, and at the same time was deemed sufficient to identify the representative segments.

4.2.2 Data used for creating long-term segments

For long term segments, data pertaining only to adult card type was chosen. The said methodology could be expanded to cover all other card types, with long-term segments created separately for each card type. As discussed previously, the eligible dataset would consist of users who have traveled on at least 50% of the weekdays.
in each month in the year of consideration. We select our sample from the dataset pertaining to year 2015, January through December. A random sample of 100,000 customers was selected and the segments were created using the January 2015 dataset. A set of long-term features, as mentioned in Chapter 3, was used for clustering.

4.3 Short-term segments

4.3.1 Correlation among features and PCA

In this section, we explore the correlation among features used for short-term segmentation—temporal and spatial, separately—and determine the ideal number of principal components to be chosen in each case. As mentioned earlier, the number of principal components are chosen based on the cumulative variance explained ratio exceeding the value of 0.9.

Temporal segments

The temporal short-term segments are created by utilizing temporal feature $TP(h)$ (Temporal Probability). The correlation among temporal features $TP(h)$, where $h$ is the hour of the day varying between 5 (5 AM) and 25 (1 AM) is shown in Figure 4-1. It can be observed that there exists a correlation among the features, and this could be mainly attributed to factors such as:

- Users making evening trips, such as at 7 PM (19) and 10 PM (22), and this positive correlation is captured in Figure 4-1.

- Some trips may be less likely, for instance trips being made at 12 AM and 1 AM, and hence this negative correlation is captured in Figure 4-1.

PCA was carried out prior to clustering. Based on the criteria outlined in Chapter 3, the first 14 principal components were chosen for clustering which together explained 92.7% of the total variance in the original data. The cumulative variance explained versus the number of components selected is presented in Figure 4-2.
Figure 4-1: Correlation among the temporal features

Figure 4-2: Cumulative variance explained versus the number of PCA components selected (temporal features)
Spatial segments

As in the case of temporal features, the correlations among the spatial features is also investigated. As is evident from Figure 4-3, in this case, the correlations are more pronounced. These correlations are indicative of the patterns of trips between zones and therefore represents of the underlying origin-destination (OD) demand. Prior to clustering, PCA was carried out (Figure 4-4). A total of 19 components were chosen, which together explain 91.6% of the total variance in the data. The data was transformed based on the components selected which were then utilized for clustering, as discussed in the subsequent section.

4.3.2 Determining the optimal number of clusters

Temporal segments

As mentioned in Chapter 3, since the number of clusters is not known in advance. In this study, we utilize the Davies-Bouldin index to determine the optimal number
Selecting number of PCA components

Figure 4-4: Cumulative variance explained versus the number of PCA components selected (spatial features)

Selecting number of clusters

Figure 4-5: Value of DB index versus number of clusters (temporal segments)
of clusters \((k)\), to be chosen. Since a lower value of this index indicates a better clustering solution (but having too many clusters may not reflect a robust solution), based on Figure 4-5, a value of \(k = 15\) was selected for the number of temporal segments created.

**Spatial segments**

Figure 4-6 shows the DB index for values of \(k\) varying between 2 and 39 for the case of segmentation based on spatial features. Clearly \(k = 24\) was found to be optimal as per the criterion chosen, and thus twenty-four short-term spatial segments were created.

**4.3.3 Temporal segments**

The fifteen temporal segments obtained were named T1 through T15, and these were analyzed further to understand their underlying characteristics. A graphical plot for each segment is presented in Figure 4-7, where for each segment the \(TP(h)\) values at the 5\(^{th}\), 50\(^{th}\) (median) and the 95\(^{th}\) percentile levels are shown. Examining these plots help understand the representative temporal travel patterns of customers who belong to each of these segments. A brief description of these segments is as follows,
- **Segment T1**: As indicated by the median value of $TP(h)$, customers in this segment usually travel between 9 AM and 10 AM and therefore travel during the morning peak period. Additional trips might occur later in the day as indicated by the 95th percentile of $TP(h)$; however, the exact time of these trips vary, and this is supported by the scattered temporal pattern of the 95th percentile value of $TP(h)$ and the comparatively lower values observed in the afternoon and evening periods. These interpretations indicate that this segment may consist of office-goers who possess little flexibility in their start times but the time of return trip varies, while some return at 4 PM or 5 PM and others return later in the day around 8 PM. Moreover, some users travel during mid-day period as well, as evident from the spikes in the 95th percentile value of $TP(h)$ during this period of the day.

- **Segment T2**: This segment consists of customers who are most likely to travel right after noon, with majority of users choosing to travel at 1 PM. Some users may travel an hour earlier or later, however unlike T1, only a small fraction of customers in this segment travel between 9 AM and 10 AM in the morning. Trips after 3 PM are highly unlikely, and therefore this segment may consist of users who do not use MTR to return back, make a one-way trip or alternatively users who use the MTR for relatively short activities in the afternoon period.

- **Segment T3**: Customers in this segment travel late in the evening, with most trips occurring around 7 PM. The likelihood of trips in the morning is very low, however there is a likelihood of additional trips occurring between 9 PM and 11 PM. It may be inferred that this segment consists of users who travel in the evening for leisure activities and is unlikely to include office-goers.

- **Segment T4**: Segment T4 resembles segment T3 to great extent, with similar scattered temporal pattern during the evening period of the day. However, the distinguishing factor is that most likely period for trips for this segment occurs
around 6 PM which is slightly earlier when compared with segment T3. Like customers in segment T3, these customers may also make additional trips later in the night after 8 PM. Therefore, we infer that segment T4 users are similar to users that belong to segment T3 except for the slight difference in the average time during which the trips take place.

- **Segment T5**: Segment T5 consists of users who are likely to travel very early in the morning (around 7 AM) and very late in the night (around 10 PM). This segment exhibits considerable spread in the time of the trips, which may indicate high temporal variability as a result of flexibility, irregular start and end times of the specific activities they may engage in, or additional non-regular trips scattered throughout the day.

- **Segment T6**: Customers in this segment resemble customers in segment T2 in that they travel between noon (12 PM) and 1 PM with high likelihood but may travel during the morning post-peak period (around 10 AM) as well, albeit with a lower probability. There is very low likelihood of travel later in the day, indicating that while these customers may use the MTR service during the morning trip, they may utilize some other mode for the return trip. Another explanation might be that customers in this segment make multiple trips in the morning thereby indicating that these customers engage in shorter activities. Analysis indicates that a large proportion of senior citizen cards belong to this segment, which may help in explaining the shorter activity duration and non work-based travel pattern.

- **Segment T7**: Customers in this segment are more likely to travel during the late afternoon to evening period, with very high likelihood of trips occurring between 7 PM and 9 PM. There is no fixed time period for additional trips during the day for the segment as a whole, as inferred from the low TP values. However, these users may make trips between 8 AM and 10 AM in the morning. Therefore, while this group might consist of customers who use MTR’s service for work-based trips, the most characteristic trait of this group is the highly
concentrated probability of trips during the evening period.

- **Segment T8**: Segments T8 and T1 are similar, with a highly probable trip in the morning and a scattered temporal pattern in the evening. However, users in segment T8 are likely to start their trips earlier in the day—between 7 AM and 9 AM (with most users traveling between 8 AM and 9 AM). Like segment T1, this segment might consist of office-goers who contribute to morning pre-peak and peak period loads. Further analysis showed that a high proportion of concessionary cards for the disabled is part of this segment.

- **Segment T9**: Unlike segments T1 and T8, customers in this segment exhibit a high variability in the time of trip made during the morning period. While the most likely time period for trips is around 11 AM, users in this segment may travel a couple of hours earlier or later. A similar trend is observed later in the day, however the likelihood of a trip after 8 PM is very small. This group may consist of office-goers who have a relatively high flexibility in start and end times, or have variable start and end times.

- **Segment T10**: The customers in this segment travel between pre-noon and evening periods, with most trips taking place around 2 PM. This pattern indicates that this segment might consists of customers engaged in non work-related trips or trips associated with shorter activities. Moreover, it is highly unlikely that the customers in this segment contribute to morning or evening peak loads.

- **Segment T11**: The customers in this segment travel around 4 PM in the afternoon with high probability. Trips in the morning may occur around 10 AM, but there is considerable scatter and hence no dominant pattern for the segment as a whole. The users in this segment travel for shorter activities than the presumed work related customer segments like T1, T8 and T9. Further analysis indicates that a high proportion of children and students belong to this segment.

- **Segment T12**: Customers in this segment are most likely to travel around 4 PM to 5 PM. The users in this segment exhibit high temporal variability. Analysis
reveals that a high proportion of children and student cards are part of this segment, much like segment T11. However, unlike segment T11, this segment shows higher variation in the time of the return trip in the afternoon and early evening period.

- **Segment T13:** The customers in this segment exhibit a highly scattered temporal travel pattern in the morning. However, in the evening there is a high probability of travel around 5 PM. This segment might consist of users that use the MTR’s service for non work-related trips or they might be part-time workers with flexible work start times and firm end times.

- **Segment T14:** Customers in this segment have a diffused trip pattern spread across the morning and afternoon periods, with highest probability of travel around 10 AM in the morning. For trips in the afternoon period, the most likely period is around 2 PM. Analysis shows this segment has a high proportion of senior citizen cards, which might help in explaining the trips made during the afternoon period.

- **Segment T15:** This segment consists of customers who exhibit a travel pattern which consists of one or more periods of travel during the afternoon and early evening period. The most likely period of travel is between 3 PM and 4 PM. There may be trips which take place earlier, between 12 PM and 2 PM, or later between 6 PM and 7 PM, albeit with a lower probability. This pattern may represent customers with a short activity duration, who may travel multiple times in the day.
Temporal distribution of travel demand
Segment T1

Segment T2

Segment T3

Segment T4

Segment T5

Segment T6
4.3.4 Spatial segments

Figure 4-8 shows the twenty-four spatial segments obtained: for each segment the average $SP$ (Spatial Probability) values were computed. We further analyze and try to understand the representative spatial travel pattern of customers who belong to each of these spatial short-term segments. The plots in Figure 4-8 include only the top five stations with the highest average SP values for each segment. These stations might be interpreted as the ‘most important’ stations for their respective segments, since these would be accessed the most (for entry or exit) by its members. A brief
description of these segments is as follows,

- **Segment S1**: Customers in this segment utilize stations on the Tsuen Wan Line primarily. The most important stations are Tsuen Wan, Tai Wo Hau, Kwai Hing, Kwai Fong and Tsim Sha Tsui. By observing the card types, it is also observed that this segment consists of a large proportion of student card holders.

- **Segment S2**: This segment consists of users who primarily utilize the stations on the Ma On Shan line, which serves the users in the eastern part of New Territories. The most important stations for this segment are Ma On Shan, City One, Sha Tin Wan, Hen On and Wu Kai Sha.

- **Segment S3**: This segment consists of users who primarily utilize stations that lie on the KTL line, west of Kowloon Tong. The most important stations for this segment include Kwun Tong, Kowloon Bay, Wong Tai Sin, Lam Tin and Choi Hung. Since these stations primarily lie on the KTL line, which serves Kowloon, therefore the users in this segment could be inferred as those who either live or work in the Kowloon region.

- **Segment S4**: Users in this segment access stations that lie on either side of the harbor, on the Tsuen Wan Line or the Island Line. The most important stations for this group are Tsim Sha Tsui, Causeway Bay, East Tsim Sha Tsui, Wan Chai and Jordan.

- **Segment S5**: Users in this segment access the border crossing stations of Lo Wu and Lok Ma Chau. Apart from the aforementioned stations, these users access other stations that lie on the East Rail Line such as Sheung Shui, Sha Tin and Fanling. Therefore, based on the above, it could be inferred that this segment consists of cross-border travelers, which is an important class of customers to consider.

- **Segment S6**: Customers in this segment access the stations on Tsuen Wan Line primarily, similar to customers in segment S1. However, unlike S1, the
customers in this segment primarily utilize that stations that lie north of Prince Edward. The most import stations for this segment include Tsuen Wan, Kwai Fong, Lai Chi Kok, Cheung Sha Wan and Sham Shui Po.

- **Segment S7**: Customers in this segment primarily utilize the stations that lie on West Rail Line in the western part of new territories. These users primarily access stations which lie west of Kam Sheung Road station. A considerable proportion of users with student cards and with disability cards access these stations, therefore this could help in catering to the specific needs of these users.

- **Segment S8**: Customers in this segment utilize some of the most crowded stations on the MTR’s network such as Central, Admiralty and Tsim Sha Tsui, which lie on either side of the harbor.

- **Segment S9**: Customers in this segment primarily utilize the stations that lie on the Tsuen Wan Line and Kwun Tong Line, especially the interchange stations. The most important stations for users that belong to this segment are Prince Edward, Mong Kok, Yau Ma Tei, Jordan and Tsim Sha Tsui.

- **Segment S10**: Customers in this segment primarily use the stations on the eastern part of Island Line. These customers travel from or to eastern and central parts of Hong Kong island. The most important stations for this segment are Chai Wan, Shau Kei Wan, Sai Wan Ho, Tai Koo and Causeway Bay. A considerable proportion of senior citizen card holders are a part of this segment and utilize these stations.

- **Segment S11**: Customers in this segment utilize stations that are on the Tsuen Wan Line and the Kwun Tong Line. These include Kwun Tong, Wong Tai Sin, and important interchange stations (between TWL and KTL) such as Prince Edward, Mong Kok and Yau Ma Tei. While their spatial travel pattern is similar to segment S9, there are subtle differences in the set of stations predominantly accessed by users in these two segments.
• Segment S12: Like customers that belong to segment S5, customers in this segment also utilize stations that lie on the East Rail Line. However, customers that belong to segment S12 also access stations that lie south of Kowloon Tong, in fact Hung Hom (south of Kowloon Tong) has the highest average SP value for this segment.

• Segment S13: Customers in this segment utilize stations on the Tsuen Wan Line and Kwun Tong Line, and although there’s some similarity with segment S11, the set of stations primarily accessed by these two segments is different. Similar to users in segment S11, these users access Kwun Tong and Wong Tai Sin, however the other important stations for this group are Kowloon Bay, Cheung Sha Wan and Sham Shui Po, which is unique to this segment.

• Segment S14: Customers in this segment primarily utilize stations on the Tsuen Wan Line and interchange stations with the Kwun Tong Line, however unlike customers in segment S9, the stations they access are more spread out spatially. The most important stations for these users are Tsuen Wan, Kwai Fong, Mong Kok, Prince Edward and Yau Ma Tei.

• Segment S15: Customers in this segment utilize stations that lie along the central part of the Hong Kong island. The most important stations for customers in this segment are Wan Chai, Sheung Wan, Causeway Bay and Fortress Hill. These stations lie along the Island Line.

• Segment S16: Customers that belong to this segment primarily travel on the stations that lie on Tsueng Kwan O Line such as Po Lam, Hang Hau, Tsueng Kwan and Tiu Keng Leng. Some stations that users in this segment utilize lie on the Island Line, therefore customers that belong to this segment might transfer between ISL and TKL lines. A large proportion of child card holders constitute this segment.

• Segment S17: Customers in this segment utilize stations on the East Rail Line and the Kwun Tong Line. Therefore, this segment may consist of some cross-
border travelers who may transfer at the Kowloon Tong interchange station between East Rail Line and Kwun Tong Line.

- **Segment S18:** Customers in this segment predominantly utilize stations that lie along the central part of Hong Kong island, on the ISL, such as Central, Admiralty, Hong Kong, Wan Chai and Causeway Bay. These customers access the busier stations along the ISL as compared to customers that belong to segment S15.

- **Segment S19:** Customers who are part of this segment primarily utilize stations that lie on the TWL and KTL, on a section that contains interchange stations like Mong Kok and Prince Edward. Other important stations include Lai Chi Kok, Cheung Sha Wan and Sham Shui Po.

- **Segment S20:** Customers in this segment primarily utilize stations on the Island line and predominantly on the western part of Hong Kong island. The key important stations for the users that belong to this segment are Sheung Wan, Kennedy Town, Sai Ying Pun, HKU and Central.

- **Segment S21:** Customers in this segment utilize the stations on the East Rail Line, but unlike segment S17, the average $SP$ values for this segment is greater for stations that lie south of Kowloon Tong station. The most important stations for this segment are Mong Kok East, Sha Tin, Tai Po Market, Sheung Shui and Lo Wu.

- **Segment S22:** Customers in this segment primarily utilize stations that lie on the Tung Chung Line, like Tung Chung, Tsing Yi, Lai King and Olympic.

- **Segment S23:** Customers that belong to this segment primarily utilize stations on the Kwun Tong Line and both the branches of the Tseung Kwan O Line, and therefore utilize stations in the eastern part of new territories. The most important stations are Tsueng Kwan O, Hang Hau, Po Lam, Kwun Tong and Lam Tin. A good proportion of users with child card belong to this segment.
• *Segment S24*: The customers in this segment primarily utilize stations on the East Rail Line, and are similar to users in segment S21. However, they access a different set of stations. For instance, Tai Wai which is the transfer station between East Rail Line and Ma On Shan line, has the highest average $SP$ value for this segment. Other stations that this segment primarily accesses include Mong Kok East, Hung Hom, Tai Wai, Sha Tin and Lo Wu.
4.3.5 Temporal stability of short-term segments

The temporal stability of both the temporal segments and the spatial segments was computed to ascertain if the segments were stable through the year. The segments obtained in May, were compared with the month of October, considered representative. For this purpose, we utilize the data from the October 2016 sample. The focus here is to ensure that there is temporal stability at least within the year in which the base segments were trained. Doing so ensures that as long as there are no major changes in the network (or major ridership changes), the base segment structure would continue to hold for subsequent months. The temporal stability is computed for the temporal segments and the spatial segments separately and the findings are presented in the subsections that follow.

**Temporal stability of temporal segments**

In Table 4.1, shows the table for computation of Correspondence Score, as described in Chapter 3. For temporal segments a Correspondence Score of 77.164% was obtained, which is a fairly high. There may be slight changes in the segment structure through months and the segments may be affected due to seasonality.
Table 4.1: Temporal stability of temporal segments as measured against October 2016 data (base data: May 2016)

<table>
<thead>
<tr>
<th>October</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
<th>T13</th>
<th>T14</th>
<th>T15</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1'</td>
<td>4494</td>
<td>107</td>
<td>682</td>
<td>464</td>
<td>2</td>
<td>468</td>
<td>26</td>
<td>13^</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>107</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>T2'</td>
<td>5</td>
<td>3629</td>
<td>25</td>
<td>1</td>
<td>523</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>685</td>
</tr>
<tr>
<td>T3'</td>
<td>13</td>
<td>0</td>
<td>8572</td>
<td>186</td>
<td>4</td>
<td>1161</td>
<td>1</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>244</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T4'</td>
<td>0</td>
<td>0</td>
<td>4248</td>
<td>0</td>
<td>143</td>
<td>0</td>
<td>141</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T5'</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3024</td>
<td>2</td>
<td>0</td>
<td>20</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>T6'</td>
<td>0</td>
<td>605</td>
<td>0</td>
<td>0</td>
<td>543</td>
<td>5395</td>
<td>513</td>
<td>0</td>
<td>11</td>
<td>474</td>
<td>227</td>
<td>515</td>
<td>72</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>T7'</td>
<td>1</td>
<td>766</td>
<td>16</td>
<td>0</td>
<td>848</td>
<td>0</td>
<td>3426</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>500</td>
<td>0</td>
<td>5</td>
<td>234</td>
<td>911</td>
</tr>
<tr>
<td>T8'</td>
<td>0</td>
<td>220</td>
<td>114</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>15</td>
<td>6443</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T9'</td>
<td>6</td>
<td>206</td>
<td>33</td>
<td>0</td>
<td>14</td>
<td>117</td>
<td>0</td>
<td>0</td>
<td>6260</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>334</td>
<td>0</td>
</tr>
<tr>
<td>T10'</td>
<td>9</td>
<td>0</td>
<td>2027</td>
<td>184</td>
<td>486</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>6282</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>625</td>
<td>0</td>
</tr>
<tr>
<td>T11'</td>
<td>19</td>
<td>671</td>
<td>26</td>
<td>0</td>
<td>162</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8670</td>
<td>5</td>
<td>163</td>
<td>2</td>
<td>217</td>
</tr>
<tr>
<td>T12'</td>
<td>2</td>
<td>5</td>
<td>69</td>
<td>0</td>
<td>161</td>
<td>3</td>
<td>818</td>
<td>1</td>
<td>91</td>
<td>0</td>
<td>96</td>
<td>5665</td>
<td>1</td>
<td>19</td>
<td>35</td>
</tr>
<tr>
<td>T13'</td>
<td>0</td>
<td>3</td>
<td>265</td>
<td>19</td>
<td>20</td>
<td>311</td>
<td>1</td>
<td>215</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3716</td>
<td>123</td>
<td>92</td>
<td>0</td>
</tr>
<tr>
<td>T14'</td>
<td>0</td>
<td>156</td>
<td>4</td>
<td>0</td>
<td>490</td>
<td>133</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>164</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>3040</td>
<td>0</td>
</tr>
<tr>
<td>T15'</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>659</td>
<td>713</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1</td>
<td>30</td>
<td>4100</td>
</tr>
</tbody>
</table>

Temporal stability of spatial segments

Table 4.2, shows the table for computation of Correspondence Score, as described in Chapter 3. For spatial segments a Correspondence Score of 83.166% was obtained, which is a fairly high number.

4.4 Long-term segments

The descriptive statistics and the density plots of the feature values for the data sample chosen is presented in Appendix B (Figure B-1 and Table B.1).

4.4.1 Correlation among features and PCA

In this section, we explore the correlation among the long-term features which were chosen for the clustering process, following which we ascertain the ideal number of PCA components to be chosen based on the cumulative variance explained ratio exceeding 0.9. Figure 4-9 shows the correlations that exist among the long-term features.

While some of these correlations exist simply because of the type of the features which were chosen, other are due to the specific way in which these long-term users may travel. We illustrate an example of each respectively:

1. $\mu_{RI_s}$ and $n(RI_s)$ are negatively correlated, and this is expected since the mean relative importance for any particular station would decrease as the user visits
Table 4.2: Temporal stability of spatial segments as measured against October 2016 data (base data: May 2016)

| October | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S14 | S15 | S16 | S17 | S18 | S19 | S20 | S21 | S22 | S23 | S24 |
|---------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 18      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 19      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 20      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 21      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 22      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 23      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 24      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 25      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 26      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 27      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 28      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 29      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
| 30      | 2593 | 13 | 7  | 0  | 0  | 12 | 7  | 1  | 0  | 0   | 194  | 0   | 0   | 0   | 51  | 1   | 7   | 0   | 0   | 0   | 4   | 0   | 0   | 0   |
more number of unique stations in a month. A similar explanation would hold true for $\mu_{RI_d}$ and $n(RI_d)$. These correlations are due to the specific choice of the features and their definition.

2. $\mu_{Distance}$ and $\sigma_{Distance}$ are positively correlated, and this implies that users who travel greater distances on the network would likely have greater deviations in distance traveled as well. This is an aggregate trend which is characteristic of the data chosen.

![Figure 4-9: Correlation among the long-term features](image)

To ensure clusters were not skewed due to correlated features, and to reduce the dimensionality of the data, PCA was carried out. Prior to PCA, the features were scaled since their ranges are different (order of magnitude). Based on the criterion identified in Chapter 3, it was determined that the first 13 principal components need to be selected, since together they retain more than 90% of the variance in the original data (Figure 4-10).
4.4.2 Determining the optimal number of clusters

Figure 4-11 shows the DB index for values of k varying between 2 and 9. Clearly k = 4 was found to be optimal as per the criterion chosen. The next optimal solution was k = 8, and therefore we further explore these two solutions to determine the characteristics of these. The prime reason to explore the eight-segment solution is to ensure richness in terms of resolution while using it for various applications as well as to ensure that the number of clusters is not too high, so that it is difficult to interpret them and to use them in applications meaningfully.

4.4.3 Four-segment configuration

The four segments obtained are named segments W through Z. The average values of the features are presented in Table 4.3 and their distributions for each segment are shown in Figure 4-12. These plots depict the density distribution of the features, separated for each segment. Interpreting the feature values for each segment can help in determining what these individual segments represent. A detailed description of these segments is presented as follows,
Selecting number of clusters

Figure 4-11: Value of DB index versus number of clusters (long-term segments)

Table 4.3: Four-segment configuration: segment-wise average feature values

<table>
<thead>
<tr>
<th>Segment</th>
<th>Composition</th>
<th>Active</th>
<th>Range</th>
<th>$\mu_{Trips}$</th>
<th>$\sigma_{Trips}$</th>
<th>$\mu_{Symmetric}$</th>
<th>$\sigma_{Symmetric}$</th>
<th>$\mu_{Homebased}$</th>
<th>$\sigma_{Homebased}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>35.8%</td>
<td>0.88</td>
<td>0.90</td>
<td>2.36</td>
<td>0.34</td>
<td>0.27</td>
<td>0.27</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>X</td>
<td>44%</td>
<td>0.91</td>
<td>0.90</td>
<td>2.07</td>
<td>0.44</td>
<td>0.64</td>
<td>0.39</td>
<td>0.00</td>
<td>0.16</td>
</tr>
<tr>
<td>Y</td>
<td>16.5%</td>
<td>0.84</td>
<td>0.90</td>
<td>1.36</td>
<td>0.48</td>
<td>0.09</td>
<td>0.19</td>
<td>0.09</td>
<td>0.23</td>
</tr>
<tr>
<td>Z</td>
<td>4.5%</td>
<td>0.88</td>
<td>0.90</td>
<td>1.71</td>
<td>0.16</td>
<td>0.65</td>
<td>0.14</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>$\mu_{Displacement}$</th>
<th>$\sigma_{Displacement}$</th>
<th>$\mu_{Distance}$</th>
<th>$\sigma_{Distance}$</th>
<th>$\mu_{Unique}$</th>
<th>$\sigma_{Unique}$</th>
<th>$\mu_{RId}$</th>
<th>$\sigma_{RId}$</th>
<th>max $RId$</th>
<th>$\mu_{RIg}$</th>
<th>$\sigma_{RIg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>1.16</td>
<td>1.44</td>
<td>20.36</td>
<td>8.26</td>
<td>3.06</td>
<td>0.92</td>
<td>0.29</td>
<td>0.08</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>1.25</td>
<td>2.40</td>
<td>17.50</td>
<td>4.37</td>
<td>2.35</td>
<td>0.49</td>
<td>0.44</td>
<td>0.15</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>6.37</td>
<td>3.10</td>
<td>10.10</td>
<td>4.40</td>
<td>2.28</td>
<td>0.44</td>
<td>0.62</td>
<td>0.19</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>2.13</td>
<td>1.00</td>
<td>13.36</td>
<td>1.09</td>
<td>2.04</td>
<td>0.00</td>
<td>0.63</td>
<td>0.54</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>$n(RId)$</th>
<th>$\max RId$</th>
<th>$\mu_{RId}$</th>
<th>$\sigma_{RId}$</th>
<th>$\mu_{RIg}$</th>
<th>$\sigma_{RIg}$</th>
<th>$\mu_{Last}$</th>
<th>$\sigma_{Last}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>17.99</td>
<td>0.36</td>
<td>0.09</td>
<td>0.11</td>
<td>12.35</td>
<td>38982.70</td>
<td>7831.94</td>
<td>70025.06</td>
</tr>
<tr>
<td>X</td>
<td>7.89</td>
<td>0.46</td>
<td>0.19</td>
<td>0.19</td>
<td>6.13</td>
<td>33934.52</td>
<td>4780.70</td>
<td>69445.08</td>
</tr>
<tr>
<td>Y</td>
<td>6.51</td>
<td>0.45</td>
<td>0.19</td>
<td>0.17</td>
<td>6.10</td>
<td>44670.89</td>
<td>7945.03</td>
<td>58829.71</td>
</tr>
<tr>
<td>Z</td>
<td>1.85</td>
<td>0.59</td>
<td>0.49</td>
<td>0.80</td>
<td>2.05</td>
<td>43296.92</td>
<td>3209.83</td>
<td>63493.90</td>
</tr>
</tbody>
</table>

84
Figure 4-12: Distribution of feature values for the four-segment solution
Description of segments

Figure 4-13 shows the split of the customers among the four segments obtained. Clearly segment X dominates the rest (44% of total users), followed by segment W (35.8%), segment Y (15.6%) and with the least number of users belonging to segment Z (4.6%).

After analyzing the averages and distributions of the features (Table 4.3 and Figure 4-12) and inferring the type of user represented by each segment, the customer profiles obtained are presented follows,

- **Segment W**: Customers in this segment have the highest average trip intensity (2.36). Therefore, these customers make the most number of trips of all the users. They also have the highest deviation in the number of trips (0.84), which may mean that these users make anywhere between 1 to 3 trips in a day, with the number of trips being skewed towards 3. These users also visit the most number of unique stations in a month (>12) with an average of 3 unique stations on any given day. This explains the relatively low values of max $RI$s and $\mu_{RI}$, implies the high number of unique OD pairs on which these users may travel, which is about 18 distinct OD pairs in a month. These users travel the longest distances ($\mu_{Distance}$). Moreover, these users make the smallest number of trips to or from
their inferred home station, and also have a high deviation in the number of these trips, which implies that on some days these users may make additional trips which do not start or end at their inferred home stations, this also explains the high deviation in number of trips per day. The value of $\mu_{\text{Symmetric}}$ is low. While, this result may be due to the additional trips, however it does indicate that these users may rely on additional modes and explains the moderately high value of $\mu_{\text{Displacement}}$ and $\sigma_{\text{Displacement}}$. Since these users travel to a large number of stations, it may be expected that they know and understand Hong Kong’s public transportation network fairly well and, therefore, may be at a relative ease to switch modes. Therefore, this segment utilizes the network extensively and, consequently, form the group that contributes appreciably to the farebox revenues. The mean time of the completion of the first trip is around 10 AM, and mean time of completion of the last trip is around 7 PM. While these averages may be skewed towards larger values due to higher values in the distribution, they indicate trip patterns akin to people who work in jobs with traditional start and end work times, thus it could be inferred that this group primarily consists of customers engaged in travel to and from work.

- **Segment X:** Like the customers in Segment W, customers in this segment also have relatively high trip intensity (2.07) and travel long distances on the system, however, the most critical distinguishing factor is the appreciably high value of $\mu_{\text{Homebased}}$ (0.9) as compared to Segment W. Therefore, users in this group make most of their trips such that they either begin or terminate at their inferred home station. This behavior naturally leads to the lowest value of $\mu_{\text{Displacement}}$ among all segments, hence, these users always return to the start location (on a day) at the end of the day using MTR’s service. This might mean that while these users may make other trips using other modes, the most important trips (arguably) are made on MTR for both to and fro legs of the journey. The trips are also fairly symmetric, which supports the fact that these users rely on MTR for their major trips. These users visit an average of 2 unique stations on any given
day and travel to about 6 unique stations in a month. Thus, we might infer that these users access a few specific stations based on their activities, which is relatively spatially constrained (when compared with Segment W). These users travel between 7-8 distinct OD pairs on average, which supports the previous conclusion. The average time of completion of the first trip is usually around 8 AM to 9 AM which is the earliest time period for the first trip among all long-term segments. Also, the mean time for completion of the last trip of the day is around 7 PM. Similar to Segment W, this segment may also consist of users who are engaged in a primary activity like a job but have limited temporal flexibility in the time of their start times.

- **Segment Y**: Customers in this segment make the fewest number of trips on the system, with average trip intensity of around 1.36. Therefore, these users may only make a single trip using MTR’s service on a day on an average. Consequently, the value of $\mu_{\text{symmetric}}$ is the lowest for this segment (0.09), and, hence, the users in this segment most probably rely on other modes. This result coupled with the fact that the value of $\mu_{\text{Homebased}}$ is moderately high (0.85) indicates that while these users may use MTR for one leg of the journey, the other leg may be through utilizing another mode. This also leads to a high mean value of displacement. These users also travel the shortest distances on the MTR network and travel among 6 unique stations in a month on average. The high value of $\sigma_{Rl_d}$ indicates that these users may travel disproportionately among the OD pairs they travel on. The users in this segment also have the lowest values of *Active* and *Range* of any group. These users travel later in the day, with average time of completion of the first trip around 12 PM and return early with average time of completion of the last trip is around 4 PM. The relative concentration of the average time of the first trip and last trip may not mean that users in this segment engage in short activities but might instead be due to the fact that these users travel only once each day. These users exhibit high deviations in their trip times and this might indicate fairly high temporal
flexibility.

- **Segment Z**: Users in this segment have the highest values of $\mu_{\text{Homebased}}$ (0.99) and $\mu_{\text{Symmetric}}$ (0.65), indicating that almost every trip begins or ends at the inferred home station and that the users travel to and from utilizing MTR’s service. These users travel to about 2 unique stations on any given day and also in a month, therefore, these users primarily travel between a single OD pair, presumably engaging in a single type of activity (example: job) utilizing the MTR and may use other modes for all other activities. Hence, these customers utilize the least number of stations on the MTR network relative to the other customer groups. This result may have strong implications for provision of personalized information since the stations these users travel to or from is known in advance with strong certainty. The average time of completion of the first trip of the day is around 10 AM and the average time of completion of the last trip of the day is around 5 PM. This segment has the lowest average deviations in time of trip for first and last trips, thereby indicating low temporal flexibility. These users might be office-goers who utilize MTR for only their work-related trips and have limited flexibility in the time of their trips. This group may be inflexible to fare promotion strategies like the Early Bird Scheme and therefore may be targeted by other means.

**4.4.4 Eight-segment configuration**

The eight segments obtained are named segments A through H. The average values of the features have been presented in Table 4.4 and their distributions for each segment have been shown in Figure 4-14. As in the case of the four-segment solution, a detailed description of these segments has been provided in the following subsection.
Table 4.4: Eight-segment configuration: segment-wise average feature values

<table>
<thead>
<tr>
<th>Segment</th>
<th>Composition</th>
<th>Active Range</th>
<th>μ_Trips</th>
<th>σ_Trips</th>
<th>μ_Symmetric</th>
<th>σ_Symmetric</th>
<th>P_Homebased</th>
<th>σ_Homebased</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11.4%</td>
<td>0.91</td>
<td>0.91</td>
<td>2.86</td>
<td>1.64</td>
<td>0.23</td>
<td>0.31</td>
<td>0.60</td>
</tr>
<tr>
<td>B</td>
<td>11.9%</td>
<td>0.98</td>
<td>0.88</td>
<td>1.75</td>
<td>0.71</td>
<td>0.14</td>
<td>0.26</td>
<td>0.65</td>
</tr>
<tr>
<td>C</td>
<td>20%</td>
<td>0.90</td>
<td>0.90</td>
<td>2.22</td>
<td>0.35</td>
<td>0.73</td>
<td>0.36</td>
<td>0.44</td>
</tr>
<tr>
<td>D</td>
<td>4.3%</td>
<td>0.88</td>
<td>0.90</td>
<td>1.76</td>
<td>0.15</td>
<td>0.70</td>
<td>0.13</td>
<td>0.59</td>
</tr>
<tr>
<td>E</td>
<td>4.4%</td>
<td>0.84</td>
<td>0.90</td>
<td>1.19</td>
<td>0.35</td>
<td>0.05</td>
<td>0.13</td>
<td>0.89</td>
</tr>
<tr>
<td>F</td>
<td>10%</td>
<td>0.89</td>
<td>0.91</td>
<td>2.07</td>
<td>0.60</td>
<td>0.46</td>
<td>0.42</td>
<td>0.83</td>
</tr>
<tr>
<td>G</td>
<td>25.2%</td>
<td>0.92</td>
<td>0.91</td>
<td>2.25</td>
<td>0.66</td>
<td>0.44</td>
<td>0.41</td>
<td>0.80</td>
</tr>
<tr>
<td>H</td>
<td>7.6%</td>
<td>0.87</td>
<td>0.90</td>
<td>1.36</td>
<td>0.51</td>
<td>0.11</td>
<td>0.22</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>μ Ditance</th>
<th>σ_Distance</th>
<th>μ Unique</th>
<th>σ Unique</th>
<th>max Rid</th>
<th>#Rid</th>
<th>σ_Rid</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5.74</td>
<td>4.20</td>
<td>6.41</td>
<td>3.39</td>
<td>12.53</td>
<td>12.53</td>
<td>12.53</td>
</tr>
<tr>
<td>C</td>
<td>0.83</td>
<td>2.80</td>
<td>13.89</td>
<td>1.05</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>D</td>
<td>1.86</td>
<td>1.86</td>
<td>2.20</td>
<td>8.15</td>
<td>1.86</td>
<td>1.86</td>
<td>1.86</td>
</tr>
<tr>
<td>F</td>
<td>5.63</td>
<td>7.59</td>
<td>4.98</td>
<td>31.90</td>
<td>5.63</td>
<td>5.63</td>
<td>5.63</td>
</tr>
<tr>
<td>G</td>
<td>1.46</td>
<td>2.29</td>
<td>15.41</td>
<td>9.47</td>
<td>1.46</td>
<td>1.46</td>
<td>1.46</td>
</tr>
<tr>
<td>H</td>
<td>5.86</td>
<td>2.90</td>
<td>15.41</td>
<td>9.67</td>
<td>5.86</td>
<td>5.86</td>
<td>5.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>n(Rid)</th>
<th>max RId</th>
<th>#Rid</th>
<th>σ_Rid</th>
<th>n(Rid)</th>
<th>μ First</th>
<th>σ First</th>
<th>E Last</th>
<th>σ Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24.75</td>
<td>0.31</td>
<td>0.07</td>
<td>0.09</td>
<td>15.64</td>
<td>37548.39</td>
<td>6700.56</td>
<td>70710.27</td>
<td>4752.67</td>
</tr>
<tr>
<td>B</td>
<td>13.49</td>
<td>0.35</td>
<td>0.10</td>
<td>0.11</td>
<td>10.56</td>
<td>46609.89</td>
<td>11904.39</td>
<td>67374.25</td>
<td>11470.65</td>
</tr>
<tr>
<td>C</td>
<td>5.96</td>
<td>0.48</td>
<td>0.23</td>
<td>0.20</td>
<td>4.88</td>
<td>33359.91</td>
<td>4216.49</td>
<td>69049.47</td>
<td>6492.79</td>
</tr>
<tr>
<td>D</td>
<td>1.84</td>
<td>0.50</td>
<td>0.49</td>
<td>0.00</td>
<td>2.05</td>
<td>35615.95</td>
<td>3018.06</td>
<td>64593.68</td>
<td>4845.14</td>
</tr>
<tr>
<td>E</td>
<td>5.31</td>
<td>0.46</td>
<td>0.22</td>
<td>0.17</td>
<td>5.23</td>
<td>62004.82</td>
<td>5380.25</td>
<td>67748.80</td>
<td>5476.99</td>
</tr>
<tr>
<td>F</td>
<td>10.75</td>
<td>0.43</td>
<td>0.13</td>
<td>0.16</td>
<td>8.34</td>
<td>36338.86</td>
<td>7389.71</td>
<td>70420.20</td>
<td>10147.05</td>
</tr>
<tr>
<td>G</td>
<td>12.85</td>
<td>0.42</td>
<td>0.13</td>
<td>0.15</td>
<td>9.15</td>
<td>34829.90</td>
<td>5496.24</td>
<td>69979.06</td>
<td>8306.28</td>
</tr>
<tr>
<td>H</td>
<td>5.82</td>
<td>0.46</td>
<td>0.21</td>
<td>0.18</td>
<td>5.99</td>
<td>34209.87</td>
<td>5415.51</td>
<td>50184.86</td>
<td>10526.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>μ First</th>
<th>σ First</th>
<th>E Last</th>
<th>σ Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>37548.39</td>
<td>6700.56</td>
<td>70710.27</td>
<td>4752.67</td>
</tr>
<tr>
<td>B</td>
<td>46609.89</td>
<td>11904.39</td>
<td>67374.25</td>
<td>11470.65</td>
</tr>
<tr>
<td>C</td>
<td>33359.91</td>
<td>4216.49</td>
<td>69049.47</td>
<td>6492.79</td>
</tr>
<tr>
<td>D</td>
<td>35615.95</td>
<td>3018.06</td>
<td>64593.68</td>
<td>4845.14</td>
</tr>
<tr>
<td>E</td>
<td>62004.82</td>
<td>5380.25</td>
<td>67748.80</td>
<td>5476.99</td>
</tr>
<tr>
<td>F</td>
<td>36338.86</td>
<td>7389.71</td>
<td>70420.20</td>
<td>10147.05</td>
</tr>
<tr>
<td>G</td>
<td>34829.90</td>
<td>5496.24</td>
<td>69979.06</td>
<td>8306.28</td>
</tr>
<tr>
<td>H</td>
<td>34209.87</td>
<td>5415.51</td>
<td>50184.86</td>
<td>10526.44</td>
</tr>
</tbody>
</table>
Figure 4-14: Distribution of feature values for the eight-segment solution

Description of segments

Figure 4-15 shows the split of the customers among the eight segments obtained. Segment G has the highest percentage of total users (25.2%) followed by segment C (25.0%), segment B (11.9%), segment A (11.4%), segment F (10.0%), segment H (7.6%), segment E (4.4%) and segment D (4.3%). After analyzing the features (Table
4.4 and Figure 4-14) and inferring the type of user represented by each segment, the analysis of the features yielded the following profiles of customers,

![Number of users in each segment](image)

Figure 4-15: Number of users in each segment for eight segment solutions

- **Segment A**: Customers in this segment have the highest trip intensity (2.85) and also the highest deviation in the number of trips per day (1.04), thereby indicating that they may make anywhere between 2 and 4 trips each day. The customers in this segment travel to the highest number of unique stations in a month (about 16), and travel on most number of distinct OD pairs (about 25). Moreover, these users visit anywhere between 2 to 4 unique stations on any given day, sometimes even more than 4. These users travel some of the longest distances on the MTR network. However, the relatively low value of $\mu_{Homebased}(0.6)$ combined with moderately high value of average displacement indicates that the users of this group may rely on modes other than the MTR. The average time of completion of the first trip is around 10 AM and the average time of completion of the last trip is 7 PM. Therefore, these users may be office-goers. The relatively high values of deviation in time of the trips indicates that users in this group may have some temporal flexibility in their trips.

- **Segment B**: Similar to segment A, customers in this segment travel to relatively high number of stations, with users in this segment traveling to around 10 unique
stations on average. Customers in this segment have a lower trip intensity (1.75) compared to Segment A, hence these users are likely to travel just once on some days. Moreover, the average value of Active (0.78) is lowest for this segment, therefore users in this segment travel on the least number of days. A low $\mu_{Homebased}$ value (0.65) denotes that few trips begin or terminate at the inferred home location of the user, however users in this segment have the highest average value of $\sigma_{Homebased}$ thereby denoting that there is considerable variance in the number of trips that begin or end at the inferred home location of the user from one day to another. The mean time of completion of first trip is around 1 PM and the average time of completion of the last trip of the day is around 6 PM. Another notable feature is that the value of $\sigma_{First}$ is highest for this segment, which indicates very high temporal variability associated with the first trip of the day. This might be due to very high degree of temporal flexibility or due to engagement in activity where the schedule that varies.

- **Segment C:** Users in this segment make around 2 trips per day, and a high average value of $\mu_{Symmetric}$ (0.73) indicates that most trips are symmetric in nature, which might imply that customers in this segment rely primarily on MTR’s service for their trips. The users in this segment travel to relatively few unique stations in a month (about 5), but a high value of $\sigma_{RL}$ (0.2) and max($RI_s$) (0.48) might imply that most of the user’s trips are concentrated along two stations, one of which is the inferred home station of the user. Therefore, these users use the MTR service primarily for one type of activity but may use it for occasional trips for other activities. The average time of completion of the first trip of the day is around 9 AM, which is lowest among all segments, therefore these users travel early during the morning peak and pre-peak periods. The average time of completion of the last trip of the day is around 7 PM. Also, the value of $\mu_{Displacement}$ is the lowest for this segment, hence these customers return to where they started in the beginning of the day. Therefore, we may infer that users in this segment may primarily use MTR’s service for work-related activity,
with some trips for other purposes.

- **Segment D:** Customers in this segment have a relatively low value of trip intensity (≈ 1.76) and have the lowest value of $\mu_{Unique}$ (2.04) and $n(RI_s)$ (2.05) denoting that these users seldom travel between more than two unique stations, and they primarily travel along one unique OD pair. This segment also possesses the lowest values of $\sigma_{Trips}$, $\sigma_{Symmetric}$, $\sigma_{Homebased}$, $\sigma_{Displacement}$, $\sigma_{Distance}$, $\sigma_{RI_d}$ and $\sigma_{RI_s}$, which indicates that users in this group make very similar trips on most days. This group has the highest value of $\mu_{Homebased}$ (0.99) and this means almost all their trips either begin or terminate at their inferred home location. The average time of completion of the first trip of the day is around 9 AM - 10 AM, and the average time of completion of the last trip of the day is around 5 PM - 6 PM. Moreover, the values of $\sigma_{First}$ and $\sigma_{Last}$ are lowest for this group, therefore denoting very little temporal variability in trip times. We conclude that these customers are office-goers with relatively low temporal flexibility, who may use MTR for one trip (either home to work or from work to home) and may rely on another mode on some days and may utilize for both legs of the journey on other days. However, these users use the MTR service almost exclusively for one type of activity and rely on other modes for all their other activities.

- **Segment E:** The average trip intensity is lowest for this segment (1.19), therefore implying that users in this segment typically use MTR service for one trip on any day. This leads to an extremely low value of $\mu_{Symmetric}$ (0.05) which is expected since they only make a trip a day. Moreover, this also leads to a very high average displacement. The users in this segment cover the least average distance relative to any other segment, and the average time of completion of the first trip is 5 PM. This shows that users in this segment either use MTR's service only for their return trips or engage in non-work activities and rely on other modes.

- **Segment F:** The customers in this group have an average trip intensity of around
2. These customers travel the longest distances in the network, however the relatively high value of $\sigma_{\text{Symmetric}}$, $\sigma_{\text{Displacement}}$ and $\sigma_{\text{Distance}}$ may imply that these users either do not make a return trip on some days, or alternatively they rely on other modes. The average time of completion of the first trip of the day is around 10 AM and the average time of completion of the last trip of the day is around 8 PM. Therefore, we conclude that this segment consists of users whose activity centers are located far from their home station, and therefore they either may not use MTR’s service to return back or may rely on other modes to make their return trips on some days. The time of completion the first and last trips indicates that these users might be engaged in work-related activity.

- **Segment G**: This segment has a relatively high value of trip intensity (2.25) and has the highest value of Active (0.92). Therefore, these users are most consistent users of the service. Similar to users in segment D, the time of completion of first trip is around 9 AM - 10 AM, however the time of completion of last trip is relatively later, around 7 PM to 8 PM. Moreover, this segment exhibits appreciably greater time variability, when compared with segment D. This might imply that either these users have greater flexibility in travel time especially for their return trip, or they may make an additional trip for other purpose on some days. They may travel, on average, to about 9 unique stations in a given month, reaffirming the assumption that users in this segment may use the MTR service for non work-related activities as well. Also, this segment possesses a relatively low $\sigma_{\text{Displacement}}$ value highlighting the high reliance of these users on MTR’s service. Therefore, based on the above factors, we conclude that users in segment G may also travel for work-related purposes, however they exhibit greater time variability and use the MTR service for more than one purpose.

- **Segment H**: Similar to segment E, this segment also has a very low average trip intensity (1.36), thus indicating that users in this segment may only make one trip on an average day using MTR’s service. Therefore, similar to segment E,
this group is also characterized by a relatively high value of $\mu_{\text{Displacement}}$ and a low value of $\mu_{\text{Symmetric}}$. The trips are relatively over short distances. Since the value of $\sigma_{\text{Homebased}}$ (0.90) is high, it implies that most trips either begin or terminate at their inferred home location. A low value of $\mu_{\text{Last}}$ and high value of $\sigma_{\text{Last}}$ may mean that these users make the return trip using other modes. Hence, this group may consist of either people who work and use another mode to return back home, or users who make leisure trips and then rely on other modes for one leg of the journey.

4.4.5 Temporal stability of long-term segments

Temporal stability of clusters (Four-segment configuration)

Based on the specific methodology identified in Chapter 3, the temporal stability of the solution is computed to ascertain whether the segments created on the January dataset holds true for other months. A higher temporal stability indicates the robustness of the solution obtained and that the segments are unaffected (as a whole) by seasonality aspects. For determining temporal stability of long-term segments, the representative months of May and October were used. The temporal stability of the clusters trained using January 2015 data was compared with May 2015 and October 2015 datasets and it was found that the Correspondence Score of 99.43% for May and 99.13% for October was obtained (refer Table 4.5), which denoted very high temporal stability of the clusters obtained.

Temporal stability of clusters (Eight-segment configuration)

The temporal stability of the clusters trained using January 2015 data was compared with May 2015 and October 2015 datasets, as in the case of four-segment solution. It was found that the Correspondence Score of 96.73% for May and 96.86% for October was obtained which denoted very high temporal stability of the clusters obtained (refer to Table 4.6).
Table 4.5: Temporal stability of four-segment solution (January to May and January to October)

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>May</td>
<td>W'</td>
<td>15913</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>X'</td>
<td>114</td>
<td>43706</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Y'</td>
<td>109</td>
<td>210</td>
<td>35249</td>
</tr>
<tr>
<td></td>
<td>Z'</td>
<td>0</td>
<td>1</td>
<td>4559</td>
</tr>
<tr>
<td>October</td>
<td>W</td>
<td>4221</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>X'</td>
<td>0</td>
<td>16014</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Y'</td>
<td>0</td>
<td>89</td>
<td>35840</td>
</tr>
<tr>
<td></td>
<td>Z'</td>
<td>0</td>
<td>264</td>
<td>473</td>
</tr>
</tbody>
</table>

Table 4.6: Temporal stability of eight-segment solution (January to May and January to October)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>May</td>
<td>A'</td>
<td>7793</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>36</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>B'</td>
<td>35</td>
<td>9677</td>
<td>2</td>
<td>98</td>
<td>297</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>C'</td>
<td>67</td>
<td>346</td>
<td>23679</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>521</td>
</tr>
<tr>
<td></td>
<td>D'</td>
<td>0</td>
<td>0</td>
<td>11167</td>
<td>363</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>E'</td>
<td>37</td>
<td>45</td>
<td>5</td>
<td>8</td>
<td>11353</td>
<td>0</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>F'</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4212</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>G'</td>
<td>26</td>
<td>453</td>
<td>13</td>
<td>151</td>
<td>472</td>
<td>0</td>
<td>24412</td>
</tr>
<tr>
<td></td>
<td>H'</td>
<td>12</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>October</td>
<td>A</td>
<td>23779</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>56</td>
<td>1033</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>B'</td>
<td>2</td>
<td>11528</td>
<td>23</td>
<td>0</td>
<td>30</td>
<td>137</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>C'</td>
<td>1</td>
<td>16</td>
<td>4413</td>
<td>2</td>
<td>15</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D'</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>3911</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>E'</td>
<td>6</td>
<td>27</td>
<td>4</td>
<td>5</td>
<td>7784</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F'</td>
<td>1</td>
<td>442</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>24261</td>
<td>429</td>
</tr>
<tr>
<td></td>
<td>G'</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11374</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>H'</td>
<td>49</td>
<td>66</td>
<td>3</td>
<td>0</td>
<td>26</td>
<td>131</td>
<td>12</td>
</tr>
</tbody>
</table>

103
Table 4.7: Cross tabulation of membership among the four-segment and eight-segment solutions

<table>
<thead>
<tr>
<th>Four segment solution</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11417</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>9207</td>
<td>152</td>
<td>2614</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>24727</td>
<td>245</td>
<td>4</td>
</tr>
<tr>
<td>Eight segment solution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>4289</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>57</td>
<td>4229</td>
<td>145</td>
</tr>
<tr>
<td>F</td>
<td>4896</td>
<td>4330</td>
<td>837</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>10242</td>
<td>14348</td>
<td>599</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>5</td>
<td>350</td>
<td>7090</td>
<td>205</td>
</tr>
</tbody>
</table>

4.4.6 Selection of configuration for further study

Thus far, we have explored both the four-segment configuration solution and the eight-segment configuration solution. Table 4.7 shows the cross tabulation of the membership among the four-segment and the eight-segment solution for the training dataset (January 2015).

From this, we observe that, roughly speaking:

- Segment A is a subset of Segment W.
- Segment C is a subset of Segment X.
- Segment D is a subset of Segment Z.
- Segment E is a subset of Segment Y.
- Segment H is a subset of Segment Y.

While both solutions are robust and may be used for further analysis, the rest of this thesis will focus on the eight-segment solution for it provides better resolution in observing travel behavior heterogeneity among the groups. One such travel behavior which may be of concern would be to look at the average time of first trip each day among the different segments. Figure 4-16 shows the trends among the different
segments. The eight-segment solution can offer a better resolution for most applications, and since there isn’t a direct hierarchical relationship between the four-segment solution and the eight-segment solution, indeed the latter captures aspects of heterogeneity that the former misses. Hence, better insight is achievable by utilizing the eight-segment solution, and moreover the eight-segment solution could be interpreted with ease, as described previously.

Figure 4-16: Average time of completion of first trip of day (and the deviation) for each segment

4.4.7 Shifts among segments

In this section we explore the shifts that take place among long-term segments in the year 2015. For this, the original sample chosen from January 2015 was tracked throughout the year 2015. Figure 4-17 shows that the relative sizes of the segments remain almost the same throughout the year, however this shouldn’t imply that there were no shifts among these segments. In fact we do find appreciably large shifts
throughout the year 2015, and an attempt has been made to identify what these shifts might imply. It is evident from Figure 4-18 that the inter-segment shifts values are above the 35% mark across each month, with the value crossing 40% for January to February and February to March. However, the fact that segment sizes remain relatively constant is an interesting feature of the segments and implies that while individual users might change their travel patterns (and thus their segments), the overall sizes of long term users remains roughly constant across the year.

Figure 4-19 shows the monthly shifts among various segments during the year 2015.

Figure 4-17: Number of customers in each long-term segment across months in the year 2015

This depiction helps in understanding the shift patterns between consecutive months. We observe that the shifts express some symmetry across months and specific characteristic patterns. While our interpretation of these shifts may be limited in scope within this study, we do make a preliminary attempt to understand and describe two specific patterns here and describe these shifts.

Based on the shifts observed in 2015 dataset, we can observe that:

- There were multiple shifts between Segment C and Segment D. The shifts in
Figure 4-18: Aggregate number of shifts as a percentage of the sample (100,000 users)

the direction of Segment D to Segment C is dominant for transitions in:

- January to February
- March to April
- May to June
- June to July
- August to September
- November to December

The shifts in the opposite direction are dominant in:

- February to March
- April to May
- May to June
- June to July
- July to August
- September to October
Segment C and Segment D both represent work-based activity, however the difference is that users in Segment C make more trips using MTR’s service on average and may travel for trips apart from their primary activity, whereas users in Segment D use the service exclusively for their primary activity. Hence, we infer that a subset of office-goers may sometimes use MTR’s service for non work-related activities in a few months, whereas they may only use MTR’s service for work related activities in other months.

- Similarly, multiple shifts were observed between Segment A and Segment G. The shifts in direction of Segment G to Segment A were dominant in:
  - January to February
  - March to April
  - August to September
  - November to December

The shifts in the opposite direction were dominant in:
  - February to March
  - April to May
  - September to October
  - October to November

While Segment A and Segment G represent users who use MTR’s service for work related trips and other activities, the prime difference is that users in Segment A travel more extensively using MTR’s service and have higher deviation related feature values, implying more variable travel pattern. Hence, these shifts may be interpreted as users who rely heavily on MTR’s service either increasing or decreasing their usage in subsequent months.
While the above interpretations are an attempt to understand and explain the shifts descriptively, with more data availability on personal data on the card users, it may become easier to explain and comprehend these shifts. Moreover, it is important to note that while the aforementioned trends are the two such trends observed in the data, they are by no means an exhaustive set of shifts observed. Information from these trends may be used by agencies along with customer experience data for better comprehension. Moreover the information from the above shifts may be used to reduce noise in comprehending attrition or churn in fare products, and other such studies thereby helping to achieve better results Stuntz (2018).

4.4.8 Home station changes

We observe from Table 4.8, and Figure 4-20 that a vast majority of customers do not change their home station, which could be expected since long-term users rarely change their homes. Moreover, it can be observed from Table 4.8 and Figure 4-21 that the number of unique inferred home stations, for a vast majority of customers for whom number of home stations is non-zero, oscillates among two or three stations. This may imply that we are either capturing their primary activity center along with home station, or these two/three inferred home stations lie within a walkable distance from the users home location. While the methodology presented here is adequate, with other data sources such as user’s personal information, it would be easier to determine the user’s home station accurately.

4.4.9 Extending results to other years

To see if the long-term segmentation scheme identified in this research could be extended to other years, a random sample of 100,000 users was chosen from the dataset pertaining to the year 2016. A similar test of temporal stability, as discussed previously, was carried out for the months of May 2016 and October 2016. The results are shown in Table 4.9. It was found that a Correspondence Score of 97.8% was obtained for the May 2016 sample and 97.4% was obtained for the October 2016
Table 4.8: Number of home station changes versus number of unique home stations

<table>
<thead>
<tr>
<th>Number of home station changes</th>
<th>Number of unique home stations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>78022</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
</tr>
</tbody>
</table>

sample. Together, these scores illustrate the generalizability of the long-term segmentation framework (and solution) beyond the training year, and therefore we can infer that the long-term segments obtained here are illustrative of the representative long-term rider across years, in the MTR system. Between the year 2015 and 2016, a few new stations were added to the MTR’s network. Hence, the fact that the segmentation scheme remains valid both before and after this network expansion event indicates that the results of long-term segmentation could be utilized to conduct before-and-after analysis to understand the impact of network expansion activities, thereby providing the agency with a strong tool to understand how long-term users might be impacted by such events. Therefore, the overarching implication here is that it may not be necessary to re-train the clusters, and, the centroids created using January 2015 data may be utilized to identify and assign segments to new long-term customers in the future, and to track segment shifts after changes in the system, thereby avoiding the time and computation costs associated with re-training clusters and to interpret them. Another important feature to note is that the sizes of the clusters remains roughly similar through the year 2016 (Figure 4-22), thereby indicating that while there may be shifts from one segment to another, the overall sizes remain relatively unchanged. This finding is similar to what was found earlier for the 2015
Table 4.9: Temporal stability of the eight-segment solution in the year 2016

<table>
<thead>
<tr>
<th>May</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>24694</td>
<td>141</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>244</td>
<td>362</td>
</tr>
<tr>
<td>B'</td>
<td>0</td>
<td>11142</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>170</td>
</tr>
<tr>
<td>C'</td>
<td>0</td>
<td>0</td>
<td>4420</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>D'</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4137</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E'</td>
<td>608</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>24059</td>
<td>7</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>F'</td>
<td>34</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>16</td>
<td>7885</td>
<td>6</td>
<td>69</td>
</tr>
<tr>
<td>G'</td>
<td>98</td>
<td>0</td>
<td>13</td>
<td>2</td>
<td>72</td>
<td>7</td>
<td>9820</td>
<td>56</td>
</tr>
<tr>
<td>H'</td>
<td>63</td>
<td>14</td>
<td>26</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>55</td>
<td>11618</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>October</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>11381</td>
<td>0</td>
<td>0</td>
<td>158</td>
<td>132</td>
<td>68</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>B'</td>
<td>1</td>
<td>23946</td>
<td>0</td>
<td>4</td>
<td>658</td>
<td>5</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>C'</td>
<td>0</td>
<td>0</td>
<td>4351</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>D'</td>
<td>15</td>
<td>120</td>
<td>0</td>
<td>9773</td>
<td>179</td>
<td>3</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>E'</td>
<td>351</td>
<td>0</td>
<td>1</td>
<td>155</td>
<td>24698</td>
<td>4</td>
<td>10</td>
<td>205</td>
</tr>
<tr>
<td>F'</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4338</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>G'</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>9</td>
<td>7809</td>
<td>0</td>
</tr>
<tr>
<td>H'</td>
<td>204</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11145</td>
</tr>
</tbody>
</table>

dataset and, therefore, we can conclude that there is little change in the aggregate sizes of the long-term segments from one-month to the next, at least for the years 2015 and 2016.

### 4.4.10 Spatial-spread based segments

For the creation of spatial-spread based segments a 1% sample (with replacement) of the January 2015 sample (which was used for the creation of long-term segments) was used. Therefore, the spatial-spread related segments created here are based on the methodology described in Chapter 3. Four segments were obtained, based on DB index criterion, and these four segments indicate segments with distinct spatial spreads. The four representative medoids were analyzed independently to understand the segments they represent, with a plot of these in Figure 4-23, where x-axis denotes the stations (ordered) and y-axis denotes the probability. The interpretation of these segments is presented as follows,
- **Segment LS1**: Customers in this segment primarily travel between two unique stations. These users make limited trips to other stations, but their overall trips are limited to about 4 to 7 stations. We could infer that these users may engage in more than one type of activity, however their primary activity dominates.

- **Segment LS2**: Most of the trips of customers in this segment are between two stations, however they make more trips to or from other stations, when compared to users in segment LS1. Therefore, we may infer that in the case of segment LS2, there are more trips for other activities as compared to LS1.

- **Segment LS3**: Customers in segment LS3 have the highest spatial spread and, therefore, make trips to or from most number of unique stations as compared to the other segments. The users in this segment may visit more than 15 unique stations in a month and therefore we may infer that these users are well aware of MTR’s network and may be heavily dependent on the MTR network.

- **Segment LS4**: Customers in this segment restrict their trips between two unique stations. Since there are just two primary stations, we may infer that users in this segment use the MTR for just one type of activity (like going to work). While these users may travel to a few more stations (up to 4), the majority of their trips are predominantly between two stations.

These four spatial-spread based segments could prove to be incredibly useful for some applications which shall be highlighted in the subsequent chapters. One thing to note here is that the above interpretations have been arrived at based on the fact that only long-term users were chosen for determining the long-term segments, for instance tourists may be expected to have a large spatial spread and, therefore, might belong to LS3, however this discussion is not within the scope of the spatial-spread analysis presented here.
(a) Shifts between January and February

(b) Shifts between February and March

(c) Shifts between March and April

(d) Shifts between April and May

(e) Shifts between May and June

(f) Shifts between June and July
Shifts between July and August

Shifts between August and September

Shifts between September and October

Shifts between October and November

Shifts between November and December

Figure 4-19: Shifts among the long-term segments in the year 2015
Figure 4-20: Distribution of users based on the number of changes in the inferred home station

Figure 4-21: Distribution of users based on the number of unique home stations, in the year 2015
Figure 4-22: Number of customers in each long-term segment across months in the year 2016

Figure 4-23: Representative medoids for the spatial-spread segments
Chapter 5

Personalized information provision

5.1 Overview

Most transit agencies provide some form of service information—static or real-time via various interfaces such as websites, smartphone applications, etc. Some agencies also use these interfaces to collect feedback or to provide promotion-related information to users. However, as mentioned previously in Chapter 2, while several other domains have largely embraced and implemented some form of personalization for their information provision systems, public transit information systems are only slowly catching up to this. Catering information to a specific user or a specific type of user can help in increasing the likelihood of the information being relevant to that particular user, thereby improving customer experience. However, in the case of public transit, most agency-led platforms for information provision still rely on the user’s self-selection in order to filter information relevant to that particular user. With public transit agencies clearly showing interest in providing a personalized experience to its riders, we feel it is useful to demonstrate how the findings of the current research could be utilized to provide personalized information, especially when the type of information being provided relies on the characteristic travel behavior of its beneficiary.

Figure 5-1 shows screenshots of the MTR’s new smartphone application developed in late 2017 which clearly shows the agency’s intent to better understand customer’s needs. This could help towards providing a personalized experience to its users. Fig-
Figure 5-1(a) shows how the agency might collect specific contextual information—in this case the user’s location information—which could be used in order to provide targeted information to users. This additional data collected might in turn help the agency to augment its current data sources, and therefore help towards better service planning in the future (as discussed in Chapter 2). In contrast to above, Figure 5-1(b) shows a more traditional way of catering specific updates and information to users, which is user self-selection, Figure 5-1(c) shows the incorporation of a feature for users to sign in using their MTR Club membership, which could potentially help towards personalization. This feature could be an apt starting point for the discussion of personalized information provision, and in demonstrating the application of personalized information provision we shall show how knowledge of a user’s past travel pattern (efficiently captured through their segments: long-term and short-term) could help in filtering information that these specific users might need and benefit from. Figure 5-1(d) shows the recently developed chatbot feature within the application. We now define the steps for providing personalized information to customers by utilizing the customer segmentation framework developed in this study.

This chapter is an attempt to show how targeted information may be facilitated by using the segmentation scheme created in this study. While information provision systems may benefit from two-way communication and active user feedback might be critical for long-term success, the framework of personalized information provision illustrated in this chapter might be an apt starting point. By relying on traditional sources of data, the framework presented here may be deployed across agencies and therefore could be generalized. Moreover, the framework could later be supplemented with additional sources of data which might be made available through sources such as the user’s smartphone (location, past trip planning efforts, etc.) and newer sources such as iBeacon. These additional “contextual” sources may prove to be incredibly useful in both selecting beneficiaries of the information as well as the crafting the specific message which might be sent to them. This may also provide insight on the user’s activities outside of the public transit system, therefore helping in planning efforts in the future. Figure 5-2 shows our attempt at charting out the information
provision system, with three primary components being the data, the processing of the data, and the actual information provision.

5.2 Framework

In this section, we demonstrate a system of providing personalized information to users by utilizing the segmentation scheme as a building block to identify users who might need certain types of information. Users who belong to the same segment share features characteristic of that segment, for instance users who belong to the same spatial short-term segment are very likely to utilize similar stations on the network, and therefore might benefit from targeted information provision pertaining to the stations they predominantly use. Therefore, these segments—temporal short-term segments, spatial short-term segments and long-term segments—may be effectively used to infer the relevance of each type of information for users that belong to these segments, and this is referred to as inferred relevance in this study, and forms the basis for selection of users. The benefit of providing targeted information is that it reduces the amount of spam or irrelevant messages that may be sent over a channel, and therefore it may lead to a better customer experience, potentially leading to better engagement. To effectively utilize the segments the following strategy is proposed to infer users who may benefit from a certain type of information (Filtering process):

Step 1: Define the scope (in terms of location and time) of the users who might need this information. The scope here indicates an informal definition of all the users who might benefit from a certain type of information.

Step 2: Identify the hours (h) and zones (z) based on the mapping defined in Chapter 3, and then utilize it to identify, in decreasing order of inferred relevance, the short-term segments which need to be provided this information (both spatial and temporal).

Step 3: Utilize features from long-term segments if deemed necessary for the specific application, and if the long-term users would benefit from it. Obtain a list of long-term segments sorted in descending order based on the inferred relevance of the
Promotion Offers

By enabling this function, MTR Mobile will provide latest train service and offer information.

Traffic News

Provide instant updates whenever a train service is expected to experience a delay (over 20 minutes), or schedule adjusted due to tropical cyclones/festive holidays.

- All lines
- Airport Express
- Disneyland Resort Line
- East Rail Line
- Island Line
- Kwun Tong Line
- Light Rail
- Ma On Shan Line
- South Island Line

(a) Collection of contextual information (here: location)

(b) Reliance on self-selection to filter information

(c) The app’s main screen with an option to log in using MTR Club credentials

(d) The newly developed chatbot interface

Figure 5-1: Screenshots of MTR’s smartphone application (Source: MTR)
information.

Step 4: Determine the level of importance of the specific information and then decide the segments this information would be provided to. For critical information, most segments may be provided this information without worrying too much about the concern of spamming users, whereas for information deemed less critical, fewer (the most relevant) user segments may be provided this information.

It is important to note that since short-term segments are trained at the end of every month, the results utilized in the above process correspond to the segments obtained in the last trained model, which might often correspond to the segments trained in the month prior to the month during which the information is sent out. Moreover, to quantify the efficacy of the information provision system, and to justify the specific choice of level of importance, it is deemed important to define specific metrics to capture the relevance of the information being provided. Since these metrics may be better defined based on the specific type of information being provided, we shall discuss these in pertinent sections. To better understand the above steps, we discuss a specific type of information provision to users—targeted information provision during incidents or service disruption—and then illustrate, using examples, how the above
process could be utilized for the provision of personalized information.

5.3 Targeted information provision during incidents

The segmentation of customers as presented in this study can help in identifying customers who may be potentially affected by an incident that disrupts service in a certain section of the system, at a certain hour of the day. Therefore, in this section we look at the provision of personalized information pertaining to a given incident to customers, on a priority basis, as inferred from their past spatio-temporal travel behavior which is captured by their short-term segments. While there exist mobile applications and websites that do provide service disruption information to customers, the customers must either self-select and choose the lines (refer Figure 5-1(b)) for which they wish to receive such information, or alternatively they may receive information pertaining to all incidents on the network by choosing to receive information via broadcast modes such as social media. Some other ways of receiving this information might be at the subway station itself via audio or visual information systems, or through social media. The approach discussed in this section might be useful for providing information to users through a dedicated agency-owned mobile application, similar to the one that exists for MTR, which allows for users to log in with their smart card information so that the agency could identify users based on their past travel behavior, and can therefore identify their corresponding segments. A benefit of this approach is that this framework can help in providing this information in an automated manner so that regardless of the customer’s self selection, they may receive this information if it is deemed suitable based on their past travel pattern. An important point to note here is that for incidents which are deemed critical, the agency might choose to provide information to all users regardless of their inferred relevance (inferred through segments). However, to contrast this with the current scenario, this framework would allow for users to effectively bypass self-selection to receive traffic news (Figure 5-1(b)). When a user self-selects, information would be provided to them regardless of the inferred relevance.
The filtering process in the case of targeted information provision during incidents would look like the following:

*Step 1:* Define the scope of the incident. For example, “an incident has occurred at 10 AM and is expected to last for two hours (until 12 PM) and it is expected to critically impact passengers who use stations A, B and C”.

*Step 2:* Short-term segments may be utilized. Based on the spatial zone affected and the temporal zone affected, the spatial and temporal segments may be sorted in descending order of mean spatial probability value \( \overline{SP}(\text{zone}(\text{station})) \) of the affected zone and the mean temporal probability value \( \overline{TP}(\text{hour}) \) of the hours of the incident. The output would consist of an ordered list of short-term spatial segments and short-term temporal segments in decreasing order of inferred relevance. Hence, spatial short-term segments with high mean SP value for the impacted zone, and temporal short-term segments with high mean TP values for the duration of the incident would be on top of the respective lists.

*Step 3:* For targeted incident information updates, in the present study, we do not distinguish between long-term and short-term users, and therefore utilize long-term segment information only for post hoc analysis (presented in a later section).

*Step 4:* Finally, the level of importance of the information is determined. In the case of incidents, the agency might decide to vary it depending upon the specific location and time of disruption. For example, an incident occurring during the peak period at a major interchange station may warrant a wider reach and therefore carry a high level of importance. The selection takes the form \((n_1, n_2)\), which implies that the first \(n_1\) spatial short-term segments and the first \(n_2\) temporal short-term segments, as identified from the ordered lists obtained in Step 2, would be informed.

As discussed previously, we define key metrics to quantify the efficacy of targeted information provision during incidents. The first metric is Effective Capture Rate (ECR), and is defined as the total number of customers who could be reached at a given level of importance. This term incorporates a correction for users who would shift (to segments which wouldn’t need the specific information) in the middle of the month since there is no way (in the current system) to determine this shift preemp-
tively, and thus does not reflect a flaw inherent in the way the segments capture user’s travel pattern. ECR is defined as follows:

\[
ECR = \left( \frac{\text{Customers affected by the incident and informed}}{\text{Customers affected by the incident} - \text{Customers who shift segments unfavorably}} \right) \times 100
\]  

(5.1)

The second metric is chosen with the aim of capturing the total reduction in spam that would be achieved through targeted information, as compared to sending information out to users under a broadcast mode (everyone receives the information regardless of the inferred relevance of the information for a user). Here, we do not apply a correction for segment shifts. This metric is termed Spam Reduction Rate (SRR) and is defined as follows:

\[
SRR = \left( 1 - \frac{\text{Customers not affected by the incident and informed}}{\text{Customers not affected by the incident}} \right) \times 100
\]  

(5.2)

Clearly, these values would also vary across incidents, and analyzing past incidents would provide better insight to the agency to choose the level of importance, which is effectively selecting the values of \(n_1\) and \(n_2\). It could be chosen based on historical knowledge and severity of the incident. In the following subsections, two illustrative examples have been provided.

5.3.1 Example I

To illustrate the four steps, an incident is analyzed. This incident occurred on 25th August, 2016 on the East Rail Line, and affected service at the following stations primarily: Sheung Shui, Fanling and Tai Wo. This incident started approximately at 11 AM and it took roughly five hours for the service to return to normalcy. While intermittent service did exist between the affected stations, it could be assumed that the disruption was serious enough to obstruct customers’ travel. We follow the steps as illustrated in the above example to select and target the users who would need this information.

Step 1: The scope is defined to be all users who utilize the affected stations, namely Sheung Shui, Fanling and Tai Wo, and they usually utilize the system between 11:00 AM and 4:00 PM.

Step 2: We utilize the short-term segments to identify relevant users. To do so, we
observe that the three affected stations lie on zone (z) 1. Moreover, once we factor in variability, we notice that we are concerned with hour (h) values that lie between 11 (11 AM) and 18 (6 PM). The spatial short-term segments are sorted in decreasing order of mean value of SP value corresponding to zone (z) 1. Similarly, the temporal short-term segments are sorted in decreasing order of mean value of TP averaged over hour (h) values between 11 and 18.

**Step 3:** Since this targeted incident information provision, we skip this step.

**Step 4:** Finally, the level of importance of the information is determined. Here, since the incident could be deemed critical since it affected the operations for a long time, inspite of some intermittent service. We select $n_1 = 5$, and $n_2 = 10$. So the spatial short-term segments chosen are: $S_i = \{S_5, S_{21}, S_{12}, S_{17}, S_{24}\}$ and the temporal short-term segments selected are: $T_i = \{T_{10}, T_{15}, T_{8}, T_{2}, T_{11}, T_{6}, T_{14}, T_{1}, T_{12}, T_{5}\}$, arranged in descending order of their inferred relevance.

If information was provided to the users as identified by the above filtering process, the following values of the key metrics, as defined earlier, would be obtained:

$$ECR = \left(\frac{4,027}{5,381 - 446} \times 100\right) = 81.60\%$$

$$SRR = \left(1 - \frac{300,131}{1,217,522} \times 100\right) = 75.35\%$$

Hence from the above calculations, we infer that effectively 81.60% of the affected users would be informed under the personalized information provision system and moreover, this would lead to reduction of about 75.35% in spam. As mentioned earlier, the choice of level of importance of the incident and consequently the values of $(n_1, n_2)$ could be chosen based on the agency’s evaluation of the type of incident.

If the agency wishes, it may tweak the values of $(n_1, n_2)$ to inform more number of customers in order to reach a wider audience risking more spam (which may be the preference if the incident is deemed severe by the agency). So, for instance in the above example, if instead of the aforementioned values of $(n_1, n_2)$, the agency decides to use $n_1 = 6$, and $n_2 = 14$, then in that case, we obtain the following values of the
key metrics:

\[
ECR = \left( \frac{4,455}{5,381 - 291} \times 100 \right) = 87.52\%
\]

\[
SRR = \left( 1 - \frac{430,174}{1,217,522} \times 100 \right) = 64.67\%
\]

Clearly, in the above scenario, we reach more number of affected customers (as reflected by the higher ECR), however, at the same time, there is an increase in the number of users who receive this information even though they were not affected by the incident.

### 5.3.2 Example II

To illustrate the usability across incidents, we consider another incident. This incident occurred on 19th June 2015 and led to suspension of service on both tracks of the East Rail Line between the stations of Fo Tan and Tai Po Market. This incident began around 4:22 PM and lasted for about 50 minutes. In this case, clearly the incident was relatively shorter in duration as compared to the incident described in the previous subsection. Given that we discretize time in calculating temporal short-term segments and that there is non negligible variability in people’s travel patterns on such a short time scale (as observed from data), we would expect ECR values to be smaller at similar levels of \((n_1, n_2)\). We follow the steps as illustrated in the above example to select and target the users who would need this information.

**Step 1:** The scope is defined to be all users who utilize the affected stations stations, namely Fo Tan, University and Tai Po Market. These users would utilize the system around 4:00 PM to 5:00 PM.

**Step 2:** We utilize the short-term segments to identify relevant users. To do so, we observe that the three affected stations lie on zone \((z)\) 1. Moreover, once we factor in variability, we notice that we are concerned with hour \((h)\) values that lie between 15 (3 PM) and 18 (6 PM). The spatial short-term segments are sorted in decreasing order of mean value of SP value corresponding to zone\((z)\) 1. Similarly, the temporal
short-term segments are sorted in decreasing order of mean value of TP averaged over hour (h) values between 15 and 18.

**Step 3:** Since this targeted incident information provision, we skip this step.

**Step 4:** Finally, the level of importance of the information is determined. Here, the incident impacted both the lines, it could be deemed fairly important. We select $n_1 = 5$ and $n_2 = 10$. So the spatial short-term segments chosen are: $S_i = \{S_5, S_{21}, S_{12}, S_{17}, S_{24}\}$ and the temporal short-term segments selected are: $T_i = \{T_{12}, T_{15}, T_{11}, T_4, T_{10}, T_9, T_5, T_{13}, T_{14}, T_7\}$, arranged in descending order of their relevance.

Upon providing information to the users as identified by the above filtering process, we find that:

$$ECR = \left( \frac{164}{292 - 78} \times 100 \right) = 76.63\%$$

$$SRR = \left( 1 - \frac{55,604}{260,861} \times 100 \right) = 78.68\%$$

Where the terms carry similar interpretation as before. In the following section, we illustrate how the segmentation scheme defined in this study could be utilized as a tool for analyzing the impact of an incident. For this, we shall utilize the long-term segments.

### 5.3.3 Post hoc analysis

After the occurrence of an incident, it might be a useful exercise to analyze its impact on customers. For this, we utilize the long-term segmentation framework, since these users are regular users who have been using the system for a long period of time and therefore much is known about their travel behavior due to the requisite quantum of travel data possessed by the agency on these users. The purpose of this analysis is to identify specific long-term segments which may have been affected due to the incident and the heterogeneity among members of these segments which is revealed based on their reaction to the incident.

For the incident presented in Example I, we find that the affected customers included
a total of 169 long-term customers. Their distribution among segments (Figure 5-3) shows an uneven trend, with a clear majority of these users belonging to long-term segment A. As we recall, customers who belong to segment A have the highest trip intensity and also a high deviation in the number of trips. This combined with the fact that these users travel to a large number of unique stations and possess a relatively low value of $\mu_{Homebased}$ strongly indicates that these users might regularly rely on other modes of travel and are fairly aware of the location. While these users might be office goers, their travel pattern indicates high temporal flexibility in their trips. Another independent study which analyzed this incident showed that there were a considerable number of customers who would enter-exit-reenter the affected stations repeatedly during the incident (Tianyou et al., 2018). While this shows that there might've been confusion among the users, it also illustrates that some of these users might have had the temporal flexibility to check whether service had been restored. Another observation from the same study showed that there were some users who did not travel using MTR on the day of the incident after initially visiting the impacted stations, and this might be due to the fact that they have access to other modes, as inferred from their segment’s characteristics as presented earlier. Although it might be noted that these users also display considerable deviation in displacement, thus they some times do not return using MTR’s service (or don’t make a return trip at all).

Post hoc analysis may also help in developing a better understanding of the customers that form these segments and therefore can help in exposing the heterogeneity in travel behavior among users within these long-term segments. Although majority of the long-term customers affected by this incident belong to Segment A, we further explore the heterogeneity among these customers (within the segment) in response to the incident. We observe the value of mean distance ($\sigma_{Distance}$) split on the basis of whether these users returned to affected stations later on the day of the incident, after the service was restored back to its normalcy. This may help to infer users within segment A who prefer to return back versus those who do not and therefore this provides useful information about the different types of travel behavior exhib-
Figure 5-3: Distribution of long-term users who were affected due to the incident

Figure 5-4: Comparison of mean distance of users who return to incident location to users who do not
ited within the segment and the characteristic features of users who exhibit varying behavior. Figure 5-4 shows that users whose mean distance is higher are more likely to return back using MTR’s service to the affected stations, as compared to those with lower value of mean distance. While this may seem intuitive, it is important to note that users with a high value of mean distance may travel on other parts of the network rather than returning on the same (and affected) portion of the network, and therefore this result is non obvious.

When a similar incident occurs in the future, the agency could provide additional information to segment A users with high values of mean distance by letting them know when the incident is resolved, on a priority basis, since these users are more likely to return back using MTR’s service on the same day. Ensuring the provision of such information may lead to better customer satisfaction and would ensure users don’t shift modes fearing that incident might not have been resolved.

5.4 Discussion

In this chapter, we developed a personalized information provision framework, which could help to provide targeted information to users. While the specific case of information provision during incidents was used to illustrate the efficacy of the framework presented here, the same could be extended to other types of information as well, by changing the filtering process. While for providing targeted information on incidents we did not distinguish between the long-term and short-term users, this may not hold true while in the case of providing information related to promotions. Here, we may filter users based on the total distance they travel, for example, to incentivize long distance commuters to switch to MTR. Clearly, this discussion highlights the potential that lies in utilizing the segmentation scheme as a building block that could effectively be used to provide personalized information.

The level of importance could be chosen based on the type of information being presented. Although a level of importance was assumed to obtain the key metrics like ECR and SRR in the examples presented in this chapter, the choice of level of
importance would ultimately reside with the agency. The agency could also choose initial values, only to be corrected by active feedback from users. For incidents in particular, the severity of the incident has a bearing over the choice of the level of importance and hence a more severe incident would have a higher level of importance and consequently higher values of $n_1$ and $n_2$ might be chosen to ensure that all potential beneficiaries of the information may be reached and in this case the issue of spam might not be very important.

Another aspect to consider would be the content of information to be provided. What would be a meaningful delay to mention as part of the message to the customers in the event of a disruption? Clearly, the answer to this question would depend on how certain the agency is with respect to the information to be provided. One way to do so would be to look at past incident records. The agency could analyze the past data on the incidents which have occurred to figure the details of the specific information to be provided. Moreover, for planning alternate arrangements, the agency could look at the characteristics of the segments of the users who might be affected—for instance, what other locations do these users travel to?; do they have access to alternative modes?; are they expected to multiple trips using the service on a given day?—and this could clearly help to improve customer experience in the face of a service disruption.

To help the agency to quickly identify the users who would be impacted as a result of the incident and therefore to facilitate information provision, a prototype dashboard (Figure 5-5) was created in this research. This dashboard allows the agency to select users that belong to relevant short-term spatial and temporal segments and finally to export their details which can be used in a downstream information dispatching tool.

The same concept could be extended in general to the case when agencies might want to observe differences in certain metrics (may be cluster features) among the various segments and therefore a generic dashboard prototype was created for long-term cluster features (Figure 5-6).
(a) Selecting the relevant spatial segments

(b) Selecting the relevant temporal segments

(c) Obtaining a list of users with specific information to export

Figure 5-5: Dashboard to select users who would be provided targeted information in the event of an incident
Figure 5-6: Dashboard for customer metrics (segmentation features)
Chapter 6

Before-and-after analysis: South Island Line

In this chapter, we shall observe and interpret the ridership changes that occur upon a major network expansion, in this case—the addition of a new line to the network. The segmentation structure identified in this study could form the lenses through which the travel behavior of users could be observed before and after such a change in the network. While we restrict ourselves to the discussion of the impact of the addition of a new line, namely the South Island Line, this chapter provides a concrete set of steps which could be utilized for observing the changes in ridership due to other changes as well—both intrinsic and extrinsic to the transit system.

6.1 Overview of the South Island Line (SIL)

MTR added a new line, named the “South Island Line” towards the end of the year 2016 (Figure 6-1), and since its operationalization several riders have been its beneficiaries. In this section, we would explore the trends of the long-term users who have begun using this new line, and more importantly analyze how the travel pattern of these users changed after the line became operational.
6.2 Analysis Framework

As mentioned previously the analysis presented in this study focuses on long-term users. The prime reason for this is that we wish to track the changes in behavior of these users upon the operationalization of the new line. Since the line opened in December 2016, we select one month before this, namely November 2016, to be indicative of the 'before' state. For the 'after' state, the month of March 2017 was chosen to ensure that any specific 'one-off' trips to or from stations that lie on the SIL (due to initial curiosity, discounts, etc.) do not affect the analysis. So, to summarize, we compare the changes by analyzing data from the following two months—November 2016 and March 2017—corresponding to before and after the network expansion event.

To analyze the changes we shall use the segments and features, identified as part of this study. A major change in the network incapacitates the applicability of short-term segmentation solution due to the addition of a new zone (i.e. the addition of a new feature) and, hence, we shall avoid the use of short-term segments inferred from March 2017, however, the short-term segments from November 2016 and the
short-term features may still be utilized in the analysis. More importantly, long-term segments are less likely to be impacted by such changes and, therefore, could be utilized if found temporally stable.

Based on the 2017 dataset, temporal stability was calculated for the long-term segments, to ensure that the segmentation structure was applicable. A Correspondence Score of around 90% was found for the January 2017 dataset (compared to January 2015 dataset), and this result indicated that indeed the long-term segmentation solution were unaffected by the addition of a new line and could be used for comparison in relation to this network expansion event. The population chosen for this comparison is the intersection of long-term users in 2016 and the set of long-term users in 2017 (January to March; data only up to March 2017 was available during this experiment). It was found that a total of 522,165 users form the population of long term users could be analyzed on this basis. Among this population of long-term users, only 35,911 users are found to make at least one trip to or from a station that lied on the SIL. Therefore, a comparatively small proportion (roughly 7.6%) of these users utilized the SIL in March 2017. Figure 6-2 shows the frequency distribution of the

Figure 6-2: Utilization of stations on the SIL for trips by long-term users
number of trips made to or from the SIL for long-term users (users with no trips have not been displayed on the plot). The median number of trips was found to be 2 and the average number of trips was found to be 6.87, with a standard deviation of 11.16 trips. We split this population at the 85th percentile, which roughly corresponds to 15 trips, in to two groups: 'high' - users who exhibit a high intensity (15 or more trips) of usage of the SIL and 'low' - users who exhibit a low intensity (fewer than 15 trips) of usage of SIL. The former set consists of a total of 4,859 users and the latter set contains 31,052 users. This allows us to analyze the top 15 percentile users of this new service (long-term users) separately. All the analyses presented in this chapter pertains to weekday travel.

6.3 Insights from short-term segments

Although we primarily utilize the long-term segmentation results in this application, we could utilize the short-term segments (November 2016) to identify the characteristics of users who adopted the SIL. Figure 6-3 shows the temporal short-term segments (for November 2016) of users who made at least one trip to or from the stations on the SIL in March 2017, split in terms of whether their usage of the SIL is low or high. It is evident that the two distributions are very similar, and therefore this indicates that a user's temporal segment in November 2016 is not have a associated with the usage of SIL in March 2017. Nevertheless, it is found that the dominant temporal segments of users who made trips to the SIL are T8, T5 and T1, in decreasing order of frequency. Recall that temporal segments T1 and T8 are characteristic office goers. While users in segment T1 usually travel between 9 AM and 10 AM in the morning, users in segment T8 usually travel a little earlier, around 8 AM to 9 AM. While a subset of these users may return early (around 4 PM to 5 PM), others may return around 8 PM. In contrast, users who belong to segment T5 have a large spatial spread, and are likely to engage in multiple trips during the day, thereby indicating a non-office goer type of activity, or alternatively office-goers with multiple trips during
Following the analysis of temporal short-term segments, we direct our focus to spatial short-term segments. Figure 6-4 shows the spatial short-term segments (for November 2016) of users who made at least one trip to or from the stations on the SIL in March 2017, split in terms of whether their usage of the SIL was low or high. It can be observed that unlike short-term temporal segments, there is a distinction among the spatial segments based on their actual usage of the SIL. In other words, the spatial short-term segment of users indeed is associated with their usage of SIL in March 2017. For users with high usage of SIL, the spatial segments S15, S8 and S10 were found to be dominant, in decreasing order of frequency. Customers that belong to the aforementioned spatial short-term segments are those that predominantly utilized (in November 2016) stations that lie around the central and eastern part of the Hong Kong island, such as Wan Chai, Sheung Wan, Causeway Bay, Central, Admiralty, Chai Wan and Shau Kei Wan. Customers who exhibit lower utilization of the stations (for entry or exit) may also belong to spatial segment S10, who use stations on the eastern part of the Hong Kong island, however, they are much more likely to belong to segments S3 and S7. Users that belong to segment S3 are those who primarily utilize stations that lie on the on the KTL line, west of Kowloon Tong such as Kwun Tong, Kowloon Bay, Wong Tai Sin, Lam Tin and Choi Hung. Users that belong to segment S7 primarily utilize the stations that lie on West Rail Line such as Tuen Mun and other stations that lie west of Kam Sheung Road station. Therefore, we observe that users who exhibit higher usage of SIL previously utilized stations that lie closer to SIL as compared to users who exhibit lower usage of SIL.

6.4 Insights from long-term segments

Following the analysis of the short-term segments, we now utilize the long-term segments. This is applicable for both before and after the network expansion and therefore allows for direct comparisons. Figure 6-5 shows the long-term segments in March 2017 and in November 2016, split in terms of the usage of the SIL. Users who utilize
Figure 6-3: November 2016 temporal short-term segments for long-term users who utilize SIL in March 2017
Figure 6-4: November 2016 spatial short-term segments for long-term users who utilize SIL in March 2017
the SIL, whether of low or high usage, predominantly belong to Groups A, G and C. Moreover, regardless of the utilization of SIL, some common trends are observed, these are as follows, Groups A and G increase in size whereas Groups B, E, F and G reduce in size after SIL is operationalized. However, segments C and D exhibit opposite trend, while they increase in size for low usage group, they reduce in size for the high usage group. Moreover, the magnitude of these size changes also varies across the low and high usage groups. The size changes for long-term segments is appreciably large, with the largest changes being the increase in the size of Group G and A, and reduction in size of Groups E and G. For users with low usage, the largest changes in size occur for the reduction in size of Group H and increase in size of Group G.

To aid in this analysis, Figure 6-6 is created to observe the predominant shifts (top ten) in long-term segments that occur between November 2016 and March 2017, split based on the usage of SIL. Now we analyze the shifts that are distinct among low usage and high usage groups. While most shifts are predominant in both cases, which we shall discuss shortly, we observe two unique shifts for high SIL usage users—for some users, a shift from Group E to Group A is observed and, hence, for these users there is a appreciable increase in trip intensity. Group E users have the lowest trip intensity, and based on their characteristics (as discussed in Chapter 4), they may rely on other modes. Hence, these long-term users may be inferred to have shifted from the use of other mode to MTR. Another shift is a shift from Group E to Group H. This would imply a slight increase in the number of trips, however all trips start or end at the inferred home location. This would imply that these users were using another mode for one way of their journey, and now use MTR for both ways.

Overall, we find most long-term users who access the stations on the SIL for entry or exit predominantly belong to Groups A, G and C. Moreover, the common shifts are Group H to Group A, Group C to Group A, and Group C to Group G. Groups A and G have the highest trip intensities, and therefore it can be inferred that long-term users who used SIL in March 2017 are found to have an increase in number of trips. For Groups A, C and G, the time of completion of the first trip is around 9 AM to
10 AM, and the time of completion of the last trip is relatively later, around 7 PM to 8 PM, and therefore the users in these segments resemble office goers. It is useful to note that there are a large number of shifts from Group H, which is a group with very low trip intensity, and therefore make few trips.

In summary, for most long-term users who utilize the SIL, the overall usage of the network has gone up (higher trip intensity). However, the specific changes are different for users with high usage of SIL versus low usage of SIL. For users whose usage is high, two unique shifts can be observed, which indicate that a subset of these users have increased the number of trips they make using MTR’s service and also that they travel solely among a few ODs and utilize MTR for both legs of their journeys.

### 6.5 Insights from home station inference

Chapter 3 describes the methodology adopted in this study to infer the home station of a particular user. This technique could help to identify the location where a particular user usually enters or exits the MTR network, and the inference procedure could prove a valuable technique for analyzing changes.

Figure 6-7 shows the proportion of users versus the number of trips, and is split based on whether their inferred home station in March 2017 lies on the South Island Line or not. It can be observed that for users whose home station lies on South Island Line are more likely to travel more to or from stations that lie on the South Island Line. While this result seems to be fairly intuitive, it provides a motivation to do the following analyses: analyzing the ridership changes for customers whose home station lies on SIL versus other customers and analyzing the previous home stations of users who now use the SIL.

For the latter analysis, we consider the distribution of customers across home stations in November 2016, and then look at users whose home station has shifted to SIL in March 2017. Based on Figure 6-8, it was found that the following five stations predominantly constituted the previous home station (in November 2016) for users
(a) Users with high usage of SIL

(b) Users with low usage of SIL

Figure 6-5: Long-term segments for long-term users (November 2016 and March 2017)
Figure 6-6: Shifts (transitions) within long-term segments (November 2016 and March 2017) split based on low or high usage of SIL
who started utilizing the stations on SIL: Causeway Bay, Hong Kong, Admiralty, Hung Hom and Wan Chai. Clearly, most of these stations lie along the central part of the Hong Kong island (see Figure 6-9), and therefore we can infer that long-term users who started utilizing the SIL line would access MTR’s network at stations located on the central part of the Hong Kong island previously (predominantly).

Figure 6-10 shows the transitions for long-term segments between the months of November 2016 and March 2017. Here, instead of splitting on the basis of usage of SIL, the transitions are split based on user’s home station (on SIL versus not on SIL). For a total of 2006 users, their home station had shifted to SIL, whereas for the remaining 33,905 users it had not shifted to SIL. It may be observed that the dominant transitions are almost similar to those observed in Figure 6-6, which may be due to the fact that users whose home station is on SIL are also the ones who exhibit a high usage, and hence any interpretation of shifts would be similar to the discussion in the previous section.

Following the above analysis, we now observe only those users who exhibit high usage of the SIL, and see the change in spatial spread of their travel pattern before-and-after the SIL started operating, and split it based on whether the user’s home station lies on the SIL or not (in March 2017). We observe that among the users with high
Figure 6-8: Number of customers versus inferred home station of users in November 2016 split by whether their inferred home station lies on SIL in March 2017 (Station codes in Table A.1)

Figure 6-9: Most frequent inferred home stations (November 2016)
Transitions in long-term segments (home on SIL)

(a) Users with inferred home station on SIL

Transitions in long-term segments (home not on SIL)

(b) Users with inferred home station not on SIL

Figure 6-10: Shifts (transitions) within long-term segments (November 2016 and March 2017) split based on whether inferred home station lies on SIL or not
Figure 6-11: Average SP values for long-term users (November 2016 and March 2017)

Table 6.1: Spatial-spread based segment shifts for users whose home station is not on the SIL in the month of March 2017

<table>
<thead>
<tr>
<th>Spatial-spread</th>
<th>March</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LS1</td>
<td>LS2</td>
<td>LS3</td>
<td>LS4</td>
</tr>
<tr>
<td>November</td>
<td>161</td>
<td>356</td>
<td>90</td>
<td>54</td>
</tr>
<tr>
<td>LS2</td>
<td>211</td>
<td>755</td>
<td>305</td>
<td>65</td>
</tr>
<tr>
<td>LS3</td>
<td>22</td>
<td>165</td>
<td>365</td>
<td>1</td>
</tr>
<tr>
<td>LS4</td>
<td>89</td>
<td>176</td>
<td>17</td>
<td>86</td>
</tr>
</tbody>
</table>
Figure 6-12: Average TP values for long-term users (November 2016 and March 2017)

Table 6.2: Spatial-spread based segment shifts for users whose home station is on the SIL in the month of March 2017

<table>
<thead>
<tr>
<th>Spatial-spread</th>
<th>March</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LS1</td>
<td>LS2</td>
<td>LS3</td>
<td>LS4</td>
</tr>
<tr>
<td>LS1</td>
<td>154</td>
<td>174</td>
<td>29</td>
<td>68</td>
</tr>
<tr>
<td>LS2</td>
<td>213</td>
<td>402</td>
<td>101</td>
<td>102</td>
</tr>
<tr>
<td>LS3</td>
<td>31</td>
<td>122</td>
<td>243</td>
<td>5</td>
</tr>
<tr>
<td>LS4</td>
<td>89</td>
<td>98</td>
<td>11</td>
<td>99</td>
</tr>
</tbody>
</table>
utilization of the SIL, the home station of 1,941 users had shifted to SIL, whereas the home station of 2,918 remaining users had not shifted to the SIL. We now utilize the spatial spread segments to analyze the change in absolute spatial travel pattern before and after the network expansion. Recall that the order of spatial spread follows LS4 < LS1 < LS2 < LS3; therefore spatial spread segment LS3 has the highest spatial spread and spatial spread segment LS4 has the lowest spatial spread. Hence spatial spread increases for the following shifts: LS4 to LS1 or LS2 or LS3, LS1 to LS2 or LS3, and LS2 to LS3.

Table 6.1 shows spatial-spread based segment shifts for users whose home station is not on the SIL in the month of March 2017. It can be observed that spatial spread increased for a total of 1033 users, or roughly 35%, whereas from Table 6.2, which shows spatial-spread based segment shifts for users whose home station is on the SIL in the month of March 2017, the corresponding number is 502 users, or roughly 26%. Therefore, it can be observed that users whose usage of SIL is high and have had their home station shift to SIL are less likely to have had an increase in their spatial spread, as compared to users whose home station has not shifted to the SIL. This implies that some users, the former, have started using the SIL as the point of entry or exit and have replaced their earlier point of entry or exit, whereas for other users, the latter, the operationalization of SIL has caused them to travel to more locations consistently, and these new trips are to or from the stations on the SIL.

6.6 Insights based on short-term features

Although the short-term segments may not used without re-training (in March 2017), the short-term features may be utilized to compare and contrast the changes in temporal and spatial travel patterns after the SIL was operationalized. Figures 6-11 and 6-12 show the average SP and average TP values for November 2016 and March 2017. It could be observed that both average SP and average TP values have undergone a higher magnitude of change for users with high usage of the SIL. To observe these changes in greater detail, the difference in average SP and average TP values are plot
From Figure 6-13, it can be observed that the changes in average SP values are minimal for users with low usage of SIL. On the other hand, there are appreciable changes for users with high usage of the SIL. The predominant changes are: average SP values for zones 16 and 14 have reduced drastically, therefore the utilization (for entry or exit) of stations that lie on the central part of Hong Kong has reduced. Similarly, a reduction was observed for zone 10 (Hung Hom) and zone 15 (western parts of Island line). A noticeable increase was found for zone 13 (Tsim Sha Tsui).
From Figure 6-14, it can be observed that similar to the spatial case, the average TP values have changed very little for users with low usage of SIL, with the most noticeable pattern being that the TP values have reduced slightly for hours 6 and 7, and have increased for the rest of the day until evening. This may be interpreted as a slight shift in the period of the day when trips are made, and perhaps the tendency for more number of trips scattered around the day. For users with high usage of the SIL, the change is more pronounced. The probability of early morning trips (6 AM - 7 AM) has greatly reduced in favor of trips in the morning peak period (8 AM - 10 AM). Similarly, the likelihood of trips in the evening (5 PM - 8 PM) and late night (9 PM - 11 PM) has greatly increased. This indicates that these users now make trips in the evening which they were previously unlikely to make, and that the trips in the morning have shifted towards the morning peak period. Hence, from the above analysis we observe that there was a appreciable change in the spatio-temporal travel pattern of users who make a high number of trips that begin or end at a station on the SIL, which is distinct from users who make fewer such trips.

6.7 Discussion

The analysis presented in this chapter allowed us to look at the changes in ridership that occurred before and after the operationalization of a new line—the South Island Line. Each component of the segmentation scheme presented in this study was utilized to yield insights, which could help the agency to understand and interpret the ridership changes. While we restricted ourselves to a single network expansion event, the above analysis technique is fairly generalizable, and in practice could be used to analyze other such changes, both intrinsic and extrinsic to MTR.
Chapter 7

Smart card attrition

7.1 Attrition in the context of this study

Based on the brief discussion in Chapter 2, we realize that it would be beneficial for the transit agency to analyze attrition. For instance, by understanding the underlying characteristics of the customers who leave the system and through inferring their needs, the transit agency might try to address the concerns of these customers. Moreover, through targeted surveys, the transit agency may collect useful feedback from these customers. Towards this goal, it is deemed beneficial to predict a customer’s tendency to stop using the service before they may actually do so, so that the agency could take steps to prevent it. However, since we lacked personal information of users and concrete fare product information in the course of this study, we tackle the problem of predicting smart card attrition instead.

Although, the problem of relevance to the transit agency is the prediction of customer attrition and not card attrition, we still try to formulate the problem of predicting card attrition. This is because, if in the future, transit smart cards may be linked to specific users, then the problem of predicting customer attrition and the problem of predicting card attrition would be similar (provided each user possesses and uses only a single transit smart card), and therefore the methodology identified here could be suitably modified to be used for addressing the customer attrition prediction problem in future.
Despite targeting card attrition as opposed to customer attrition, there are still issues which need to be addressed, and these are:

- Since in reality an individual user may possess and/or use multiple transit smart cards, we adopted reasonable safeguards to try to ensure that the cards chosen in the dataset for attrition prediction have a high likelihood of being their owner’s primary (and ideally only) smart card, in order to minimize the errors in the model. For this, only those cards were used in the dataset which demonstrate frequent usage for at least a year. We recognize that long-term users satisfy the above criterion, and therefore only long-term users (those who attrition and those who don’t) are used in the dataset. Through this we believe that the likelihood of the cards being their respective owner’s primary (or only) fare card is increased.

- The smart card dataset utilized in this study contains anonymized identification numbers, and therefore to ensure that the same cards are tracked longitudinally, the following steps were taken:
  - Only those cards were included where the card type did not change throughout their usage. Therefore cards that show transitions, for instance of the type: Adult to Senior Citizen, were not included in the dataset for building the prediction model.
  - The smart card database had a transaction ID associated with each use of the card. For ensuring card IDs were not reassigned, only those cards were retained where the transaction IDs remained non-decreasing with time.

- Given that we do not have information on the users who possess the cards, we define attrition under a rather narrow definition:

  In this study, we define attrition to be the event when the usage of a long-term user card reduces such that the value of Active falls below the value of 0.5. Note that once this happens, the user card cannot
be assigned a long-term segment. Moreover, after this, the card usage should not increase, hence value of Active should remain below 0.5, and at the same time it should be consistently above 0.

This ensures, to an extent, that cards which simply disappeared from the database due to factors such as loss of smart card, or other technical issues are not included. Hence, although the above definition of attrition may not be rigorous, this allows us to formulate the problem under the given constraints (data).

In an effort to train a prediction model, a dataset was created for each month which contained both cards which attrition in that particular month and cards that do not attrition in the year of observation. The cards that show attrition in the base dataset consisted of all cards which were deemed to be long-term user cards in the year 2015, however were not so anymore in the year 2016. Figure 7-1 shows two separate instances of long-term user cards’ Active values for the year 2016. While these cards were long-term user cards in 2015, and therefore their Active value was consistently above the 0.5 mark in 2015, in the year 2016 this value dropped below 0.5, and therefore did attrition in the year 2016.

7.2 Problem formulation

The customer attrition problem is formulated in the following manner:

1. All long-term user smart cards as determined from the 2015 dataset, including but not restricted to the base sample for creation of long-term segments, were considered for observation of customer attrition. Smart cards whose usage dropped to zero (unobserved in the dataset) in the year 2016 were not considered for issues stated in the previous section.

2. A binary indicator variable called attrition was created and its value set to 1 for cards whose Active value dropped below 0.5, and stayed consistently below
Figure 7-1: Two long-term user cards (2015) which show reduction in usage in 2016
0.5 (for nine months) and its value was set to 0 for others.

3. To predict whether a card would attrition, we would need to predict the value of dependent variable attrition. The month when the Active value first dropped below 0.5, for a particular card, was termed the month of attrition. The symbol $N$ would denote the month of attrition. And therefore, to predict attrition, data equivalent to two prior months before month of attrition was utilized. For this, we utilized the long-term features corresponding to month $N - 1$ and month $N - 2$.

4. Since from the above assumptions, we start predicting cards that attrition starting in March 2016, this effectively implies that these cards had an Active value less than 0.5 for the next nine months, i.e. until the year 2016. For cards that did attrition in April 2016, this would imply these cards had an Active value less than 0.5 (and above 0) in each month up to January 2017. Since we had data only up to March 2017 at the time of this study, we trained customer attrition models for the months of March 2016 to June 2016.

5. The prediction dataset would be imbalanced since the number of cards that attrition would be significantly outnumbered by those that do not attrition. To reduce the computation time, and to mimic a real dataset for statistical reasons, the cards that did attrition in a given month and roughly five times of this—cards that did not attrition in that year—were included in the dataset for training.

The data was split into training and test sets, and then the models were trained and validated. We utilized a few different techniques in this study. The first of these was Classification and Regression Trees (CART), which we used for what was a classification task in this case. The two classes here are attrition and not attrition. The details of this technique can be found in Breiman et al. (1984). The second technique used in this study was the binary logit model, and the details of the technique could be found in Cox (1958). The third technique utilized was the penalized logistic regression de-
developed in the study by Friedman et al. (2010), and the details of the implementation could be found in their paper. A multilayer perceptron was also utilized.

### 7.3 Predicting attrition

Based on the assumptions and the specific formulation of the problem as presented in the previous sections, we now look at predictive models that could be utilized to predict attrition. Since it may be assumed that cards that attrition in any given month of the year may possess unique characteristics due to seasonal variations in travel patterns across the year, a predictive model was built that would be utilized for each month of possible prediction. Therefore, unique predictive models for each month were created for four months—March, April, May and June 2016. A training dataset was created for each of these months for the creation of month-specific predictive models for detecting attrition. Following this, a single model was created for the entire dataset spanning all the months considered in this study. For this, the month specific training datasets were combined. The results for both types of models have been presented in the following subsections. Figure 7-2 shows the number of cards that attrition in each month, for the months considered in this study.

![Figure 7-2: Number of long-term user cards which attrition in each month](image)

Figure 7-2: Number of long-term user cards which attrition in each month
7.3.1 Month-specific attrition model

For predicting the month-specific attrition models, a classification tree algorithm was utilized. The data was split into a 80:20 ratio for the training and test sets respectively. The results for each month have been presented in Figures 7-3, 7-4, 7-5 and 7-6. The package rpart was utilized in R to create the aforementioned models, with minbucket parameter set to 25, implying that the minimum number of observations at each leaf node of the decision tree was 25, and this worked well to ensure that the model was not overfit. Across the models, it is found that the variables Range and Active play a crucial role in distinguishing cards that attrition and those that do not.

From Table 7.1, it may be observed that the accuracy values are consistently

![Decision tree plot](a) Decision tree plot

![ROC plot](b) ROC plot

Figure 7-3: Decision tree solution for March 2016 data

![Decision tree plot](a) Decision tree plot

![ROC plot](b) ROC plot

Figure 7-4: Decision tree solution for April 2016 data
above 98% for all models, which is a good sign except that one must look out for 
_curse of imbalance_ that clearly plagues this dataset (like most other commonly found 
attrition datasets), and therefore sensitivity, which is defined as the proportion of 
cards that attrition and are predicted to do so, is also calculated for each model. 
Sensitivity might be an important value to capture since in the problem of attrition, 
it may be more crucial to correctly identify cards that may attrition at the expense of 
erroneously classifying a few cards which might not attrition (thereby reducing accu-
ricy). The sensitivity remains above 90% for each model, and is hence an indication 
that predictive attrition models indeed could prove helpful in detecting cards which 
would attrition in the near future. The next subsection discusses the creation of a 
single attrition model for all months considered in this study.
### Table 7.1: Results for monthly models

<table>
<thead>
<tr>
<th></th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>0</td>
<td>1760</td>
<td>1487</td>
<td>1407</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>16</td>
<td>338</td>
<td>277</td>
</tr>
<tr>
<td></td>
<td></td>
<td>256</td>
<td>11</td>
<td>306</td>
</tr>
<tr>
<td>Accuracy</td>
<td>98.73%</td>
<td>98.88%</td>
<td>98.52%</td>
<td>98.11%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>95.48%</td>
<td>93.26%</td>
<td>91.10%</td>
<td>96.53%</td>
</tr>
</tbody>
</table>

### Table 7.2: Logit model estimation results

| Coefficient    | Estimate | Std. Error | t value | Pr(>|t|) | Significance |
|----------------|----------|------------|---------|-------|--------------|
| (Intercept)    | 3.05E+00 | 6.82E-02  | 4.46    | <2e-16| ***          |
| $\mu_{\text{RI}}$ [N=1] | -6.54E+00 | 1.24E-01 | -52.803 | <2e-16| ***          |
| $\phi_{\text{Trips-N}}$ [N=1] | -8.32E-01 | 1.25E-01 | -6.494 | 5.34E-11| ***          |
| $\sigma_{\text{Trips-N}}$ [N=1] | 2.42E-02 | 9.01E-02 | 2.692 | 0.00799| **           |
| $\phi_{\text{Range-N}}$ [N=1] | 6.84E-01 | 1.61E-01 | 4.258 | 2.06E-05| ***          |
| $\sigma_{\text{Prox-N}}$ [N=1] | 3.60E-05 | 3.70E-06 | 9.571 | <2e-16| ***          |
| $\mu_{\text{Last-N}}$ [N=1] | 1.68E-05 | 3.59E-06 | 4.746 | 2.07E-08| ***          |
| $\sigma_{\text{Last-N}}$ [N=1] | -3.44E-05 | 3.57E-06 | -9.641 | <2e-16| ***          |
| max $RI$ [N=1] | 1.89E-05 | 4.07E-06 | 4.605 | 4.13E-06| ***          |
| $\mu_{\text{R}}$ [N=1] | -2.71E-01 | 5.76E-01 | -0.47 | 0.63656|             |
| $\sigma_{\text{R}}$ [N=1] | 8.15E-02 | 6.64E-03 | 12.276 | <2e-16| ***          |
| $\mu_{\text{R}}$ [N=1] | -5.11E-01 | 7.48E-01 | -6.824 | 8.84E-12| ***          |
| $\sigma_{\text{R}}$ [N=1] | -4.77E-02 | 1.13E-02 | -4.219 | 2.45E-05| ***          |
| $\phi_{\text{Symmetric-N}}$ [N=1] | 1.78E+00 | 1.05E+01 | 11.506 | <2e-16| ***          |
| $\phi_{\text{Homebased-N}}$ [N=1] | -1.27E-01 | 1.20E-01 | -1.056 | 0.29101|           |
| $\sigma_{\text{R}}$ [N=1] | -5.09E-01 | 5.06E-01 | -9.622 | <2e-16| ***          |
| $\sigma_{\text{Homebased-N}}$ [N=1] | -8.55E-01 | 1.91E-01 | -4.468 | 7.90E-06| ***          |
| $\phi_{\text{Displacement-N}}$ [N=1] | -1.63E+00 | 8.90E-02 | -18.316| <2e-16| ***          |
| $\phi_{\text{Distance-N}}$ [N=1] | -1.01E-01 | 4.28E-02 | -2.465 | 0.01379|            |
| $\phi_{\text{Unique-N}}$ [N=1] | 2.46E-01 | 1.29E-01 | 1.911 | 0.05598|           |
| $\phi_{\text{Unique-N}}$ [N=1] | -1.41E+00 | 8.85E-02 | -15.979 | <2e-16| ***          |
| $\phi_{\text{Trips-N}}$ [N=2] | -4.09E+00 | 1.18E+00 | -34.182 | <2e-16| ***          |
| $\phi_{\text{Trips-N}}$ [N=2] | -1.31E+00 | 1.29E+01 | -10.157 | <2e-16| ***          |
| $\phi_{\text{Trips-N}}$ [N=2] | -4.06E+01 | 8.54E-02 | -5.612 | 2.00E-08| ***          |
| $\phi_{\text{Range-N}}$ [N=2] | -2.55E+00 | 1.90E-01 | -13.322 | <2e-16| ***          |
| $\phi_{\text{Range-N}}$ [N=2] | -2.39E+00 | 3.55E-01 | -6.592 | 0.01454|           |
| $\phi_{\text{Last-N}}$ [N=2] | -2.44E+05 | 3.05E+00 | -6.023 | 1.71E-09| ***          |
| $\phi_{\text{RI}}$ [N=2] | -5.34E+00 | 4.05E+01 | -13.184 | <2e-16| ***          |
| $\phi_{\text{RI}}$ [N=2] | 2.93E+00 | 5.68E-01 | 5.158 | 2.49E-07| ***          |
| $\phi_{\text{RI}}$ [N=2] | 1.01E-01 | 6.61E-03 | 15.254 | <2e-16| ***          |
| $\phi_{\text{RI}}$ [N=2] | 2.97E+00 | 4.72E-05 | 6.281 | 3.37E-10| ***          |
| $\phi_{\text{RI}}$ [N=2] | -3.17E+00 | 1.12E+02 | -2.822 | 0.09696|            |
| $\phi_{\text{RI}}$ [N=2] | -4.44E-01 | 7.28E-01 | -6.158 | 2.9e-16| ***          |
| $\phi_{\text{RI}}$ [N=2] | 1.66E+00 | 1.53E+01 | 12.763 | <2e-16| ***          |
| $\phi_{\text{RI}}$ [N=2] | -3.33E-01 | 1.19E+00 | -2.851 | 0.00435| **           |
| $\phi_{\text{RI}}$ [N=2] | 8.03E+00 | 1.93E+01 | -4.477 | 7.56E-06| ***          |
| $\phi_{\text{RI}}$ [N=2] | -1.59E+00 | 2.18E+01 | -7.322 | 2.48E-13| ***          |
| $\phi_{\text{Displacement-N}}$ [N=2] | 4.55E-02 | 9.25E-03 | 4.919 | 8.76E-07| ***          |
| $\phi_{\text{Distance-N}}$ [N=2] | 8.46E+03 | 7.96E-03 | 1.010 | 0.31271|            |
| $\phi_{\text{Distance-N}}$ [N=2] | -7.15E+03 | 4.32E+03 | -1.655 | 0.0979| ***          |
| $\phi_{\text{Distance-N}}$ [N=2] | -1.39E+02 | 4.92E+03 | -2.85 | 0.00645| **           |
| $\phi_{\text{Distance-N}}$ [N=2] | 5.14E-01 | 1.30E-01 | 3.97 | 7.19E-05| ***          |
| $\phi_{\text{Distance-N}}$ [N=2] | -1.63E+00 | 8.90E-02 | -18.316 | <2e-16| ***          |
Table 7.3: Results for unified model

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 1760 11</td>
</tr>
<tr>
<td>1</td>
<td>16 338</td>
</tr>
</tbody>
</table>

Accuracy 98.73%
Sensitivity 95.48%

Figure 7-7: ROC curve for the logit model
Table 7.4: Coefficient estimates for penalized logistic regression model

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.55E+01</td>
</tr>
<tr>
<td>Active [N-1]</td>
<td>-6.16E+00</td>
</tr>
<tr>
<td>μ_{Trips} [N-1]</td>
<td>-6.17E-01</td>
</tr>
<tr>
<td>σ_{Trips} [N-1]</td>
<td>4.85E-01</td>
</tr>
<tr>
<td>Range [N-1]</td>
<td>2.93E-05</td>
</tr>
<tr>
<td>σ_{First} [N-1]</td>
<td>1.26E-05</td>
</tr>
<tr>
<td>n_{Last} [N-1]</td>
<td>3.14E-05</td>
</tr>
<tr>
<td>σ_{Last} [N-1]</td>
<td>1.58E-05</td>
</tr>
<tr>
<td>max R_{i,j} [N-1]</td>
<td>8.50E-01</td>
</tr>
<tr>
<td>μ_{Ri,j} [N-1]</td>
<td>-2.63E+00</td>
</tr>
<tr>
<td>σ_{Ri,j} [N-1]</td>
<td>6.82E-02</td>
</tr>
<tr>
<td>n(R_{i,j}) [N-1]</td>
<td>8.96E-02</td>
</tr>
<tr>
<td>μ_{Ri,j} [N-1]</td>
<td>-4.63E+00</td>
</tr>
<tr>
<td>σ_{Ri,j} [N-1]</td>
<td>-4.56E+00</td>
</tr>
<tr>
<td>n(R_{i,j}) [N-1]</td>
<td>-1.05E+00</td>
</tr>
<tr>
<td>μ_{Symmetric} [N-1]</td>
<td>1.13E+00</td>
</tr>
<tr>
<td>σ_{Symmetric} [N-1]</td>
<td>-6.93E-02</td>
</tr>
<tr>
<td>n_{Homebased} [N-1]</td>
<td>-6.42E+01</td>
</tr>
<tr>
<td>μ_{Homebased} [N-1]</td>
<td>-9.93E-01</td>
</tr>
<tr>
<td>μ_{Displacement} [N-1]</td>
<td>2.52E-03</td>
</tr>
<tr>
<td>σ_{Distance} [N-1]</td>
<td>5.48E+03</td>
</tr>
<tr>
<td>n_{Distance} [N-1]</td>
<td>1.33E+00</td>
</tr>
<tr>
<td>n(R_{i,j}) [N-2]</td>
<td>8.96E-02</td>
</tr>
<tr>
<td>max R_{i,j} [N-2]</td>
<td>1.33E+00</td>
</tr>
<tr>
<td>μ_{Ri,j} [N-2]</td>
<td>-3.61E+00</td>
</tr>
<tr>
<td>σ_{Ri,j} [N-2]</td>
<td>-5.95E+00</td>
</tr>
<tr>
<td>n(R_{i,j}) [N-2]</td>
<td>9.54E-04</td>
</tr>
<tr>
<td>μ_{Symmetric} [N-2]</td>
<td>4.87E+00</td>
</tr>
<tr>
<td>σ_{Symmetric} [N-2]</td>
<td>-2.49E+01</td>
</tr>
<tr>
<td>n_{Homebased} [N-2]</td>
<td>-7.05E+01</td>
</tr>
<tr>
<td>μ_{Homebased} [N-2]</td>
<td>-1.38E+00</td>
</tr>
<tr>
<td>μ_{Displacement} [N-2]</td>
<td>3.28E-02</td>
</tr>
<tr>
<td>σ_{Distance} [N-2]</td>
<td>9.94E-03</td>
</tr>
<tr>
<td>n_{Distance} [N-2]</td>
<td>1.48E-02</td>
</tr>
<tr>
<td>μ_{Unique} [N-2]</td>
<td>3.96E-01</td>
</tr>
<tr>
<td>σ_{Unique} [N-2]</td>
<td>-1.41E+00</td>
</tr>
</tbody>
</table>
7.3.2 Unified attrition model

The overall dataset thus contains 4,165 long-term cards which did attrition in the year 2016, and 20,835 cards which did not. With this combined dataset, we now train a classifier that can reasonably predict user cards which would attrition. Similar features, namely long-term clustering feature values for two months prior to attrition, are utilized for training this classifier. An 80-20 split was created, with the larger share chosen to train the classifier and the smaller share chosen to test it. The first model trained used a multilayer perceptron (MLP) for the purposes of this classification task. The configuration chosen had rectified linear unit as the activation function, with four layers (two hidden, with 100 neurons in each hidden layer). Although this model achieved an accuracy of only 54.14%, however a sensitivity of 96.05% was obtained.

To explore the specific factors which determine attrition, the binary logit model was selected as the classifier of choice. For this, the R function glm is invoked, with ‘attrition’ being the variable to be predicted and a class weight of 10 was assigned to all the
positive attrition samples \( \text{attrition} = 1 \). The results of the logit model are presented in Table 7.2. The corresponding accuracy and sensitivity for the test dataset are presented in Table 7.3 and the ROC curve has been shown in Figure 7-7. Clearly, the model shows high degree of accuracy (98.73%), sensitivity (95.48%) and area under curve of 0.8601, and therefore, inspite of the model’s high chi-squared value (13840, Hosmer and Lemeshow goodness of fit test) due to its disaggregate nature, the model is further analyzed, albeit with caution, since shows good predictive accuracy.

While most variables are found to be significant, a penalized binary logistic regression was trained with elastic net penalty, for determining which variables are truly significant predictors. The R package \texttt{glmnet} was used for training the model and five-fold cross validation is used. Misclassification error was the chosen criterion and Figure 7-8 shows the value of misclassification error with respect to values of \( \lambda \), and the value of \( \lambda_{\text{min}} \) was found to be 0.00033. The coefficient values obtained are shown in Table 7.4.

Based on the significant errors and the signs of coefficients of the variables found across the unified models, the following predominant trends of cards which may be prone to attrition are found—the cards show usage of the service over fewer days and the trip intensity reduces. These aforementioned trends indicate a gradual decrease of use of system. This is expected when the card is about to attrition. Moreover, there are fewer trips to or from the inferred home station and trips among fewer OD pairs, and moreover the number of unique stations reduces. While above is a preliminary attempt at interpreting what might happen when a card is about to attrition, having additional sources of data could prove extremely useful in understanding the exact causes of customer attrition.

### 7.4 Potential pitfalls

While the models developed in this chapter show potential for use by the agency, it is important to summarize some pitfalls associated with them:
• We are tracking card attrition, since the data on the actual users is unknown, and although some safeguards are in place to ensure that the cards being tracked are the primary cards of their respective users, it may still be the case that the users of these cards have alternative cards and they have only switched their preferred/primary card.

• The model was developed for just four months, and it would be important to expand the model to all months of the year, when the data is available, to ensure that the results found can be truly extended to an annual model. Therefore although the results look promising, analyzing the annual model might be necessary to proceed further.

• Currently, there is no way to address the issue of cards whose usage drops to zero, and therefore the prediction models may incorrectly predict these cards in new data.

• The lack of causal factors in the model severely limit the use of the models developed in this study for purposes beyond prediction.

This chapter was an attempt to develop a predictive card attrition model. While being severely limited in terms of its potential uses, it is important to understand that the predictive attrition models created in this chapter serve as a proof of concept for future studies which may benefit from the findings presented within this study. With data combined from other sources (such as fare pass data, customer satisfaction survey data, socio economic data, etc.), a lot of error minimizing assumptions used in this study may be relaxed, and therefore the models build with this enriched source of data might have a better enhanced predictive capabilities and interpretive value. Customer satisfaction is a high priority for every agency, and therefore it may be deemed useful to predict when a customer is about to leave the system. In this regard, newer sources such as GPS data from transit smartphone applications may also benefit the prediction of attrition—by looking at user’s travel pattern outside the system, their reliance on alternative modes could be inferred. Similarly, information
from auto-balance refill systems or other membership based systems could provide information on whether a user has changed their residence. Apart from its benefit in detecting and possibly preventing customer attrition, the models developed here could also be utilized in de-noising predictions of fare product usage shifts and related research such as in Stuntz (2018).

While additional analysis might be useful to distinguish among users who voluntarily left the system versus users who were forced to leave the system (due to change in their homes/jobs, etc.), this analysis has been skipped and could be done when additional data sources, as mentioned previously, might be made available. However, it is important to note that if the home station inference as done in this study is validated, it could be utilized to see if the user’s inferred home station changed after they reduced their usage (or just before this happened). Such an analysis could provide useful insights about whether a change in home location is what forced a user to shift travel modes. No causal factors were available to analyze what caused customers to reduce the usage of their primary card. As mentioned previously, in future, it may be useful to tap into other sources of data such as customer feedback surveys.
Chapter 8

Conclusion

This chapter summarizes the segmentation scheme presented in this study and discusses the main findings. It also discusses limitations and key elements we have omitted in the presentation of the study. The chapter concludes with a discussion of research directions for the future. Section 8.1 summarizes the findings in this current study. Section 8.2 describes the main contributions of this study and Section 8.3 presents some of the limitations and the avenues for future research.

8.1 Summary

The research presented in this study used transit smart card data from Hong Kong’s MTR system and accomplished the following:

- A two-tiered segmentation scheme was created to understand the spatial and temporal characteristics of users, and to identify the dominant customer segments that use the Hong Kong MTR system.

- A thorough analysis was conducted to interpret the specific characteristics of the segments identified in this study, and the temporal stability of these segments was computed to ensure that the results were relevant across time periods.

- Using customer segmentation as the building block, it was demonstrated how targeted information could be provided to users who may find it relevant. Illus-
trative examples were used to show how information in the event of an incident (disruption) could be provided in a personalized way.

- The segmentation framework was used to understand the impact of changes in the network, through a before and after study. A specific instance—the operationalization of the South Island Line was chosen as the case study.

- Models for predicting card attrition were developed by using the features created for the purpose of segmentation. It was observed that the models developed in this study could accurately predict attrition in the near future.

8.2 Contributions

The customer segmentation approach presented in this study takes the form of a unique tiered structure, which accounts for the differences in data availability on different types of transit customers. Therefore, a two-tiered segmentation structure is created to capture a customer’s spatial and temporal travel behavior, which helps in drawing meaningful comparisons among the customer segments and to analyze the impact of changes on them. This framework could be a powerful tool in the hands of a transit agency to conduct an in-depth analysis of the ridership changes over time and to understand the needs of different segments. Since the framework only relies on transit smart card data, we claim the methods presented in this study to be highly generalizable to other transit agencies which employ smart card mode of payment.

The customer segmentation framework was utilized to demonstrate meaningful applications. The provision of personalized information in the context of transit was facilitated through the use of the segmentation framework. Moreover, its applicability in analyzing the impact of network changes on the spatial and temporal changes of users was explored through a case study, and an in-depth analysis was made possible due the specific structure of the segmentation scheme. Moreover, the problem of predicting card attrition was explored, and the results, while preliminary, show great promise for further development.
8.3 Limitations and avenues for future research

Some omissions, limitations and avenues of future research are described as follows:

- While the approach presented in this study could be classified as data-driven it should be noted that domain knowledge was utilized during the process especially for the creation of features that were utilized for clustering.

- While long-term segments were only created for users who hold the adult type smart card, long-term segments may be obtained for users who use other card types such as children, senior citizens, etc. The methodology would remain similar to the one adopted for adult smart cards.

- Although weekday travel pattern was analyzed in this study, the framework could be easily extended to weekends. While there exist several ways of creating segments from the data, we suggest separating weekday and weekend data to create separate segments, since this would serve the applications (as illustrated in this study) well.

- While the study relied on adopting a trip-based approach due to lack of data on user's travel on other modes outside the MTR network, if data from all transit agencies is available in the future, it could help in considering the user's specific activities, rather than focusing solely on trips. Moreover, if data from traditional travel surveys would be available, it would help in enriching the dataset, and could help to infer trip purpose, actual destination, actual home location and other such user details.

- The specific application of providing information to users in the event of an incident could not be validated in real-time due to lack of infrastructure to provide such alerts by utilizing the framework presented in this study. The real-time efficacy of this system could be validated in the future.

- For personalized information provision, data from new sources such as user's smartphone application could help in establishing two-way communication, and
to obtain the user’s preferences on the types of information they would like to receive (and other such feedback). Moreover, a recommender system could supplement the present framework to ensure relevant information is provided to customers, and new interfaces such as chatbots could be explored.

- The predictive attrition model developed in this study could be enriched with customer experience surveys. This could help in understanding the specific features of the service which cause users to leave, so that the agency could pursue corrective action.

- Insights from the segmentation framework could be utilized while devising Travel Demand Management schemes in the future for transit.

- Additional data sources (such as iBeacon, WiFi, GPS, etc.) could help to understand the specific route choice of users in the stations, and this information could be integrated within the framework presented in this study (by adding dummy stations, etc.) providing better insight into passenger’s spatial travel behavior.
Appendix A

MTR Network
<table>
<thead>
<tr>
<th>Station Name</th>
<th>Location Code</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central</td>
<td>CEN</td>
<td>1</td>
</tr>
<tr>
<td>Admiralty</td>
<td>ADM</td>
<td>2</td>
</tr>
<tr>
<td>Tsim Sha Tsui</td>
<td>TST</td>
<td>3</td>
</tr>
<tr>
<td>Jordan</td>
<td>JOR</td>
<td>4</td>
</tr>
<tr>
<td>Yau Ma Tei</td>
<td>YMT</td>
<td>5</td>
</tr>
<tr>
<td>Mong Kok</td>
<td>MOK</td>
<td>6</td>
</tr>
<tr>
<td>Shek Kip Mei</td>
<td>SKM</td>
<td>7</td>
</tr>
<tr>
<td>Lok Fu</td>
<td>LOF</td>
<td>9</td>
</tr>
<tr>
<td>Wong Tai Sin</td>
<td>WTS</td>
<td>10</td>
</tr>
<tr>
<td>Diamond Hill</td>
<td>DIH</td>
<td>11</td>
</tr>
<tr>
<td>Choi Hung</td>
<td>CHH</td>
<td>12</td>
</tr>
<tr>
<td>Kowloon Bay</td>
<td>KOB</td>
<td>13</td>
</tr>
<tr>
<td>Ngau Tau Kok</td>
<td>NTK</td>
<td>14</td>
</tr>
<tr>
<td>Kwun Tong</td>
<td>KWT</td>
<td>15</td>
</tr>
<tr>
<td>Prince Edward</td>
<td>PRE</td>
<td>16</td>
</tr>
<tr>
<td>Sham Shui Po</td>
<td>SSP</td>
<td>17</td>
</tr>
<tr>
<td>Cheung Sha Wan</td>
<td>CSW</td>
<td>18</td>
</tr>
<tr>
<td>Lai Chi Kok</td>
<td>LCK</td>
<td>19</td>
</tr>
<tr>
<td>Mei Foo</td>
<td>MEF</td>
<td>20</td>
</tr>
<tr>
<td>Lai King</td>
<td>LAK</td>
<td>21</td>
</tr>
<tr>
<td>Kwai Fong</td>
<td>KWF</td>
<td>22</td>
</tr>
<tr>
<td>Kwai Hing</td>
<td>KWH</td>
<td>23</td>
</tr>
<tr>
<td>Tai Wo Hau</td>
<td>TWH</td>
<td>24</td>
</tr>
<tr>
<td>Tsuen Wan</td>
<td>TSW</td>
<td>25</td>
</tr>
<tr>
<td>Sheung Wan</td>
<td>SHW</td>
<td>26</td>
</tr>
<tr>
<td>Wan Chai</td>
<td>WAC</td>
<td>27</td>
</tr>
<tr>
<td>Causeway Bay</td>
<td>CAB</td>
<td>28</td>
</tr>
<tr>
<td>Tin Hau</td>
<td>THH</td>
<td>29</td>
</tr>
<tr>
<td>Fortress Hill</td>
<td>FOH</td>
<td>30</td>
</tr>
<tr>
<td>North Point</td>
<td>NOP</td>
<td>31</td>
</tr>
<tr>
<td>Quarry Bay</td>
<td>QUB</td>
<td>32</td>
</tr>
<tr>
<td>Tai Koo</td>
<td>TAK</td>
<td>33</td>
</tr>
<tr>
<td>Sai Wan Ho</td>
<td>SWH</td>
<td>34</td>
</tr>
<tr>
<td>Shau Kei Wan</td>
<td>SKW</td>
<td>35</td>
</tr>
<tr>
<td>Heng Fa Chuen</td>
<td>HFC</td>
<td>36</td>
</tr>
<tr>
<td>Chai Wan</td>
<td>CHW</td>
<td>37</td>
</tr>
<tr>
<td>Lam Tin</td>
<td>LAT</td>
<td>38</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>HOK</td>
<td>39</td>
</tr>
<tr>
<td>Kowloon</td>
<td>KOW</td>
<td>40</td>
</tr>
<tr>
<td>Olympic</td>
<td>OLY</td>
<td>41</td>
</tr>
<tr>
<td>Tsing Yi</td>
<td>TSY</td>
<td>42</td>
</tr>
<tr>
<td>Tung Chung</td>
<td>TUC</td>
<td>43</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>HOKA/ITCI</td>
<td>44/61</td>
</tr>
<tr>
<td>Kowloon</td>
<td>KOWA/ITCI</td>
<td>45/62</td>
</tr>
<tr>
<td>Tsing Yi</td>
<td>TSYA</td>
<td>46</td>
</tr>
<tr>
<td>Airport</td>
<td>AIR</td>
<td>47</td>
</tr>
<tr>
<td>Yau Tong</td>
<td>YAT</td>
<td>48</td>
</tr>
<tr>
<td>Tiu Keng Leng</td>
<td>TIK</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station Name</th>
<th>Location Code</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tseung Kwan O</td>
<td>TKO</td>
<td>50</td>
</tr>
<tr>
<td>Hang Hau</td>
<td>HAH</td>
<td>51</td>
</tr>
<tr>
<td>Po Lam</td>
<td>POA</td>
<td>52</td>
</tr>
<tr>
<td>Sunny Bay</td>
<td>SUN</td>
<td>54</td>
</tr>
<tr>
<td>Disneyland Resort</td>
<td>DIS</td>
<td>55</td>
</tr>
<tr>
<td>AsiaWorld Expo</td>
<td>AWE</td>
<td>56</td>
</tr>
<tr>
<td>LOHAS Park</td>
<td>LHP</td>
<td>57</td>
</tr>
<tr>
<td>Hung Hom</td>
<td>HUH</td>
<td>64</td>
</tr>
<tr>
<td>Mong Kok East</td>
<td>M KK</td>
<td>65</td>
</tr>
<tr>
<td>Kowloon Tong</td>
<td>KOT</td>
<td>66</td>
</tr>
<tr>
<td>Tai Wai</td>
<td>TAW</td>
<td>67</td>
</tr>
<tr>
<td>Sha Tin</td>
<td>SHT</td>
<td>68</td>
</tr>
<tr>
<td>Fo Tan</td>
<td>FOT</td>
<td>69</td>
</tr>
<tr>
<td>Racecourse</td>
<td>RAC</td>
<td>70</td>
</tr>
<tr>
<td>University</td>
<td>UNI</td>
<td>71</td>
</tr>
<tr>
<td>Tai Po Market</td>
<td>TAP</td>
<td>72</td>
</tr>
<tr>
<td>Tai Wo</td>
<td>TWO</td>
<td>73</td>
</tr>
<tr>
<td>Fanling</td>
<td>FAN</td>
<td>74</td>
</tr>
<tr>
<td>Sheung Shui</td>
<td>SHS</td>
<td>75</td>
</tr>
<tr>
<td>Lo Wu</td>
<td>LOW</td>
<td>76</td>
</tr>
<tr>
<td>Lok Ma Chau</td>
<td>LMC</td>
<td>78</td>
</tr>
<tr>
<td>East Tsim Sha Tsui</td>
<td>ETS</td>
<td>80</td>
</tr>
<tr>
<td>Sai Ying Pun</td>
<td>SYP</td>
<td>81</td>
</tr>
<tr>
<td>HKU</td>
<td>HKU</td>
<td>82</td>
</tr>
<tr>
<td>Kennedy Town</td>
<td>KET</td>
<td>83</td>
</tr>
<tr>
<td>Ho Man Tin</td>
<td>HOM</td>
<td>84</td>
</tr>
<tr>
<td>Whampoa</td>
<td>WHA</td>
<td>85</td>
</tr>
<tr>
<td>Ocean Park</td>
<td>OCP</td>
<td>86</td>
</tr>
<tr>
<td>Wong Chuk Hang</td>
<td>WCH</td>
<td>87</td>
</tr>
<tr>
<td>Lei Tung</td>
<td>LET</td>
<td>88</td>
</tr>
<tr>
<td>South Horizons</td>
<td>SOH</td>
<td>89</td>
</tr>
<tr>
<td>Che Kung Temple</td>
<td>CKT</td>
<td>96</td>
</tr>
<tr>
<td>Sha Tin Wai</td>
<td>STW</td>
<td>97</td>
</tr>
<tr>
<td>City One</td>
<td>CIO</td>
<td>98</td>
</tr>
<tr>
<td>Shek Mun</td>
<td>SHM</td>
<td>99</td>
</tr>
<tr>
<td>Tai Shui Hang</td>
<td>TSH</td>
<td>100</td>
</tr>
<tr>
<td>Heng On</td>
<td>HEO</td>
<td>101</td>
</tr>
<tr>
<td>Ma On Shan</td>
<td>MOS</td>
<td>102</td>
</tr>
<tr>
<td>Wu Kai Sha</td>
<td>WKS</td>
<td>103</td>
</tr>
<tr>
<td>Austin</td>
<td>AUS</td>
<td>111</td>
</tr>
<tr>
<td>Nam Cheong</td>
<td>NAC</td>
<td>112</td>
</tr>
<tr>
<td>Tsuen Wan West</td>
<td>TWW</td>
<td>114</td>
</tr>
<tr>
<td>Kam Sheung Road</td>
<td>KSR</td>
<td>115</td>
</tr>
<tr>
<td>Yuen Long</td>
<td>YUL</td>
<td>116</td>
</tr>
<tr>
<td>Long Ping</td>
<td>LOP</td>
<td>117</td>
</tr>
<tr>
<td>Tin Shui Wai</td>
<td>TIS</td>
<td>118</td>
</tr>
<tr>
<td>Siu Hong</td>
<td>SIH</td>
<td>119</td>
</tr>
<tr>
<td>Tuen Mun</td>
<td>TUM</td>
<td>120</td>
</tr>
</tbody>
</table>
Figure A-1: MTR network map (Source: MTR, Fall 2017)
Appendix B

Descriptive statistics and density plots of long-term features
Table B.1: Descriptive statistics for the features selected for long-term segmentation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Median</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
<td>0.909</td>
<td>0.885</td>
<td>0.111</td>
</tr>
<tr>
<td>$\mu_{Trips}$</td>
<td>1</td>
<td>11.789</td>
<td>10.789</td>
<td>2</td>
<td>2.046</td>
<td>0.546</td>
</tr>
<tr>
<td>$\sigma_{Trips}$</td>
<td>0</td>
<td>6.057</td>
<td>6.057</td>
<td>0.547</td>
<td>0.577</td>
<td>0.325</td>
</tr>
<tr>
<td>Range</td>
<td>0.451</td>
<td>0.935</td>
<td>0.483</td>
<td>0.903</td>
<td>0.901</td>
<td>0.054</td>
</tr>
<tr>
<td>$\mu_{First}$</td>
<td>20074.818</td>
<td>89476.454</td>
<td>69401.636</td>
<td>34356.885</td>
<td>37530.031</td>
<td>9661.430</td>
</tr>
<tr>
<td>$\sigma_{First}$</td>
<td>1.788</td>
<td>30571.506</td>
<td>30569.718</td>
<td>5773.196</td>
<td>6293.333</td>
<td>4970.941</td>
</tr>
<tr>
<td>$\mu_{Last}$</td>
<td>22212.666</td>
<td>91623.238</td>
<td>69410.571</td>
<td>69036.041</td>
<td>67909.551</td>
<td>8721.802</td>
</tr>
<tr>
<td>$\sigma_{Last}$</td>
<td>3.691</td>
<td>30159.841</td>
<td>30156.149</td>
<td>8123.897</td>
<td>8869.977</td>
<td>5081.647</td>
</tr>
<tr>
<td>max $RI_d$</td>
<td>0.031</td>
<td>1</td>
<td>0.968</td>
<td>0.425</td>
<td>0.422</td>
<td>0.160</td>
</tr>
<tr>
<td>$\mu_{RI_d}$</td>
<td>0.010</td>
<td>1</td>
<td>0.989</td>
<td>0.111</td>
<td>0.146</td>
<td>0.129</td>
</tr>
<tr>
<td>$\sigma_{RI_d}$</td>
<td>0</td>
<td>0.474</td>
<td>0.474</td>
<td>0.127</td>
<td>0.134</td>
<td>0.077</td>
</tr>
<tr>
<td>$\mu_{RI_s}$</td>
<td>20074.818</td>
<td>89476.454</td>
<td>69401.636</td>
<td>34356.885</td>
<td>37530.031</td>
<td>9661.430</td>
</tr>
<tr>
<td>$\sigma_{RI_s}$</td>
<td>1.788</td>
<td>30571.506</td>
<td>30569.718</td>
<td>5773.196</td>
<td>6293.333</td>
<td>4970.941</td>
</tr>
<tr>
<td>$\mu_{Last}$</td>
<td>22212.666</td>
<td>91623.238</td>
<td>69410.571</td>
<td>69036.041</td>
<td>67909.551</td>
<td>8721.802</td>
</tr>
<tr>
<td>$\sigma_{Last}$</td>
<td>3.691</td>
<td>30159.841</td>
<td>30156.149</td>
<td>8123.897</td>
<td>8869.977</td>
<td>5081.647</td>
</tr>
<tr>
<td>max $RI_s$</td>
<td>0.062</td>
<td>0.5</td>
<td>0.4375</td>
<td>0.444</td>
<td>0.424</td>
<td>0.070</td>
</tr>
<tr>
<td>$\mu_{RI_s}$</td>
<td>0.020</td>
<td>0.5</td>
<td>0.479</td>
<td>0.142</td>
<td>0.165</td>
<td>0.102</td>
</tr>
<tr>
<td>$\sigma_{RI_s}$</td>
<td>0</td>
<td>0.244</td>
<td>0.244</td>
<td>0.153</td>
<td>0.148</td>
<td>0.055</td>
</tr>
<tr>
<td>$\mu_{Last}$</td>
<td>22212.666</td>
<td>91623.238</td>
<td>69410.571</td>
<td>69036.041</td>
<td>67909.551</td>
<td>8721.802</td>
</tr>
<tr>
<td>$\sigma_{Last}$</td>
<td>3.691</td>
<td>30159.841</td>
<td>30156.149</td>
<td>8123.897</td>
<td>8869.977</td>
<td>5081.647</td>
</tr>
<tr>
<td>max $RI_s$</td>
<td>0.062</td>
<td>0.5</td>
<td>0.4375</td>
<td>0.444</td>
<td>0.424</td>
<td>0.070</td>
</tr>
<tr>
<td>$\mu_{RI_s}$</td>
<td>0.020</td>
<td>0.5</td>
<td>0.479</td>
<td>0.142</td>
<td>0.165</td>
<td>0.102</td>
</tr>
<tr>
<td>$\sigma_{RI_s}$</td>
<td>0</td>
<td>0.244</td>
<td>0.244</td>
<td>0.153</td>
<td>0.148</td>
<td>0.055</td>
</tr>
<tr>
<td>$\mu_{Last}$</td>
<td>22212.666</td>
<td>91623.238</td>
<td>69410.571</td>
<td>69036.041</td>
<td>67909.551</td>
<td>8721.802</td>
</tr>
<tr>
<td>$\sigma_{Last}$</td>
<td>3.691</td>
<td>30159.841</td>
<td>30156.149</td>
<td>8123.897</td>
<td>8869.977</td>
<td>5081.647</td>
</tr>
<tr>
<td>max $RI_s$</td>
<td>0.062</td>
<td>0.5</td>
<td>0.4375</td>
<td>0.444</td>
<td>0.424</td>
<td>0.070</td>
</tr>
<tr>
<td>$\mu_{RI_s}$</td>
<td>0.020</td>
<td>0.5</td>
<td>0.479</td>
<td>0.142</td>
<td>0.165</td>
<td>0.102</td>
</tr>
<tr>
<td>$\sigma_{RI_s}$</td>
<td>0</td>
<td>0.244</td>
<td>0.244</td>
<td>0.153</td>
<td>0.148</td>
<td>0.055</td>
</tr>
<tr>
<td>$\mu_{Last}$</td>
<td>22212.666</td>
<td>91623.238</td>
<td>69410.571</td>
<td>69036.041</td>
<td>67909.551</td>
<td>8721.802</td>
</tr>
<tr>
<td>$\sigma_{Last}$</td>
<td>3.691</td>
<td>30159.841</td>
<td>30156.149</td>
<td>8123.897</td>
<td>8869.977</td>
<td>5081.647</td>
</tr>
</tbody>
</table>
ItLast
0.25 0.50
yLast
0.00 0.006
n(RId)
25 50 75
maxRId
0 0.25 0.50 0.75 1.0
x(RId)
0.00 0.02 0.04 0.06 0.08
m(RId)
0 5 10 15 20 25
Density vs. maxRL

Density vs. PRL

Density vs. cRL

Density vs. n( RL)

Density vs. Parameter

Density vs. Parameter
Figure B-1: Density plots of long-term features used in this study


